

Computational Democratization of Network Medicine for Drug Repurposing

Cumulative Dissertation

SUBMITTED TO
Faculty of Mathematics, Informatics, and Natural Sciences
OF
University of Hamburg
Hamburg, Germany

2026

FOR THE DEGREE OF
Doctor rerum naturalium
Dr. rer. nat.

SUBMITTED BY **Andreas Maier**
M.Sc.

AUTHOR

Andreas MAIER

SUPERVISOR

Prof. Jan BAUMBACH

INSTITUTE

Computational Systems Biomedicine, Hamburg, Germany

REVIEWERS OF THE DISSERTATION

FIRST REVIEWER	Prof. Dr. Jan BAUMBACH
SECOND REVIEWER	Prof. Dr. Richard RÖTTGER
THIRD REVIEWER	Prof. Dr. Gunnar KLAU

EXAMINATION COMMITTEE

CHAIR	Prof. Dr. Matthias RAREY
VICE CHAIR	Prof. Dr. Fabian KERN
EXAMINERS	Prof. Dr. Jan BAUMBACH
	Prof. Dr. Richard RÖTTGER

DATE OF DEFENSE

11 September 2026

CONTENTS

List of Figures	v
List of Tables	ix
Abstract	x
Kurzzusammenfassung	xii
Declaration of Authorship	xiv
Acknowledgments	xv
Publication Record	xvi
Acronyms	xviii
1 Introduction	1
1.1 The Grand Challenge: A Crisis in Drug Development	2
1.2 Mechanism-focused Drug Repurposing as a Strategic Solution	4
1.3 Systems and Network Medicine as Methodological Foundations	5
1.3.1 Systems Medicine	5
1.3.2 Network Medicine	8
1.3.2.1 Biological Networks	8
1.3.2.2 Databases and Datasets	10
1.3.2.3 Integrated Network Databases for Drug Repurposing	15
1.3.3 Network Medicine-based Drug Repurposing	16
1.4 Potential and Research Gap in Network-based Repurposing	18
1.4.1 Drug Repurposing - The State of the Art	18
1.4.2 Advantages and Potential of Network-based Methods	22
1.4.3 The Network-driven Drug Repurposing Workflow	22
1.4.3.1 Disease Gene Selection	23
1.4.3.2 Disease Module Identification	24
1.4.3.3 Drug Ranking	27
1.4.3.4 <i>In Silico</i> Validation	29
1.5 Motivation	31
1.6 Aims and Contributions: Improving Access to Drug Repurposing Hypotheses	35
2 Integrated Network-driven Disease Module and Drug Repurposing Candidate	
Identification	38
2.1 Introduction	38
2.2 Results	40
2.3 NeDRex-Web	40
2.3.1 Summary	40
Publication A - NeDRex-Web	42
2.4 Drugst.One	50
2.4.1 Summary	50

	Publication B - Drugst.One	51
3	Disease Gene Identification through Epistasis Detection	61
3.1	Introduction	61
3.2	Results	63
3.3	NEeDL & Epistasis Disease Atlas	64
3.3.1	Summary	64
	Publication C - NEeDL & Epistasis Disease Atlas	66
4	Disease Mechanism Validation	85
4.1	Introduction	85
4.2	Results	85
4.3	DIGEST	86
4.3.1	Summary	86
	Publication D - DIGEST	88
5	Analysis of Tool Integration and Community Reach	102
5.1	Methodological Challenges in Web-Based Success Quantification	104
6	Discussion	107
6.1	Disease Module and Drug Repurposing Candidate Identification	107
6.1.1	NeDRex-Web	107
6.1.2	Drugst.One	108
6.2	Seed Gene Identification	110
6.3	<i>In Silico</i> Validation	111
6.4	Summary	112
7	Conclusion and Outlook	116
7.1	Summary of Presented Results	116
7.2	Conclusion	117
7.3	Outlook	118
	Bibliography	120
A	Appendix	134
A.1	Thesis Supplementary Material	134
A.2	Publication Related Material	138
A.2.0	Licenses and Permissions for Re-publication	138
A.2.0.1	Publication Permissions	138
A.2.0.2	Figure Attribution Statement	139
A.2.1	Authors' Contribution Statements for Supporting Publications	140
A.2.2	Publication Supplements	142
	Supp. A NeDRex-Web	143
	Supp. B Drugst.One	166
	Supp. C NEeDL & EDA	199
	Supp. D DIGEST	244

LIST OF FIGURES

Figure 1.1	The general goal of drug development is the identification of a drug that can treat a certain disease (A) . In drug repurposing (DR), only drugs and pharmacoactive components, that are already known, are screened to identify such links. (B) In a network-based setting, the disease-gene association layer and the drug-target interaction layer can be combined, to identify shared genetic links between existing drugs and known disorders. (C) Multiple reasons limit the power of direct disease-gene-drug link analysis. Firstly, most, if not all of them, are already well researched. Secondly, traditional disease definitions do not fit to the systems medicine approach and the genetic background might not be fully known, both leading to inadequate disease genes. Using the protein-protein interaction (PPI) layer in between drugs and diseases can help to buffer these issues and identify the disease modules that define the mechanistic basis of the disorder which can be targeted by one or multiple drugs (D)	19
Figure 1.2	The network-based drug repurposing (DR) workflow commonly consists of four distinct steps: I) Disease associated genes are identified, e.g., through differential gene expression analysis, knowledge annotated in databases or knowledge graphs, expert knowledge, or any other way. II) The disease genes are taken as input for disease module detection, where specialized algorithms attempt to identify more genes within the same context. The goal is to include genes that either were missing from the initial gene set through methodological limitations or a lack of knowledge and to recruit additional potential drug targets indirectly affecting the disease genes. III) Given the disease module, drug prioritization algorithms create a ranked list of drug candidates. IV) In order to filter out unreliable predictions, <i>in silico</i> validation can be employed on modules and drug candidate results. Statistical and annotation-based approaches are the most common ones to estimate result quality, apart from analytical validation of the results by including external sources about disease genes and drug indications.	23

Figure 1.3	Summary overview of the network-based drug repurposing (DR) workflow in A) and which steps are covered by the presented publications in B). A) For a more nuanced depiction, step I), II), and IV) have annotated sub-categories presenting different solutions to the presented steps. In II) “Network integration” relates to enabling disease mechanism exploration by integrating gene information with network data, enabling connection-based assessments. “Detection algorithms” then specifically hints towards purposefully integrated functions and algorithms assisting this mechanism identification step leveraging the network information. In B) this fine-grained split is utilized to highlight intrinsic capabilities of the presented tools as solid lines, and the expanded and available functionalities through synergistic integration of others of the presented projects. NeEDL and the Epistasis Disease Atlas provide a novel solution for <i>de novo</i> disease gene identification based on network-medicine assumptions and provides a user interface (UI) for interactive exploration of results, integrated with network information. The referral to Drugst.One expands the on-site functions with the whole chain of network-based drug repurposing (DR) functionalities, from disease module detection to drug prioritization. DIGEST provides the missing <i>in silico</i> validation tools for generated disease modules. NeDRex-Web is separated from the other three solutions to highlight the effort of it being a fully integrated and self-sufficient end-to-end solution attached only to the NeDRex infrastructure, in contrast to the other published projects. 36
Figure 2.1	The NeDRex-Web workflow (Reprinted from publication A [250], ©2024 IEEE): NeDRex-Web covers all steps of the network-based drug repurposing (DR) pipeline end-to-end. (A) Users start an analysis by selecting a starting point, e.g. finding genes that represent a disease and analyze them. (B) For this, users can decide between exploring freely through manual network filtering and guided network link analysis steps, or by applying network medicine algorithms. (C) Detailed and informative result overviews provide researchers with options for additional steps, such as the <i>in silico</i> validation of results, exploration through the interactive network explorer, the export or consultation of linked resources for external follow-ups or simply (D) the option to iteratively deepen or refine their hypothesis. All generated results are stored for the user in their personal history. . . 41

Figure 2.2	Drugst.One (Reprinted from publication B [251] with permission from Oxford University Press): Drugst.One can help to transform any existing website (A) dealing with some list of genes or proteins by adding drug repurposing (DR) features. (B) With as little as three lines of code, two imports and the use of the <drugst-one> tag in the html body for actually rendering the component, Drugst.One provides a full suite of easy to use and highly customizable functionalities for (C) simple context exploration of annotated information, network-driven analyses, disease mechanotyping, drug prioritization and in general disease mechanism or drug repurposing (DR) hypothesis generation. 50
Figure 3.1	NeEDL and EpistasisDiseaseAtlas (Reprinted from publication C with permission from Oxford University Press): To be able to assess mutation combinations that could explain complex phenotypes with a depth and statistical significance that has been unseen before, NeEDL provides a heuristic approach based on the combination of network medicine information with local-search-driven epistatic interaction set refinement. So-called single nucleotide polymorphism (SNP)-single nucleotide polymorphism (SNP) interaction networks are created on biological information, such as protein-protein interactions (PPIs) combined with the single nucleotide polymorphisms (SNPs) association to genes, to limit the searchable space of the local-search algorithm. Local-search was prepared and implemented to not only run on traditional binary hardware but also quantum computing (QC) hardware, preparing for once it is readily available. Results are stored and provided through the Epistasis Disease Atlases (EDAs), a first-of-its-kind online resource for promising human epistatic interaction sets, options for refinement, and the direct possibility to start drug repurposing (DR) analyses. 64
Figure 4.1	DIGEST (Reprinted from publication D [273] with permission from Oxford University Press): Overview of <i>in silico</i> validation functions provided. (a) Gene or disease sets, clusterings, or subnetworks can (b) be evaluated with or without a reference to compare to regarding their functional and genetic coherence, respectively. For this, information from NeDRex, KEGG, Gene Ontology, and others is integrated allowing the support of multiple identifier (ID) spaces, the functional and genetic annotation, and the establishment of a background signal for empirical p-value calculation. (c) Results are persistently reachable through a stable URL and present options to deeply analyze the composition of the input selection of diseases or genes based on their functional annotation to assess mechanistic hypotheses and refine or expand the search for additional evaluation. 87

Figure 5.1	Webtool Reach: Cumulative Unique Interactions: Apache vhost logs are processed by splitting all tools into individual logs (where necessary), applying some filters to remove non-genuine requests as far as possible, and then plotting unique Internet Protocol (IP) interactions aggregated over all components of each tool (A) and split into the individual components statistics, where possible, (B) over time. The colors are in line with Figure 1.3, where purple corresponds to Drugst.One related services, turquoise to NeDRex-Web, red to the Epistasis Disease Atlas (EDA) and blue to DIGEST. The tools in gray are 23 other publicly hosted web tools on the same web server for comparison. In July 2023 the server was first set up, hence no older logs exist, even though some tools predate this timepoint.	105
Figure 5.2	Webtool Usage: Monthly Total Requests (Hits): Logged requests serve as a proxy for general continued interest and recurring usage. This figure highlights the absolute monthly usage statistics (in number of requests logged) as a proxy for general continued interest in and usage of the tool, regardless of it being new or recurrent users: (A) Aggregated monthly hits per tool, showing Drugst.One's steadily high request volume. (B) Component-specific monthly hits, revealing the high traffic levels sustained by backends and APIs compared to front-facing websites.	106
Figure A.1	Web Tool Reach: Cumulative Unique Interactions - Unfiltered: This figure was created equivalently to Figure 5.1, but the completely unfiltered server logs were used as source of the statistics.	137
Figure A.2	Web Tool Usage: Monthly Total Requests (Hits) - Unfiltered: This figure was created equivalently to Figure 5.2, but the completely unfiltered server logs were used as source of the statistics.	137

LIST OF TABLES

Table 1.1	Common resources for nodes in network medicine; lexicographically ordered by database name.	11
Table 1.2	Common resources for edges/relationships in network medicine; lexicographically sorted by database name.	13
Table 1.3	Summary of different methodology classes for targeted drug repurposing (DR) including their core principles, essential data, and key challenges.	21
Table 1.4	Overview of resources and methods for disease gene selection. . . .	24
Table 1.5	Overview of common methods to identify disease mechanisms, including disease module identification algorithms, graph algorithms, and integrative tools.	26
Table 1.6	Overview of drug prioritization and ranking methodologies.	28
Table 1.7	Overview of strategies for the <i>in silico</i> validation of computational hypotheses.	31
Table 5.1	Reach and usage metrics for the developed network medicine tools (as of end of April 2026). Data excludes known web crawlers and a list of 45,398 exclusion Internet Protocol (IP) addresses. The Metric / Component column describes the specific component of a tool that is being assessed and how; “Total” in that context refers to a collective evaluation of components of the current tool or if specifically mentioned, a specific version of the tool. As unique Internet Protocol (IP) addresses are used as a metric, result sets of individual components of the same tool can and often do record the same Internet Protocol (IP) addresses, making it impossible to simply add up final counts of unique Internet Protocol (IP) addresses over all components to receive a total.	102
Table A.1	Updated list of tools verifiably integrating Drugst.One (in alphabetical order). Excluded are four tools that are not publicly available or not published and one page that used it in the past, but removed it since.	134

ABSTRACT

Finalizing the first human genome assembly in 2003 gave rise to the hope that diseases, even complex ones, would finally become fundamentally understandable and treatable. Yet, nearly a quarter-century into the postgenomic era, humanity is still far from achieving universal treatment adequacy. A fundamental “translational gap” remains, leaving patients of many disorders, especially complex conditions, underserved due to outdated disease definitions, limited causal understanding of diseases, a lack of actionable biomarkers, and a critical efficiency crisis in *de novo* drug development. Today, bringing a new compound to market requires over a decade of research and investments exceeding \$1 billion. One major reason for these immense costs is that approximately 90% of drug candidates fail in clinical trials. While drug repurposing (DR), the identification of new therapeutic options for existing drugs, offers a cost-effective alternative, historical successes have largely been serendipitous. To move beyond accidental discoveries and enable a targeted approach to underserved conditions, it is essential to elucidate the mechanistic background of diseases for precise targeting with drugs. Systems medicine provides an adequate, conceptual paradigm for this holistic view of the human organism. A practical computational implementation of systems medicine is called network medicine, abstracting biological entities and their complex relationships into a graph of nodes and edges to allow for the systematic analysis of molecular interplay and associated relations. Through this framework, network medicine identifies disease modules, i.e. functional units of multiple genes involved in a disorder, to prioritize therapeutic candidates and uncover the systemic perturbations driving complex phenotypes.

However, the potential of network-based DR is hindered by three primary bottlenecks that prevent the translation of computational hypotheses into clinical reality. First, a **lack of democratization** persists because existing network tools are often too complex or fragmented for domain experts –such as clinicians and pharmacologists– to use effectively. Second, a **seed gene knowledge gap** remains; inconsistent disease definitions and the “missing heritability” of complex traits can make it difficult to identify appropriate genetic starting points for network analyses. Finally, an ***in silico* validation gap** exists, as the field lacks robust, quantifiable metrics to assess the mechanistic relevance of identified disease modules and repurposing hypotheses before moving to expensive experimental verification.

This thesis addresses these three open challenges by presenting four novel, synergistic software solutions that together cover the entire *in silico* network-driven DR value chain while prioritizing accessibility and integrability. **NEeDL** provides a network medicine-founded framework for identifying higher-order epistatic interactions from genome-wide association study (GWAS) data for both classical binary and emerging quantum computing

(QC) architectures. Through the associated **Epistasis Disease Atlas (EDA)** NEEeDL results for eight complex disorders are made publicly available, making it the first epistasis candidate set resource and offering a novel source of disease-associated genes from genetic data. **NeDRex-Web** offers a fully functional, browser-based environment for deep network-driven mechanism exploration and drug prioritization, eliminating the technical barriers of standalone software for interactions with the NeDRex database and adding functionalities like facilitating tissue-specific analyses. **Drugst.One** is –predominantly– a plug-and-play JavaScript solution that enables any biomedical platform to integrate a comprehensive kit of network medicine and DR tools with minimal code, greatly increasing the distribution of these methods. Finally, **DIGEST** introduces a novel statistical approach for the *in silico* validation of disease modules, using functional and genetic coherence metrics to increase the confidence in computational findings.

These projects facilitate the democratization of network medicine by providing intuitive, web-native interfaces. By collectively covering the full network-driven workflow from seed gene identification to statistical validation, these tools allow domain experts to harvest their expert knowledge in identifying DR candidates. At the same time, this work significantly lowers the entry barrier for network-driven research, hopefully accelerating the transition from *in silico* hypothesis to real-world patient benefit.

KURZZUSAMMENFASSUNG

Die Fertigstellung des ersten menschlichen Genom-Assemblys im Jahr 2003 weckte die Hoffnung, dass Krankheiten, selbst komplexe, endlich grundlegend verstehbar und heilbar werden würden. Doch fast ein Vierteljahrhundert nach Beginn der postgenomischen Ära ist die Menschheit noch weit von einer flächendeckenden adäquaten Gesundheitsversorgung entfernt. Es besteht weiterhin eine fundamentale „translationale Lücke“, die dazu führt, dass viele Patienten, insbesondere komplexer Erkrankungen, aufgrund veralteter Krankheitsdefinitionen, eines begrenzten kausalen Verständnisses von Krankheiten, eines Mangels an klinisch nutzbaren Biomarkern und einer kritischen Effizienzkrise in der *De-novo*-Arzneimittelentwicklung unterversorgt bleiben. Heute erfordert die Markteinführung eines neuen Wirkstoffs über ein Jahrzehnt Forschung und Investitionen von mehr als einer Milliarde US-Dollar. Ein Hauptgrund für diese immensen Kosten ist, dass etwa 90 % der Wirkstoffkandidaten in klinischen Studien scheitern. Während das Drug Repurposing (DR) – die Identifizierung neuer therapeutischer Optionen für bestehende Medikamente – eine kostengünstige Alternative bietet, beruhten historische Erfolge bisher weitgehend auf Zufallsfinden. Um über zufällige Entdeckungen hinauszugehen und einen gezielten Ansatz für unterversorgte Krankheiten zu ermöglichen, ist es unerlässlich, den mechanistischen Hintergrund von Krankheiten aufzuklären, um diese präzise mit Wirkstoffen adressieren zu können. Die Systemmedizin bietet hierfür ein geeignetes konzeptionelles Paradigma für eine ganzheitliche Betrachtung des menschlichen Organismus. Eine praktische computergestützte Implementierung der Systemmedizin ist die Netzwerkmedizin, die biologische Entitäten und ihre komplexen Beziehungen in einem Graphen aus Knoten und Kanten abstrahiert, um eine systematische Analyse molekularer Wechselwirkungen und zugehöriger Beziehungen zu ermöglichen. In diesem Rahmen identifiziert die Netzwerkmedizin Krankheitsmodule – d. h. funktionelle Einheiten von mehreren Genen, die an einer Krankheit beteiligt sind –, um therapeutische Kandidaten zu priorisieren und die systemischen Störungen aufzudecken, die komplexen Phänotypen zugrunde liegen.

Das Potenzial des netzwerkgestützten Drug Repurposing (DR) wird jedoch durch drei zentrale Engpässe behindert, welche die Übertragung computergestützter Hypothesen in die klinische Realität verhindern. Erstens gibt es einen **Mangel an Demokratisierung**, da bestehende Netzwerk-Tools für Fachexperten – wie Kliniker und Pharmakologen – oft zu komplex oder verteilt sind, um sie effektiv nutzen zu können. Zweitens besteht eine sogenannte „**Seed-Gene**“ **Wissenslücke**; inkonsistente Krankheitsdefinitionen und die „Missing Heritability“ komplexer Merkmale erschweren es, geeignete genetische Ausgangspunkte für Netzwerkanalysen zu identifizieren. Schließlich gibt es eine **In-silico-Validierungslücke**, da es dem Feld an robusten, quantifizierbaren Metriken fehlt, um die

mechanistische Relevanz identifizierter Krankheitsmodule und Repurposing-Hypothesen zu bewerten, bevor diese einer kostspieligen experimentellen Verifizierung unterzogen werden.

Diese Dissertation adressiert diese drei offenen Herausforderungen durch die Vorstellung von vier neuen, synergetischen Softwarelösungen, die gemeinsam die gesamte *In-silico*-Wertschöpfungskette des netzwerkbasierten DR abdecken und dabei Barrierefreiheit und Integrierbarkeit priorisieren. **NEeDL** bietet ein auf der Netzwerkmedizin basierendes Framework zur Identifizierung epistatischer Interaktionen höherer Ordnung aus GWAS-Daten, sowohl für klassische binäre als auch für aufstrebende Quantencomputing-Architekturen. Über den zugehörigen **EpistasisDiseaseAtlas (EDA)** werden NEeDL-Ergebnisse für acht komplexe Krankheiten öffentlich zugänglich gemacht, was ihn zur ersten Ressource für Epistasie-Kandidatensätze macht und eine neuartige Quelle krankheitsassoziierter Gene aus genetischen Daten bietet. **NeDRex-Web** bietet eine voll funktionsfähige, browserbasierte Umgebung für die tiefgehende netzwerkgestützte Mechanismus-Exploration und Wirkstoff-Priorisierung. Es beseitigt die technischen Barrieren eigenständiger Software für die Interaktion mit der NeDRex-Datenbank und bietet Funktionen wie die Erleichterung gewebespezifischer Analysen. **Drugst.One** ist allem voran eine Plug-and-Play-JavaScript-Lösung, die es jeder biomedizinischen Plattform ermöglicht, ein umfassendes Kit an Netzwerkmedizin- und Drug-Repurposing-Tools mit minimalem Code-Aufwand zu integrieren, was die Verbreitung dieser Methoden erheblich steigert. Schließlich führt **DIGEST** einen neuen statistischen Ansatz für die *In-silico*-Validierung von Krankheitsmodulen ein, der funktionelle und genetische Kohärenzmetriken nutzt, um das Vertrauen in computergestützte Ergebnisse zu stärken.

Zusammengenommen fördern diese Projekte die Demokratisierung der Netzwerkmedizin durch intuitive, webbasierte Schnittstellen. Indem sie gemeinsam den gesamten Workflow von der Identifizierung der Seed-Gene bis zur statistischen Validierung abdecken, ermöglichen diese Tools es Fachexperten, ihr Fachwissen bei der Identifizierung von DR-Kandidaten voll auszuschöpfen. Gleichzeitig senkt diese Arbeit die Einstiegshürde für die netzwerkgestützte Forschung erheblich und beschleunigt hoffentlich den Übergang von der *In-silico*-Hypothese zum realen Nutzen für den Patienten.

EIDESSTATTLICHE VERSICHERUNG

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Dissertationsschrift selbst verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Sofern im Zuge der Erstellung der vorliegenden Dissertationsschrift generative Künstliche Intelligenz (gKI) basierte elektronische Hilfsmittel verwendet wurden, versichere ich, dass meine eigene Leistung im Vordergrund stand und dass eine vollständige Dokumentation aller verwendeten Hilfsmittel gemäß der Guten wissenschaftlichen Praxis vorliegt. Ich trage die Verantwortung für eventuell durch die gKI generierte fehlerhafte oder verzerrte Inhalte, fehlerhafte Referenzen, Verstöße gegen das Datenschutz- und Urheberrecht oder Plagiate.



Andreas Maier

May 13, 2026

ACKNOWLEDGMENTS

First and foremost, I would like to thank **Prof. Jan Baumbach**, my supervisor, for giving me the chance, support, and environment to grow and do my doctoral work at CoSy.Bio. After all this time, I am still very grateful for the opportunity to come to Hamburg and, together with my peers, kick off the CoSy.Bio group.

Without my supportive family and friends, I would not have made it to where I am today. Coming from a non-academic background, I was blessed to be primed with an interest in science, crafts, and nature by my parents and (late) grandparents from the very beginning. They fueled my curiosity about how the world works, which has all culminated in my current passion for a diverse field of sciences, and for that, I am forever grateful. A big thank you also goes to my closest friends for sticking by me. Even though we lead very dissimilar lives in different places, I cherish our friendship like little else. A special thank you to my brother, who shows me that you can always reinvent yourself and start new, passionate journeys if you just go for it. And thank you to the more recent people in my life who choose to spend a lot of time with me and have the patience to continue doing so.

A general but heartfelt thank you goes out to many of my peers in CoSy.Bio, especially the management team, for always being available and offering help. While our line of work is not always the easiest mentally, this environment –and especially the people in it– helped me again and again to find new heights and the energy to push through hardship. I especially want to acknowledge my colleagues who have become more than just co-workers to me, spending extended time together outside of work at concerts, on camping trips, sports, board games, or enjoying other fun activities that enrich my life.

I am even grateful for the unavoidably less pleasant situations and experiences over these past years, as they taught me a lot and helped me learn and grow.

PUBLICATION RECORD

CORE PUBLICATIONS OF THIS DISSERTATION

A	Maier*, A., Rafiei, M., Anastasi, E., Zolotareva, O., Skelton, J., Elkjaer, M. L., ... [et al.] (2024): NeDRex-Web: An Interactive Web Tool for Drug Repurposing by Exploring Heterogeneous Molecular Networks. <i>IEEE BIBM</i> , IEEE Computer Society, 429–434. doi: 10.1109/BIBM62325.10822390 .
B	Maier*, A., Hartung*, M., Abovsky, M., Adamowicz, K., Bader, G. D., Baier, S., ... [et al.] (2024): Drugst.One – A plug-and-play solution for online systems medicine and network-based drug repurposing. <i>Nucleic Acids Res.</i> , gkae388. doi: 10.1093/nar/gkae388 .
C	Hoffmann*, M., Poschenrieder*, J. M., Incudini*, M., Baier*, S., Fritz*, A., Maier*, A., ... [et al.] (2024): Network medicine-based epistasis detection in complex diseases: ready for quantum computing. <i>Nucleic Acids Research</i> , 52(17), 10144–10160. doi: 10.1093/nar/gkae697 .
D	Adamowicz, K., Maier, A., Baumbach, J., Blumenthal, D. B. (2022): Online <i>in silico</i> validation of disease and gene sets, clusterings or subnetworks with DIGEST. <i>Brief. Bioinform.</i> , 23(4), bbac247. doi: 10.1093/bbac247 .

* (shared) first authors

SUPPORTING PUBLICATIONS OF THIS DISSERTATION

Detailed individual author contribution statements for all co-authored supporting publications listed below are provided collectively in [Appendix A.2.1](#).

[a] Adamowicz, K., Arend, L., **Maier, A.**, Schmid, J. R., Kuster, B., Tsoy, O., Zolotareva, O., Baumbach, J., Laske, T. (2023): Proteomic meta-study harmonization, mechanotyping and drug repurposing candidate prediction with ProHarMeD. *npj Syst Biol*, 23(4), bbac247, doi: [10.1038/s41540-023-00311-7](https://doi.org/10.1038/s41540-023-00311-7).

[b] Oubounyt, M., Adlung, L., Patroni, F., Wenke, N. K., **Maier, A.**, Hartung, M., Baumbach, J., Elkjaer, M. L. (2023): Inference of differential key regulatory networks and mechanistic drug repurposing candidates from scRNA-seq data with SCANet. *Bioinformatics*, 39(11), btad644, doi: [10.1093/bioinformatics/btad644](https://doi.org/10.1093/bioinformatics/btad644).

[c] Lobentanzer, S., Aloy, P., Baumbach, J., . . . , **Maier, A.**, . . . , Saez-Rodriguez, J. (2023): Democratizing knowledge representation with BioCypher. *Nat. Biotechnol.*, 1–4, doi: [10.1038/s41587-023-01848-y](https://doi.org/10.1038/s41587-023-01848-y).

[d] Sarkar, S., Lucchetta, M., **Maier, A.**, Abdrabbou, M. M., Baumbach, J., List, M., Schaefer, M. H., Blumenthal, D. B. (2023): Online bias-aware disease module mining with ROBUST-Web. *Bioinformatics*, 39(6), btad345, doi: [10.1093/bioinformatics/btad345](https://doi.org/10.1093/bioinformatics/btad345).

[e] Sadegh, S., Skelton, J., Anastasi, E., **Maier, A.**, Adamowicz, K., Möller, A., Kriege, N. M., Kronberg, J., Haller, T., Kacprowski, T., Wipat, A., Baumbach, J., Blumenthal, D. B. (2023): Lacking mechanistic disease definitions and corresponding association data hamper progress in network medicine and beyond. *Nat. Commun.*, 14(1), 1662, doi: [10.1038/s41467-023-37349-4](https://doi.org/10.1038/s41467-023-37349-4).

[f] Delgado-Chaves, F. M., Spindler, L., Firoozbakht, F., **Maier, A.**, Hartung, M., ... [et al.] (2024): DrugRepoChatter: A Drug Repurposing Expert Chatbot Curated by the REPO4EU Consortium. *DrugRepo*, 1(2), doi: [10.58647/DRUGREPO.24.2.0014](https://doi.org/10.58647/DRUGREPO.24.2.0014).

[g] Süwer, S., Bagemihl, K., Baier, S., Dicunta, L., List, M., Baumbach, J., **Maier, A.**, Delgado-Chaves, F. M. (2025): Conversational no-code and multi-agentic disease module identification and drug repurposing prediction with ChatDRex. *Preprint*, doi: [arXiv.2511.21438](https://doi.org/10.21203/rs.3.rs-1511111/v1).

[h] Spindler, L. M., Kersting, J., Manz, Q., Hartung, M., **Maier, A.**, Mamdouh, Z. M., Casas, A. I., Baumbach, J., List, M. (2026): Drugst.One DREAM–Drug repurposing through expert annotation and modification. *Br. J. Pharmacol.*, 1–14, doi: [10.1111/bph.70495](https://doi.org/10.1111/bph.70495).

[i] Firoozbakht, F., Süwer, S., Elkjaer, M. L., Handy, D. E., **Maier, A.**, Li, J., Lancashire, L., Loscalzo, J., Baumbach, J. (2026): NetMedGPT – A network medicine foundation model for extensive disease mechanism mining and drug repurposing. *Preprint*, doi: [10.64898/2026.01.04.697552](https://doi.org/10.64898/2026.01.04.697552).

ACRONYMS

AI	Artificial Intelligence
CDN	Content Delivery Network
DR	Drug Repurposing
EDA	Epistasis Disease Atlas
EHR	Electronic Health Record
EMA	European Medicines Agency
FDA	U.S. Food And Drug Administration
GNN	Graph Neural Network
GWAS	Genome-wide Association Study
ICD	International Classification Of Diseases
ID	Identifier
IP	Internet Protocol
KG	Knowledge Graph
LLM	Large Language Model
ML	Machine Learning
NNT	Number Needed To Treat
PMDA	Pharmaceuticals And Medical Devices Agency
PPI	Protein-protein Interaction
QC	Quantum Computing
SNP	Single Nucleotide Polymorphism
UI	User Interface
WHO	World Health Organization

*

*I watch the shadows fall on horizons we echo
We are the difference
A congregation of like-minded hues and separation
An introversion of life
Am I broken, lost, battered, and tossed?
I was the one on the outside
Searching for peace I would never find
Black roses loom in the dark
Never to fall apart (To fall apart)
We were perfectly sewn over imperfect bones
We made a home in the shadow of madness*

— LORNA SHORE

"In Darkness" (2025)

Written by A. O'Connor, W. Ramos, and Lorna Shore

INTRODUCTION

According to data from the World Health Organization (WHO), at the start of the 21st century non-communicable diseases, diseases not transmissible between individuals, were responsible for 60% of global deaths, increasing to 70% just over a decade later [1, 2]. This trend persists despite the milestone assembly of the human genome in 2003, which many believed would lead us into an era of near-universal treatability [3, 4]. In the years since, the anticipated clarity of the post-genomic era has been replaced by an appreciation for the immense complexity of human biology, proving that genetic information is only one piece of a much larger mechanistic puzzle. With advancements in understanding this complexity of the human organism and with next-generation and now third-generation sequencing machines broadly available, genetics and genomics has undoubtedly had a sizable impact on medical research since [5, 6]. Still, nearly a quarter of a century into genetic and genomic research, immense limitations remain with regard to the causal definition of many diseases and, thus, the availability of effective treatment options. For example, recently, the long-standing theory of amyloid beta being causal and therefore the best target to treat Alzheimer's disease was challenged, with the impact on the research field still being assessed [7, 8]. But it is possible that this misinterpretation of correlation as causality slowed down the research for effective treatments for this neurodegenerative disorder [7].

In general, *de novo* drug development requires more time and resources than ever before, giving the impression that all "trivial" diseases and therapeutics have been identified [9]. At the same time, commonly prescribed therapeutics suffer from low efficacy and a high number needed to treat (NNT), a measure of the number of people who must be given a drug for one person to benefit from it [10–12]. Even though NNTs are a gross simplification of a complex situation, they reveal a fundamental problem in how diseases are currently understood and defined, and how these definitions are used to prescribe drugs that may have limited or no effect on the disease or, if anything, the symptom [13].

This thesis will assess how drug repurposing (DR), i.e. identifying new applications for already existing drugs, can help overcome or at least mitigate the current drug development crisis by democratizing DR research. It will specifically show how combining DR with systems and network medicine can help to elucidate the molecular mechanisms of disorders to find drugs that act on these mechanisms. Knowledge of the molecular mechanism targeted by a drug further allows clinicians to screen patients for specific, treatable perturbations and to avoid administering the drug to those where the perturbation is absent. All of these approaches and their consequences have the power to lower the observed high NNT. The focus of this thesis is to provide accessible and integrated tools to the community, enabling not only computational specialists but especially clinical experts, like pharmacologists, to contribute to the computational search for new drug

potential.

To this end, this first chapters details the current issues and challenges of drug development before introducing how systems and network medicine paradigms can help to approach these problems. As a practical solution, mechanism-focused DR is introduced, including the underlying concepts and computational tools that can help address these shortcomings. This chapter of the thesis concludes by presenting the open challenges that must be addressed to realize this potential and the specific steps taken as part of this doctoral research to move the field forward.

1.1 THE GRAND CHALLENGE: A CRISIS IN DRUG DEVELOPMENT

Despite the promise of the genomic era drug development faces a crisis of efficiency. It has never been more resource consuming than now to develop a new therapeutic option for a specific condition by identifying a previously unused chemical molecule or creating a completely new compound. Typically, 10–15 years and \$1.0–\$2.6 billion are required from starting research on a new target to drug approval [14, 15]. The U.S. Food and Drug Administration (FDA) outlines the required steps from start to approval, excluding the post-market safety monitoring, as: 1. Discovery and Development of a new drug; 2. Preclinical Research, establishing basic drug safety in lab and/or animal models; 3. Clinical Research, testing the drug in clinical trials; 4. FDA Review for potential approval of the drug [16]. The “Clinical Trial Phases” of the “Clinical Research” step (3) are the main hurdle for most efforts: Phase 1, establishing the safety and dosage of the drug in healthy humans; Phase 2, establishing efficacy and side effects in diseased patients; Phase 3, further studying the efficacy and identifying and adverse reactions in a larger patient group [17]. According to the FDA the amount of drugs not making it to the next phase of trials is 30%, 67%, and 65–70%, respectively, collectively causing more than 90% of drug candidates to be dropped some time during this clinical research step. The reasons can include issues concerning drug delivery, dosing feasibility, and other translational roadblocks from initial preclinical successes to trial efficacy [18, 19]. In the current drug development pipeline, drugs that show high efficacy, meaning strong performance in controlled clinical trial settings, often fail to show the same level of effectiveness in real-world clinical practice [20].

A major reason for this lack of effectiveness is that many drugs target surrogate markers and treat symptoms yet fail to treat the actual disease. For example, many drugs focusing on poorly understood conditions, like psychiatric, neurological, or pain-related disorders, achieve only symptomatic or progression-altering treatment [21–24]. They can temporarily improve the quality of life of patients or slow disease progression, but do not address the underlying disorder [25]. Realistically, complete cure can often not be expected except in many infectious and some cancer-related diseases with currently existing knowledge and techniques [26, 27]. While targeted gene therapy holds the promise of curative treatment for a growing number of conditions, effective disease management to reduce disease burden for patients remains the current standard of care and the best especially medication can achieve [28–30].

Trials often measure drug efficacy based on surrogate markers that may not directly correlate with disease progression or perceived severity [10]. A study has shown on a list of 41 drugs, that in only around 15% of the cases approved drugs have associated genetic markers allowing for stratification of highly-responding groups, hence, highly-effective treatment [13]. Of those, only

around 50% are estimated to qualify for patient stratification [13]. In many other cases, conditions might be multifactorial and too complex to expect high effect sizes from simple genetic markers [31]. While 41 drugs are hardly representative it quantifies the general issue.

Even if drugs appear to show potential in early trials, failures before reaching the market, or the lack of effectiveness in real-world patients, can often be linked to trial design [32, 33]. The reliance on animal models, currently already decreasing, leads to issues in translation to humans, as the response to drugs or the actual condition can vary greatly between humans and other animals [33–35]. However, even simple trials and approval processes are lagging behind in acknowledging and accounting for sex differences, even though pharmacodynamic differences between biologically male and female patients are well established in many disorders [36]. After passing all required clinical trials, the drug needs to be proven to have a large enough effect on a given condition, with a small effect size being sufficient in case no other medication for a serious condition exists [37].

The current landscape reveals an inherent tension in regulatory oversight. On one hand, stricter approval guidelines are imperative to ensure patient safety and the approval of only effective treatments. Simultaneously, as approval processes have become more stringent, many established therapies remain in use. Unless they cause major complications like thalidomide did, many older drugs stay on the market even if they wouldn't meet today's regulatory requirements [38]. Aspirin, for example, could fail the latest safety regulations. For an over-the-counter drug, its 1% risk of gastrointestinal bleeding is very high relative to its comparatively low effectiveness when measured against other non-steroidal anti-inflammatory drugs like ibuprofen [39]. Recent studies highlight the ineffectiveness of paracetamol in the treatment of low back pain, going against clinical care recommendations and disease management guidelines [40]. Similarly, digoxin, a drug historically used in treating cardiovascular disorders, shows a reduction in hospitalizations but carries a high risk of overdose [41, 42].

So, while these drugs remain in use, the same stringent guidelines necessary for patient safety can simultaneously hamper significant progress in bringing new, affordable, and effective drugs to patients. This is especially true for more precision-oriented medications focusing on causal treatments of underserved conditions [22, 43]. The most extreme case of neglected conditions are rare diseases. While more than 200 million people are estimated to live with a rare disease, each individual disease affects a small part of the population, depending on the definition, fewer than 200,000 or even 2,000 people [44, 45]. Naturally, investing billions of dollars into developing a drug for a rare disease where, in the eyes of a profit-oriented company, "worst case" only 2,000 people are affected and paying for a remedy is not of highest interest [25]. Especially not, when additional regulatory hurdles complicate the approval of such DR drugs and no adequate return of investment can be expected [46, 47]. Even financial incentives designed to address this gap, such as the Orphan Drug Act, have had limited success. While the Orphan Drug Act led to the development and patenting of drugs targeting rare diseases, the extended market exclusivity it grants often limits affordability during the initial seven-year period, meaning even if the act incentivizes development of new orphan drugs, it is not directed to the primary benefit of patients [48, 49].

All these issues accumulate and negatively influence clinical outcomes. Because the molecular understanding of how a medication acts on a disorder is often missing, well-founded biomarkers that could differentiate responders from non-responders cannot be identified [50]. This means

initial selection of medication is heavily based on the symptoms instead of a balanced and holistic view of the patient's condition through molecular patient stratification methods. Medication is then adjusted if it fails to improve the condition or its symptoms [31, 51]. This "trial-and-error" approach leads to an unreasonably high NNT for common drugs, which are often overprescribed, helping only a small fraction of patients while exposing the majority to the risk of adverse drug reactions [10]. For example, the ten drugs generating the highest gross revenue in the US have an NNT between four (Humira, Enbrel, Remicade for treating arthritis, psoriasis, and Crohn's disease, respectively) and 20 (Crestor, Advair Diskus for high cholesterol and asthma, respectively), or even 25 (Nexium against heartburn) [12].

The root cause of these issues is mostly a lack of molecular understanding of disease-drug relationships, limiting not only the future possibility of personalized medicine but also the effectiveness of targeted medicine today. Fully personalized medicine with individually tailored solutions is already possible on a small scale for certain groups of disorders, but affordability remains a big hurdle [52–54]. So-called "precision medicine" has the potential to serve as an intermediate step by enabling precise patient stratification and treatment selection for common conditions, if, in addition to causal and effective treatment solutions, availability and affordability of the treatments are addressed. Current drug development pipelines do not need to be fully discarded, but could be significantly improved by incorporating this causal-relationship level to improve translation from preclinical models and clinical trials to real-world treatment effectiveness.

1.2 MECHANISM-FOCUSED DRUG REPURPOSING AS A STRATEGIC SOLUTION

Drug repurposing (DR; also called drug repositioning, and more rarely drug redirecting or drug reprofiling) attempts to address the issues of long drug development times, availability, and affordability by identifying new indications for already researched and known chemical compounds. Drugs can be already indicated and approved for other conditions, all trials so far might have failed, or the drug has been withdrawn from its original use case, but information about them has been gathered before. Reusing already investigated drugs helps to reduce development pipeline times from 10–15 to 3–12 years and promises a cost reduction of 60%–90% in comparison to developing a novel drug from scratch [55–58]. Specifically, certain steps in the drug-to-market process can be skipped, as baseline information about dosage and safety profiles in healthy humans might already have been established in earlier attempts. This can allow the approval efforts of a DR candidate to start with Phase 2 clinical trials [59]. Translational success rates can also be higher, as established safety profiles and existing pharmacokinetic and toxicologic data help to exclude unsuitable candidates early in the process [60, 61]. Overall, DR can be seen as a complementary approach to traditional drug development, utilizing all available information and compounds to bring suitable treatment candidates to underserved conditions.

DR is not a new idea. There are many well-known historic examples where a new purpose for an existing drug was found. For instance, galantamine was identified in the 1950s as an anti-paralytic agent for myopathies but is now one of the few drugs used to treat Alzheimer's disease [62]. In 1961, thalidomide was withdrawn from its original indication as a sedative for pregnant women due to severe effects on embryonic development, yet it is now used to treat patients with multiple myeloma [63] and leprosy [64]. A very prominent example is sildenafil, better known as Viagra, which was originally used for the treatment of cardiovascular issues such as angina pectoris

[65]. Similarly, the desirable side effect of hair growth promotion induced by minoxidil created a new purpose for this antihypertensive drug [66, 67]. Recently, the appetite reduction effects of semaglutide, marketed first as Ozempic and approved for diabetes management, caused mass off-label use of the drug for weight-loss purposes, leading to shortages for patients depending on it [68]. As a result, semaglutide was tried and approved with higher dosage under the name Wegovy as a weight-loss drug for specific cases of obesity [69]. While many more examples of successfully repurposed drugs exist, the key difference is that compared to the modern approach these historical discoveries were largely serendipitous, occurring mostly without a prior understanding of the drug's mode of action and without deliberate searching for them.

The transition to purpose-driven DR could help address several unmet medical needs, such as finding effective treatment options for the hundreds of millions of people worldwide living with a rare disease [25, 45]. Complex multifactorial diseases require novel, yet-unidentified treatments or combination therapies [70, 71]. Understanding or at least identifying the molecular cause and the underlying disease mechanism, is a first step to inform ways to target the disease directly instead of its symptoms [30]. Similarly, infectious disease outbreaks like COVID-19 have reminded humanity that newly emerging pathogens require rapid therapeutic responses, especially for vulnerable groups. DR has the potential to facilitate all that, by making the search for novel and targeted treatments more affordable whenever resources are limited [72–75]. For pharmaceutical companies DR can even offer opportunities to prolong patent life and market exclusivity for certain drug applications [76, 77]. Thus, mechanistically driven DR provides a powerful potential to reduce the burden on healthcare systems, identify more and new causative treatment options, and mitigate the current drug development crisis, if adequate methodologies are used.

1.3 SYSTEMS AND NETWORK MEDICINE AS METHODOLOGICAL FOUNDATIONS

The following sections detail the methodological foundations on which mechanism-based DR can be built. The discussion will lead from high-level paradigms to the practical data resources. It begins by introducing *systems medicine* to establish the conceptual framework of viewing diseases holistically. It then explains how *network medicine* serves as the computational implementation of these principles. Subsequently, the chapter covers the fundamentals of *biological networks* and provides a critical assessment of the data sources required to build them. This includes an analysis of both general-purpose *databases and datasets* and those specifically oriented towards DR, highlighting the inherent challenges and limitations within these foundational resources. Finally, this critical perspective culminates in an examination of current *network medicine-driven approaches*, which serves to identify the specific potential of publicly available tools and their limitations upon which the subsequent motivation for this dissertation is built.

1.3.1 Systems Medicine

A central idea of modern medical research is systems medicine, with the concept of investigating the human organism as one system, viewing tissues and organs as connected units rather than distinct entities. Systems medicine is the scientific foundation for a concept called P4 medicine. P4 medicine describes the clinical implications of systems medicine and a vision for healthcare that is **predictive, preventive, personalized, and participatory** rather than trying to treat disorders the same way in every patient or even outright protect against potential diseases in the first place

[78]. Systems medicine lies the foundation to achieve this healthcare revolution, by aiming to identify and treat conditions correctly, that until now have not been sufficiently understood or gone ill-diagnosed. Reasons for diagnostic challenges may include a main symptom occurring in a certain organ resulting in the condition being classified as organ-specific, yet there might be other affected organs and the cause might not even be found in one organ alone. This is comparable to classifying carbon monoxide poisoning as a lung-related issue, which presumably was the case before the molecular cause was studied and described throughout the 18th and 19th centuries [79]. Therefore, systems medicine's focus lies on identifying systemic causes and finding treatments whenever possible for disorders that are caused by behavioural, environmental, and/or genetic factors, causing microbiological perturbations in multiple organs, tissues, or even the whole organism. This of course is not the case for every known disease, yet can help advance in the understanding and potential treatment of until now underserved diseases.

To prevent and predict diseases with personalized impact on a large scale, as P4 medicine envisions it, systems medicine relies on data-driven approaches to understand or at least uncover the link between genotype and phenotype. While the central dogma of molecular biology describes how genetic information is translated into functional proteins, the complexity of regulatory mechanisms means that a systemic genetic fault does not always produce a systemic effect. Through processes like differential gene expression, a defective protein may only be expressed in specific tissues or at specific times [80–82]. Consequently, the resulting phenotype can be highly organ-specific, such as an enlarged liver in glycogen storage disease type I, which results from a defective glucose-6-phosphatase or glucose-6-translocase, in type Ia or Ib respectively, even though the underlying genetic cause is present in every cell of the organism [83]. On the other hand, in cases of purely genetic disorders, essential mechanisms can be impacted through a single mutation causing systemic effects. In monogenic cases of autoimmune disorders like monogenic lupus, a systemic autoinflammatory disease, autoimmunity and inflammation can affect multiple organs like skin, the central nervous system, eyes, and more [84, 85]. Even environmentally induced changes to an organism often only show very local effects. Mouse models have in many cases shown tissue-specific responses when exposed to toxins like arsenic, lead, polluted air, and others [86].

A significant hurdle for systems medicine is the general use of non-causal disease definitions. As these classifications often predate modern genetic research, diseases are frequently named for observable symptoms (e.g., hypertension), the affected organ (e.g., hepatitis), an eponym (e.g., Parkinson's disease), or a geographical origin (e.g., Lyme disease), rather than relating to their underlying cause. This limited precision means a single disease name can describe a phenotype caused by multiple distinct molecular disorders, thereby lacking the necessary granularity for cause-driven research [87, 88]. Consequently, even modern classification systems like the International Classification of Diseases (ICD) often retain this symptomatic and organ-based structure to serve practical healthcare logistics, such as insurance billing and specialist referrals. Given the underlying structure of current healthcare systems that are not yet fulfilling the P4 goals, this way of working is fully understandable and in many cases wanted and even beneficial. Patients with similar diseases are aggregated with specialists which increases the chance of them having seen this or a very similar case throughout their professional life, in turn increasing the chance of successful treatment. So while healthcare institutions are sitting on a treasure trove of data, building on such non-mechanistic definitions limits the new insights possible to extract from

it, with regard to research of disease causes.

Alongside the non-mechanistical definitions of diseases, systems medicine is facing additional data-related challenges. More and more medical and omics data is generated and subsequently new and larger datasets become available. Omics is the umbrella term for any analysis including results of our genome sequence, like genomics, proteomics, metabolomics, and others. Harmonizing, aggregating, connecting, or integrating all that data is not straightforward, as e.g. each genomics and proteomics dataset individually might require different sorts of pre-processing like batch effect corrections, based on the used method of data generation [31, 89]. Further, there is no clear consensus of how to name certain conditions or which disease classification system of the plethora of available ones to use, yet the choice can have a large impact on the quality of the integrated information and details can be lost in noise upon translation e.g. from ICD to the more mechanistic OMIM [90]. This makes it hard to retrieve conditions-specific case and control samples in sufficient quality allowing to differentiate between the foreground and background signal in the data and derive a mechanistic description of the condition or identifying several mechanistic subtypes having the same phenotype altogether.

In summary, when researching diseases on an omics level, like systems medicine is attempting, many of these existing disease definitions can contribute to diluting any available signal in the data [90]. Additionally, gaps in the causal understanding or mechanistic definition of diseases leads to the observed high NNT and low effectiveness in clinical treatment. This can be due to either lacking the molecular or genetic understanding of the disorder entirely or due to the current understanding being a combination of molecular causes of multiple subtypes and thus lacking precision [91, 92]. The lack of sufficient understanding is exemplified in diseases like type 1 diabetes, where the known genetic foundation cannot fully explain observed heritability patterns [93, 94]. Through large-scale genome-wide association study (GWAS) studies thousands of mutations can be screened and evaluated for their association, yet knowledge gaps remain. It is intuitive to assume that complex diseases can arise from a very specific combination of mutations. In fact, this effect has been called epistasis since many years and it assumes that a specific set of mutations has to co-occur to cause a specific phenotype and explain its heritability patterns, with only one of the mutations not present destroying most of the signal again [95, 96]. The hope exists that epistatic effects can explain complex diseases, shining a new light onto the cause and potential treatments, but to this day detecting epistatic effects is a very complex problem. On the other end of disease complexity, especially in rare diseases with very distinct genetic mutations a high agreement between symptom-based diagnosis and genetics-based stratification can be observed. Yet even here, the big hope that system medicine brings is that with more mechanistic definitions of diseases medicine could move to more precise ways of identifying treatment options and thus heavily reducing the NNT [92].

Even though systems medicine is pushing a more systemic view on organisms, the scope matters when looking for specific molecular explanations. A tissue-level examination of a disease can still be preferable to the system-level, as tissue, cell type, or even cellular compartment specific processes dictate the actual molecular processes taking place. Further, not all regions in our body can be addressed in the same way by drugs, as for example the blood brain barrier in a healthy individual only allows certain molecules to pass into the brain [97]. These layers of complexity necessitate a computational framework that can abstract these intricate relationships into a tractable and analyzable model.

1.3.2 *Network Medicine*

Network medicine resembles a computational application of systems medicine and aims to mitigate data integration challenges. Molecular biological or clinical data are abstracted into simple relationships between entities, like disorders, genes, proteins, and drugs. The relationships are modeled as links (edges) between different entities (nodes). Edges can hold specific information about the link, whereas nodes describe the specific entity.

The foundational concept of network medicine was introduced by Goh *et al.* [98] with the creation of the diseasome, a comprehensive map of human disorders and their shared genetic roots. This pioneering work mapped the relationships between diseases by linking those that have common genetic associations. The resulting human disease network revealed not only expected clusters of similar diseases but, more significantly, uncovered surprising connections between disorders that appeared symptomatically unrelated. This suggested that seemingly distinct and generally unrelated conditions could share underlying genetic or mechanistic pathways. The human disease network demonstrated that a network perspective could generate novel hypotheses about disease relationships, establishing the core principles of the emerging field of network medicine [99, 100].

Using this approach, integrating extensive available knowledge into networks such as interactions between proteins, genes that encode for proteins, drugs that target proteins, genes that are associated with diseases, and drugs that are indicated for diseases allows us to explore the drug-disease space. Particularly, discovering transitive links between drugs and disorders through shared genes can highlight potential causal drug candidates.

Essential for network medicine is thus the integration of diverse data types into a single network, which comes with fewer problems than integrating the original experimental data all together, yet is still impacted by the previously discussed disagreement between many common disease classification systems [90].

The ultimate potential of network medicine lies in its ability to construct an abstracted, simplified model of the complex biological systems that run life. This abstraction is not intended to be a perfect representation of molecular reality. Its strength lies in representing relevant information with sufficient fidelity to make complex systems tractable. The primary goal is to generate novel, testable hypotheses by identifying meaningful patterns and connections within this network context. This means that it is crucial to recognize, however, that these *in silico* findings are not conclusions in themselves. Their interpretation and the investigation of their biological meaning, cause, and impact requires subsequent and rigorous validation using other appropriately selected methods, ultimately including experimental verification.

1.3.2.1 *Biological Networks*

As biological networks form the basis of network medicine and the work presented in this thesis, understanding how biological networks are constructed and what information they can represent is fundamental.

Network structures, regardless of the field of application, can be formally described as a graph

$$G = (V, E)$$

, where V is a set of nodes, also known as vertices, describing all the different entities or objects in a network and E is a set of edges connecting the nodes.

Networks are usually referred to as *homogeneous networks* when they only have one type of nodes and edges. In contrast, *heterogeneous networks* (or *non-homogeneous networks*) are those where V can be partitioned into distinct classes or types of nodes [101]. Each type usually comes with a distinct set of attributes. For example, the starting network of the diseasome project was heterogeneous, consisting of disorder and gene node types. The final projection, the human disease network itself, is a homogeneous network as only disorder nodes and the edges between them remain.

In most cases where biological networks are used to extract new knowledge heterogeneous networks are used. In modern bioinformatics, complex, multi-relational heterogeneous networks are often referred to as *biomedical knowledge graph (KG)*. Within the scope of this thesis, the terms “heterogeneous network” and “knowledge graph” will be used interchangeably. The individual layers that make up these KGs can be broken down into two types of nodes and one connecting edge type. The information layer used for the diseasome was merely the disease-gene association layer, linking disorders and genes with an edge representing information about known association. The following nodes and information layers are the most relevant ones for this thesis, but naturally many more exist:

Nodes:

- **Diseases:** A disease (or disorder) node describes a unique pathological condition, often based on some existing disease classification system.
- **Genes:** Gene nodes depict the approximately 20,000—25,000 units of genetic information that encode human proteins. Mutations in these genes can therefore alter the resulting protein structure or affect its expression levels.
- **Proteins:** Proteins are the functional machinery of the cell and are associated with specific functions under particular conditions. Critically, they are also the direct molecular targets for the vast majority of therapeutic drugs, forming the primary bridge between pharmacology and cellular function.
- **Drugs:** Drug nodes represent pharmacologically active molecules that exert an effect on the human system, typically by acting on proteins to modulate their function.

Edges:

- **Protein-Protein Interactions (PPI):** A PPI represents a physical or functional interaction between two proteins, thus also often called the interactome. The definition of an interaction can range from a stable protein complex to a transient association, and the evidence can vary from experimental validation to computational prediction. This definition-level ambiguity introduces uncertainty into other networks upon integration, with profound implications for analysis. Further, PPIs evidence-basis can be split into the direct, physical binding events between two proteins (often detected by methods like yeast two-hybrid), and co-complex interactions, which includes proteins that are stable members of the same multi-protein

complex (often detected by affinity purification-mass spectrometry). Thus, when analyzing connected components in the interactome the interpretation depends heavily on which types of interactions are being considered. This issue will be discussed further in the context of source databases (Section 1.3.2.2). While the term interactome can include not only PPIs but also interactions of proteins with, e.g. ligands and small molecules, the terms interactome and PPI are used interchangeably in this thesis.

- **Drug-Target Interactions:** A drug-target interaction describes which drug acts on which protein. This can include details about the mode of action, such as inhibition or activation.
- **Disease-Gene Associations:** A disease-gene association refers to a relationship between a gene and a disease, implying that alterations in the sequence of a gene or its regulation are linked to the observed phenotype classifying as the disorder. These associations often stem from GWAS, differential expression analysis, or other omics-driven methods. It is crucial to acknowledge that these associations can carry varying levels of evidence and potential biases. For instance, associations from GWAS often identify statistically linked variants, while causal genes and mechanisms remain unknown. Associations from differential expression analysis are correlational and may reflect downstream effects rather than primary causes.
- **Disease-Drug Indications:** A drug-disease indication represents a known, investigational, or approved treatment option, where an edge signals that a drug has a beneficial effect on a disease. Knowledge sources are not exclusively clinical research but can include off-label use reports.
- **Gene-Protein Coding:** This information layer links genes to the proteins they encode. It is fundamental for integrating disparate data types, for example, to connect a protein-based drug-target interaction layer with a gene-based disease association layer. For simplicity, this translation will often be handled implicitly in this thesis.

As mentioned, nodes and edges can have attributes that store additional information about the specifics of an entity or a relationship, respectively. These attributes can be numeric, for example, to store weights representing binding affinities in a drug-target interaction layer or quantifiable effects on expression, which allows for rapid searching and ranking based on specific properties.

While this work focuses mainly on non-weighted edges and considers edge directionality only implicitly, both represent important properties that edges can encode. In gene regulatory networks, for example, the directionality of an edge is critical as it represents that one entity regulates another, a relationship that does not necessarily hold true for the inverse direction. Binding affinities of drugs to their targets or gene and protein expression information in diseased cases could be represented as edge weights rather than edge properties and exploited further.

1.3.2.2 Databases and Datasets

Numerous databases exist that contain the knowledge relevant for the nodes and edges described previously. These resources can be broadly categorized by their level of data aggregation. Primary data sources consist of scientific publications and raw experimental data, which form the basis for the information found in secondary data sources. Tertiary data sources, in turn, can be classified as aggregating information from a mix of primary and secondary sources. This section will focus on prominent secondary and tertiary databases, first examining those that provide specific

information layers, such as node definitions or edge types, before discussing clearly tertiary, integrated, multi-layer KGs in a subsequent subsection.

It is critical to consider the definitions and evidence criteria upon which a database builds its information collection process, as the meaning and reliability of the extracted data can change significantly from one source to another. For instance, the definition of a PPI can differ dramatically. A database like BioGRID [102], which aggregates experimentally verified interactions from individual publications, provides a different view of the interactome than STRING [103], which integrates experimental data with computationally predicted interactions and text-mining evidence. This distinction is not trivial; it has fundamental effects on the network's topology and the biological interpretation of any analysis performed on it. This variance must be taken into account when merging information layers into a large heterogeneous biological network and when interpreting the knowledge extracted from it.

Nodes

Selecting databases for node information is often guided by a central consideration: the choice of the main identifier (ID) space for each node type. This choice is a foundational methodological decision, particularly for diseases, as different disease categorization systems were built for different purposes. Examples of commonly used sources and ID spaces can be found in Table 1.1. The ICD, for example, is an administrative classification optimized for billing and epidemiology, often grouping diseases by clinical presentation rather than molecular cause. In contrast, OMIM provides a mechanistically oriented catalog of genetic disorders but has limited coverage of complex diseases. This creates a significant challenge for data integration. The implications of this decision can range from small, as in the case of drug nodes, where mapping between ID spaces of different drug ID spaces is close to 1:1, to extremely impactful on downstream results, as it is the case for disease nodes. This is due to the different ways diseases can be defined or classified, making some disease classification systems incompatible regarding how much information is lost when translating from one ID space to another, which often follows an n:m mapping scheme [90]. A selection of relevant resources providing information for different node types within network medicine can be found in Table 1.1.

Table 1.1: Common resources for nodes in network medicine; lexicographically ordered by database name.

Database	Node Type	Description
ChEMBL [104]	Drug	ChEMBL stores information about small bioactive molecules including structure and bioactivity information.
Disease Ontology [105]	Disorder	Disease Ontology (DO) aims to provide a semantics-based, well-defined ontology of diseases for biomedical data integration while linking to other existing disease classification systems.
DrugBank [106]	Drug	A comprehensive database combining detailed drug data with drug target, action, and interaction information.

Continued on next page...

Table 1.1 continued from previous page

Database	Node Type	Description
DrugCentral [107]	Drug	A resource collecting drug information from regulatory agencies such as the FDA, European Medicines Agency (EMA), and Pharmaceuticals and Medical Devices Agency (PMDA) of Japan
Ensembl [108]	Gene, Protein	A genome platform providing annotation for genes, proteins, and transcripts across multiple species.
ICD [109]	Disorder	A WHO-coordinated classification system used globally for health statistics, epidemiology, and clinical reimbursement.
MONDO [110]	Disorder	Mondo Disease Ontology: An ontology that harmonizes multiple disease definitions by creating a coherent, logical structure of disease relationships.
NCBI [111]	Gene	A comprehensive public repository for gene-specific information, including sequence, function, and variation data.
OMIM [112]	Disorder	Online Mendelian Inheritance in Man: A comprehensive, curated catalog of human genes and genetic disorders, focusing on Mendelian inheritance.
SNOMED CT [113]	Disorder	A comprehensive, multilingual clinical terminology designed for standardized use in electronic health records (EHRs).
UMLS [114]	Disorder	Unified Medical Language System: A metathesaurus of clinical and biomedical terminologies designed to enable interoperability between systems.
UniProt [115]	Protein	A comprehensive, high-quality resource for protein sequence and functional information.

Edges

As edges represent the relationships that enable network medicine, the choice of their source has a profound impact on any subsequent analysis. Differences can be a matter of completeness. DrugBank, for instance, contains more extensive drug-target interaction data than DrugCentral. They can also be a matter of scope, as DisGeNET provides genetic associations for a broader range of complex diseases than the Mendelian-focused OMIM. Furthermore, licensing models can play a role; the more comprehensive nature of DrugBank and DisGeNET is supported by commercial licenses, although both offer free academic access. The most significant differences, however, are often found in the evidence criteria for an interaction. For PPIs, the source of evidence can range from computational prediction and text-mining to direct experimental validation. This variety in definition leads to immense differences in the resulting PPI networks, affecting both their analytical potential and inherent noise levels. A more extensive list of some commonly used sources and differences between their origin of evidence for a relationship can be found in Table 1.2. This underscores the importance of understanding the data source, as the validity of conclusions drawn from the network depends on the nature of the evidence it represents.

To clarify the standardized terms for the Main Evidence Types (ordered by occurrence) used in Table 1.2, the following definitions apply:

- **Integrated:** Indicates that the database combines data from multiple sources and/or of multiple evidence types into a cohesive resource. The specific types or methods are listed in parentheses.
- **Aggregation:** Data collected and compiled from other secondary databases, no large transformational changes to data applied.
- **Literature Curation:** Data manually extracted from peer-reviewed scientific literature (primary sources) by expert curators.
- **Experimental:** Data derived directly from laboratory experiments (primary sources e.g., yeast two-hybrid, affinity purification-mass spectrometry).
- **Prediction:** Data generated by computational algorithms that predict relationships *de novo* (e.g., based on protein structure, domain interactions, or sequence homology).
- **Text-mining:** Data automatically extracted from the text of scientific publications using natural language processing techniques.
- **Inference:** Data derived through logical deduction from other data types, such as inferring interactions via orthology mapping between species.
- **Regulatory:** Information sourced from official regulatory agency filings (e.g., from the FDA, EMA, or PMDA). Special case of *Aggregation*.
- **Sequence Data:** Foundational data from genomic or proteomic sequencing efforts.

Table 1.2: Common resources for edges/relationships in network medicine; lexicographically sorted by database name.

Database	Edge Types	Description	Main Evidence Type(s)
APID [116]	Protein-Protein-Interaction	Provides PPI networks for over 400 species, focusing on experimentally validated physical interactions.	Integrated (Aggregation, Experimental)
BioGRID [102]	Protein-Protein-Interaction	A repository of protein and genetic interactions for all major model organisms, offering datasets with varying levels of validation.	Integrated (Aggregation, Literature Curation)
ChEMBL [104]	Drug-Target-Interaction	A database of bioactive molecules with drug-like properties, linking them to their genomic and proteomic targets.	Literature Curation

Continued on next page...

Table 1.2 continued from previous page

Database	Edge Types	Description	Main Evidence Type(s)
CTD [117]	Drug-Disease Indication, Disease-Gene Associations	The comparative toxicogenomics database: Stores data on chemical-gene, chemical-disease, and gene-disease relationships, using MeSH and OMIM identifiers.	Integrated (Literature Curation, Aggregation, Inference)
DGIdb [118]	Drug-Target Interaction	The Drug-Gene Interaction Database aggregates drug-gene interactions from multiple curated sources.	Integrated (Literature Curation, Aggregation)
DISEASES [119]	Disease-Gene Associations	A daily-updating resource for disease-gene associations, with confidence scores assigned to each relationship.	Integrated (Text-mining, Literature Curation, Experimental, Aggregation)
DisGeNET [120]	Disease-Gene Associations	A comprehensive discovery platform for exploring the genetic basis of human diseases and their associated variants.	Integrated (Aggregation, Literature Curation, Experimental, Inference)
DrugBank [106]	Drug-Target Interaction, Drug-Disease Indication	A comprehensive database combining detailed drug data with drug target, action, and interaction information.	Integrated (Experimental, Literature Curation, Regulatory)
DrugCentral [107]	Drug-Target Interaction, Drug-Disease Indication	A resource for drug information that also includes data on veterinary applications.	Integrated (Aggregation, Regulatory, Inference)
IID [121]	Protein-Protein Interaction	Integrated Interactions Database contains human and model organism interactions.	Integrated (Aggregation, Experimental, Prediction)
IntAct [122]	Protein-Protein Interaction	An open-source database and analysis toolkit for molecular interaction data partially driven by user-based data submission.	Integrated (Aggregation, Experimental, Literature Curation, Inference)

Continued on next page...

Table 1.2 continued from previous page

Database	Edge Types	Description	Main Evidence Type(s)
OMIM [112]	Disease-Gene Association	Online Mendelian Inheritance in Man: A catalog of human genes and genetic disorders, with a focus on the relationship between genotype and phenotype.	Literature Curation
STRING [103]	Protein-Protein Interaction	A database of known and predicted PPIs for thousands of organisms.	Integrated (Aggregation, Experimental, Text-mining, Prediction)
UniProtKB [115]	Gene-Protein Coding	A comprehensive resource for protein sequence and functional information, providing the mapping from genes to proteins.	Integrated (Sequence Data)

1.3.2.3 Integrated Network Databases for Drug Repurposing

Large integrated network databases represent a key class of tertiary data sources for network medicine. They combine a multitude of node and edge types by aggregating information from tens or even hundreds of mostly secondary data sources. For such a resource to be useful in DR, it must contain drug and disease nodes and provide information layers that connect them. The nature of the intermediate nodes and edges along these connection paths forms the basis from which mechanistic hypotheses for DR can be derived. While frameworks like BioCypher make it easier than ever to combine the previously discussed data sources for nodes and edges into a custom KG by reusing community developed adapters, actively maintaining and updating one still remains a larger task [123]. The following section highlights some of the most prominent integrated databases relevant for DR.

OmniPath

OmniPath is an extensive resource that integrates information from more than 100 sources, providing a comprehensive collection of protein interactions, gene regulatory networks, and drug-target interactions [124]. Monthly updates from all resources ensure currency of information. OmniPath allows users to explore this data in tabular format via a web interface¹, provides a Cytoscape plugin and for programmatic access Python, and R-lang libraries.

Het.io

While OmniPath provides an exhaustive repository of signaling pathways ideal for deep mechanistic investigation, the Het.io KG was purpose-built for large-scale, systematic evaluation of drug-disease pairs [125]. Het.io integrates high-quality information from 29 databases, covering 11 node and 24

1 <https://omnipathdb.org>

edge types. Its web interface² and Neo4j graph browser³ are designed for path-based queries and network queries, respectively. Multiple Python libraries exist for programmatic access. Originating from Project Rephetio, this KG enabled the systematic computation of treatment probabilities for over 200,000 drug-disease pairs [125].

OpenTargets

Backed by EMBL-EBI, Pfizer, and other big organizations, the OpenTargets platform serves the primary purpose of identifying and validating novel therapeutic targets [126]. Integrating over 40 data sources, its web user interface (UI)⁴, allows researchers to start with a disorder, identify its most strongly associated genes, and then explore known drugs that target those gene products. Programmatic access is provided via a GraphQL API.

NeDRex

Similarly, the NeDRex platform⁵ and its underlying database, NeDRexDB, were designed with a specific DR focus as part of the REPO-TRIAL [127] and later REPO4EU [128] project [129]. This weekly-updating resource, which is co-maintained by the author of this thesis, has grown from 10 databases to 28, now comprising 11 node and 19 edge types. It provides multiple access points, including a Neo4j browser for direct graph queries, a documented API⁶ for computational experts, and a Cytoscape plugin for users preferring a graphical interface. The NeDRex platform is distinguished by its integration of drug-repurposing-specific algorithms, which are accessible through the API and have been used to identify novel repurposing candidates [129].

While all these resources provide invaluable data repositories (and sometimes DR specific features) their practical application in DR research is often limited by a reliance on generic interfaces (e.g., a Neo4j browser) or expert-oriented standalone software (e.g., Cytoscape) requiring installation and configuration. These resources often lack either the specific functionalities tailored for repurposing workflows or the intuitive, web-native UIs needed to lower the barrier to entry. Consequently, a gap remains between the availability of integrated data and its accessibility to the domain experts, such as clinicians and pharmacologists, who are best positioned to guide the discovery process. This highlights a critical need for purpose-built, fully online tools integrating network visualization and analysis, that are accessible to a broad scientific audience, and explicitly support DR research.

1.3.3 *Network Medicine-based Drug Repurposing*

The main goal in DR is to find relevant links between existing drugs and diseases (Figure 1.1(A)). Following the systems medicine approach, this requires identifying the systemic or mechanistic causes of a disease, such as genetic variants, and the systemic targets of available drugs, typically

² <https://het.io>

³ <https://neo4j.het.io>

⁴ <https://platform.opentargets.org>

⁵ <https://nedrex.net>

⁶ <https://api.nedrex.net>

proteins (Figure 1.1(B)). Drugs and diseases can be connected by their shared underlying genetic mechanisms when integrating it with the information that proteins are encoded by genes. In network medicine, this is translated into KGs containing drug-(protein-)target interactions and disease-gene associations. Analyzing, direct causal relationships, such as a single drug targeting a single protein encoded by a single mutated gene, causing a single phenotype, represent the simplest cases and thus a logical consideration, those however are already well-studied [31].

A feasible starting point for novel DR hypothesis generation is the analysis of comorbidity data, which describes the statistically significant co-occurrence of two or more diseases in the same patient on a population level [130–132]. Such co-occurrence may suggest that one disease is a risk factor for the other, or that both share a common underlying cause [133]. With this information, known genetic associations between comorbid disorders can be included as well as drugs that are known to treat the disorders. This allows for the generation of hypotheses where drugs known to treat one disorder might be transferable to a comorbid one [132, 134, 135]. In practice, however, this approach is often limited by the nature of the available data. Comorbidity information is typically derived from clinical records using classification systems like the ICD, making it difficult to integrate with molecular data (e.g., disease-gene associations) on a large scale without losing significant signal and detail [90].

For polygenic and other complex, multifactorial diseases, the challenge is greater. These conditions often arise not from single mutations, but from the impairment of multiple or entire molecular pathways [31]. In these cases expanding the molecular background of drug-protein-gene-disorder connections to drug-protein-protein-gene-disorder paths by adding a PPI layer, interaction and regulatory effects on the omics level are better represented. While one could use the PPI to find transitive links between drugs and diseases (similar to the human disease network), simply projecting all possible connections introduces significant noise. Differentiating mechanistically relevant combinations from the vast number of irrelevant or even false positive paths becomes a serious challenge.

This challenge warrants the introduction of the idea of disease modules. This core concept of network medicine states that the proteins associated with a specific disease are not randomly scattered across the interactome. Instead, they tend to cluster in a localized, interconnected neighborhood, forming a disease module that encapsulates the molecular mechanisms driving the disorder [136, 137]. A perturbed disease module can manifest as the disease phenotype. Thus, different perturbations to the same module or a change in environmental factors can lead to varied phenotypes, even if the underlying molecular cause, and as such the potential molecular target for effective treatment, remain the same [138, 139].

Crucially, this hypothesis provides a reasonable molecular explanation for the low effectiveness and high NNT observed in clinical practice. When two molecularly distinct diseases contribute to, or produce a similar or identical phenotype, they still may be driven by perturbations in entirely separate disease modules. Treatment targeting one module would be ineffective for the other, even if the symptoms appear identical. Prescribing medication based on phenotype alone would inevitably lead to treatment failure in one of these patient groups, thus elevating the NNT.

Type 1 diabetes and maturity-onset diabetes of the young (MODY) pose such an example, where only in 1974 it was identified that the same diabetic symptoms are caused by two distinct disorders [140]. In 1990 the causal difference was identified. T1D is caused by an autoimmune reaction destroying insulin producing beta cells and MODY arises from a mutated receptor gene responsible

for the regulation of insulin release. Whereas T1D can be treated giving insulin injections, MODY can be handled by using sulfonylureas [140]. Even worse, when misclassifying an infant with Dravet Syndrome as standard epilepsy, standard sodium channel blocker treatment leads to an increase in seizures, rather than a reduction [141]. This is caused by a mutation in *SCN1A* responsible for sodium channels in inhibitory neurons, leading to a dysregulation of the nervous system that can be further amplified by mentioned treatment. This reason was only understood after uncovering the genetic relationships [141, 142].

Therefore, a primary focus of network-driven DR is the identification of disease modules to uncover and understand disease-driving mechanisms [143]. The logical next step is to identify drugs that specifically target these modules, creating a mechanistic link between a compound and a therapeutic indication (Figure 1.1(C)). This task, however, is complicated by the fact that the precise molecular mode of action for some approved drugs remains poorly understood [144]. To compensate for gaps in knowledge and to explore alternative or indirect mechanistic relationships, this also warrants the integration of a PPI layer.

The potential for DR is immense, especially once an accurate disease mechanism is identified. Mechanism-related units can be described as disease modules. Specific techniques for their identification will be discussed in Section 1.4.3.2 in detail, but range from manual curation to the use of purpose-built algorithms, on top of existing knowledge or data. Once a disease mechanism is clear, individual drugs targeting the module can be investigated (Figure 1.1(D)). Alternatively, drug combination approaches or polypharmacology or polytarget can be employed, using a combination of drugs or one single multi-target one, respectively, to target the mechanism from multiple angles simultaneously [145, 146]. This can allow for lower individual drug dosages, potentially limiting adverse side- and off-target effects while achieving a combined therapeutic impact on the condition of interest.

1.4 POTENTIAL AND RESEARCH GAP IN NETWORK-BASED REPURPOSING

In the following, specific implementations of DR are discussed, to assess the state of the art and highlight their limitations. Based on this, the motivation of this work will cover how the limitations of current processes can be addressed and how this thesis contributes to that.

1.4.1 Drug Repurposing - The State of the Art

As introduced earlier, until recently DR research has been mostly serendipitous. This section aims to give an outline of the current systematic and deliberate computational methods in this field while highlighting the ongoing limitations.

DR promises to address many current issues of the drug development system. Diverse computational approaches for the targeted identification of DR candidates exist [147, 148]. The approaches mainly differ in their starting point. Disease-centric DR efforts focus on identifying repurposing candidates starting at a specific disease, whereas drug-centric approaches try to identify new indications for a specific drug. There are also target-centric approaches that identify potentially addressable disorders and usable drugs given promising molecular targets. Again

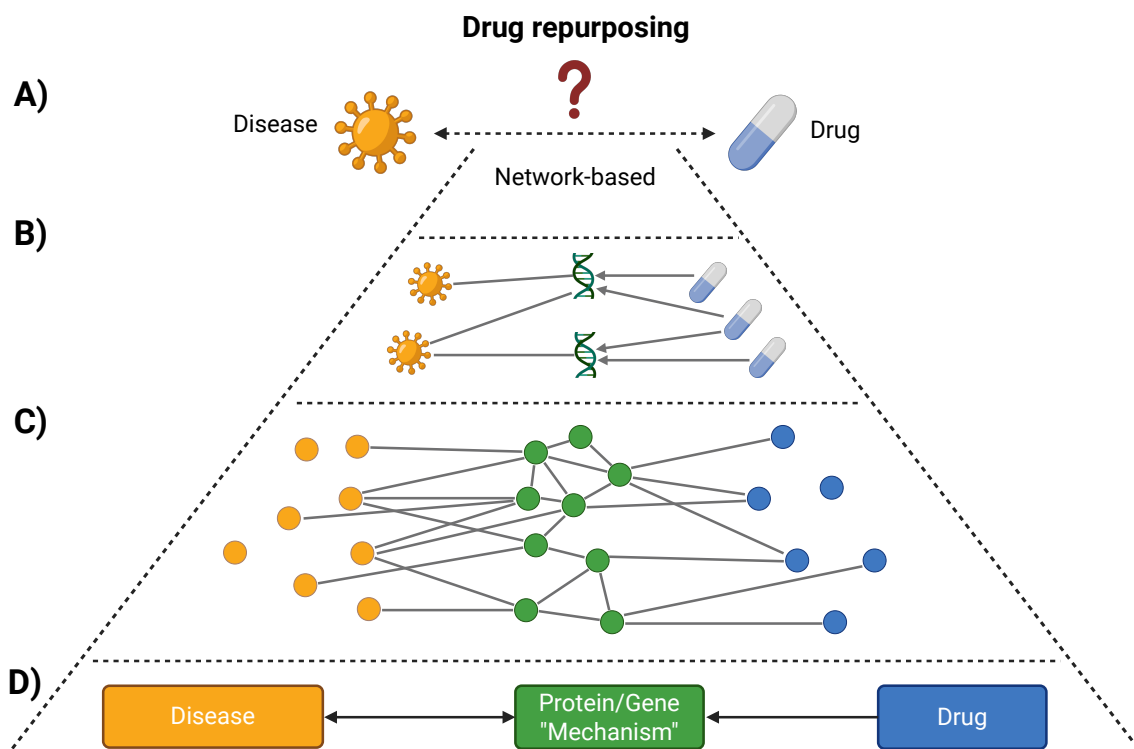


Figure 1.1: The general goal of drug development is the identification of a drug that can treat a certain disease (A). In DR, only drugs and pharmacologically active components, that are already known, are screened to identify such links. (B) In a network-based setting, the disease-gene association layer and the drug-target interaction layer can be combined, to identify shared genetic links between existing drugs and known disorders. (C) Multiple reasons limit the power of direct disease-gene-drug link analysis. Firstly, most, if not all of them, are already well researched. Secondly, traditional disease definitions do not fit to the systems medicine approach and the genetic background might not be fully known, both leading to inadequate disease genes. Using the PPI layer in between drugs and diseases can help to buffer these issues and identify the disease modules that define the mechanistic basis of the disorder which can be targeted by one or multiple drugs (D).

others are direction agnostic and, for example, assess the relevance of a given drug-disease pair.

In all cases, targeted and mechanism-driven DR is achieved by integrating molecular data of diseases and drugs into research pipelines. Advances in omics technologies enable systematic discovery of molecular mechanisms driving phenotypes as well as drug targets and responses. While DR approaches can be differentiated by the underlying data they build on, overlaps can be large and in the following approaches will instead be roughly classified based on different underlying methodologies [147–149]. The most prominent approaches are based on signatures, text-mining, artificial intelligence (AI) (including machine learning (ML)), docking, and knowledge (including networks). A summary overview is provided in Table 1.3.

Docking-based approaches belong to the first directed methods employed but are still proving successful to this day. Sometimes combined with other approaches, docking-based methods are driven by molecular structure of drugs and targets [150, 151]. The goal is the identification of chemical compounds that fit a known, specific binding pocket, often with the aim of a higher binding affinity than an unwanted-interaction that is to be prevented [150]. This is often highly relevant for pandemic preparedness, as the host proteins the virus uses in its replication cycle can

be identified fairly quickly and temporary allosteric inhibition of certain targets can limit the viruses replication abilities [150, 152]. DRDOCK [153] and ACID [154] are former, and MTIOpenScreen [155] and DrugRep [156] are currently available online services, providing access to DR focused resources, lowering the level of entry for virologists and clinical experts once a target or ligand of interest is identified. Natural difficulties in this field include proteins with intrinsically or highly unstructured regions, making the docking process a moving target.

Signature-based approaches are omics-driven and try to match molecular signatures of disease and drugs [157]. Most commonly, differential gene expression informs about changes in expression in a disorder. Resources such as CMAP or LINCS store information about expression changes in cell lines, given chemical compounds; drugs with a complementary signature can be identified from those resources [158]. The hypothesis is that a chemical component that reverses the signature of a disorder has a chance of counteracting its effects [157, 159]. Integrated online solutions for signature-based DR seem rare but L1000CDS2 allows the upload of up- and down-regulated genes and some parametrization to search for relevant compounds [160]. While cell lines can give great insight into which compounds could help to address certain conditions, problems like drug delivery limit seamless translation from model to organism level.

Text-based approaches attempt to extract and recombine existing knowledge from scientific literature [161, 162]. The intuition is that not all available knowledge and hypotheses might be covered in structured databases yet and thus using information directly from publications can help generate novel DR hypotheses or angles [163, 164]. Resources like STRING for PPI and DISEASES for disease-related information, fit this category, as they partially build on information generated from text-mining efforts. Those resources usually employ some sort of statistical measures like CoCoScore to assess the accuracy and relevancy of extracted information [165]. With recent advancements of large language models (LLMs), this statistical approach might see some change and the idea of text-mining-based knowledge extraction, recombination, and hypothesis generation is more feasible than ever [166]. Additionally, LLMs have the power to raise text-based information from serving as a knowledge source to an accessible hypothesis generation machine. This works already in a zero-shot fashion (without providing additional resources) by just interrogating LLMs regarding their trained-in knowledge that they can recombine on demand [167]. Without additional evaluation, grounding, reflection, and validation the validity of predicted results can be rather low, as hallucinations and inaccuracies are still a general issue in modern LLMs [168, 169]. A published example is DrugReAlign, making DR information extraction from data accessible through natural language using an LLM chat bot [170].

Knowledge-based approaches use curated information about drugs, diseases, and molecular pathways, typically representing population-level knowledge rather than individual-level omics data. Knowledge can be organized using graph structures, following the network medicine idea, making it KG- or **network-based** DR [129, 171–173]. The quality but especially meaning of the content of a KG heavily rely on the integration process and the integrated sources; too loose filtering criteria can lead to noisy networks; too rigid ones can limit the potential for new findings [174]. Additionally, the exact meaning and definition of “proximity” or “closeness” within a network and especially what proximity in a network means for biological components is not well defined and heavily context dependent. Protein-target and protein-protein information layers have a different meaning of distance than for example metabolic network information or co-expression information

layers, while all of them could be integrated into one large knowledge graph [175–177]. Specialized network algorithms, often combined with manual refinement based on expert knowledge, are used to identify the mechanistic foundations of diseases and potential DR candidates [129, 178]. Web applications, like BalestraWeb [179], the previously mentioned Het.io [125], OpenTargets [126], and others provide mainly searchable graph databases, without any additional DR focused features. Projects like ParmOmics on the other hand combine signature and network-based methods, to provide a dedicated resource for DR hypothesis generation [180].

AI and ML approaches, as in many other medical fields, are also applied to predict DR candidates [181, 182]. These models are trained on any raw or preprocessed data type –including patient-specific omics, population-level KGs, and full-text information from sources like EHRs or a combination thereof– to maximize predictive power [183–185]. While AI-driven approaches are incredibly promising due to the recent advances in foundation models, the quality of an AI model is heavily correlated with the quality of training data; if the data contains biases, the model will learn them. In addition, prediction and model explainability and the interpretability of results often lack behind. Nonetheless, this challenge is increasingly being addressed [181, 185–187]. Accessible AI-supported DR solutions appear to be uncommon, which potentially stems from the necessity of powerful GPUs to run large AI models. Still, projects like RepurposeDrugs include their ML-based predictions into a searchable web application [188]. The field of graph neural network (GNN) specifically focuses on training AI models on data organized in graph-structure. These methods form a future-oriented approach to network medicine and are already in use [185, 189, 190]. Using trained GNNs requires the model weights and architecture to be published and adequate computing hardware to run the model, overall requiring decent domain knowledge.

Table 1.3: Summary of different methodology classes for targeted DR including their core principles, essential data, and key challenges.

Methodology	Core Principle	Primary Data Inputs	Key Challenges	Representative Example(s)
Docking	Structure-based binding/affinity prediction	3D protein/ligand structures	Target flexibility, scoring function accuracy	MTIOpenScreen [155], DrugRep [156]
Signatures	Reversing disease-associated expression patterns	Transcriptomics (e.g., LINCS L1000)	Cell line-to-patient translation, context-specificity	L1000CDS2 [160]
Text-based (Mining & LLM)	Knowledge extraction, synthesis, and recombination of unstructured text	Scientific literature, clinical notes	Semantic ambiguity, factual grounding, hallucinations	DrugReAlign [170]

Continued on next page...

Table 1.3 continued from previous page

Methodology	Core Principle	Primary Data Inputs	Key Challenges	Representative Example(s)
AI/ML	Complex pattern recognition and link prediction	Heterogeneous data (omics, KGs, EHRs)	Interpretability (“black box”), data quality/bias	TxGNN [185], DeepLPI [191]
Network-based	Proximity and topology in integrated biological networks	Knowledge Graphs (PPIs, drug–target interactions, gene–disease associations)	Network incompleteness and noise, defining “proximity”	NeDRex [129], Het.io [125]

1.4.2 Advantages and Potential of Network-based Methods

Network-based DR generally extracts knowledge from KGs. These KGs can be designed, structured, and filtered according to specific research needs. Generally given a disease of interest, its associated genes are identified, translated into protein space, disease modules are identified, DR candidates prioritized, and evaluated adequately afterwards.

The main advantage over many other techniques, especially AI approaches for DR, is inherent interpretability and explainability. All intermediate results are simply expansions or reductions of the previous networks. Paths through the network from the initial disorder to the final drug candidates can help interpret its potential. Additionally, this method can easily follow an expert-in-the-loop paradigm, that allows expert knowledge to be infused into the process at any step, e.g. to add non-accounted for disease-related proteins, or removing irrelevant targets from the disease module, before drug prioritization [129].

Further, it effectively solves data integration issues, as any information can be reduced to nodes or edges and any additional information can be stored in attributes of these entities and used for manual or automated use in the DR workflow.

However, despite the promise of these network-based computational methods, a series of fundamental bottlenecks prevent their widespread adoption and translational success. These gaps exist not only at the level of specific scientific challenges but also in the overarching issues of usability and accessibility, creating a barrier between promising tools and the researchers who could benefit most. After introducing a conventional workflow for network-driven DR, potential solutions for these challenges will be discussed in Section 1.6.

1.4.3 The Network-driven Drug Repurposing Workflow

As previously discussed, knowledge- and network-driven methods are amongst the most versatile and interpretable DR options. Source information can come from any sort of data type and can be

represented in a network format. Here I will introduce the most crucial steps in a disease-centric, network-driven DR workflow (Figure 1.2). Namely, these are the disease gene selection, module detection, drug ranking, and the *in silico* validation [129]. Notably, every step comes with its own challenges with regard to information quality and uncertainty, which can propagate from step to step making *in silico* validation steps highly relevant [90, 192].

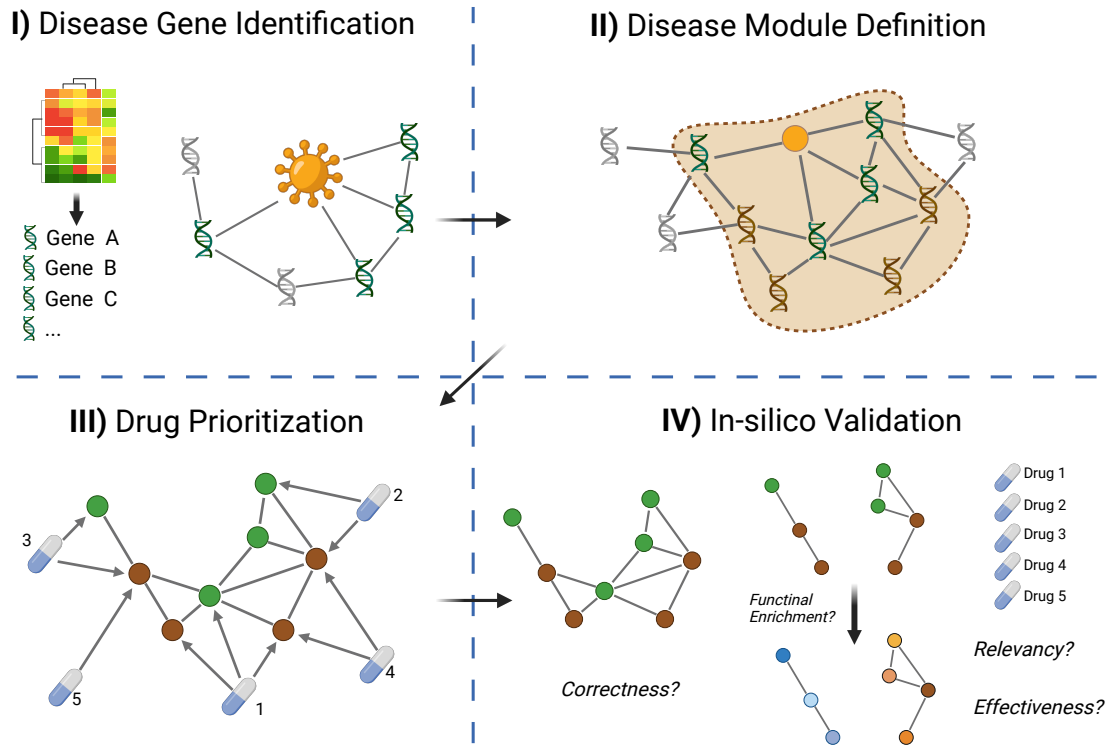


Figure 1.2: The network-based DR workflow commonly consists of four distinct steps: **I)** Disease associated genes are identified, e.g., through differential gene expression analysis, knowledge annotated in databases or knowledge graphs, expert knowledge, or any other way. **II)** The disease genes are taken as input for disease module detection, where specialized algorithms attempt to identify more genes within the same context. The goal is to include genes that either were missing from the initial gene set through methodological limitations or a lack of knowledge and to recruit additional potential drug targets indirectly affecting the disease genes. **III)** Given the disease module, drug prioritization algorithms create a ranked list of drug candidates. **IV)** In order to filter out unreliable predictions, *in silico* validation can be employed on modules and drug candidate results. Statistical and annotation-based approaches are the most common ones to estimate result quality, apart from analytical validation of the results by including external sources about disease genes and drug indications.

1.4.3.1 Disease Gene Selection

The first step in this specific pipeline is usually the identification of the genetic background given a certain condition (Figure 1.2(I)). This is often called *disease gene selection*, *disease gene identification*, or *seed gene identification*. Seed, in this context, refers to these disease genes serving as the starting point, so the seeds for subsequent steps in a DR workflow. In a systems medicine context, disease genes often come from population-level sources. A short summary of the different methodologies

will be presented in the following, accompanied by [Table 1.4](#).

The overall issue remains that many diseases are not understood on a mechanistic and genetic basis yet. A second, already highlighted issue is the lack of common disease definitions. This means that, depending on which of the disease-gene association datasets or databases is used, these relationships can be highly dissimilar, due to different sources and different definitions of the identically named disease resulting in highly diverse seed sets [90, 193].

One current solution to this problem is to work with integrated knowledge bases, being aware of the sources and the meaning of the information, or integrating multiple sources, to cover a wider field of disease genes, if necessary. Still, this can result in false positive (too many) or false negative (missing) disease genes in the seed set. These false hits can be accounted for and mitigated by the subsequent module identification step

A second solution is to base disease-associated and disease-relevant gene selection on patient-level or at least cohort-level experimental data [90]. For example, by selecting differentially expressed genes of clearly defined case-control cohorts, the initial list of disease associated genes has the potential to provide a very clear case-specific signal.

While both of these solutions exist in the form of well-used databases (presented earlier) and tools (like DESeq2 for differential expression analysis [194]), integrated solutions with the goal of DR are largely lacking.

Table 1.4: Overview of resources and methods for disease gene selection.

Tool / Resource	Category	Focus / Method
OMIM & GWAS Catalog	Databases	Curated records of established gene-disease associations and single-variant genetic studies.
DisGeNET & Open Targets	Integrated Knowledge Base	Aggregates information from multiple sources to provide a wider, weighted field of disease genes.
DESeq2	Experimental Analysis	Identifies differentially expressed genes from specific case-control cohorts to provide case-specific signals.

1.4.3.2 Disease Module Identification

Based on a selection of disease-defining genes, in a second step, a disease module (or mechanism) can be identified ([Figure 1.2\(II\)](#)). The goal is to contextualize the initially identified disease genes within the interactome, the PPI network, in order to determine how they collectively contribute to the specific disorder. Again, the source of the interactome has a large impact, potentially the largest among all workflow steps, as the definition of what classifies as relevant PPI differs heavily between

data sources [192]. Table 1.5 contains a summary of all the algorithms presented in the following.

Using purpose-built network medicine algorithms allows identifying additional impactful or relevant genes based on the connectivity to the original set of disease-describing genes. The idea is that the current disease understanding still has gaps, which can render the identified seed genes “incomplete” and insufficient for the task of identifying drug candidates. But the seeds can be used to paint the context and the presumed gaps can be filled by following the disease module hypothesis and identifying connected components adding key connections between the disease genes, to enrich this disease definition with high-likelihood candidates. And even in case these methods identify genes that have no direct impact on the disease through their interaction with the actual disease module, they might still serve as valuable treatment targets through their downstream effect in the molecular pathways of the disease.

As so far, there is no gold standard annotation of available, existing, or even validated disease modules, a wide array of available module detection algorithms exists. Many come with their individual, sound logic but produce widely different results.

DIAMOnD, the first module identification algorithm, for example, iteratively expands a selection of genes by identifying genes with the highest “connectivity significance”, factoring in the number of connections of a gene with the current module in comparison to any other gene in the interactome [195]. This usually results in one large connected mechanism. Contrary to that, DOMINO explicitly aims to locate and construct connected components, in which a subset of the seed genes is enriched, in many cases resulting in multiple disconnected disease mechanisms [196]. The multi-Steiner tree algorithm (MuST) creates minimal scoring trees connecting all seeds and collecting additional relevant ones on the way. Because the Steiner tree problem is NP-hard and multiple optimal solutions can exist especially in an unweighted network, a heuristic approach was taken to identify multiple sub-optimal trees, subsequently merged into one solution [197].

An effort for the identification of only the most significant, but not necessarily directly connected genes, was undertaken with ROBUST [198]. It is an evolution of the MuST algorithm, but prioritizes seed genes over interactome-derived ones. Since, an improved version, combatting research-bias introduced degree-bias in interactomes, has been released with “ROBUST bias-aware” [199]. In similar top-down fashion, TOPAS iteratively identifies a minimum amount of relevant connection genes between all seeds [200]. In contrast to ROBUST, this approach always results in a fully connected subnetwork.

Besides those network medicine specific algorithms, general graph algorithms are used in module detection research. This includes classical ranking and centrality measures like TrustRank [201], an improved version of Google’s PageRank algorithm [202], Harmonic Centrality [203], a variant of Closeness Centrality [204] that works on unconnected graphs, Betweenness Centrality [205], and more.

Two solutions that attempt to close the gap between disease gene selection and disease module creation are KeyPathwayMiner and BiCoN. KeyPathwayMiner accepts an indicator matrix, specifying which genes are active, e.g., derived from gene expression data, and uses this information to identify all maximally connected pathways containing at most a configurable number of non-active genes [206]. Similarly, BiCoN accepts raw gene expression data and identifies

subnetworks in the interactome that maximize the inter-patient-group expression difference [207]. This can, in cases where sufficient experimental data is present, help to reduce the uncertainty of disease definitions and aggregated information, discussed in the disease gene identification step [90].

Table 1.5: Overview of common methods to identify disease mechanisms, including disease module identification algorithms, graph algorithms, and integrative tools.

Tool	Methodological Category	Key Characteristic
DIAMOnD [195]	Local Expansion	Iteratively expands the module by adding genes with the highest “connectivity significance” to seeds.
DOMINO [196]	Subnetwork Construction	Constructs connected components enriched for seed genes, potentially yielding multiple disconnected modules.
MuST [197]	Steiner Tree Heuristic	Merges multiple sub-optimal Steiner trees to provide a minimal connectivity solution.
ROBUST [198]	Steiner Tree Evolution	Prioritizes the inclusion of seed genes and highly significant connector genes.
ROBUST bias-aware [199]	Bias-Corrected ROBUST	Improved version designed to mitigate degree-bias introduced by research-bias in interactome data.
TOPAS [200]	Iterative Top-Down	Identifies a minimum number of connection genes between seeds to ensure a fully connected subnetwork.
Betweenness Centrality [205]	General Graph Algorithm	Identifies nodes acting as crucial bridges by measuring shortest paths passing through them.
Closeness Centrality [204] & Harmonic Centrality [203]	General Graph Algorithm	Rank nodes by average shortest path distance. Closeness Centrality cannot deal with disconnected components, Harmonic Centrality can.
TrustRank [201] & PageRank	General Graph Algorithm	Propagates a signal from seeds across the network to identify nodes most influenced by the disease signature.
KeyPathwayMiner [206]	Integrative Enrichment	Identifies maximally connected pathways containing at most a configurable number of non-active genes.

Continued on next page...

Table 1.5 continued from previous page

Tool	Methodological Category	Key Characteristic
BiCoN [207]	Patient-Omics Clustering	Uses expression data to identify subnetworks maximizing differences between patient groups.

1.4.3.3 Drug Ranking

Drug ranking, or drug prioritization, is the final additive algorithmic step in identifying promising therapeutic candidates. Given a disease module or a set of disease-associated genes, the goal is to score and rank all drugs in the KG based on their relationship to these disease-describing subnetworks (Figure 1.2(III)). Existing methods for this task can be broadly divided into three categories: general network scoring and ranking algorithms, purpose-built prioritization methods, and neural network approaches. A condensed overview of all following methods will be found in Table 1.6.

While sophisticated, purpose-built algorithms have been developed for disease module identification, drug prioritization has historically relied more on the application of general-purpose algorithms from network science. In contrast to more complex methods, these scoring-based options typically operate by calculating a measure of relevance or proximity for each drug relative to the disease module, often followed by a cutoff to select the top candidates for further review. A common strategy also in this step involves the use of centrality measures, which assess a node's importance within a network's topology [208]. Betweenness centrality [209], harmonic centrality, as the adjusted version of closeness centrality [203], and degree centrality [210], which simply ranks nodes by their number of connections, can also be used. Degree centrality is naturally highly sensitive to the research bias inherent in well-studied proteins [211].

Other methods focus on propagating a signal from the disease module (or disease genes) through the network. The same network propagation algorithms as for module identification can be used. For example, TrustRank, which is based on Google's PageRank algorithm, starts with the selected genes as a source of "trust" and iteratively spreads their trust score across the network, highlighting –in this case– drugs that are most influenced by this signal [202]. Similarly, diffusion-based approaches like "Random Walks with restart" start at the disease module to empirically score other nodes based on how frequently they are "visited" [212]. This technique has been successfully applied in several DR studies [213, 214].

To expand the field beyond general-purpose algorithms, a small number of methods has been developed specifically for DR. One prominent example is the Network Proximity algorithm, proposed by Guney *et al.* [215]. This method calculates the distance in the network from a drug's targets to the disease module while accounting for biological rationale, such as the principle that a drug does not need to target every protein in a module to be effective. It establishes statistical significance by comparing the observed distance to a random background distribution. This approach has been influential in subsequent studies [216]. SAveRUNNER, which builds on Network Proximity, additionally assesses drug similarity to score candidates that may not be directly connected to the module, but are structurally or functionally similar to those that are [217].

More recently, methods like PANACEA have further refined this by, when available, propagating disease-specific scores through the network to improve prioritization [218].

As mentioned earlier, with recent advances in AI, GNNs are increasingly being applied to DR [219, 220]. These models excel at learning complex patterns from network data and are often used for so-called “link prediction” tasks, such as predicting novel drug-target interactions or drug-disease indications. While powerful, these approaches are best suited for large-scale, systematic prediction studies rather than on-demand, exploratory research. This is because they typically require training on a specific, fixed node-set and can not easily be transferred to the dynamic, user-driven queries common in exploratory research workflows [221, 222].

Table 1.6: Overview of drug prioritization and ranking methodologies.

Method / Algorithm	Primary Mechanism	Key Advantage / Description
Network Proximity [215]	Geometric Distance	Calculates the topological distance between drug targets and the disease module relative to a random background.
Random Walk with Restart [212]	Network Diffusion	Empirically scores nodes based on how frequently they are “visited” starting from a disease module.
TrustRank [202] & PageRank	Signal Propagation	Propagates a signal of “trust” from seeds across the network to identify the most influenced drugs.
SAveRUNNER [217]	Proximity & Similarity	Enhances proximity scores by incorporating drug-drug structural and functional similarity metrics.
PANACEA [218]	Score Propagation	Designed for oncology; propagates disease-specific scores through the interactome to improve ranking.
Closeness Centrality [204] & Harmonic Centrality [203]	General Graph Algorithm	Rank nodes by average shortest path distance. Closeness Centrality cannot deal with disconnected components, Harmonic Centrality can.
Betweenness Centrality [209]	General Graph Algorithm	Identifies drugs or targets acting as crucial bridges (shortest paths) within the network.
Degree Centrality [210]	General Graph Algorithm	Simple ranking by the number of connections; however, highly sensitive to research bias.

Continued on next page...

Table 1.6 continued from previous page

Method / Algorithm	Primary Mechanism	Key Advantage / Description
GNNs (Link Prediction)	Deep Learning	Excels at learning complex latent patterns for predicting novel drug-disease or drug-target interactions.

1.4.3.4 *In Silico Validation*

Lastly, one essential topic is the validation of the resulting predictions. In the absence of a ground truth or gold standard *in silico* approaches need to rely on alternative validation strategies and criteria to assess the reliability of predicted results (Figure 1.2(IV)). This essential step helps to counteract uncertainty introduced in previous steps at least by flagging results that appear more random than case-focused. The goal is that this corrective measure can facilitate the assessment of relevance of certain predictions and findings and thus help with DR research translation by filtering all hypotheses to the ones with the highest chance to advance to experimental verification of generated hypotheses. An overview of all in the following presented methodologies can be found in Table Table 1.7.

The expectation for *in silico* validation methods in network-based DR is that disease modules or a list of ranked DR candidates are assessed regarding their actual relevancy or validity, best case in a quantifiable and interpretable way. Different options are assessed in the following, where some methods might only be applicable to module evaluation or to DR candidate evaluation, but not both.

Naturally, one simple way to assess if disease modules or DR candidate predictions are sensible is by putting them in context with primary resources for current scientific and clinical information, a process sometimes called analytical validation [148, 223]. This can give a first impression about the relevance or general usability of methods and some results. Specifically for novel exploration and discoveries this approach alone has limited power.

In network-driven methods not every information stored in the available KG is necessarily also used in the DR candidate prediction process. Left out relationships, e.g., information about known drug-disease indications is often not taken into account [224]. Therefore, such links can be used as a control mechanism by evaluating if DR approaches were able to re-identify them, making it a specific case of analytical validation. This validation strategy can be used to systematically test DR prediction methods, but also can give a general sense of importance of individual results. Equivalently, if for the disease module creation the initial list of disease associated genes was filtered or thresholded in any way, excluded genes can be used to assess the added module genes [225].

This information can be expanded with other aggregated information sources, like clinical trial databases [224]. It allows to not only check drug-disease connections that are established and approved but includes ones that are under research or trial, showing that other scientists came up with similar or the same treatment hypotheses. Other valuable information is off-label or coincidental use of drugs in combination with patients' disease history. This type of data allows us to observe and potentially validate predicted drug-disease indications using actual responses to fortify a hypothesis based on empirical and external information [224, 226].

Obviously, the so-far mentioned methods only compare to existing knowledge and thus are predominantly inadequate to highlight and evaluate relevance of completely novel hypotheses for disease modules or DR candidates.

For that, transitive information can be included to “evaluate by association”. Often, multiple drugs have similar effects on the human system by binding to the same binding pockets of the same proteins. This information can be used to assess the likelihood of achieving a wanted effect, based on the existence of similar drugs or drugs with the same targets being indicated for the disease of interest. Similar evidence can be found by using information about disorder comorbidity. Comparing the disease modules of one disease with the disease genes of a highly comorbid one, can help to find evidence solidifying the modules relevancy. As comorbidity of diseases or disease subtypes often stems from a shared genetic background, in cases of high comorbidity some information about one disease might be transferable to the other [227–229]. Even further, predictions can be compared through additional information, like drug side effects and disease symptoms, that both are just different descriptions for phenotypes present in an organism. Using phenotype vectors as a proxy to assess overlaps between a drug and a disease can hint towards mechanistic similarity, where the disorder has this mechanism perturbed and a drug is targeting and affecting the underlying mechanism in a similar way [230, 231]. This of course does not directly mean that drugs can treat diseases whenever a shared phenotype signature is present, but this information can help assess the mechanistic relevancy and relationship between the two and potentially fill in gaps in knowledge about their underlying mechanisms. Disease modules can be evaluated similarly. Information about the pathways proteins play a role in is well curated as is information about the effect and resulting disease of a perturbed pathway [232].

The natural progression from the latter approach is enrichment-based validation strategies. With functional enrichment analyses, disease modules or connected components within them, can be evaluated based on their main functional roles within the human organism and as such, be related back to the disease in question. Resources like Gene Ontology [233], g:Profiler [234], DAVID [235], and others are essential resources for functional, compartmental, and other annotation. For example, in cancer this method can highlight if all hallmarks of cancer are explained adequately by the identified disease module and also classify functional components of disease modules of other complex disorders [236, 237].

Data-driven evaluation approaches support the testing of hypotheses regarding newly identified key genes in a disorder even further. Verifying the expression of all proteins of the hypothesized module in patient data, can help reduce false positive observations, at least for cases where sufficient or fitting data is available [238]. Similarly, yet with weaker impact, aggregated tissue context and specificity information about protein expression can be used to evaluate hypotheses about how the expression of disease module genes fits within the mainly affected tissues of the disorder. If, for example, all genes of a Parkinson’s disease module are expressed in lung or heart but not in brain tissue, this should raise questions of relevancy and accuracy [227]. Small inconsistencies and incompatibility with existing tissue-level information can be explained by dysregulation as a result of the disorder, which could be dedicatedly investigated by previously discussed data-driven methods [239, 240].

Finally, statistical evaluation and quantification by empirical p-value calculation are additional valid strategies [241]. Here, the relevance of a resulting disease module or DR candidate list is measured based on the probability that it could have been obtained randomly, given an adequate background model. These can be freely selected as long as they represent the background adequately and therefore can be adjusted to evaluate different assumptions and methods which makes it a well suitable approach. In the case of network-based DR, the creation of random modules to compare the actual results to could, for example, focus on fulfilling similar topological properties, like size and number of connected components. Creation of comparison lists for DR candidates can additionally include prior knowledge of the disease module, as that limits the list of available drug targets. Alternatively, other interpretations of reference-based sampling to establish a random background can be applied. NeDRex, for example, uses a list of reference drugs, known to be indicated for the disease at hand. This list needs to be provided by the user, and based on it the predicted DR candidates and disease modules are statistically evaluated and a p-value assessing the significance of the result sets is provided [129].

Table 1.7: Overview of strategies for the *in silico* validation of computational hypotheses.

Strategy	Validation Type	Description
Analytical Validation	Contextual & Literature	Manually or computationally checking findings against primary scientific and clinical resources.
Indication Re-identification	Benchmark-driven	Uses “left out” relationships (e.g., known drug indications) to assess if a method can re-identify established links.
Indication Expansion	Clinical Context	Cross-references findings with clinical trial databases or evidence from off-label/coincidental drug use.
Transitive Evidence	Evaluation by Association	Assesses likelihood by comparing phenotype signatures (symptoms vs. side effects) or comorbidity comity.
Enrichment Analysis	Functional	Uses resources like GO, DAVID, or g:Profiler to verify if modules align with biological hallmarks or pathways.
Patient Data Integration	Data-driven	Verifies protein expression levels in patient data or affected tissues to filter false positives.
Empirical P-value	Statistical	Quantifies relevance by comparing results to randomized models matching relevant properties (e.g. size and topology).

1.5 MOTIVATION

The advantages and translational potential of network-driven DR are hampered by the accessibility and availability of current solutions. This applies to the previously introduced workflow from disease gene identification to *in silico* validation as a whole, as well as each step individually. The

following section outlines several key bottlenecks that form the basis of this dissertation and motivate the work presented herein.

Bottleneck 1: A lack of democratization

Despite the plethora of available tools, the general availability, usability, and explainability of DR research remain limiting factors. While network-driven methods promise interpretable approaches and results, there is a very limited offer of implementations available that try to provide general access to such integrated network-driven DR workflows.

This stems firstly from an issue with fragmentation and accessibility of individual network medicine algorithms, as it is the case for many other fields within bioinformatics. Data scientists and bioinformaticians might be able to use the tool well, but the most valuable user group are pharmacologists and clinical experts of a specific disease. These specialists could use their expert knowledge in combination with computational tools to gain new insights into the mechanistic understanding of diseases and identify mechanistically plausible candidates for DR trials.

Limitations in usability can arise from complex installation processes and requirements, as the publication of tools only through a GitHub repository is common in bioinformatics, without any user-friendly way to set-up, test, or run them easily from a UI, e.g. through an intuitive web application. Even if an interface exists, many resources just fulfill one part of the DR pipeline (e.g. standalone KG, AI model) where again technical expertise is required to piece together a whole DR workflow that facilitates guided DR candidate prediction while also allowing for exploratory freedom for the experts.

Collections of disease genes are usually available through databases where specific information can be extracted, downloaded, and often also explored within an adjacent platform. DR options, especially network medicine driven ones, are frequently too specific to be part of said platforms. For other methods of disease gene identification, like BiCoN that combines disease gene and module identification by extracting information from differential expression analysis and interactome data, online exploratory resources exist [207]. The same is true for KeyPathwayMiner, yet in both cases the options for deep exploratory research, hypothesis generation, and interoperability are limited, and thus staying behind their potential of serving as essential tools for systems medicine motivated DR without secondary integration [206].

Algorithms solely focusing on module detection have similar issues. Methods like DIAMOnD [195] or ROBUST [198], only exist as executables or code on GitHub. As introduced, this significantly limits the potential user base to those that have sufficient computational knowledge or a team with such skills. DOMINO, as a positive example, was deployed with its own web interface⁷, allowing simple and isolated execution of the tool [242].

As drug prioritization algorithms are mostly lent from general network theory, online solutions targeting DR research are rare. Some generally drug identification-focused platforms like BalestraWeb [179] or PROMISCUOUS [243] provided basic search engines for drug-target relationships. By now PROMISCUOUS, PROMISCUOUS 2.0 [244], as well as other drug-target

⁷ <https://domino.cs.tau.ac.il/>

interaction web resources namely DT-Web [245] and DASPfind [246] are seemingly out of commission, probably due to a lack of use of these resources. Available platforms with more extensive features, such as the previously mentioned Hetionet⁸ [125] and OpenTargets⁹ [126], followed in their footsteps, providing a multitude of options to query their respective resources for drug target related research. But, even those tools do not cover the whole chain from a start selection of a disease and multiple disease associated genes over module and candidate prediction to validation steps, since they have a different focus.

Some integrated solutions that cover more of this workflow do exist. First to mention is CoVex¹⁰, an integrated solution for coronavirus disease treatment candidate prediction using a network medicine approach [197]. While it unites module detection and drug prioritization in one easily accessible web application, it is limited to this very specific use case, namely SARS-CoV-2 research, and lacks options for *in silico* result validation.

CADDIE¹¹, a cancer-focused full-conversion and expansion of CoVex, supports many more different types of disorders under the cancer umbrella [247]. It added useful result evaluation options, like cell line study information of cancer drugs, to provide tools for filtering and highlighting proposed drug candidates. While CADDIE presents an all-around complete application for drug-repurposing for cancer disorders, it is not available outside this specific group of diseases and also not easily extensible to that, as integrated data is heavily cancer focused and not available for other diseases to the same degree.

ROBUST-bias-aware¹², the recent successor to the ROBUST module detection tool, with help of results of this thesis, was released with a more integrated web interface, allowing to not only create disease modules in isolation, but also adding DR candidate prediction, and result validation [199].

A completely separate approach of interaction was taken with NeDRexApp of the NeDRex platform [129]. Like similar resources, it leverages the information stored in the underlying database (NeDRexDB) through a REST API (NeDRexAPI) and provides a UI within Cytoscape [248]. Cytoscape is a well-known standalone application within the network medicine and biological network space providing an extensive toolkit and ecosystem for network exploration and manipulation. NeDRexApp is a Cytoscape plugin and serves as a wrapper of the NeDRexAPI to offer functions from disease gene selection over module detection to drug ranking and even *in silico* validation options using empirical p-value calculation. While the solution does provide a full implementation of a generalistic network-driven DR workflow, there are some drawbacks. First, the local installation of Cytoscape and NeDRexApp is required, increasing the hurdle of entrance to this tool in comparison to readily available web applications. Further, some resource-demanding tasks, like storing larger KGs and layout computation of those networks are taking place on the user's machine. In cases of large interactomes, this task can require multiple gigabytes of memory and several minutes for computation.

⁸ <https://het.io>

⁹ <https://www.opentargets.org/>

¹⁰ <https://exbio.wzw.tum.de/covex/explorer>

¹¹ <https://apps.cosy.bio/caddie/>

¹² <https://robust-web.net/>

Web tools have the option to outsource these computations to powerful servers or to prepare the available networks and information in a way that the retrieval of any information is as efficient and optimized as can be. Naturally, implementing a completely new web interface and backend service for network-driven DR demands substantial effort, know-how, development resources, and time, in comparison to building on a 20-year-long battle-tested and constantly improved system like Cytoscape, that comes with a complete toolkit of network functions.

Bottleneck 2: The seed gene knowledge gap

One continuous point of debate, and the first to introduce significant uncertainty if not careful, is the selection of seed genes, the disease associated genes that are the starting point. Aggregated information can suffice in many cases and help to discover new DR hypotheses. In complex and not well understood disorders, on the other hand, a lack of knowledge or understanding can render the rest of the pipeline useless, as an initial selection might not capture the actual relevant disease genes. Subsequent steps in the pipeline cannot compensate for it and this issue is further amplified by the continued issue of disease classification heterogeneity which introduces noise by simply converting between different disease ID spaces [90].

Data-driven disease gene identification is in many cases more robust than population level-aggregated information, which is especially relevant for precision and personalized medicine. For now, available and previously mentioned solutions can build on omics expression data, but genetics appears to be a blind-spot. Information on genetic impact is implicitly available through disease-gene association databases that incorporate GWAS data but this aggregation can lose the nuance and signal these individual studies could provide. Even then, standard GWAS approaches still fail to explain large portions of the heritability of complex diseases. Epistasis research focuses specifically on this problem but data and hypotheses about mutation and gene sets of epistasis hypotheses are not centrally available. Through the strongly overlapping nature of the previously introduced hypotheses about epistasis (a combination of mutations that causes a certain phenotype, but not in isolation) and disease modules (disease associated proteins forming connected components in the interactome) exploring the mechanisms of disorders using epistasis candidates to build precision disease modules through network medicine approaches seems logical.

Even though computational power keeps reaching new heights, epistasis detection stays an incredibly complex task such that no comprehensive analyses were able to go past variant set sizes of more than three. This, and the absence of any publicly available resource that provides insight into promising epistasis candidate sets, limits the implementation of a DR pipeline building on and evaluating epistasis hypotheses.

Bottleneck 3: Lack of in silico validation strategies

Finally, tools for *in silico* validation of network-driven DR hypotheses are not generally available. With existing tools and data, numerous module detection and DR hypotheses can be generated, even without the availability of many integrated solutions. Not all hypotheses can be experimentally tested. This is why, *in silico* validation provides a necessary filter mechanism to remove hypotheses

that cannot hold up.

As mentioned earlier, tools like g:Profiler or DAVID can be used manually to conduct enrichment-based assessments of module coherence with regard to our understanding of a disease. Naturally, they do not come with a quantification of importance assessing the initial question of DR potential though, as they were never intended for that line of research.

The bigger picture

Looking at the three main bottlenecks just presented as a collection of issues of different stages of a network-based DR pipeline (Figure 1.2) their impact becomes even clearer. The **seed gene knowledge gap** often causes the pipeline to start with noisy and incomplete input. These deficiencies are then inherited by subsequent steps, leading to the construction of potentially low-confidence disease modules and, ultimately, to drug predictions with similarly compromised reliability. The **in silico validation gap** makes it exceedingly difficult to computationally distinguish the true biological signal from this background noise in a meaningful or quantifiable manner. Finally, the **democratization gap** prevents domain experts, who possess the biological intuition needed to manually curate and filter these noisy results, from effectively engaging with the tools. This creates a vicious cycle of low-confidence predictions that are difficult to validate, leading to a low rate of translation from computational hypothesis to clinical reality. This dissertation seeks to interrupt that cycle by presenting a collection of tools that respectively address each of these bottlenecks.

1.6 AIMS AND CONTRIBUTIONS: IMPROVING ACCESS TO DRUG REPURPOSING HYPOTHESES

The goal of this dissertation is to address the outlined bottlenecks and, to this end, presents four novel, integrated software tools that aim to help a broad range of users –including domain experts, biomedical researchers, clinicians, and computational biologists– conduct computational DR pipelines by providing integrated and integrable solutions. The whole chain of network-driven DR steps (Figure 1.3(A)) is represented and how the newly developed solutions cover them is illustrated in Figure 1.3(B). Synergistic and complementary benefits between the presented work, enabled by a focus on developing integrable solutions, is highlighted.

To summarize: making network-based DR pipelines accessible to experts in the field is hampered by the potential of the overall accessibility and integration of fragmented tools combined with room for improvement in every step of the pipeline to reduce the propagation of uncertainties. The general workflow integrates, as presented earlier: 1. The identification of disease or condition associated genes as seeds (Figure 1.3(A)I); 2. The detection of disease modules based on the given seeds (Figure 1.3(A)II); 3. The prioritization of drugs, targeting the identified disease modules (Figure 1.3(A)III); Figure 1.3. The *in silico* validation of results (Figure 1.3(A)IV).

The following goals are set to improve this current state of DR research conducted on a network medicine basis and beyond. The contributions made to fulfill these goals are shortly highlighted to introduce the work done and the final structure of this thesis:

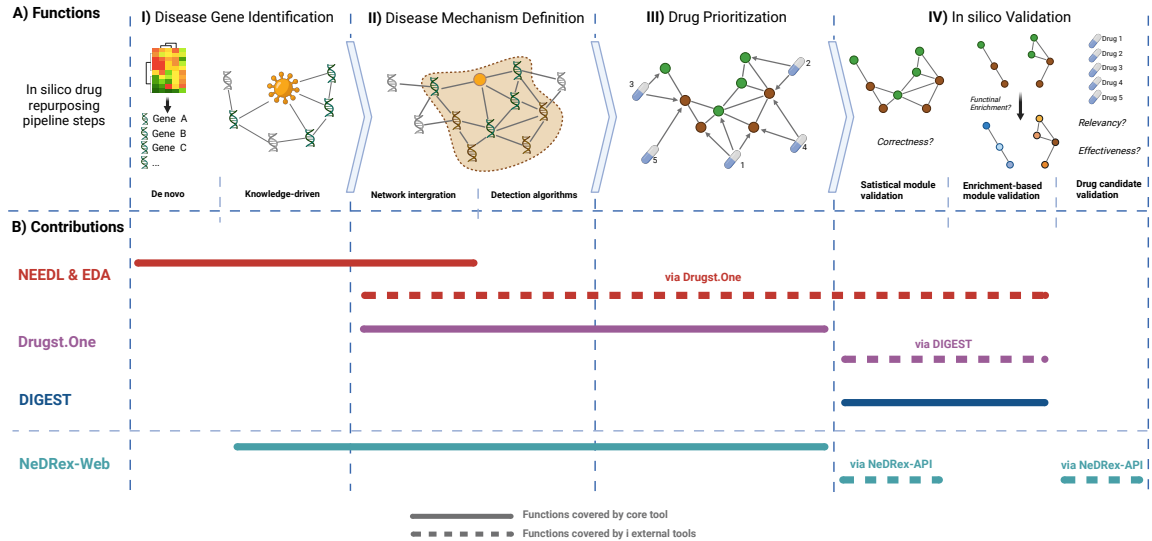


Figure 1.3: Summary overview of the network-based DR workflow in A) and which steps are covered by the presented publications in B). A) For a more nuanced depiction, step I), II), and IV) have annotated sub-categories presenting different solutions to the presented steps. In II) “Network integration” relates to enabling disease mechanism exploration by integrating gene information with network data, enabling connection-based assessments. “Detection algorithms” then specifically hints towards purposefully integrated functions and algorithms assisting this mechanism identification step leveraging the network information. In B) this fine-grained split is utilized to highlight intrinsic capabilities of the presented tools as solid lines, and the expanded and available functionalities through synergistic integration of others of the presented projects. **NeEDL** and the **Epistasis Disease Atlas** provide a novel solution for *de novo* disease gene identification based on network-medicine assumptions and provides a UI for interactive exploration of results, integrated with network information. The referral to **Drugst.One** expands the on-site functions with the whole chain of network-based DR functionalities, from disease module detection to drug prioritization. **DIGEST** provides the missing *in silico* validation tools for generated disease modules. NeDRex-Web is separated from the other three solutions to highlight the effort of it being a fully integrated and self-sufficient end-to-end solution attached only to the NeDRex infrastructure, in contrast to the other published projects.

GOAL 1: Integrated and accessible end-to-end solutions for network-based drug repurposing

CONTRIBUTION 1: Addressing the Democratization Gap with NeDRex-Web and Drugst.One.

This work presents NeDRex-Web and Drugst.One, a comprehensive platform and a plug-and-play solution, respectively, that make the end-to-end repurposing pipeline accessible to non-computational experts.

GOAL 2: Integrating network medicine and epistasis research to facilitate identification of novel data-based disease genes for a new, robust drug repurposing angle

CONTRIBUTION 2: Breaking the Seed Gene Gap with NeEDL and the Epistasis Disease Atlas.

This thesis introduces NeEDL, a novel framework leveraging network medicine to overcome the

combinatorial barrier in epistasis detection, and the Epistasis Disease Atlas (EDA), providing a new and more robust source of seed genes in the first openly accessible repository for epistasis analysis results integrated with DR research tools.

■ **GOAL 3:** Development of novel, and integrable purpose built validation algorithms for predicted disease modules

■ **CONTRIBUTION 3:** Closing the Validation Gap with DIGEST

This dissertation presents DIGEST, a purpose-built statistical tool that provides novel metrics to assess the mechanistic and genetic coherence of computational hypotheses, adding a necessary layer of confidence to network-driven DR research.

To achieve these three goals, the following chapters will present the novel contributions designed to address each bottleneck. **Chapter 2** will introduce NeDRex-Web and Drugst.One, two resources developed to address the democratization gap by providing comprehensive and accessible tools to the community. **Chapter 3** will then address the seed gene knowledge gap by presenting the NEeDL algorithm and the Epistasis Disease Atlas, the first online resource for exploring higher-order epistasis results. Subsequently, **Chapter 4** will detail DIGEST, a novel method for *in silico* validation that closes the final validation gap. Finally, this work evaluates the collective impact of these tools as an interconnected ecosystem, utilizing community adoption and usage analytics, presented in **Chapter 5**, as a quantitative proxy for the success of the overarching democratization effort. Finally, the dissertation will conclude with a discussion in **Chapter 6** that assesses the contributions made with regard to the identified goals. Concluding remarks will be made in **Chapter 7** as well as an outlook on future work.

INTEGRATED NETWORK-DRIVEN DISEASE MODULE AND DRUG REPURPOSING CANDIDATE IDENTIFICATION

2.1 INTRODUCTION

Towards the end of my bioinformatics studies, my journey into network medicine and DR began by learning about the then soon-to-be-published NeDRex platform. A previous project, named CoVex, a Coronavirus explorer developed at the early stages of the SARS-CoV-2 pandemic to research DR candidates, was the first interaction of such an integrated network-based drug repurposing tool [197]. Following the same underlying idea, a generalistic platform for any disease was built. NeDRex was designed to have a large heterogeneous KG at its core, that can be queried and explored both manually and algorithmically. For that, disease module and drug ranking algorithms were integrated and made available through the so-called NeDRexAPI, alongside many other functions for accessing the information stored in the NeDRexKG. For an accessible UI at that time a Cytoscape plugin was chosen and developed using NeDRexAPI to query data, run algorithmic tasks, and to load the results into Cytoscape. While this NeDRexApp worked as intended, it had some significant drawbacks. Firstly, Cytoscape needs to be installed on the researcher's device, as it is a standalone application. This further means that after networks are downloaded through the NeDRexAPI any visualization, layouting, and exploration is happening locally on the user's machine and using local resources. While this is not an issue for small networks, loading large interactomes leads to significant memory footprints of 16GB and higher, which means the researcher's machine requires significant memory resources and decent bandwidth internet to load all the data in the first place.

Recognizing the potential of network-driven DR, but seeing the issues in accessibility, I aimed to develop a web application building on the NeDRex knowledge base and general toolkit but optimizing it for on-demand access requiring only any modern browser. At the same time, one objective was to maintain the freedom that Cytoscape offers with regard to exploration and expert-in-the-loop focus, while helping and supporting computationally less experienced users as best as possible.

The goal was to develop and provide the first integrated web resource for network-driven DR, called **NeDRex-Web**. At the time, any type of available DR-oriented web-resources were scarce. Generalistic resources that allow for the identification of potential drug candidates do exist, but are quite limited in functionalities, with very little freedom of operation and exploratory research for the expert. Examples are CROSSBAR¹ [249], a tool for analytically exploring and searching paths within a large KG, and PROMISCUOUS 2.0² [244], which was mainly focused on drugs, their

¹ <https://crossbar.kansil.org/>

² Unavailable by now

targets, and side effects. More comparable to NeDRex are the BioThings Explorer³ and Hetionet⁴ [125] as they specifically enable targeted identification of links between drugs and diseases, but in contrast lack purpose-built and integrated algorithmic solutions with a focus on DR to have a larger impact in this field.

Online available resources, that give more power to the user, include the previously mentioned CoVex⁵, which was one of the first attempts of a unified interface focusing on identifying DR candidates for COVID19 specifically, by integrating host-virus interaction datasets [197]. Further, representing a major and very special group of disorders, a cancer-specific DR application in the form of CADDIE⁶ was developed in an equivalent way [247]. It focuses on the integration of data and information specifically relevant to cancer, like using genetic variant files as a starting point and cell line drug response information for anti-cancer drugs. While all those resources aim to provide tools for the identification of novel treatment approaches in some disorders, centrally available resources that are easy to access and provide a rich toolkit of DR functionalities for a more general case are missing.

One of the most labor intensive but equally central parts of any of the existing solutions is an interactive and robust way to intuitively explore networks to navigate and understand these complex molecular relationships. While many implement a network explorer, the quality and depth vary greatly as each solution is developed mostly from scratch, and it is a sizable endeavor for developers to deal with the implementation of such a feature. Examples are BiCoN⁷ [207] and DOMINO web-server⁸ [242] but this extends even to databases for interactome data like STRING [103] or IID⁹ [121]. While JavaScript libraries for networks and even biological networks exist in the form of cytoscape.js, it still requires significant additional work, like the handling of input and output formatting, the integration of several data sources to enrich the network with information and, depending on the goal, the integration of network algorithms, that are executed asynchronously on a backend server.

Based on this, we additionally identified the need for a reusable solution that simplifies the implementation of a network visualizer and certain base functionalities for DR. That project was labeled **Drugst.One** and aimed for bringing DR to any platform that deals with protein networks or merely generates a list of disease associated genes. This way, it has the potential to enable DR research, where insights about diseases are generated. The hope was that by providing an easy to use and easy to integrate solution, adding DR to existing resources would become “a no-brainer”. This would lead to distribution and increased access to DR, not only within the existing community but also outside, democratizing the research for DR candidates and providing access to the experts that can make a difference.

3 <https://explorer.biothings.io>

4 <https://het.io>

5 <https://exbio.wzw.tum.de/covex/explorer>

6 <https://apps.cosy.bio/caddie/>

7 <https://exbio.wzw.tum.de/bicon/>

8 <https://domino.cs.tau.ac.il/>

9 <https://iid25.ophid.utoronto.ca/>

2.2 RESULTS

Here I present **NeDRex-Web (Publication A)** [250] and **Drugst.One (Publication B)** [251], two fully web-based, integrated end-to-end solutions for network-based DR that I (co-)developed and published to make DR research more accessible. The two applications are not solving the same problem twice, but rather target different niches. While NeDRex-Web provides a fully integrated extensive toolkit with, visualization options, exploratory functions, and elaborate history and result sharing system allowing to dive deep into data on this one platform, Drugst.One focuses more on essential functions, simple overviews, and integrateability to extend existing solutions. The projects can be seen as complementary solutions to enable general access but also to improve distribution and thus visibility of DR aimed at clinical experts that never worked on DR and long standing pharmacological researchers alike.

2.3 NEDREX-WEB

2.3.1 Summary

With NeDRex-Web I address the lack of a comprehensive and general online DR application that provides an end-to-end solution for network-based seed selection, disease module and drug candidate prediction, including p-value-based *in silico* result validation through the original NeDRexAPI (Figure 1.3(IV) & Figure 2.1). It is a completely cohesive, standalone tool, requires no installation, and is targeted towards clinical and pharmacological experts that want to deep-dive into exploring mechanistic relationships between genes, drugs, and diseases, or disease modules and pharmacological relationships. It is mostly decoupled from the original NeDRex platform, which makes it robust to any changes to the main platform down the line, while syncing the database on a weekly basis to provide up-to-date and consistent information. Furthermore its toolbox was expanded past the already existing functionalities of NeDRex, integrating additional module detection algorithms, the explicit option to use tissue-level filtered interactomes, purpose-built connectivity searches, more documentation of the integrated tools, and support for disease comorbidity-based research of shared mechanisms, amongst others. General utility for hypothesis generation, mechanism, and DR candidate identification are presented in several use cases. NeDRex-Web is fully and freely available at web.nedrex.net¹⁰.

¹⁰ <https://web.nedrex.net>

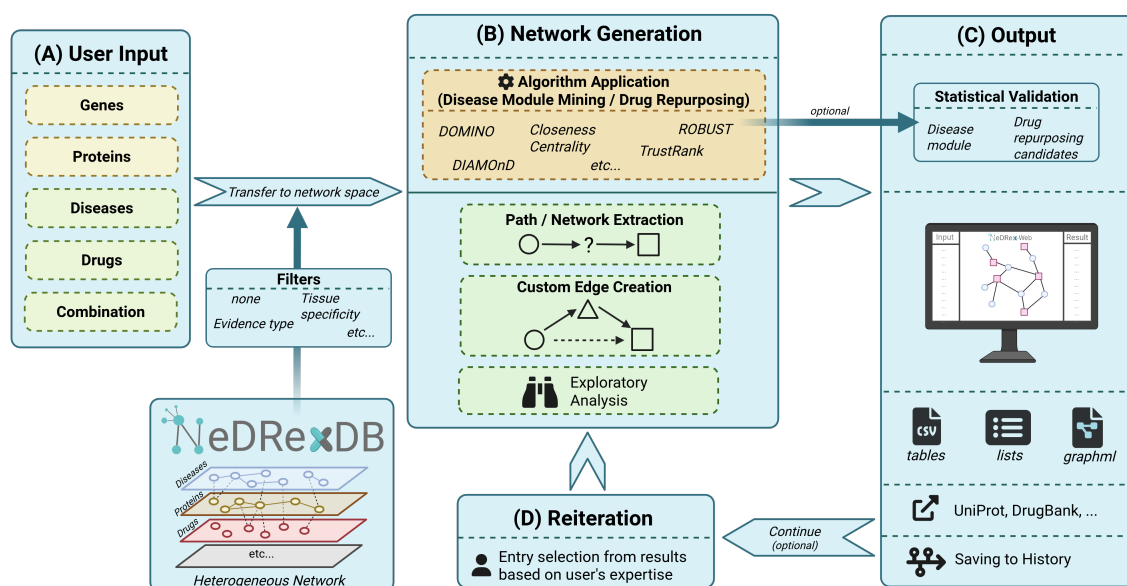


Figure 2.1: The NeDRex-Web workflow (Reprinted from publication A [250], ©2024 IEEE): NeDRex-Web covers all steps of the network-based DR pipeline end-to-end. **(A)** Users start an analysis by selecting a starting point, e.g. finding genes that represent a disease and analyze them. **(B)** For this, users can decide between exploring freely through manual network filtering and guided network link analysis steps, or by applying network medicine algorithms. **(C)** Detailed and informative result overviews provide researchers with options for additional steps, such as the *in silico* validation of results, exploration through the interactive network explorer, the export or consultation of linked resources for external follow-ups or simply **(D)** the option to iteratively deepen or refine their hypothesis. All generated results are stored for the user in their personal history.



Andreas Maier*, Mahdie Rafiei, Elisa Anastasi, Olga Zolotareva, James Skelton, Maria L Elkjaer, Ana I Casas, Cristian Nogales, Harald HHW Schmidt, Tim Kacprowski, David B Blumenthal, Anil Wipat, Sepideh Sadegh, and Jan Baumbach

“NeDRex-Web: An Interactive Web Tool for Drug Repurposing by Exploring Heterogeneous Molecular Networks”

Published in: 2024 *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*
Volume: 12, Issue: 4, pp. 429–434. Published December 2024.

DOI: [10.1109/bibm62325.2024.10822390](https://doi.org/10.1109/bibm62325.2024.10822390)

Complete Reference:

Maier, A., Rafiei, M., Anastasi, E., *et al.* (2024). NeDRex-Web: An Interactive Web Tool for Drug Repurposing by Exploring Heterogeneous Molecular Networks. *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 12(4), 429–434. <https://doi.org/10.1109/bibm62325.2024.10822390>

Authors' Contribution: *I planned and implemented NeDRex-Web while collecting input from clinical experts on expected functionalities and from software developers and computational experts on usability to improve the system accordingly. The implementation included the development of robust and maintainable backend and database solutions, allowing for auto updates from NeDRexDB on a weekly basis, an extensible and asynchronously running tool-execution system, and a user history storage and management module. I also wrote most of the paper, apart from some of the use cases that were constructed by co-authors and then improved and/or updated me author for publication. I further improved the paper based on feedback from the co-authors.*

Code Availability: The source code for the whole project of NeDRex-Web can be found at github.com/repotrial/NeDRex-Web.

NeDRex-Web: An Interactive Web Tool for Drug Repurposing by Exploring Heterogeneous Molecular Networks

1st Andreas Maier

Institute for Computational Systems Biology
University of Hamburg
Hamburg, Germany
andreas.maier-1@uni-hamburg.de

2nd Mahdie Rafiei

Institute for Computational Systems Biology
University of Hamburg
Hamburg, Germany
mahdie.rafiee@uni-hamburg.de

3rd Elisa Anastasi

School of Computing
Newcastle University
Newcastle upon Tyne, UK
Elisa.Anastasi@newcastle.ac.uk

4th Olga Zolotareva

Institute for Computational Systems Biology
University of Hamburg
Hamburg, Germany
olga.zolotareva@uni-hamburg.de

5th James Skelton

School of Computing
Newcastle University
Newcastle upon Tyne, UK
James.Skelton@newcastle.ac.uk

6th Maria L. Elkjaer

Department of Neurology
Odense University Hospital
Odense, Denmark
maria.louise.elkjaer@uni-hamburg.de

7th Ana I. Casas

Department of Pharmacology and Personalised Medicine
Maastricht University
Maastricht, the Netherlands
AnaIsabel.CasasGuijarro@uk-essen.de

8th Cristian Nogales

Department of Pharmacology and Personalised Medicine
Maastricht University
Maastricht, the Netherlands
cristian.nogales@univie.ac.at

9th Harald H. W. Schmidt

Department of Pharmacology and Personalised Medicine
Maastricht University
Maastricht, the Netherlands
hschmidt@ppmlab.net

10th Tim Kacprowski

Braunschweig Integrated Centre of Systems Biology
Technical University of Brunswick
Brunswick, Germany
t.kacprowski@tu-braunschweig.de

11th David B. Blumenthal

Department Artificial Intelligence in Biomedical Engineering
Friedrich-Alexander-Universität Erlangen-Nürnberg
Erlangen, Germany
david.b.blumenthal@fau.de

12th Anil Wipat

School of Computing
Newcastle University
Newcastle upon Tyne, UK
anil.wipat@ncl.ac.uk

13th Sepideh Sadegh*

Institute for Computational Systems Biology
University of Hamburg
Hamburg, Germany
ssadegh@health.sdu.dk

14th Jan Baumbach*

Institute for Computational Systems Biology
University of Hamburg
Hamburg, Germany
jan.baumbach@uni-hamburg.de

Abstract—Finding new indications for approved drugs is a promising alternative to the often very lengthy and expensive process of *de novo* drug development. Systems medicine has brought forth several different approaches to tackle this important task. We recently published NeDRex, a network medicine tool for the identification of disease modules and drug repurposing. NeDRex-Web (<https://web.nedrex.net>) brings existing and new features of the NeDRex platform to a user-friendly and research-oriented web application, enabling online exploration

of large heterogeneous molecular networks. Focusing mainly on drug repurposing, NeDRex-Web implements customizable disease module identification and drug prioritization workflows to support users of diverse backgrounds in their research. Users are assisted during every step of their analysis, including the definition of relevant input sets, the selection from various algorithms for module identification or drug prioritization, and the prioritization of the results by their statistical significance. A guided connectivity search provides an easy way to identify

links between node sets of interest and can be used to create user-specific induced networks.

Index Terms—drug repurposing, knowledge graph, disease mechanism identification, interactive network explorer, interactome.

I. INTRODUCTION

Since the conclusion of the human genome project in 2004, the biomedical field realized that relationships between genes, diseases, and drugs are more intricate than previously anticipated. By modeling these complex relationships with graph-theoretical methods, the field of network medicine unfolded. Heterogeneous networks can be constructed by integrating information about the associations amongst different biological entities. These relationships are extracted from results of now broadly available experiments including genome-wide association studies, chromatin immunoprecipitation sequencing, differential gene expression analysis, and many more.

The *de novo* drug identification and development pipeline is a time-consuming, costly, and highly demanding process [1]. Thus, drug repurposing aims to be the key strategy to overcome this recurrent translational roadblock. Historically, there are multiple examples of drugs where observed side effects, or initially unknown off-target events, led to the identification of therapeutic effects in new disease indications. For instance, sildenafil, a well-known drug commonly prescribed for erectile dysfunction, was originally developed for the treatment of cardiovascular diseases, i.e. angina pectoris [2]. Similarly, galantamine, originally an anti-paralytic agent used to treat myopathies, is now one of the few approved drugs used for treating Alzheimer’s disease [3].

To identify drugs that have repurposing potential for a given disease, network medicine approaches exploit information that can be deduced from large heterogeneous biological networks. Several online services following this paradigm exist: Hetionet [4] integrates different resources into one heterogeneous network, but its exploration functionality is limited to queries for paths connecting entries of interest, and only static drug repurposing in the form of precomputed lists for 1,538 drug compounds and 136 diseases is supported. The BioThings explorer [5] mainly serves as a wrapper for general network requests, without providing any options for the interactive exploration of the unique network. CoVex [6] and CADDIE [7] which are more flexible than BioThings and CROssBAR, but are limited to COVID-19 and cancer, respectively. CADDIE in particular, includes cancer-specific resources, like drug-response data from cellline studies and mutation frequency data, which helps users gain insight into that specific medical field. Drugst.One [8], on the other hand, a more recent and general version, is as flexible but targeted towards simplicity and its use as an integrative component to other websites instead of providing a solution for complex exploratory queries and large networks. Other network-based drug repurposing resources such as PROMISCUOUS 2.0 [9] have limited options for user-specific customization, only providing options to filter precomputed results for a fixed set of diseases. In cases,

where additional information about a disease or drug could be leveraged, these approaches become unattractive due to the missing flexibility.

With NeDRex-Web, we overcome the limitations of existing web services and provide a unique combination of a disease-centric, network-based drug repurposing application, and a general network browser for complex exploratory analysis. Its user-friendly and research-oriented interface supports highly customizable drug repurposing and exploratory analysis, allowing users to engage their expert knowledge in every step. NeDRex-Web is freely available, to allow any user to use it for their research. Furthermore, a web-based solution where all computationally demanding tasks are run remotely removes the requirements for any specific operating system or hardware. Specifically, the NeDRex-Web knowledge base contains comprehensive information on protein, gene, drug, disorder and pathway nodes, interconnected by twelve different association edge types (Figure 1). Overall, it contains 439,712 nodes and 9,111,193 edges, necessitating various pre-processing and optimization steps to deliver a seamless network exploration experience. All queries and results are stored as a users-specific research history thus allowing NeDRex-Web for reloading previous results or exporting the whole network for offline analysis or getting the unique URL to quickly share the results with colleagues.

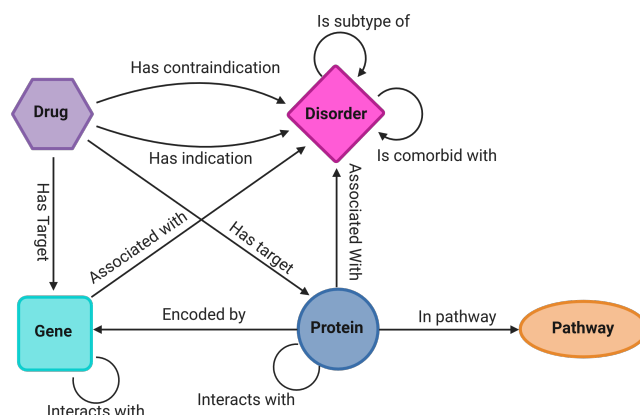


Fig. 1. **The NeDRex-Web metagraph.** NeDRex-Web integrates five node types, connected by 12 edge types. The inferred gene-gene, disorder-protein, and drug-gene edges are derived by mapping genes to proteins and *vice versa*.

NeDRex-Web is driven by the knowledge graph of the network medicine platform NeDRex [10]. In addition to providing interactive online access to NeDRexDB, the following key-functions are new in the web tool compared to the stand-alone NeDRexApp:

- Integration of additional module identification methods (KeyPathwayMiner [11], DOMINO [12], ROBUST [13])
- Support of tissue-specific interactomes
- Provision of user-specific research history
- Entry selection through keyword search with auto-complete suggestions

- On-demand comparison of results to ClinicalTrials.gov entries
- Provision of comorbidity information to e.g. identify shared underlying disease mechanisms

II. METHODS

In the following, the three ways of starting any exploratory research with NeDRex-Web are highlighted. Information about implementation details and used frameworks can be found in the supplementary material (Section 1.9).

A. The NeDRex-Web Knowledge Graph

NeDRex-Web uses the knowledge graph from an upgraded version of NeDRexDB [10] with additional integrated databases. The knowledge graph supports five node types (gene, protein, drug, disorder, and pathway) (Figure 1, Table S2) connected by over 9 million edges of 12 types (Table S3). It integrates 13 different publicly available sources (Table S2), which provide additional attributes for the node and edge entries (updates from the public data sources are executed weekly). More information about the data integrated in the knowledge graph is found in Supplementary Material (Section 2). NeDRex-Web uses a relational database to store all the entries and their attributes. The full network structure is stored in-memory, which allows the in-memory execution of all basic network and connection-based queries. NeDRex-Web allows the user to set filters based on, e.g., disease-gene association score, protein-protein interaction evidence type, or tissue-specific expression. The latter allows the construction of tissue-specific interactomes, which can be used in downstream disease module identification and drug repurposing analyses.

B. Exploration Routes Supported by NeDRex-Web

NeDRex-Web provides various means to explore the integrated knowledge graph. An overview of the workflow is shown in Figure 2. For entry-level users without a background in network medicine, easy to use step-by-step pipelines are developed for the identification of repurposable drug candidates and to establish connections between entries of interest within the knowledge graph. Besides these guided options, NeDRex-Web allows the unrestricted exploration to follow individual exploratory analysis approaches. In the following, an overview of these three options are provided. More detailed explanation as well as information about the different input and output types of the tool can be found in the Supplementary Material (Section 1).

Quick Drug Repurposing:

- 1) A quick drug repurposing query typically starts with a set of genes, called a seed list. Such a set can for example contain genes differentially expressed in a disease of interest. Alternatively, a disease or a set of disease types can be selected by names, ICD-10 codes or other supported identifiers (Table S2) which then is used to generate a seed set based on disease-associated genes (Figure 2A).

- 2) The seed list is used to extract disease modules from the corresponding, potentially tissue specific interactome. This helps to identify missing genes that describe the disease mechanism or to find additional interaction partners with the disease genes that can serve as drug targets. Available module identification algorithms are DIAMOnD [14], DOMINO [12], KeyPathwayMiner [11], ROBUST [13], and MuST [7] (Figure 2B, top). An alternative approach is provided by BiCoN [15], allowing the unsupervised extraction of the disease module directly from expression data. A detailed explanation of the algorithms is given in the Supplementary Material (Section 1).
- 3) Drug-target information is added to establish links between the disease module and drugs. TrustRank [16] or Closeness Centrality [17] can be chosen to extract and rank drug-repurposing candidates (Figure 2B).
- 4) The statistical relevance of the gene modules and ranked list of drugs can be estimated by computing empirical *P*-values for the results (Figure 2C, top). A detailed explanation of the different statistical validation methods is given in the Supplementary Material (Section 1.9.3)
- 5) In the result overview, the seeds and module genes as well as a table of ranked drugs are provided. Drugs that are known to be approved or investigated for the treatment of the queried disease are highlighted, using data from ClinicalTrials.gov [18]. The connections between all genes and drugs are represented by an interactive network visualization, combining gene-gene interaction and drug-target information (Figure 2C).

Guided Connectivity Search:

- 1) The Guided Connectivity Search allows to identify links between sets of any entries. For instance, two sets of diseases may be used as so-called source and target sets to investigate mechanistic similarities between them (Figure 2A).
- 2) The type of connection is defined by selecting one of the proposed paths. For instance, a shared molecular mechanism between the disease groups under investigation can be researched with the disease – gene – disease path. Additionally, it is possible to filter edges and nodes by relevant attributes and define an exclusion or inclusion list for connecting nodes (Figure 2B).
- 3) The result overview presents the extracted links by providing tables for both input disease groups and the connector nodes, in this example the shared genes associations. The connectivity network is visualized for interactive exploration of the result (Figure 2C). In a drug repurposing context, the shared mechanism might be now identified and enriched by adding drug-target information. This allows to investigate drugs that may be repositioned from one disease group to the other (Figure 2D).

Free Exploration:

- 1) The Free Exploration option allows advanced users to

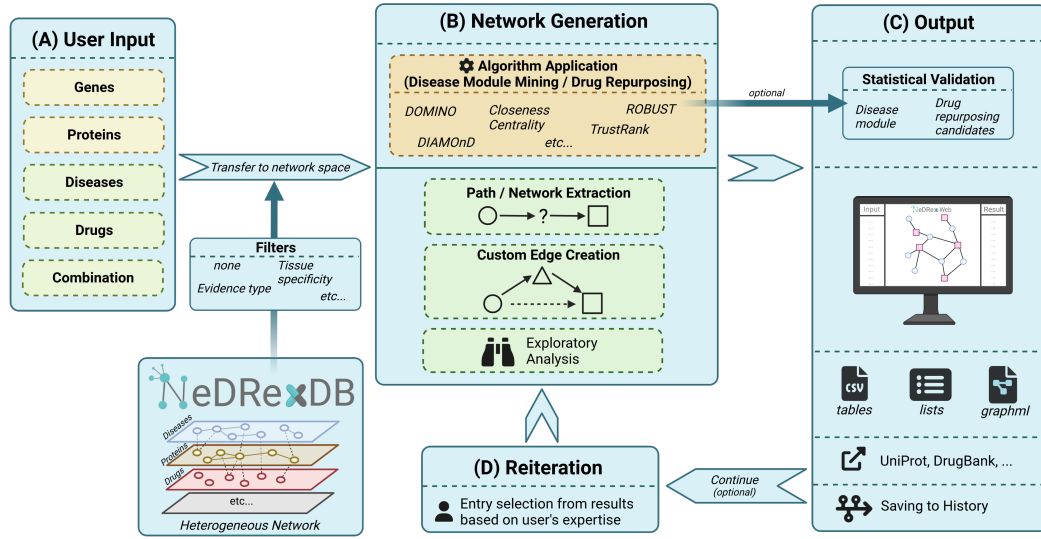


Fig. 2. **The NeDRex-Web workflow.** The procedure of result extraction from the knowledge graph includes four conceptual phases. (A) Input definition: The user-driven selection of nodes or edges. (B) Network generation: The integration of user-input into the network space and application of methods for knowledge extraction. (C) Output presentation: Result representation through a variety of visualizations, tables and exports. (D) Iterative analysis: A result network can be fully or partially used as input for subsequent analysis steps. NeDRex-Web profits from the the expert knowledge that can be employed at steps A, B, and D.

freely explore the NeDRex-Web knowledge graph. A starting network is defined through the selection of node and edge types. For instance, for a drug-centric repurposing approach, one could select i) drug nodes, limited to the drugs to be investigated, ii) drug-protein edges, iii) protein-protein interactions, and iv) disorder-protein associations (Figure 2A, B).

- 2) The created network not only integrates all directly targeted proteins and associated diseases, but also contains diseases indirectly affected through the drugs' effects on the interactome. The diseases and all other entries can be investigated using the network browser or the node and edge lists, allowing them to be filtered or sorted, e.g., based on their network degrees (Figure 2C).
- 3) A new network can be derived by selecting node and edge entries to focus on in the next steps (Figure 2D). For instance, in a drug-centric use case, it might be interesting to add information about drugs indicated for the identified diseases (drug-disorder edges) or to run a module identification method to further investigate the mechanism targeted by the drug.

III. RESULTS

To exemplify the functionalities available in NeDRex-Web and their utility, we report the results of two case studies. Another use case, a detailed exploration of heterogeneous brain lesions in patients with brain lesions in progressive multiple sclerosis (Section 3.3), as well as full-sized figures and detailed information for reproducibility can be found in the Supplementary Material (Section 3). The following cases

were produced Oct. 8 2024 but as the database auto-updates weekly, future networks might differ slightly.

A. Drug Repurposing for Cystic Fibrosis

Cystic fibrosis (CF) is a monogenic, recessive disorder caused by mutations in the cystic fibrosis transmembrane conductance regulator (*CFTR*) gene. Utilizing NeDRex-Web, we identified 11 genes associated with CF and employed them as seed genes for module identification using the DIAMOnD algorithm. The resulting disease module, with a statistically significant empirical P -value (precision-based $P < 0.04$), encompasses 27 additional genes. Many of these genes encode ribosome subunits (*RPSA*, *RPS5*, *RPS7*, *RPS11*, *RPS12*, *RPS14*, *RPS18*, *RPS23*, *RPL6*, *RPL7*, *RPL9*, *RPL10A*, *RPL11*, *RPS16*, *RPS20*, *RPL12*, *RPL18*, *RPL24*, *RPL30*, *RPLP0*) and ubiquitination-associated proteins *WWP2* and *PRKN*. In CF, ribosomal contribution to the folding of the missense *CTCF* protein has been postulated, with a beneficial effect on *CTCF* folding specifically linked to the knockdown of *RPL12* [19]. Ubiquitination, crucial for cellular homeostasis due to its role in marking proteins for degradation or other modifications, has been identified as a potential mitigator of the CF phenotype through its associated proteins [20].

Subsequently, we applied the TrustRank drug prioritization algorithm to identify drugs targeting the uncovered module, generating a ranked list of 62 drugs (empirical, rank-agnostic P -value of $P < 0.05$) (Figure 3A and full-sized in S3). Notably, amiloride (rank 2) [21] and quercetin (rank 14) are already undergoing clinical trials for CF treatment. Triamterene, structurally similar to amiloride, emerged as the most promising therapeutic option. Etanercept (rank 5)

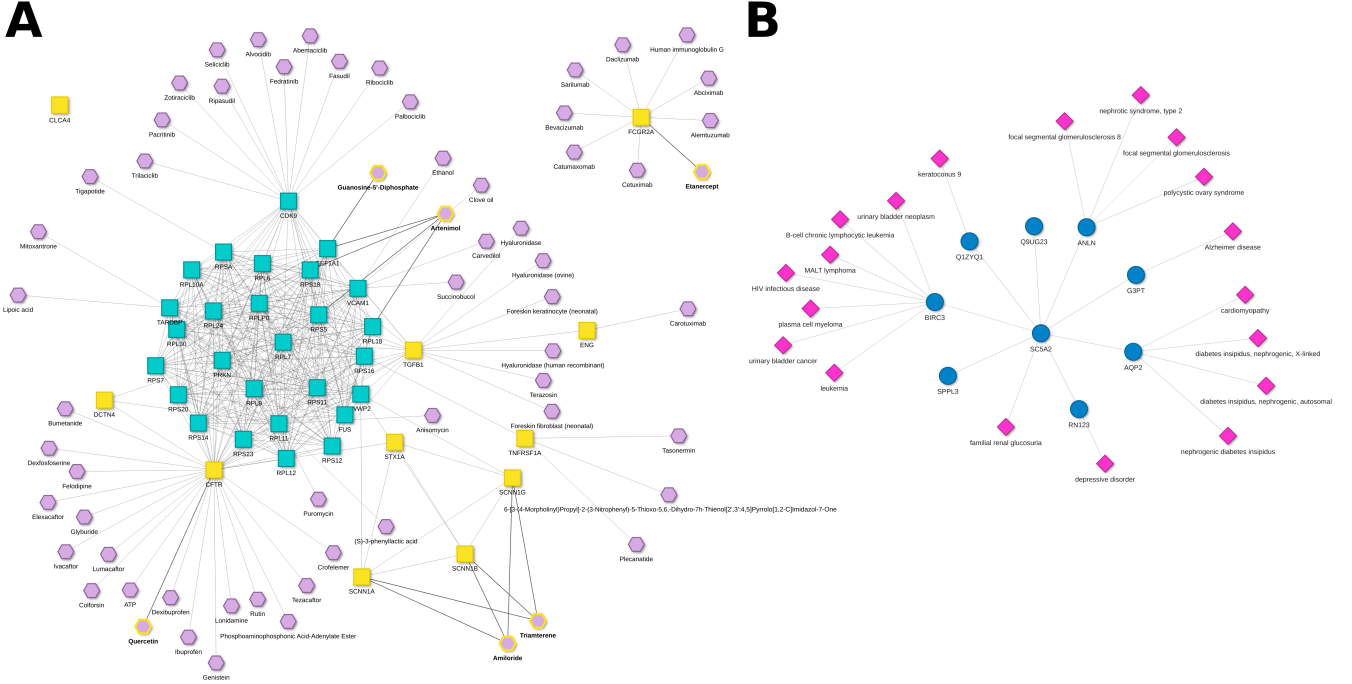


Fig. 3. The results of two different application cases: **A)** The resulting subnetwork from the disease module and drug repurposing pipeline performed for Cystic Fibrosis (CF). Yellow nodes depict the seeds associated with the CF disease in the NeDRexDB. Turquoise nodes are genes incorporated by the DIAMoND module (P -value threshold 10^{-35}), and purple nodes represent top drugs targeting all module genes after TrustRank drug prioritization. Drugs highlighted with a yellow border are drugs identified to have shown potential or might be candidates for use in CF treatment, namely amiloride, artenimol, etanercept, guanosine-5-diphosphate, quercetin, and triamterene. **B)** First neighbour PPI network of the *SGLT2* co-transporter (*SC5A2*, UniProt P31639) and associated disease phenotypes. On the right, Aquaporin 2 (*AQP2*) appeared directly connected to *SC5A2* and associated to cardiomyopathy. These results, supported also by previous research, suggest *AQP2* as a potential mechanism that could explain the positive outcomes of *SGLT2* inhibitors in heart failure.

has demonstrated efficacy in improving the prognosis of CF patients [22]. Guanosine-5-diphosphate (rank 19) inhibits interactions between G proteins and their effectors, leading to an increase in cAMP-activated Cl^- currents in airway epithelial cells of CF patients [23]. In a similar vein, artenimol (rank 9) and its derivatives have shown potential as anti-fibrotic therapeutics [24].

B. Exploratory Analysis of Positive Outcomes of *SGLT2* Inhibitors in Heart Failure

Sodium-glucose cotransporter 2 inhibitors (*SGLT2*Is) are glucose-lowering drugs widely used clinically in the treatment of type 2 diabetes mellitus. *SGLT2*Is increase urinary glucose by inhibiting its reabsorption in the proximal tubules. *SGLT2*I intake has shown unexpected beneficial effects, reducing cardiovascular mortality and heart failure hospitalizations among patients [25]. However, the mechanism by which *SGLT2*Is improve cardiovascular outcomes is currently unknown. Here, we make use of the exploratory analysis capabilities of NeDRex-Web to generate hypotheses that could explain this mechanism.

*SGLT2*Is target the sodium-glucose cotransporter (*SC5A2*), for which we extracted all experimentally validated first neighbours in the interactome using the Guided Connectivity Search. The resulting network was then expanded with the diseases connected via protein-disorder edges (Figure 3B and

full-sized in S4). These diseases included cardiomyopathy, which is an umbrella term for diseases affecting the muscles of the heart that can lead to heart failure. In the expanded network, cardiomyopathy is associated with *SC5A2* via Aquaporin 2 (*AQP2*), a water channel protein expressed in renal collecting ducts. While previous animal studies have shown that upregulated *AQP2* is associated with heart failure [26], *SGLT2*I treatment significantly suppresses its expression [27]. Our analysis hence resulted in the hypothesis that the beneficial effects of *SGLT2*Is could be explained through *AQP2* downregulation.

IV. CONCLUSIONS

NeDRex-Web is a powerful and user-friendly web-interface to interactively explore a large heterogeneous biomedical knowledge graph. With options to tailor queries to a specific use case by engaging expert knowledge, NeDRex-Web facilitates expert-in-the-loop solutions to determine drug repurposing candidates. Through the freedom to vary the starting point of workflows, network-based analysis customization, and result processing and download, NeDRex-Web integrates well into an active, hypothesis-driven research workflow.

V. DISCUSSION AND OUTLOOK

Knowledge graph-based approaches, as demonstrated in this work, provide valuable foundational frameworks for initial

research exploration, while acknowledging the necessity of subsequent in-depth investigations. While the development of NeDRex-Web prioritized an accessible user interface, this effort revealed an inherent tension between comprehensive functionality and ease of use, a common challenge in academic software development. Looking ahead, the emergence of large language models in natural language processing suggests a promising direction for resolving this trade-off. Future iterations of research tools may leverage conversational interfaces to bridge the gap between analytical flexibility and user accessibility.

DATA AVAILABILITY

The NeDRexDB knowledgebase supporting the findings of this study are available via <https://api.nedrex.net/>. Use of the NeDRex-Web is governed by an End User License Agreement (available at <https://web.nedrex.net/#about>). NeDRex-Web's source code is publicly available in GitHub: <https://github.com/repotrial/NeDRex-Web>. Supplementary material is available here: <https://github.com/repotrial/NeDRex-Web/blob/master/material/NeDRex-Web%20-%20BIBM2024%20-%20Supplements.pdf>.

ACKNOWLEDGMENT

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777111. This publication reflects only the authors' view and the European Commission is not responsible for any use that may be made of the information it contains. (J.B., H.H.W.S., A.I.C., S.S., A.M.) This project is funded by the European Union under grant agreement No. 101057619. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Health and Digital Executive Agency (HADEA). Neither the European Union nor the granting authority can be held responsible for them. This work was also partly supported by the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract No. 22.00115. (A.I.C., H.H.W.S., J.B.) This work was supported by the German Federal Ministry of Education and Research (BMBF) within the framework of "CLINSPECT-M" (grant 161L0214A). (O.Z., J.B.) M.L.E. is grateful for financial support from Lundbeckfonden (no. R347-2020-2454). A.I.C. was supported by the DFG Walter Benjamin Program (ref. DFG CA 2642/1-1) and Förderprogramm der Corona-Stiftung im Stifterverband. Figure 1 and 2 are created with BioRender.com.

REFERENCES

- [1] S. Morgan, P. Grootendorst, J. Lexchin, C. Cunningham, and D. Greyson, "The cost of drug development: A systematic review," *Health Policy*, vol. 100, no. 1, pp. 4–17, Apr. 2011.
- [2] H. A. Ghofrani, I. H. Osterloh, and F. Grimminger, "Sildenafil: from angina to erectile dysfunction to pulmonary hypertension and beyond," *Nat. Rev. Drug Discov.*, vol. 5, no. 8, pp. 689–702, Aug. 2006.
- [3] H. A. Mucke, "The case of galantamine: Repurposing and late blooming of a cholinergic drug," *Future Sci. OA*, vol. 1, no. 4, Nov. 2015.
- [4] D. S. Himmelstein et al., "Systematic integration of biomedical knowledge prioritizes drugs for repurposing," *eLife*, vol. 6, Sep. 2017.
- [5] J. Xin et al., "Cross-linking BioThings APIs through JSON-LD to facilitate knowledge exploration," *BMC Bioinf.*, vol. 19, no. 1, pp. 1–7, Feb. 2018.
- [6] S. Sadeh et al., "Exploring the SARS-CoV-2 virus-host-drug interactome for drug repurposing," *Nat. Commun.*, vol. 11, no. 1, pp. 1–9, Jul. 2020.
- [7] M. Hartung et al., "Cancer driver drug interaction explorer," *Nucleic Acids Res.*, vol. 50, no. W1, pp. W138–W144, May 2022.
- [8] A. Maier et al., "Drugst.One — a plug-and-play solution for online systems medicine and network-based drug repurposing," *Nucleic Acids Res.*, vol. 52, no. W1, pp. W481–W488, May 2024.
- [9] K. Gallo et al., "PROMISCUOUS 2.0: a resource for drug-repositioning," *Nucleic Acids Res.*, vol. 49, no. D1, pp. D1373–D1380, Jan. 2021.
- [10] S. Sadeh et al., "Network medicine for disease module identification and drug repurposing with the NeDRex platform," *Nat. Commun.*, vol. 12, no. 1, pp. 1–12, Nov. 2021.
- [11] M. List et al., "KeyPathwayMinerWeb: online multi-omics network enrichment," *Nucleic Acids Res.*, vol. 44, no. W1, pp. W98–W104, Jul. 2016.
- [12] H. Levi, R. Elkon, and R. Shamir, "DOMINO: a network-based active module identification algorithm with reduced rate of false calls," *Mol. Sys. Biol.*, vol. 17, no. 1, p. e9593, Jan. 2021.
- [13] J. Bernett et al., "Robust disease module mining via enumeration of diverse prize-collecting Steiner trees," *Bioinformatics*, vol. 38, no. 6, pp. 1600–1606, Mar. 2022.
- [14] S. D. Ghiassian, J. Menche, and A. L. Barabási, "A Disease Module Detection (DIAMOND) Algorithm Derived from a Systematic Analysis of Connectivity Patterns of Disease Proteins in the Human Interactome," *PLoS Comput. Biol.*, vol. 11, no. 4, Apr. 2015.
- [15] O. Lazareva et al., "BiCoN: network-constrained biclustering of patients and omics data," *Bioinformatics*, vol. 37, no. 16, pp. 2398–2404, Aug. 2021.
- [16] Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen, "Combating Web Spam with TrustRank," in *Proc. 2004 VLDB Conf.*, pp. 576–587, 2004.
- [17] A. Bavelas, "Communication Patterns in Task-Oriented Groups," *J. Acoust. Soc.*, vol. 22, no. 6, p. 725, Jun. 1950.
- [18] D. A. Zarin, T. Tse, R. J. Williams, and S. Carr, "Trial Reporting in ClinicalTrials.gov - The Final Rule," *N Engl J Med*, vol. 375, no. 20, pp. 1998–2004, Nov. 2016.
- [19] K. E. Oliver et al., "Slowing ribosome velocity restores folding and function of mutant CFTR," *The Journal of Clinical Investigation*, vol. 129, no. 12, pp. 5236–5253, Dec. 2019.
- [20] V. Tomati et al., "Genetic inhibition of the ubiquitin ligase Rnf5 attenuates phenotypes associated to F508del cystic fibrosis mutation," *Scientific Reports*, vol. 5, no. 1, p. 12138, Jul. 2015.
- [21] R. P. Tomkiewicz et al., "Amiloride Inhalation Therapy in Cystic Fibrosis: Influence on Ion Content, Hydration, and Rheology of Sputum," *Am. Rev. Respir. Dis.*, vol. 148, no. 4, pp. 1002–1007, Dec. 2012.
- [22] S. Visser, M. Martin, and D. J. Serisier, "Improvements in cystic fibrosis lung disease and airway inflammation associated with etanercept therapy for rheumatoid arthritis: a case report," *Lung*, vol. 190, no. 5, pp. 579–581, Oct. 2012.
- [23] E. M. Schwiebert, N. Kizer, D. C. Gruenert, and B. A. Stanton, "GTP-binding proteins inhibit cAMP activation of chloride channels in cystic fibrosis airway epithelial cells," *Proc. Natl. Acad.*, vol. 89, no. 22, pp. 10623–10627, 1992.
- [24] D. Dolivo, P. Weathers, and T. Dominko, "Artemisinin and artemisinin derivatives as anti-fibrotic therapeutics," *Acta Pharm. Sin. B.*, vol. 11, no. 2, pp. 322–339, Feb. 2021.
- [25] R. Cardoso et al., "SGLT2 inhibitors decrease cardiovascular death and heart failure hospitalizations in patients with heart failure: A systematic review and meta-analysis," *EClinicalMedicine*, vol. 36, no. 100933, p. 100933, Jun. 2021.
- [26] S. Nielsen et al., "Congestive heart failure in rats is associated with increased expression and targeting of aquaporin-2 water channel in collecting duct," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 94, no. 10, pp. 5450–5455, May 1997.
- [27] S. Chung et al., "Empagliflozin contributes to polyuria via regulation of sodium transporters and water channels in diabetic rat kidneys," *Front. Physiol.*, vol. 10, p. 271, 2019.

2.4 DRUGST.ONE

2.4.1 Summary

Drugst.One is a plug-and-play solution bringing network medicine functionalities and DR to any existing web tool (Figure 2.2). This aims to drastically lower the hurdle of development of such features as with only three lines of code and some configuration, a tailored instance can be integrated. Detailed documentation and a dedicated configuration creation tool help with the integration process. Additionally examples of Drugst.One integrations into several different frameworks provided through publicly accessible repositories help with framework-dependent configuration differences. A dedicated “Standalone” page on Drugst.One website allows use of the plugin in a neutral environment now only through manual interaction. Drugst.One Standalone accepts GET and POST requests carrying content and configuration for the on-site Drugst.One plugin. That allows external tools to present users the options to open a configured and filled Drugst.One instance without integrating it directly into their website. It further provides direct links to external resources (such as g:Profiler¹¹, NDEx¹², and DIGEST¹³) forwarding information from the visualized network and interconnecting resources on demand to increase the exploratory depth and interoperability further. The more than 20 projects using Drugst.One already show the versatility and adoption of this plugin. Use cases further outline the general relevance of the implemented tools and functions. The landing page including configuration and integration instructions is available at drugst.one¹⁴.

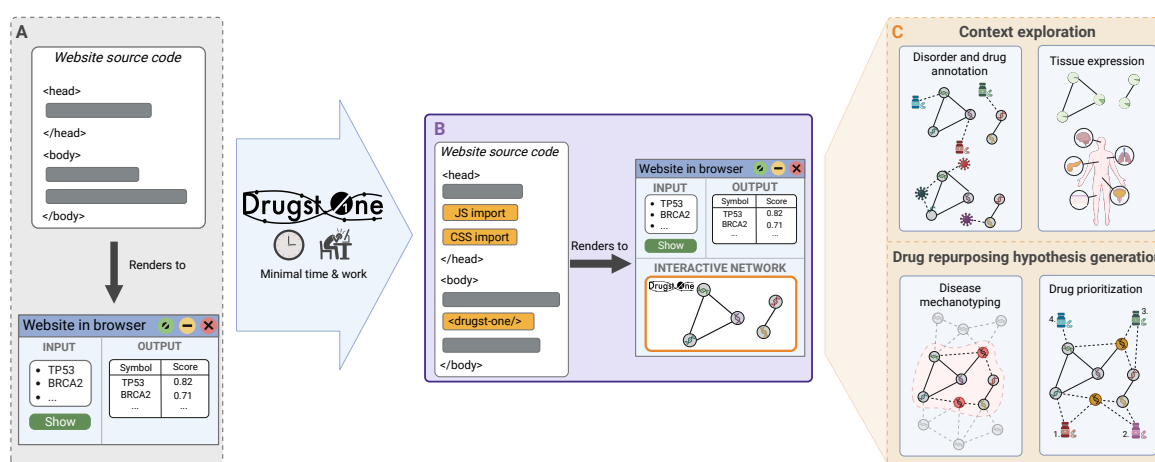


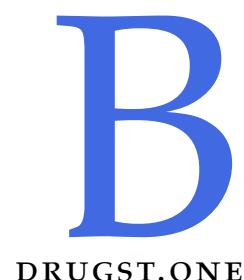
Figure 2.2: Drugst.One (Reprinted from publication B [251] with permission from Oxford University Press): Drugst.One can help to transform any existing website (A) dealing with some list of genes or proteins by adding DR features. (B) With as little as three lines of code, two imports and the use of the `<drugst-one>` tag in the html body for actually rendering the component, Drugst.One provides a full suite of easy to use and highly customizable functionalities for (C) simple context exploration of annotated information, network-driven analyses, disease mechanotyping, drug prioritization and in general disease mechanism or DR hypothesis generation.

11 <https://biit.cs.ut.ee/gprofiler/gost>

12 <https://www.ndexbio.org/iquery>

13 <https://digest-validation.net>

14 <https://drugst.one>



Andreas Maier*, **Michael Hartung***, Mark Abovsky, Klaudia Adamowicz, Gary D. Bader, Sylvie Baier, David B. Blumenthal, Jing Chen, Maria L. Elkjaer, Carlos Garcia-Hernandez, Mohamed Helmy, Markus Hoffmann, Igor Jurisica, Max Kotlyar, Olga Lazareva, Hagai Levi, Markus List, Sebastian Lobentanzer, Joseph Loscalzo, Noel Malod-Dognin, Quirin Manz, Julian Matschinske, Miles Mee, Mhaned Oubounyt, Chiara Pastrello, Alexander R. Pico, Rudolf T. Pillich, Julian M. Poschenrieder, Dexter Pratt, Nataša Pržulj, Sepideh Sadegh, Julio Saez-Rodriguez, Suryadipto Sarkar, Gideon Shaked, Ron Shamir, Nico Trummer, Ugur Turhan, Rui-Sheng Wang, Olga Zolotareva, and Jan Baumbach

“Drugst.One — A plug-and-play solution for online systems medicine and network-based drug repurposing”

Published in: *Nucleic Acids Research*, 2024, Vol. 52, No. W1, W481–W488.
(Web Server Issue). Published May 2024.

DOI: [10.1093/nar/gkae388](https://doi.org/10.1093/nar/gkae388)

Complete Reference:

Maier, A., Hartung, M., Abovsky, M., Adamowicz, K., Bader, G. D., *et al.* (2024). Drugst.One — A plug-and-play solution for online systems medicine and network-based drug repurposing. *Nucleic Acids Research*, 52(W1), W481–W488. <https://doi.org/10.1093/nar/gkae388>

Authors' Contribution: *I worked on all parts of this project equally with Michael Hartung, the shared first author. This includes the planning, prototyping, implementation, and deployment of all associated software components, built on an initial proof-of-concept developed by co-author Julian Matschinske. Equivalently the responsibility of introducing the project to collaborators and helping them include Drugst.One in their projects if necessary, was split equally as were the conceptualization and writing of the publication under consideration of input and feedback from co-authors. Only the Python package implementation was supervised by Michael Hartung predominantly.*

Code Availability: The source code for all components of Drugst.One, including the website, the backend, the plugin, the plugin building and distribution system, the Python package, integration examples, and more are available at github.com/drugst-one/.

Drugst.One — a plug-and-play solution for online systems medicine and network-based drug repurposing

Andreas Maier^{1,*†}, Michael Hartung^{1,†}, Mark Abovsky^{2,3}, Klaudia Adamowicz¹, Gary D. Bader^{4,5,6,7,8}, Sylvie Baier⁹, David B. Blumenthal¹⁰, Jing Chen¹¹, Maria L. Elkjaer^{12,13,14}, Carlos Garcia-Hernandez¹⁵, Mohamed Helmy^{16,17,18,19,20,21}, Markus Hoffmann^{9,22,23}, Igor Jurisica^{2,3,24,25}, Max Kotlyar^{2,3}, Olga Lazareva^{26,27,28}, Hagai Levi²⁹, Markus List⁹, Sebastian Lobentanzer³⁰, Joseph Loscalzo³¹, Noel Malod-Dognin¹⁵, Quirin Manz⁹, Julian Matschinske^{1,9}, Miles Mee^{4,5,6}, Mhaned Oubounyt¹, Chiara Pastrello^{2,3}, Alexander R. Pico³², Rudolf T. Pillich¹¹, Julian M. Poschenrieder^{1,9}, Dexter Pratt¹¹, Nataša Pržulj^{15,33,34}, Sepideh Sadegh^{1,9,35,36}, Julio Saez-Rodriguez³⁰, Suryadipto Sarkar¹⁰, Gideon Shaked²⁹, Ron Shamir²⁹, Nico Trummer⁹, Ugur Turhan¹, Rui-Sheng Wang³¹, Olga Zolotareva^{1,9} and Jan Baumbach^{1,37}

¹Institute for Computational Systems Biology, University of Hamburg, Hamburg, Germany

²Division of Orthopaedic Surgery, Schroeder Arthritis Institute, Toronto, Canada

³Data Science Discovery Centre for Chronic Diseases, Krembil Research Institute, Toronto, ON M5T 0S8, Canada

⁴Department of Molecular Genetics, University of Toronto, Toronto, ON, Canada

⁵The Donnelly Centre, University of Toronto, Toronto, ON, Canada

⁶Department of Computer Science, University of Toronto, Toronto, ON, Canada

⁷Princess Margaret Cancer Centre, University Health Network, Toronto, ON, Canada

⁸The Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Toronto, ON, Canada

⁹Data Science in Systems Biology, TUM School of Life Sciences, Technical University of Munich, Munich, Germany

¹⁰Department Artificial Intelligence in Biomedical Engineering (AIBE), Friedrich-Alexander University Erlangen-Nürnberg (FAU), 91052 Erlangen, Germany

¹¹Department of Medicine, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA

¹²Department of Neurology, Odense University Hospital, Odense, Denmark

¹³Institute of Clinical Research, University of Southern Denmark, Odense, Denmark

¹⁴Institute of Molecular Medicine, University of Southern Denmark, Odense, Denmark

¹⁵Barcelona Supercomputing Center (BSC), 08034 Barcelona, Spain

¹⁶Vaccine and Infectious Disease Organization (VIDO), University of Saskatchewan, Canada

¹⁷School of Public Health, University of Saskatchewan, Canada

¹⁸Department of Computer Science, University of Saskatchewan, Canada

¹⁹Department of Computer Science, Lakehead University, Canada

²⁰Department of Computer Science, Idaho State University, USA

²¹Bioinformatics Institute (BII), A*STAR, Singapore

²²Institute for Advanced Study, Technical University of Munich, Germany

²³National Institute of Diabetes, Digestive, and Kidney Diseases, Bethesda, MD 20892, USA

²⁴Departments of Medical Biophysics and Computer Science, University of Toronto, Toronto, Canada

²⁵Institute of Neuroimmunology, Slovak Academy of Sciences, Bratislava, Slovakia

²⁶Division of Computational Genomics and Systems Genetics, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany

²⁷Junior Clinical Cooperation Unit Multiparametric methods for early detection of prostate cancer, German Cancer Research Center (DKFZ), Heidelberg, Germany

²⁸European Molecular Biology Laboratory, Genome Biology Unit, 69117 Heidelberg, Germany

²⁹Blavatnik School of Computer Science, Tel-Aviv University, Tel-Aviv, Israel

³⁰Heidelberg University, Faculty of Medicine, and Heidelberg University Hospital, Institute for Computational Biomedicine, Bioquant, Heidelberg, Germany

³¹Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA

³²Institute of Data Science and Biotechnology, Gladstone Institutes, 1650 Owens Street, San Francisco, 94158 California, USA

³³Department of Computer Science, University College London, London WC1E 6BT, UK

³⁴ICREA, Pg. Lluís Companys 23, 08010 Barcelona, Spain

Received: February 1, 2024. Revised: April 8, 2024. Editorial Decision: April 25, 2024. Accepted: April 29, 2024

© The Author(s) 2024. Published by Oxford University Press on behalf of Nucleic Acids Research.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License

(<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

³⁵Department of Clinical Genetics, Odense University Hospital, Odense, Denmark

³⁶Clinical Genome Center, Department of Clinical Research, University of Southern Denmark, Odense, Denmark

³⁷Computational Biomedicine Lab, Department of Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark

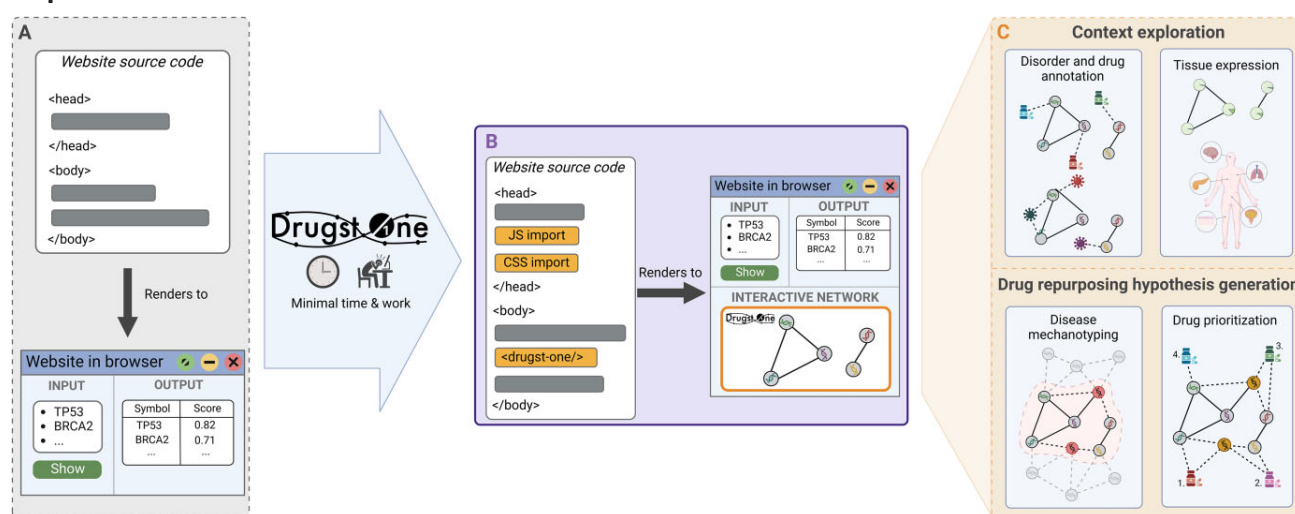
*To whom correspondence should be addressed. Tel: +49 40 42838 7635; Email: andreas.maier-1@uni-hamburg.de

†The first two authors should be regarded as Joint First Authors.

Abstract

In recent decades, the development of new drugs has become increasingly expensive and inefficient, and the molecular mechanisms of most pharmaceuticals remain poorly understood. In response, computational systems and network medicine tools have emerged to identify potential drug repurposing candidates. However, these tools often require complex installation and lack intuitive visual network mining capabilities. To tackle these challenges, we introduce Drugst.One, a platform that assists specialized computational medicine tools in becoming user-friendly, web-based utilities for drug repurposing. With just three lines of code, Drugst.One turns any systems biology software into an interactive web tool for modeling and analyzing complex protein-drug-disease networks. Demonstrating its broad adaptability, Drugst.One has been successfully integrated with 21 computational systems medicine tools. Available at <https://drugst.one>, Drugst.One has significant potential for streamlining the drug discovery process, allowing researchers to focus on essential aspects of pharmaceutical treatment research.

Graphical abstract



Introduction

In recent years, rapid technological advancements and unmet medical needs have fueled the development of computational tools that leverage systems biology methodologies to decipher complex biomedical data (1). These tools frequently target the identification of specific proteins or genes in a given disease context, such as marker genes indicative of disease progression (2,3). The visualization of these results in a biomedical network context can greatly improve their interpretability, allowing us to better understand the underlying disease mechanisms and the interrelationships among the identified entities (4). This principle applies to a variety of biomedical fields, including oncology (5,6), virology (7) and disease subtype identification and patient stratification through differential gene expression analysis (8,9). Rendering these intricate cellular processes as graphs aids researchers in tailoring more precise pharmaceutical treatments, minimizing side effects (10,11) and opening prospects for novel therapeutic and diagnostic strategies, such as mechanistic drug repurposing (12).

Key challenges in the development of systems biology platforms include the integration of comprehensive biomedical data and the creation of flexible graphical user interfaces for data analysis, prioritization, and visualization. Stand-alone

software such as Cytoscape (13) visualizes biological networks but necessitates local installation for each user. To circumvent this, developers often provide online solutions dependent solely on browser compatibility. However, this presents additional hurdles for researchers who may lack sufficient web-development skills and need to establish and maintain an infrastructure, including a server hosting a database and a website. Beyond network visualization, the collection, harmonization, integration, and incorporation of diverse biomedical data demand a significant time investment (14). Moreover, the database should be maintained and regularly updated, a chore that is often not addressed by bioinformatics tools that primarily provide a result overview with a limited set of features. Thus, if network exploration is not neglected due to the additional workload, unique solutions are being developed from scratch, resulting in network visualizers and explorers of varying quality (7,8,15).

We developed Drugst.One to reduce software engineering overhead, bundle development capacities, and to standardize and simplify network analysis and visual network exploration for biomedical web tools. With minimal programming effort, Drugst.One can turn any gene or protein-based systems biology tool into a powerful online toolkit for network integration and visualization, as well as mechanistic drug

repurposing. Drugst.One is a customizable plug-and-play solution for web-application developers in need of a feature-rich network explorer coupled with a biomedical protein-drug-disease network data warehouse. With as little as three lines of code, Drugst.One can be added to any biomedical web tool, highlighting opportunities for drug repurposing and elucidating disease mechanisms. Incorporating multiple state-of-the-art databases (Supplementary Table S2) to complement visualized data, Drugst.One provides an intuitive interface for applying algorithms for exploratory network analyses, drug target, and drug repurposing candidate identification and prioritization. Weekly updates guarantee the relevance of its database for frequently changing data. Currently, Drugst.One is integrated into 21 systems medicine software resources (Supplementary Table S1), including mirDIP (Supplementary Figure S5) (16), pathDIP (17) and WikiPathways (18). In this article, we describe the functionality of Drugst.One and demonstrate its utility on the basis of two studies – on drug repurposing for inflammatory bowel disease (IBD) and on exploring the smooth muscle cell (SMC) proliferation pathway.

Materials and methods

Database

Drugst.One integrates 14 data sources, forming a large, heterogeneous database (Supplementary Table S2). Basic entities in Drugst.One are proteins/genes, drugs, and diseases. The available ID spaces for gene or protein entities are HGNC (19), UniProt (20), Ensemble (21), and Entrez (22). For drugs Drugst.One uses DrugBank and for disorders MONDO (23) identifiers. To describe links between the different entities Drugst.One integrates four different relational layer types, namely protein-protein, protein-drug, protein-disorder, and drug-disorder data.

Using the secondary database NeDRexDB (6), which is updated on a weekly basis, all datasets from NeDRexDB are automatically updated weekly. While all protein-protein and drug-target interaction datasets included in NeDRex are available individually, the NeDRex datasets represent a combination of all individual data sources. Some data sources have restrictive reuse licenses attached, e.g., for use in a commercial scenario. In Drugst.One, we provide both, licensed and openly available datasets, but the access to licensed data has to be unlocked with a configuration parameter.

Input and output

The required input is a list of proteins or genes in HGNC, UniProt, Ensembl or Entrez ID space. On demand, these entities are integrated into the interactome and automatically annotated with clinically relevant information, e.g. targeting drugs or known disease associations. Exploratory functions allow the visualization of known drug indications and disease associations as well as an overlay for tissue-specific expression information. For most information-enriching functions, Drugst.One provides several data sources to choose from (Supplement 3.1). Additionally, network algorithms allow to find and rank directly and indirectly associated drugs.

Algorithms

Drugst.One originates from the network-based drug repurposing platforms CoVex (7) and CADDIE (5), developed for

the application in SARS-CoV-2 and cancer, respectively. While they provide disease-specific information, both tools share underlying principles and algorithms (Supplementary Table S3). These tested and published methods form the Drugst.One algorithmic toolkit (Supplement 3.2). Module identification algorithms provide means to identify additional potential drug targets from the interactome to enrich the mechanistic context. In a second step, drugs that are directly or indirectly linked can be ranked. This allows the assessment of the compound's potential to be repurposed using network-based algorithms. Although both steps work automatically, users can infuse their expert knowledge by adjusting input gene sets. Users can choose among seven drug prioritization and drug target identification algorithms to rank small molecules directly or indirectly targeting disease proteins, thus serving as potential drug repurposing candidates (Supplement 3.2). Information on the execution time scaling behavior of module identification and drug prioritization algorithms can be found in supplementary material (Supplementary Figure S3 and S4, respectively). Overrepresentation analysis using g:Profiler (24) or functional coherence validation using DIGEST (25) on all loaded proteins in the network can be run with one click.

Results

Integration and customization

The Drugst.One ecosystem is a multi-component platform consisting of a website, the web plugin, a server, a content delivery system (CDS) and a Python package (Figure 1, Supplement 1). Drugst.One is free and open to all users and no login is required.

The web plugin can be added to any webpage by importing one JavaScript and one stylesheet file from the <https://cdn.drugst.one> distribution server, and by adding the 'drugst-one' HTML tag to the source code of any system medicine tool's website (Supplementary Figure S2). Features can be customized to a high degree through JSON configuration strings that are passed as attributes. This includes default states of on/off toggles, the network, and the node and edge groups that define the network style. The plugin is responsive to changes during runtime, allowing developers to add buttons or other controls to the host page, for example switching between networks. For seamless integration of the rendered plugin into any website, styling and coloring are controllable by adding specific CSS variables to the website stylesheet (Supplementary Figure S1). To assist developers in the integration process, the Drugst.One website provides conclusive documentation of available parameters, features, and styles. It further offers an interactive configuration page at <https://drugst.one/playground> where configuration options are categorized, and the replication of a configured Drugst.One instance is achieved by simple copy-pasting of the generated code snippets to the developers' websites. This low-code approach allows bioinformatics researchers to provide the community with an interactive mechanism mining web tool within hours or even minutes instead of days. The lightweight Drugst.One JavaScript library connects to the Drugst.One data warehouse server, which handles all the computationally expensive work like data annotation, mapping, and asynchronous algorithm execution.

Alternatively, a standalone integration of Drugst.One is provided at <https://drugst.one/standalone>, which can be

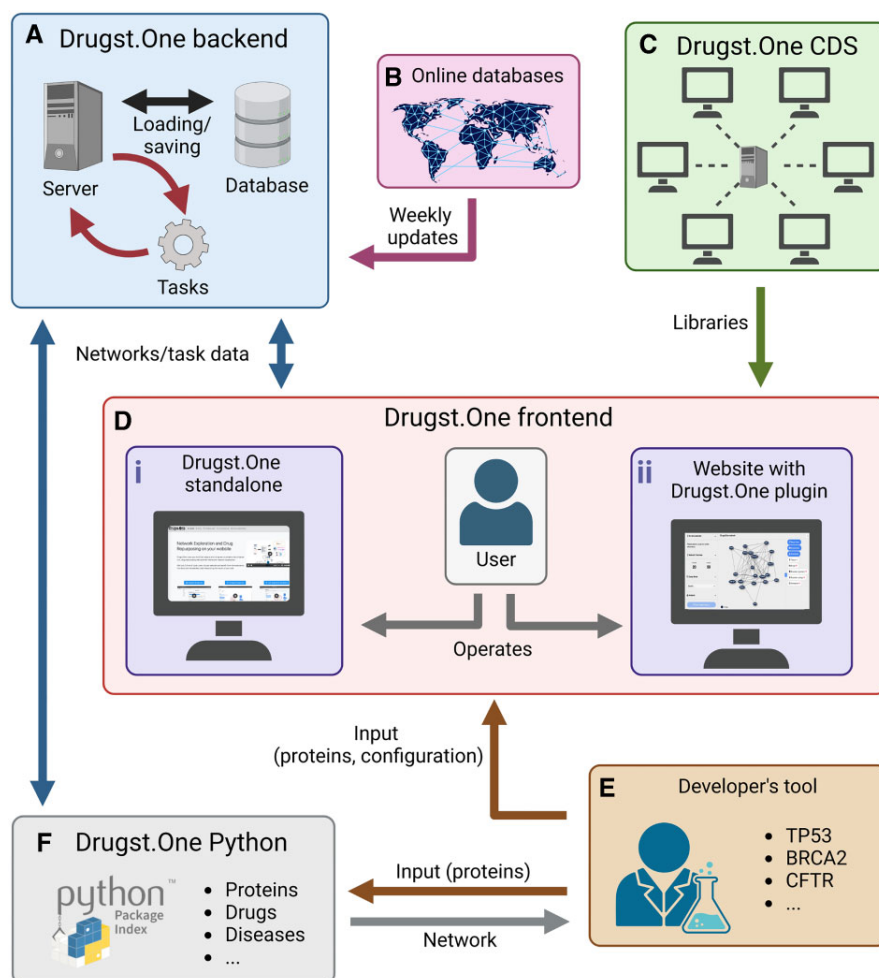


Figure 1. The Drugst.One ecosystem: the Drugst.One server (A) updates weekly from online databases (B), executes computationally demanding tasks, and provides data to the Drugst.One plugin (D i and D ii). The frontend is loaded from the content delivery system (CDS), (C), receives the network data from the developer integrating Drugst.One (E) and presents it to the user. Drugst.One can also be accessed programmatically through a Python package (F).

accessed and customized using URLs or POST-based requests. This way, results from any website or even a command line tool can be redirected to Drugst.One through a simple web service request (Supplement 2). Detailed documentation about all Drugst.One integration options can be found at <https://drugst.one/doc>.

Integration examples

Drugst.One plugin integration with BiCoN

BiCoN (8) is a systems medicine tool for simultaneous patient stratification and disease mechanism identification, i.e. network-based endotyping. BiCoN uses a molecular interaction network as input and identifies two subgroups of patients along with a subnetwork that is enriched for differentially expressed genes between the two groups. These subnetworks can serve as composite biomarkers but may also be enriched for putative drug targets. Since BiCoN also features a web version (<https://exbio.wzw.tum.de/bicon>), we integrated the Drugst.One plugin for enhancing the result presentation by interactively visualizing the identified subnetworks. This allows users to explore possible drug repurposing candidates

targeting the newly identified disease mechanisms, which can subsequently be experimentally validated.

Drugst.One link-out from WikiPathways

WikiPathways (26) is a widely used, community-driven platform for exploring molecular pathways. It allows users to upload, edit, browse, and download a constantly growing pool of pathway datasets. Pathway data can be used to identify and understand key players in metabolism, which is critical for understanding rare or common diseases such as COVID-19 (26). Thus, pathways allow for the prediction of drug target and drug repurposing candidates and are commonly used in the development of new disease treatments (27). When inspecting individual pathways on the WikiPathways platform, users now have the option to forward the pathway genes to the Drugst.One standalone version by clicking a 'Query Drugst.One' link now provided by WikiPathways, located in the search menu of the 'Participants' table. The link redirects the user to the Drugst.One website, visualizing pathway genes, drugs directly targeting them, and offering the complete toolset of Drugst.One. In the following, we give an example

of the Drugst.One usage for exploration of the smooth muscle differentiation and proliferation pathway (WP1991).

WikiPathway WP1991 describes the mechanism behind smooth muscle cell (SMC) differentiation and proliferation. The WikiPathways web interface now incorporates a button to export the pathway genes into Drugst.One (using the magnifier glass in the table showing the proteins participating in the pathway, state 29.01.24) and visualizing their interactions with drugs in Drugst.One directly. To gain a general overview of the complications (symptoms, comorbidities, etc.) associated with dysfunctional SMC development and their drivers, Drugst.One allows for extending the WikiPathways-exported network by the corresponding disorders and their associated pathway genes. Several disease nodes appear in the network, mainly representing various cardiovascular disorders (CVDs), e.g. cardiomyopathy, coronary artery disease, and aortic valve disease (Figure 2). Further, genes with a strong link to vascular function are present: NKX2-5 is associated with heart and cardiac muscle development (28); KLF4 is involved in cytoskeletal stabilization of vascular SMCs (29); MYOCD is critical for cardiovascular development and vascular SMC autophagy regulation (30). This leads to the assumption that the mechanism vascular SMC proliferation at least partially overlaps with WP1991, the pathway of SMC. The importance of SMCs for proper vascular functionality (31,32) and thus to atherosclerosis, hypertension, myocardial infarction, and other cardiovascular diseases was reported before (33–35). An isolated subnetwork community of proteins is formed by the three myocyte enhancer factors MEF2A, MEF2C, and MEF2D. Besides their obvious cardiovascular implications, these factors play a role in neurological processes (36). An impact of SMCs on status epilepticus was shown in mouse models (37), and a connection between migraine and SMC dysfunction was suggested as well (38,39).

Drugst.One allows for the projection of (gene) expression data from GTEx on the proteins in the network. The relative expression of these genes appears to be quite high in arteries, indicating again a relevancy of these genes in vascular SMC. Further, the expression also seems to be elevated in organs that have to perform physical motion, like heart, lung, bladder and skeletal muscles, but with observable fluctuations between the tissues in the relative expression of genes like CCND2.

With one mouse click, we import drug target information for drug repurposing candidate prediction. Despite SMCs relation to cardiovascular diseases, no corresponding CVD drugs have been identified. Mainly anti-cancer drugs (e.g. sunitinib, erlotinib, midostaurin and ruxolitinib) targeting calcium/calmodulin-dependent protein kinase II delta (CAMK2D), which is associated with cancer growth (40), are found. Notably, however, CAMK2D also plays a role in calcium signaling, which is essential for the upkeep of SMC function (41,42). Hence, this may explain the observed cardiovascular side effects of CAMK2D-targeting drugs. According to SIDER (43), sunitinib may cause hypertensive symptoms and corresponding studies suggest that midostaurin has cardiotoxic effects (44). Algorithms integrated in Drugst.One can extend the search space by looking for indirectly connected drugs. The selection menu offers a function to automatically add all displayed proteins to the selection, serving as the starting point (seeds) of subsequent searches. The harmonic centrality algorithm (Supplement 3.2.3) was used to extend the network by the ten drugs with the highest score, including indirectly (transitively) connected drugs from the NeDRex

database. Through this search, the tyrosine kinase inhibitor nintedanib, which has shown promising effects in pulmonary arterial smooth muscle cells and intestinal smooth muscle cells (45,46), can be identified.

This shows the identification potential of mechanism-associated drugs through the network-based drug repurposing functions Drugst.One incorporates. Whereas before only drugs primarily used in cancer were present through direct association with SMC pathway participants, Drugst.One suggested more relevant options for this context.

Due to the weekly automatic updates of the Drugst.One data warehouse, analyses performed on the Drugst.One website may not always be fully reproducible. For instances where long-term reproducibility is prioritized over the most current data, we offer an alternative at <https://stable.drugst.one/>, which receives updates only on the first day of each year. Similarly, a Drugst.One plugin is available that connects to this stable database.

More Drugst.One use cases

Other exemplary case studies using Drugst.One on more diseases can be found in the supplementary material section 4. We recreated a drug repurposing example case proposed by Sadeh et al. 2021 on inflammatory bowel disease (6) using the same seed genes (Supplementary Table S4) and matching algorithms (Supplementary Table S5) and through Drugst.One we were able to identify even more promising treatment options than the original study did (Supplementary Table S6). Using the integrated version of Drugst.One in the web tool mirDIP, we explored the IBD case from the angle of microRNA targets. A short report about ROBUST-Web, a web application for module detection, that uses Drugst.One as primary tool for visualization can be found in section 5 of the supplementary material.

Discussion

Biomedical research generates a wealth of data that could inform the development of novel therapies or treatments. However, despite this potential, a significant portion of the analyses conducted in this field fail to translate into clinical trials, leading to major issues in the effectiveness of public health research (47). To this end, Drugst.One has the potential to help transform specifically omics-based research results into actionable hypotheses with potential clinical impact by closing the gap between disease mechanism mining and hypothesis generation for drug repurposing. Drugst.One offers a community-driven solution to streamline the knowledge distributed over many online resources for multi-omics analyses and other biomedical tools (48) to turn the results of biomedical analyses into concrete candidate drug targets and drug repurposing hypotheses. Still, we emphasize that the drug target and drug repurposing predictions are merely candidates and supervision with expert knowledge is still required before experimental validation. Drugst.One delivers explainable indications based on established biological data like expression and known disease associations or drug indications, however, the interpretation of their application in the case-specific context is up to the user. Therefore, we designed Drugst.One to be operated with maximal transparency and allow optional user input for every step of the analysis.

With the infrastructure and the resources being provided, Drugst.One helps to find a community-wide solution for

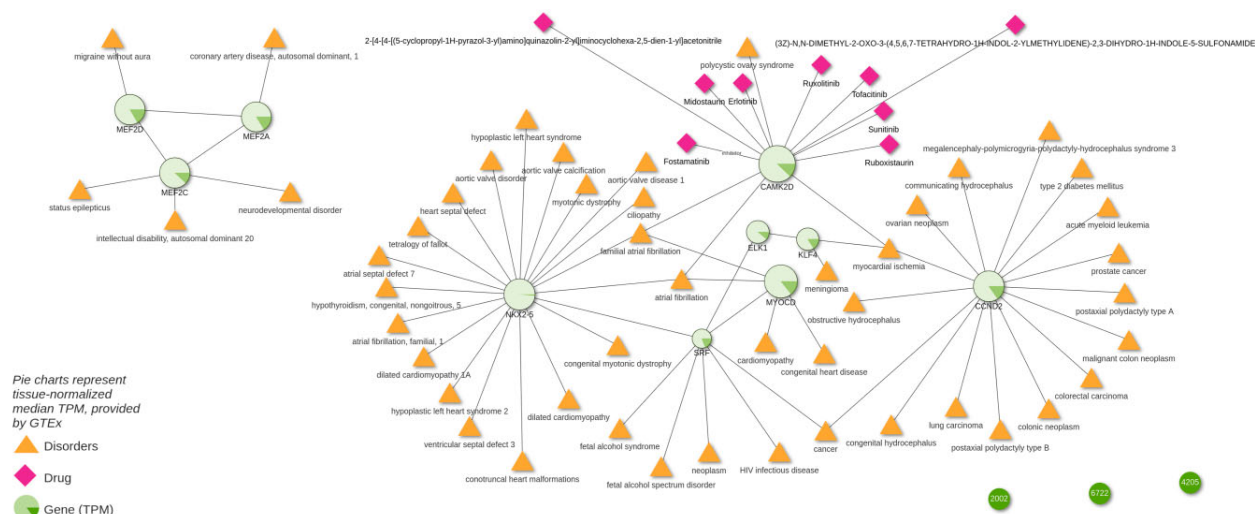


Figure 2. Participants of WikiPathway WP1991 displayed in Drugst.One. Adjacent diseases and drugs are enabled, as well as diseases linked to drugs targeting this smooth muscle cell (SMC) and differentiation pathway. To investigate if there is the chance that WP1991 also represents vascular SMC proliferation, normalized median expression values for 'Artery - Aorta' are overlaid as pie charts, where 360° represent the maximum observed transcripts per million (TPM) in the selected tissue and all other TPMs are exponentially scaled.

standardization and streamlining the visualization of explainable disease modules and their pharmacological implications. Drugst.One provides various interfaces to be highly accessible and customizable by all members of the community while maintaining up-to-date database information and network analysis algorithms. Smooth integration into most biomedical websites and tools is confirmed by 21 resources already integrating Drugst.One. For future developers who wish to customize Drugst.One before its integration, an interactive web interface provides copy-paste-able code for customized plugin integration with their own website. An endpoint for developers who want to link out from any of their websites, apps, or command line tools is provided by Drugst.One as well. By default, all endpoints connect to the database that is updated weekly. However, we provide adjusted versions or configuration options to use the stable database, which is updated only once a year, to facilitate long-term reproducibility of results.

Drugst.One complies with community standards regarding data management as defined by the FAIRness principles (49). Download links for any data shown in Drugst.One are provided at any step, whether it is a table with drug target and drug candidates or the visualized network with all activated extensions like expression information. Export to current community standards is supported via exporting compatible .graphml files, which can be loaded directly into, e.g. Cytoscape (13). To further increase reproducibility and interoperability, concrete plans to implement community standards, such as the export of Drugst.One networks to NDEx (50,51) are made (Supplementary Table S7). In summary, Drugst.One offers an important service to the systems medicine research community to tackle the widely recurring problem of web-based disease mechanism mining and drug repurposing candidate prediction by capturing the results of biomedical assays.

Data availability

The authors declare that all data supporting the findings of this study are available publicly and their integration is described accordingly within the pa-

per and its supplementary information file. All code of the Drugst.One platform is available on GitHub (<https://github.com/drugst-one>) and on Zenodo (plugin: <https://doi.org/10.5281/zenodo.11073055>, backend: <https://doi.org/10.5281/zenodo.11073096>, Python package: <https://doi.org/10.5281/zenodo.11073116>, integration examples: <https://doi.org/10.5281/zenodo.11073112>, Django template: <https://doi.org/10.5281/zenodo.11073104>, runtime evaluation: <https://doi.org/10.5281/zenodo.11073073>).

Supplementary data

Supplementary Data are available at NAR Online.

Acknowledgements

The Graphical abstract and Figure 1 and Supplementary Figures S1, S2 and S5 were created with BioRender.com.

Author contributions: A.M. and M.H. developed and implemented the Drugst.One ecosystem, and wrote the manuscript. J.B. and O.Z. supervised the project and provided continuous feedback and ideas during the development. The rest of the Drugst.One Initiative integrated the Drugst.One ecosystem into their platforms, tested, provided helpful feedback, and revised the manuscript. All authors agreed to the final version of the manuscript.

Funding

REPO-TRIAL: this project has received funding from the European Union's Horizon 2020 research and innovation programme [777111]; this publication reflects only the authors' view and the European Commission is not responsible for any use that may be made of the information it contains; RePo4EU: this project is funded by the European Union [101057619]; views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Health and Digital Executive Agency (HADEA). Neither the European

Union nor the granting authority can be held responsible for them; Swiss State Secretariat for Education, Research and Innovation (SERI) [22.00115]; German Federal Ministry of Education and Research (BMBF) within the framework of ‘CLINSPECT-M/-2’ [F031L0214A, 161L0214A, 16LW0243K]; Technical University Munich – Institute for Advanced Study, funded by the German Excellence Initiative; Intramural Research Programs (IRPs) of the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK); Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) [422216132]; J.B. was partially funded by his VILLUM Young Investigator Grant [13154]; European Research Council (ERC) Consolidator Grant [770827]; Spanish State Research Agency AEI 10.13039/501100011033 [PID2019-105500GB-I00]; I.J. was supported in part by funding from Natural Sciences Research Council [NSERC #203475], Canada Foundation for Innovation [CFI #225404, #30865]; Ontario Research Fund [RDI #34876, RE010-020]; IBM and Ian Lawson van Toch Fund; S.L. has received funding from the European Union’s Horizon 2020 research and innovation programme [965193] for DECIDER. Funding for open access charge: Horizon Europe project Repo4EU.

Conflict of interest statement

J.S.R. reports funding from GSK, Pfizer and Sanofi and fees from Travers Therapeutics and Astex Pharmaceuticals.

References

- Hufsky, F., Lamkiewicz, K., Almeida, A., Aouacheria, A., Arighi, C., Bateman, A., Baumbach, J., Beerenwinkel, N., Brandt, C., Cacciabue, M., *et al.* (2021) Computational strategies to combat COVID-19: useful tools to accelerate SARS-CoV-2 and coronavirus research. *Brief. Bioinform.*, **22**, 642–663.
- Gordon, D.E., Jang, G.M., Bouhaddou, M., Xu, J., Obernier, K., O’Meara, M.J., Guo, J.Z., Swaney, D.L., Tummino, T.A., Hüttenhain, R., *et al.* (2020) A SARS-CoV-2-human protein-protein interaction map reveals drug targets and potential drug-repurposing. *Nature*, **583**, 459–468.
- Zolotareva, O. and Kleine, M. (2019) A survey of gene prioritization tools for mendelian and complex human diseases. *J. Integr. Bioinform.*, **16**, 20180069.
- Hütter, C.V.R., Sin, C., Müller, F. and Menche, J. (2022) Network cartographs for interpretable visualizations. *Nat. Comput. Sci.*, **2**, 84–89.
- Hartung, M., Anastasi, E., Mamdouh, Z.M., Nogales, C., Schmidt, H.H.H.W., Baumbach, J., Zolotareva, O. and List, M. (2022) Cancer driver drug interaction explorer. *Nucleic Acids Res.*, **50**, W138–W144.
- Sadegh, S., Skelton, J., Anastasi, E., Bennett, J., Blumenthal, D.B., Galindez, G., Salgado-Albarrán, M., Lazareva, O., Flanagan, K., Cockell, S., *et al.* (2021) Network medicine for disease module identification and drug repurposing with the NeDRex platform. *Nat. Commun.*, **12**, 6848.
- Sadegh, S., Matschinske, J., Blumenthal, D.B., Galindez, G., Kacprowski, T., List, M., Nasirigerdeh, R., Oubounyt, M., Pichlmair, A., Rose, T.D., *et al.* (2020) Exploring the SARS-CoV-2 virus-host-drug interactome for drug repurposing. *Nat. Commun.*, **11**, 3518.
- Lazareva, O., Canzar, S., Yuan, K., Baumbach, J., Blumenthal, D.B., Tieri, P., Kacprowski, T. and List, M. (2020) BiCoN: Network-constrained biclustering of patients and omics data. *Bioinformatics*, **37**, 2398–2404.
- Zolotareva, O., Khakabimamaghani, S., Isaeva, O.I., Chervontseva, Z., Savchik, A. and Ester, M. (2021) Identification of differentially expressed gene modules in heterogeneous diseases. *Bioinformatics*, **37**, 1691–1698.
- Nelissen, E., van Hagen, B.T.J., Argyrousi, E.K., van Goethem, N.P., Heckman, P.R.A., Paes, D., Mulder-Jongen, D.A.J., Ramaekers, J.G., Blokland, A., Schmidt, H.H.H.W., *et al.* (2022) Soluble guanylate cyclase stimulator riociguat improves spatial memory in mice via peripheral mechanisms. *Neurosci. Lett.*, **788**, 136840.
- Casas, A.I., Hassan, A.A., Larsen, S.J., Gomez-Rangel, V., Elbatreek, M., Kleikers, P.W.M., Guney, E., Egea, J., López, M.G., Baumbach, J., *et al.* (2019) From single drug targets to synergistic network pharmacology in ischemic stroke. *Proc. Natl. Acad. Sci. U.S.A.*, **116**, 7129–7136.
- Goh, K.-I., Cusick, M.E., Valle, D., Childs, B., Vidal, M. and Barabási, A.-L. (2007) The human disease network. *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 8685–8690.
- Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.
- Lobentanz, S., Aloy, P., Baumbach, J., Bohar, B., Carey, V.J., Charoentong, P., Danhauser, K., Doğan, T., Dreoj, J., Dunham, J., *et al.* (2023) Democratizing knowledge representation with BioCypher. *Nat. Biotechnol.*, **41**, 1056–1059.
- Levi, H., Elkon, R. and Shamir, R. (2021) DOMINO: a network-based active module identification algorithm with reduced rate of false calls. *Mol. Syst. Biol.*, **17**, e9593.
- Hauschild, A.-C., Pastrello, C., Ekaputeri, G. K.A., Bethune-Waddell, D., Abovsky, M., Ahmed, Z., Kotlyar, M., Lu, R. and Jurisica, I. (2023) MirDIP 5.2: tissue context annotation and novel microRNA curation. *Nucleic Acids Res.*, **51**, D217–D225.
- Pastrello, C., Kotlyar, M., Abovsky, M., Lu, R. and Jurisica, I. (2024) PathDIP 5: improving coverage and making enrichment analysis more biologically meaningful. *Nucleic Acids Res.*, **52**, D663–D671.
- Pico, A.R., Kelder, T., van Iersel, M.P., Hanspers, K., Conklin, B.R. and Evelo, C. (2008) WikiPathways: pathway editing for the people. *PLoS Biol.*, **6**, e184.
- Shows, T.B. and McAlpine, P.J. (1978) The catalog of human genes and chromosome assignments. A report on human genetic nomenclature and genes that have been mapped in man. *Cytogenet. Cell Genet.*, **22**, 132–145.
- The UniProt Consortium (2017) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **45**, D158–D169.
- Cunningham, F., Allen, J.E., Allen, J., Alvarez-Jarreta, J., Amode, M.R., Armean, I.M., Austine-Orimoloye, O., Azov, A.G., Barnes, I., Bennett, R., *et al.* (2022) Ensembl 2022. *Nucleic Acids Res.*, **50**, D988–D995.
- Maglott, D., Ostell, J., Pruitt, K.D. and Tatusova, T. (2007) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **35**, D26–D31.
- Mungall, C.J., McMurtry, J.A., Köhler, S., Balhoff, J.P., Borromeo, C., Brush, M., Carbon, S., Conlin, T., Dunn, N., Engelstad, M., *et al.* (2017) The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Res.*, **45**, D712–D722.
- Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H. and Vilo, J. (2019) g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.*, **47**, W191–W198.
- Adamowicz, K., Maier, A., Baumbach, J. and Blumenthal, D.B. (2022) Online in silico validation of disease and gene sets, clusterings or subnetworks with DIGEST. *Brief. Bioinform.*, **23**, bbac247.
- Martens, M., Ammar, A., Riutta, A., Waagmeester, A., Slenter, D.N., Hanspers, K., A Miller, R., Digles, D., Lopes, E.N., Ehrhart, F., *et al.* (2021) WikiPathways: connecting communities. *Nucleic Acids Res.*, **49**, D613–D621.
- Akhurst, R.J. and Hata, A. (2012) Targeting the TGFβ signalling pathway in disease. *Nat. Rev. Drug Discov.*, **11**, 790–811.

28. Lints,T., Parsons,L., Hartley,L., Lyons,I. and Harvey,R. (1993) Nkx-2.5: a novel murine homeobox gene expressed in early heart progenitor cells and their myogenic descendants.. *Development*, **119** 2, 419–431.
29. Liu,Y., Zheng,B., Zhang,X., Juan Nie,C., hui Li,Y. and Wen,J. (2013) Localization and function of KLF4 in cytoplasm of vascular smooth muscle cell.. *Biochem. Biophys. Res. Commun.*, **436**, 162–168.
30. Shi,D., Ding,J., Xie,S., Huang,L., Zhang,H., Chen,X., Ren,X., Zhou,S., He,H., Ma,W., Zhang,T., *et al.* (2022) Myocardin/microRNA-30a/Beclin1 signaling controls the phenotypic modulation of vascular smooth muscle cells by regulating autophagy. *Cell Death Dis.*, **13**, 121.
31. Steucke,K.E., Tracy,P.V., Hald,E.S., Hall,J.L. and Alford,P.W. (2015) Vascular smooth muscle cell functional contractility depends on extracellular mechanical properties. *J. Biomech.*, **48**, 3044–3051.
32. Jaminon,A., Reesink,K., Kroon,A. and Schurgers,L. (2019) The role of vascular smooth muscle cells in arterial remodeling: focus on calcification-related processes. *Int. J. Mol. Sci.*, **20**, 5694.
33. Zhuge,Y., Zhang,J., Qian,F., Wen,Z., Niu,C., Xu,K., Ji,H., Rong,X., Chu,M. and Jia,C. (2020) Role of smooth muscle cells in cardiovascular disease. *Int. J. Biol. Sci.*, **16**, 2741–2751.
34. Liu,M. and Gomez,D. (2019) Smooth muscle cell phenotypic diversity. *Arterioscler. Thromb. Vasc. Biol.*, **39**, 1715–1723.
35. Allahverdian,S., Chaabane,C., Boukais,K., Francis,G.A. and Bochaton-Piallat,M.-L. (2018) Smooth muscle cell fate and plasticity in atherosclerosis. *Cardiovasc. Res.*, **114**, 540–550.
36. Chaudhary,R., Agarwal,V., Kaushik,A.S. and Rehman,M. (2021) Involvement of myocyte enhancer factor 2c in the pathogenesis of autism spectrum disorder. *Heliyon*, **7**, e06854.
37. Cozart,M.A., Phelan,K.D., Wu,H., Mu,S., Birnbaumer,L., Rusch,N.J. and Zheng,F. (2020) Vascular smooth muscle TRPC3 channels facilitate the inverse hemodynamic response during status epilepticus. *Sci. Rep.*, **10**, 812.
38. Napoli,R., Guardasole,V., Zarra,E., Matarazzo,M., D'Anna,C., Saccà,F., Affuso,F., Cittadini,A., Carrieri,P.B. and Saccà,L. (2009) Vascular smooth muscle cell dysfunction in patients with migraine. *Neurology*, **72**, 2111–2114.
39. Napoli,R., Guardasole,V., Zarra,E., De Sena,A., Saccà,F., Ruvolo,A., Grassi,S., Giugliano,S., De Michele,G., Cittadini,A., Carrieri,P.B. and Saccà,L. (2013) Migraine attack restores the response of vascular smooth muscle cells to nitric oxide but not to norepinephrine. *World J Cardiol.*, **5**, 375–381.
40. He,Q. and Li,Z. (2021) The dysregulated expression and functional effect of CaMK2 in cancer. *Cancer Cell Int.*, **21**, 326.
41. Hill-Eubanks,D.C., Werner,M.E., Heppner,T.J. and Nelson,M.T. (2011) Calcium signaling in smooth muscle. *Cold Spring Harb. Perspect. Biol.*, **3**, a004549.
42. Adelstein,R.S. and Sellers,J.R. (1987) Effects of calcium on vascular smooth muscle contraction. *Am. J. Cardiol.*, **59**, 4B–10B.
43. Kuhn,M., Letunic,L., Jensen,L.J. and Bork,P. (2016) The SIDER database of drugs and side effects. *Nucleic Acids Res.*, **44**, D1075–D1079.
44. Giudice,V., Vecchione,C. and Selleri,C. (2020) Cardiotoxicity of novel targeted hematological therapies. *Life*, **10**, 344.
45. Tsutsumi,T., Nagaoka,T., Yoshida,T., Wang,L., Kuriyama,S., Suzuki,Y., Nagata,Y., Harada,N., Kodama,Y., Takahashi,F., *et al.* (2019) Nintedanib ameliorates experimental pulmonary arterial hypertension via inhibition of endothelial mesenchymal transition and smooth muscle cell proliferation. *PLoS One*, **14**, e0214697.
46. Kataria,J., Kerr,J., Lourenssen,S.R. and Blennerhassett,M.G. (2022) Nintedanib regulates intestinal smooth muscle hyperplasia and phenotype in vitro and in TNBS colitis in vivo. *Sci. Rep.*, **12**, 10275.
47. Drolet,B.C. and Lorenzi,N.M. (2011) Translational research: understanding the continuum from bench to bedside. *Transl. Res.*, **157**, 1–5.
48. Luo,J., Wu,M., Gopukumar,D. and Zhao,Y. (2016) Big Data application in biomedical research and health care: a literature review. *Biomed. Inform. Insights*, **8**, 1–10.
49. Wilkinson,M.D., Dumontier,M., Aalbersberg,I. J.J., Appleton,G., Axton,M., Baak,A., Blomberg,N., Boiten,J.-W., da Silva Santos,L.B., Bourne,P.E., *et al.* (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data*, **3**, 160018.
50. Pratt,D., Chen,J., Pillich,R., Rynkov,V., Gary,A., Demchak,B. and Ideker,T. (2017) NDEx 2.0: a clearinghouse for research on cancer pathways. *Cancer Res.*, **77**, e58–e61.
51. Pillich,R.T., Chen,J., Churas,C., Liu,S., Ono,K., Otasek,D. and Pratt,D. (2021) NDEx: accessing network models and streamlining network biology workflows. *Curr. Protoc.*, **1**, e258.

DISEASE GENE IDENTIFICATION THROUGH EPISTASIS DETECTION

3.1 INTRODUCTION

While network medicine approaches promise a reasonable way to bring the systems medicine idea into practical research application, big issues remain unsolved. One central problem is the identification of reliable and representable disease genes that serve as a starting point of a network-driven exploration or prediction.

In general, methods used to identify disease genes can be categorized into two main groups: Knowledge-driven and data-driven. Data-driven approaches extract insights about specific condition-gene relationships *de novo* from data and measurements representative for patient groups with the relevant condition, whereas knowledge-driven approaches focus on resources containing direct or aggregated results from data-driven insights.

Knowledge-driven approaches, such as LLM-supported, literature-grounded search engines like Consensus.AI¹ and perplexity.ai² can provide valuable tools for identifying potential seed sets of disease-associated genes distilled from information found in publications on PubMed, GoogleScholar, and more, but suffer from potential study or visibility bias and hallucinations. More structured, already aggregated information can be found in many different databases all with their individual focus or sources. OMIM focuses specifically on linking disease pheno- and genotypes, similarly to DisGeNET, a database that is no longer openly available. Apart from generalistic databases, more focused resources exist, e.g., specifically for cancer, the most prevalent mutation-driven disorder, namely COSMIC [252], IntOGen [253], or NCG [254]. Additionally, previously mentioned network-based databases and KGs, such as Hetionet [125], OpenTargets [126], and NeDRex [129], can be leveraged by using disease-gene association edges as the required information.

One problem that usually arises when trying to integrate information from these or other databases with additional datasets and disease information is the incompatibility between certain disease ID spaces. As the origin and reason for most disease classification systems is widely diverse they are in many cases highly incompatible; translating between them can destroy any signal that is present in the data [90]. Systems like ICD are developed for disease classification and reimbursements in the healthcare system, therefore they are highly biased towards symptoms and treatments, rather than for example the OMIM ID spaces, that try to mechanistically classify disorders [255, 256]. There are projects like MONDO or HPO [257], that try to at least provide the option to map between those ID spaces facilitating data integration on a technical level. This

¹ <https://consensus.ai>

² <https://perplexity.ai>

further means, that, based on the database disease genes stem from, the definition of a certain disease might differ strongly, such that the resulting disease genes can only be used while taking into account their source.

One conclusion of Sadegh *et al.* [90], when assessing the impact of disease ID spaces on network medicine research was that integrating knowledge-driven disease information with experimental data-grounded information can mitigate or avoid issues introduced by disease definition mismatches. This can easily be done with omics driven approaches. A standard approach is differential analysis, e.g. in genomics differential gene expression using DESeq [258], DESeq2 [194], or limma [259]. This way, genomic changes between groups of samples, often diseased and healthy, can be quite clearly identified. Differentially expressed genes can give a good hint about which molecular processes are perturbed in the diseased condition, but gene expression is for the most part still just a phenotype of an underlying cause.

Additionally, the use of omics-driven approaches for disease gene identification is mainly limited to computationally-savvy researches, as mostly approaches are published in the form of code or an executable, with some exceptions like ExpressAnalyst³ and iDEP.96⁴ for differential expression analysis. While providing the expected results in the form of differential expressed gene sets, these tools are often not integrated into other resources that would allow the exploration of the results within the same graphical user interface.

Genetics-driven approaches, like genome-wide association study (GWAS) try to go a step further and link single nucleotide variants (SNVs; also called single nucleotide polymorphisms (SNPs)), meaning point mutations in the genome, to the observed phenotype. While there are some cases, where one variant alone can fully explain the cause of the phenotype, like SNP rs334 in Sickle Cell Anemia, in many cases establishing causal links is not that trivial. This for once comes from the pleiotropic properties of many genes, where a mutation of a single gene has effects on multiple molecular processes and thus phenotypes. Additionally, there are mutations classified not as causally related but as disease predisposing or disease modifying mutations [260, 261]. For example, Gaucher disease, first classified in 1882, is a very rare disorder that can have a wide range of phenotypes with the exact disease modifying causes still being not completely understood [261].

The entirety and complexity around causal interactions and influences within our genome is yet to be fully understood. Still, it seems surprising that no undisputed combination of more than three SNPs has been identified yet that causes a phenotype, while being non-causal individually. This field of research is, as previously introduced, known as epistasis detection and it is driven by the observation that in many genetic disorders the heritability pattern cannot be explained well with current knowledge. As this turns the single SNP associations of GWAS into a SNP combination association problem, each SNP-set co-observed in patients would require evaluation regarding their probability of being jointly responsible for a condition, when observed in a single human. This of course can, depending on the condition, not only be two SNPs but many. As there are still no undisputed and replicated, higher-order cases of epistatic interactions causing human phenotypes, it is unclear how many SNPs could be in epistasis and how to best evaluate, if they are.

³ <https://expressanalyst.ca>

⁴ <http://bioinformatics.sdstate.edu/idep96/>

One major resource for SNP information is NCBI's dbSNP, holding a staggering 1.1 billion unique SNPs⁵ [262]. The combinatorial explosion of the search-space arising from having to test all SNP pairs, triplets, quadruples, and so on, is apparent. The 1000 Genomes project uncovered that compared to a reference genome, the median human genome contains around 4.31 million autosomal SNPs and identified a combined number of 84.7 million unique SNPs [263]. In this more than ten year old project alone, all possible sets of ten SNPs available would lead to just shy of 2×10^{79} individual sets, a number so large that it equates to 19% of the number of atoms in the observable universe. Unsurprisingly, this is impossible to compute in any reasonable time with current and possible even near future technology. Therefore, existing tools limit the search space to SNPs present in the large-scale GWAS-datasets required for such analyses, often staying between 500 thousand and one million unique SNPs. Still, tools like Potpourri [264], LinDen [265], MACOED [266], and BiologicalEpistasis [267] are limited to the computation of only very limited combinatorial sizes, mainly only pairs of SNPs. Only PoCos [268] is able to produce SNP sets larger than that in a reasonable time, which is possible through the use of a heuristic approach to the problem. Newer approaches trying to leverage the increasing capabilities of deep neural networks show some potential, but still leave the general problem unsolved [269]. Another general limitation is that it is still unclear which statistical method provides the best quantitative or qualitative measure, due to the lack of any verified and undisputed case of larger numbers of SNPs in epistasis.

Current epistasis detection tools come with the hurdle of requiring substantial processing power which often is on the level of super-computers. Especially when trying to compute higher order set-sizes, meaning sets with more than two SNPs, existing methods reach current computational limits, rendering systematic computation of epistasis candidate sets impossible. This is probably one reason, why resources for general GWAS results exist, like e.g. the GWAS Catalog⁶, but there is no comprehensive, precomputed online database or resource for human epistasis candidate sets. In absence of such a resource that provides probable epistasis sets, a direct integration as the starting point for DR research is not easily possible, limiting the possibility to research. To solve this, a novel epistasis detection method is required, allowing for deeper computation, to be able to build and publish epistasis candidate sets in a first-of-its-kind platform and connect it to DR research.

3.2 RESULTS

While having an immense potential to revolutionize the causal definition of genetically driven diseases, epistasis detection fails to make significant progress. To advance the research of systematic, larger epistasis set computation as the basis for creating a human epistasis set resource and solving the established *Bottleneck 2: The seed gene knowledge gap*, we introduce the **EpistasisDiseaseAtlas**, driven by **NeEDL (Publication C)** [212], a novel epistasis detection approach, leveraging network-medicine ideas to adequately reduce the search space in a mechanistically founded way (Figure 3.1). By prioritizing the search for interacting SNPs that are located within or near genes that are themselves close together in the human protein-protein interaction network, NeEDL introduces a biologically informed heuristic. This reduction of the searchable space is combined with the local-search algorithm, which iteratively searches for promising epistatic interaction sets. Further, with QC hardware on the verge of becoming practically usable, preparing our approach

⁵ State: April 2026

⁶ <https://www.ebi.ac.uk/gwas>

for this technology was an essential goal, as certain computations can be achieved with QC solutions that cannot reasonably be computed on traditional hardware. As local-search can be reduced to a QUBO problem, which can be solved through quantum annealing (a specific version of QC) this quantum-ready version was implemented. In an attempt to open up the detection and analysis of SNP sets in epistasis to the general research community, the **Epistasis Disease Atlas (EDA)** is a first-of-its-kind online resource providing access to all findings of processing eight large GWAS datasets and allowing the recomputation of individual set scores was a major goal. Additionally, functions for the seamless exploration and hypothesis generation regarding mechanistic relationships between the genes of the SNPs, coming from the underlying SNP-SNP interaction network, was a core concern to advance the idea of relevant disease gene identification, serving as a starting point for potential DR angles. The integration with DR is achieved by using the previously presented **Drugst.One** infrastructure. Instead of directly integrating the plugin into EDA, we decided to just redirect to the Drugst.One Standalone page, as the EDA explorer page is already informative enough. Still, this way it allows the transfer of genes linked to the SNPs of the candidate set to Drugst.One and thus can serve as the start of a network-based analysis, such as DR.

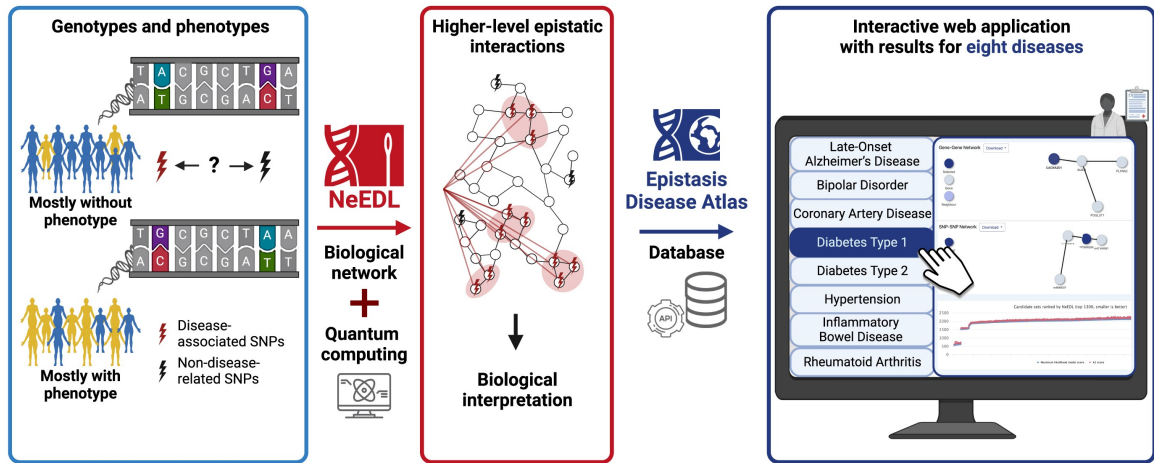


Figure 3.1: NeEDL and EpistasisDiseaseAtlas (Reprinted from publication C with permission from Oxford University Press): To be able to assess mutation combinations that could explain complex phenotypes with a depth and statistical significance that has been unseen before, NeEDL provides a heuristic approach based on the combination of network medicine information with local-search-driven epistatic interaction set refinement. So-called SNP-SNP interaction networks are created on biological information, such as PPIs combined with the SNPs association to genes, to limit the searchable space of the local-search algorithm. Local-search was prepared and implemented to not only run on traditional binary hardware but also QC hardware, preparing for once it is readily available. Results are stored and provided through the EDAs, a first-of-its-kind online resource for promising human epistatic interaction sets, options for refinement, and the direct possibility to start DR analyses.

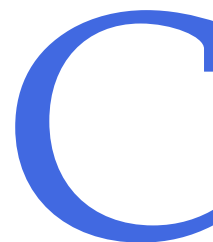
3.3 NEEDL & EPISTASIS DISEASE ATLAS

3.3.1 Summary

NeEDL provides a new solution to the detection of epistatic interaction sets from GWAS data. It is guided by network medicine essentially prefiltering the search space which allows the evaluation of higher-order SNP sets, expanding the possible set size to on average five,

compared to a size of two for existing tools. The implementation of the detection algorithms as the optimization problem of a local-search further reduced the computational demand, increased the statistical significance of produced SNP-sets, and made it QC ready. In the paper the successful implementation for QC hardware was demonstrated promising computation of even higher-order epistatic interaction sets once required quantum hardware becomes scalable with regard to the available q-Bits. The results for eight diseases are openly available in the EpistasisDiseaseAtlas⁷, namely: Late-onset Alzheimer's Disease, Bipolar Disorder, Coronary Artery Disease, Hypertension, Type-1 Diabetes, Type-2 Diabetes, Rheumatoid Arthritis, and Inflammatory Bowel Disease. It provides an integrated network explorer for the SNP interactions, allowing set refinement and on-demand recomputation of SNP-set scores, as well as the derived gene-gene network and thus the connection to network-based DR by using Drugst.One. This way, it closes the gap of using epistasis hypothesis-based SNP-sets as a starting point for DR research.

⁷ <https://epistasis-disease-atlas.com/home>



NEEDL & EPISTASIS DISEASE ATLAS

Markus Hoffmann*, **Julian M Poschenrieder***, **Massimiliano Incudini***, **Sylvie Baier***, **Amelie Fritz***,
Andreas Maier*, **Michael Hartung***, **Christian Hoffmann***, **Nico Trummer***, Klaudia Adamowicz,
Mario Picciani, Evelyn Scheibling, Maximilian V Harl, Ingmar Lesch, Hunor Frey, Simon Kayser, Paul
Wissenberg, Leon Schwartz, Leon Hafner, Aakriti Acharya, Lena Hackl, Gordon Grabert, Sung-Gwon
Lee, Gyuhyeok Cho, Matthew E Cloward, Jakub Jankowski, Hye Kyung Lee, Olga Tsoy, Nina Wenke,
Anders Gorm Pedersen, Klaus Bønnelykke, Antonio Mandarino, Federico Melograna, Laura Schulz,
Héctor Climente-González, Mathias Wilhelm, Luigi Iapichino, Lars Wienbrandt, David Ellinghaus,
Kristel Van Steen, Michele Grossi, Priscilla A Furth, Lothar Hennighausen, Alessandra Di Pierro, Jan
Baumbach, Tim Kacprowski, Markus List, and David B Blumenthal

“Network medicine-based epistasis detection in complex diseases: ready for quantum computing”

Published in: *Nucleic Acids Research*, 2024, Vol. 52, No. 17, pp. 10144–10160.

Published September 2024.

DOI: [10.1093/nar/gkae697](https://doi.org/10.1093/nar/gkae697)

Complete Reference:

Hoffmann, M., Poschenrieder, J. M., Incudini, M., Baier, S., Fritz, A., Maier, A., *et al.* (2024). Network medicine-based epistasis detection in complex diseases: ready for quantum computing. *Nucleic Acids Research*, 52(17), 10144–10160. <https://doi.org/10.1093/nar/gkae697>

Authors' Contribution: *I was involved in early discussions on this project and rejoined at a later point in time for the management, coordination, and conceptualization of the Epistasis Disease Atlas and additionally directly responsible for the implementation of the backend service. I further revised the manuscript and contributed to the EDA focused sections.*

Code Availability: The source code for the core algorithm of NeEDL can be found at github.com/daisybio/NeEDL. Source code for the frontend of the EDA at github.com/daisybio/NeEDL-frontend and the backend at github.com/daisybio/NeEDL-api.

Network medicine-based epistasis detection in complex diseases: ready for quantum computing

Markus Hoffmann^{1,2,3,*}, Julian M. Poschenrieder^{1,4,†}, Massimiliano Incudini^{5,†}, Sylvie Baier^{1,†}, Amelie Fritz^{6,7,†}, Andreas Maier^{4,†}, Michael Hartung^{4,†}, Christian Hoffmann^{4,†}, Nico Trummer¹, Klaudia Adamowicz⁴, Mario Picciani⁸, Evelyn Scheibling¹, Maximilian V. Harl^{1,9}, Ingmar Lesch¹, Hunor Frey¹, Simon Kayser¹, Paul Wissenberg¹, Leon Schwartz¹, Leon Hafner^{1,2}, Aakriti Acharya^{10,11}, Lena Hackl¹², Gordon Grabert^{10,11}, Sung-Gwon Lee^{3,12}, Gyuhyeok Cho¹³, Matthew E. Cloward¹⁴, Jakub Jankowski¹⁵, Hye Kyung Lee¹⁶, Olga Tsoy¹⁷, Nina Wenke⁴, Anders Gorm Pedersen⁶, Klaus Bønnelykke⁷, Antonio Mandarino¹⁵, Federico Melograna^{16,17}, Laura Schulz¹⁸, Héctor Climente-González¹⁹, Mathias Wilhelm^{18,20}, Luigi Iapichino¹⁸, Lars Wienbrandt²¹, David Ellinghaus²¹, Kristel Van Steen^{16,17}, Michele Grossi²², Priscilla A. Furth²³, Lothar Hennighausen^{2,3}, Alessandra Di Pierro^{5,†}, Jan Baumbach^{4,24,†}, Tim Kacprowski^{9,10,†}, Markus List^{1,25,*} and David B. Blumenthal^{18,*}

¹Data Science in Systems Biology, School of Life Sciences, Technical University of Munich, Freising, Germany

²Institute for Advanced Study (Lichtenbergstrasse 2 a) Technical University of Munich, D-85748 Garching, Germany

³National Institute of Diabetes and Digestive and Kidney Diseases, Bethesda, MD 20892, USA

⁴Institute for Computational Systems Biology, University of Hamburg, Germany

⁵Dipartimento di Informatica, Università di Verona, Strada le Grazie 15 - 34137 Verona, Italy

⁶Department of Health Technology, Section for Bioinformatics, Technical University of Denmark, DTU, 2800 Kgs. Lyngby, Denmark

⁷Copenhagen Prospective Studies on Asthma in Childhood (COPSAC), Herlev and Gentofte Hospital, University of Copenhagen, Copenhagen, Denmark

⁸Computational Mass Spectrometry, Technical University of Munich, Freising, Germany

⁹Department of Health Sciences and Technology, Neuroscience Center Zürich (ZNZ), Swiss Federal Institute of Technology (ETH Zürich), Zürich 8092, Switzerland

¹⁰Division Data Science in Biomedicine, Peter L. Reichertz Institute for Medical Informatics, Technische Universität Braunschweig and Hannover Medical School, Rebenring 56, 38106 Braunschweig, Germany

¹¹Braunschweig Integrated Centre of Systems Biology (BRICS), Technische Universität Braunschweig, Rebenring 56, 38106 Braunschweig, Germany

¹²School of Biological Sciences and Technology, Chonnam National University, Gwangju, Korea

¹³Department of Chemistry, Gwangju Institute of Science and Technology, Gwangju, Korea

¹⁴Department of Biology, Brigham Young University, Provo, UT, USA

¹⁵International Centre for Theory of Quantum Technologies, University of Gdańsk, 80-309 Gdańsk, Poland

¹⁶BIO3 - Systems Genetics; GIGA-R Medical Genomics, University of Liège, Liège, Belgium

¹⁷BIO3 - Systems Medicine; Department of Human Genetics, KU Leuven, Leuven, Belgium

¹⁸Leibniz Supercomputing Centre of the Bavarian Academy of Sciences and Humanities (LRZ), Garching b. München, Germany

¹⁹RIKEN Center for Advanced Intelligence Project, Tokyo, Japan

²⁰Munich Data Science Institute (MDSI), Technical University of Munich, Garching, Germany

²¹Institute of Clinical Molecular Biology, Christian Albrechts University of Kiel, Kiel, Germany

²²European Organization for Nuclear Research (CERN), Geneva 1211, Switzerland

²³Departments of Oncology & Medicine, Georgetown University, Washington, DC, USA

²⁴Computational BioMedicine Lab, University of Southern Denmark, Denmark

²⁵Biomedical Network Science Lab, Department Artificial Intelligence in Biomedical Engineering, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

*To whom correspondence should be addressed. Email: markus.hoffmann@nih.gov

Correspondence may also be addressed to Markus List. Email: markus.list@tum.de

Received: November 22, 2023. Revised: July 12, 2024. Editorial Decision: July 15, 2024. Accepted: August 1, 2024

© The Author(s) 2024. Published by Oxford University Press on behalf of Nucleic Acids Research.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License

(<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

Correspondence may also be addressed to David B. Blumenthal. Email: david.b.blumenthal@fau.de

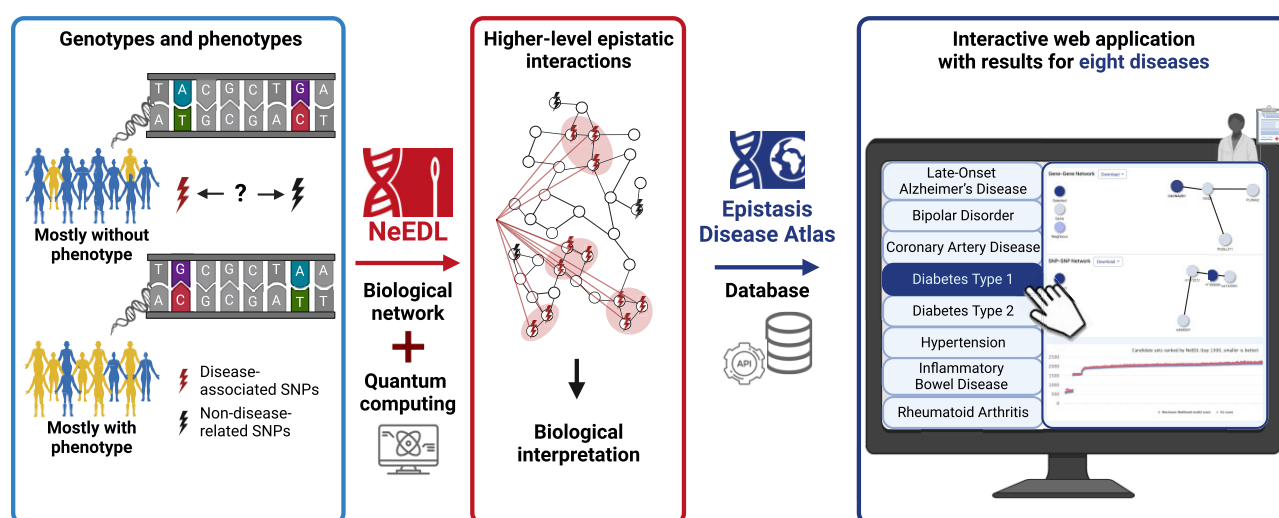
[†]The first eight authors should be regarded as Joint First Authors.

^{*}The last five authors should be regarded as Joint Last Authors.

Abstract

Most heritable diseases are polygenic. To comprehend the underlying genetic architecture, it is crucial to discover the clinically relevant epistatic interactions (EIs) between genomic single nucleotide polymorphisms (SNPs) (1–3). Existing statistical computational methods for EI detection are mostly limited to pairs of SNPs due to the combinatorial explosion of higher-order EIs. With NeEDL (network-based epistasis detection via local search), we leverage network medicine to inform the selection of EIs that are an order of magnitude more statistically significant compared to existing tools and consist, on average, of five SNPs. We further show that this computationally demanding task can be substantially accelerated once quantum computing hardware becomes available. We apply NeEDL to eight different diseases and discover genes (affected by EIs of SNPs) that are partly known to affect the disease, additionally, these results are reproducible across independent cohorts. EIs for these eight diseases can be interactively explored in the Epistasis Disease Atlas (<https://epistasis-disease-atlas.com>). In summary, NeEDL demonstrates the potential of seamlessly integrated quantum computing techniques to accelerate biomedical research. Our network medicine approach detects higher-order EIs with unprecedented statistical and biological evidence, yielding unique insights into polygenic diseases and providing a basis for the development of improved risk scores and combination therapies.

Graphical abstract



Introduction

Genome-wide association studies (GWAS) aim to identify genetic single nucleotide polymorphisms (SNPs) that are individually associated with a phenotype such as a disease (1–3). Thousands of individual SNPs have been associated with diseases since the early 1990s, yet they account only for a fraction of the investigated traits' heritability (4,5). The hypothesis is that a significant proportion of the missing heritability can be explained by epistatic SNP interactions that are jointly but not individually associated with the phenotype (6). However, no undisputed cases of epistasis in humans are known. Therefore, scalable detection tools for epistatic interactions that yield interpretable, high-quality results are needed to advance the understanding of possible genetic causes of diseases.

Detecting biologically plausible epistatic candidate SNP sets is difficult: Firstly, there is no consensus on the choice of the formal model in order to make this problem algorithmically accessible. Secondly, it is often unclear if predicted cases of epistasis are statistical artifacts since they are hardly interpretable. Thirdly, comprehensively testing higher-order interactions is computationally intractable due to the combinatorial explosion of the search space. Existing epistasis detection tools, including Potpourri (9), LinDen (10), PoCos (11),

MACOED (12) and BiologicalEpistasis (13) hence do not scale to large datasets and are mostly restricted to pairwise interactions (Supplementary Table S1). Finally, easily accessible resources to browse and explore pre-computed epistasis candidates interactively are lacking. To overcome these challenges, we present NeEDL (network-based epistasis detection via local search), an epistasis detection tool leveraging quantum computing and network medicine to identify biologically interpretable candidate sets of epistatic interaction between SNPs (see Figure 1). NeEDL unifies various GWAS input formats and initially filters datasets for relevant SNPs (Supplementary Figure S1A). NeEDL then offers different statistical epistasis models (8) for associating higher-order SNP sets with phenotypes, prioritizing SNP sets using a biologically-informed SNP-SNP interaction (SSI) network. Our hypothesis is that SNPs affecting the same or functionally related proteins are more likely to be involved in relevant epistatic interaction and less likely to be statistical artifacts. Thus, SNPs are mapped to genes using dbSNP (7) and connected if they affect the same protein or a functionally associated protein (e.g., in a protein-protein interaction (PPI) network, Figure 1A). Within the SSI network, NeEDL uses local search with multi-start and simulated annealing to find connected subgraphs of a user-specified size that are

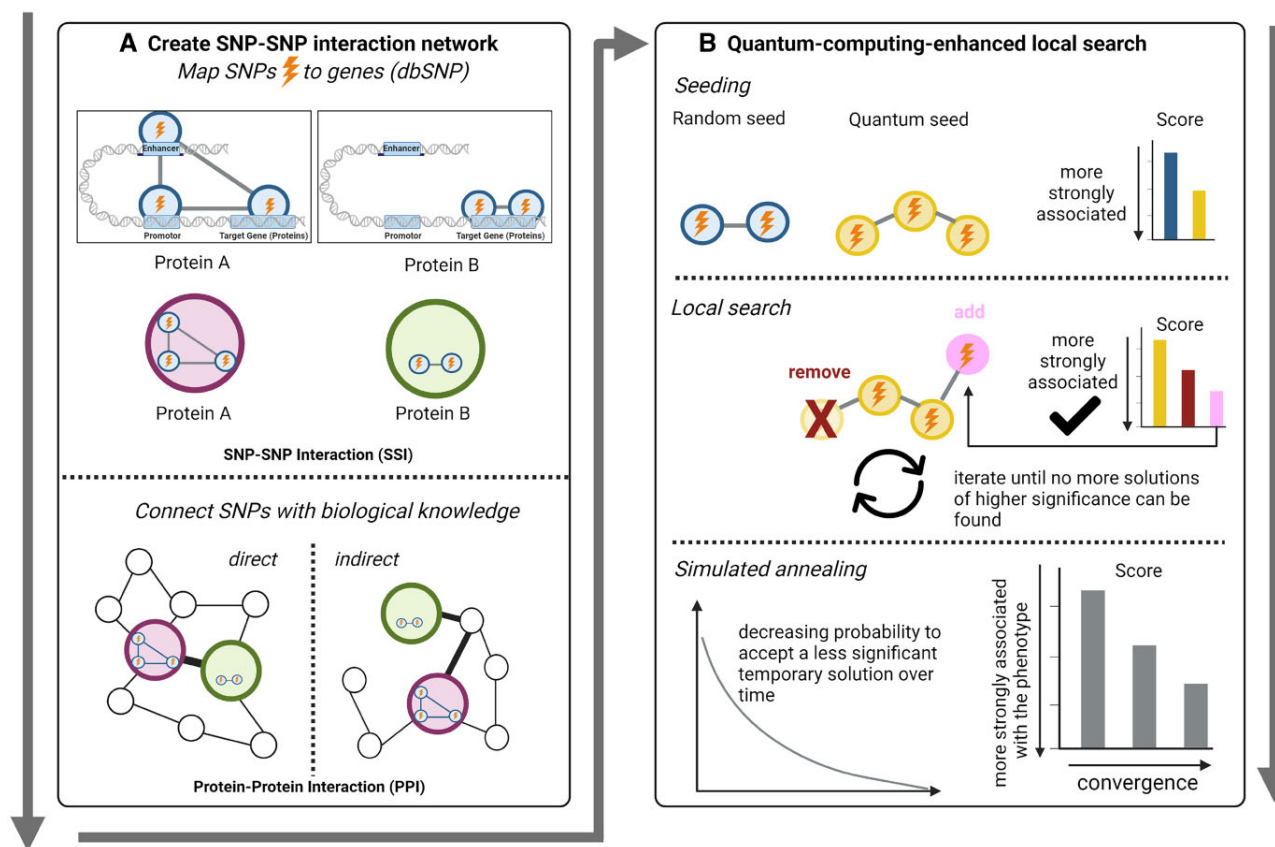


Figure 1. Overview of the methodology. **(A)** SNPs are mapped to genes via SNP-gene links obtained from dbSNP (7) and are then connected in an SNP-SNP interaction (SSI) network if they are mapped to the same protein or to adjacent proteins in a protein-protein interaction (PPI) network (i.e., directly when the proteins interact with each other or indirectly connected when the SNPs are linked through other proteins). **(B)** Next, NeEDL picks either random seeds or seeds selected by a quantum-computing optimization algorithm. It then uses local search (i.e., either adding new SNPs from the direct neighborhood in the SSI network or removing SNPs that worsen the score) to find SNP sets of a user-specified maximum size that induce connected subgraphs in the SSI network and are locally optimal w. r. t. statistical association of the higher-order genotype with the investigated phenotypic trait. To quantify association strength, NeEDL implements various statistical epistasis models suggested in the literature (8). Since local search can get stuck due to the requirement of constant improvement in each step, we use simulated annealing, which allows us to accept a less significant intermediate solution with decreasing probability over time.

locally optimal w. r. t. the selected statistical epistasis model (Figure 1B). Focusing on SNP sets inducing connected subgraphs in the SSI network not only increases the likelihood of uncovering biologically meaningful cases of epistasis but also dramatically reduces the size of the search space. Finally, NeEDL fully integrates quantum computing (QC) algorithms to yield high-quality initial solutions for the local search to reduce NeEDL's runtime further once QC hardware becomes more readily available. Our work hence adds epistasis detection to the growing array of bioinformatics problems that can be tackled using QC, where prior studies have focused on applications such as molecular docking (14), genome assembly (15), DNA assembly (16), DNA reconstruction (17) and drug design (18).

We benchmark NeEDL against frequently used epistasis detection tools on GWAS data for eight human diseases—Late-Onset Alzheimer's Disease (LOAD), Bipolar Disorder (BD), Coronary Artery Disease (CAD), Diabetes Type 1 (T1D), Diabetes Type 2 (T2D), Hypertension (HT), Inflammatory Bowel Disease (IBD) and Rheumatoid Arthritis (RA) (see Supplementary Tables S2–S10 for an overview over sample numbers and SNP numbers and Datasets for information on datasets). Our validation shows that NeEDL markedly out-

performs the currently used epistasis detection tools in terms of the discovered candidate SNP sets' associations with the diseases (i.e. the statistical score). The top-scoring candidate SNP sets are further supported by (i) significant biological associations with the respective diseases, and (ii) in the case of LOAD, by replication in independent cohorts (i.e., discovery with NeEDL in the TGen cohort and replication in the UK Biobank cohort). We make the results of NeEDL for the eight diseases, which we obtained in over two million CPU hours, widely accessible through the Epistasis Disease Atlas (<https://epistasis-disease-atlas.com>, Supplementary Figure S1B) via an application programming interface (API), an R and Python package, an FTP server, and an interactive, feature-rich web application.

Materials and methods

Implementation details

NeEDL is implemented in C++. It employs the Boost Graph Library version 1.71.0 (19) and iGraph version 0.9.8 (20) for the construction and handling of graphs. We use OpenMP (21) for the parallelization of the initialization of NeEDL (i.e. read

the input data, map the SNPs to genes, construct the SSI network, and get random start seeds) and the local search with simulated annealing. We further included the following external dependencies: CMake v. 2.6 or higher, Doxygen, Catch version 2.11.0, CLI11 version 1.9.0 and Eigen version 3.3.7 (<http://eigen.tuxfamily.org>). We provide an installer script that installs most external dependencies (excluding CMake, Doxygen, and OpenMP). The dockerized version of NeEDL can be directly pulled and executed from: <https://hub.docker.com/r/bigdatainbiomedicine/needl>.

We conducted computation on the high-performance computing systems provided by the Leibniz Supercomputing Center of the Bavarian Academy of Sciences and Humanities (LRZ). Each MACOED task was run using a single core (2.6 GHz nominal frequency) on LRZ's Large Memory Teramem cluster (HP DL580 shared memory system) since MACOED has extremely high memory requirements (approx. 1.3 terabytes of RAM). For each NeEDL task, we used 28 threads on 14 physical cores (2.6 GHz nominal frequency) on LRZ's CoolMUC-2 cluster (28-way Haswell-EP nodes).

Since MACOED uses a randomized optimization technique (ant colony optimization), we ran it 100 times on each dataset to minimize the impact of random bias. All obtained SNP sets were used for downstream evaluation. Also, NeEDL includes a randomized subroutine, namely, the seeding of the initial solutions for the local search. To account for this, we ran NeEDL's local search with multi-start and introduced a global time limit of 12 h.

The behavior of NeEDL can be influenced through various parameters, including (i) seeding procedure (default: random seeding), (ii) maximal size of SNPs in a candidate SNP set (default: 10), (iii) maximum iteration of lookups inside one local neighborhood (default: 300), (iv) statistical model used for the local search (default: maximum likelihood model) and (v) SNP filters inside the network (e.g. minor allele frequency, maximum marginal association, linkage disequilibrium cutoff, default: None). For further parameters, please check the GitHub repository and the manual of NeEDL.

Data format converter, data cleaning and data filtering

We faced challenges in using different genotype and phenotype data formats with NeEDL and other epistasis detection tools due to the lack of available data processing tools for converting these formats to the JSON format required by NeEDL. Furthermore, most epistasis detection tools require formatted, pre-cleaned, and pre-filtered datasets as input. To automate preprocessing, cleaning, filtering, and conversion into a joint machine-readable format, we developed epiJSON which supports VCF, PED/MAP, TPED/TFAM, BED/BIM/FAM and other formats and follows recommendations laid out by Marees *et al.* (22): (i) excluding samples with missing phenotypes; (ii) excluding SNPs with more than 20% of missing data across individuals and individuals with more than 20% of missing data across SNPs; (iii) checking an individual's assigned sex using X chromosome data, and removing those with a discrepancy; (iv) considering the number of samples within the given dataset to determine a suitable minor allele frequency threshold; (v) removing variants that fail the Hardy–Weinberg test at a threshold recommended for binary traits of $1e^{-10}$ in controls and $1e^{-6}$ in cases; (vi) excluding individuals with heterozygosity that is too high or too low, in-

dicating sample contamination or inbreeding and (vii) excluding SNPs with any missing data across individuals to ensure that the dataset has no missing values. Cleaning processes and filtering steps can be adjusted with parameters following the guidelines in the GitHub repository.

Datasets

Eight datasets were included in our benchmark: LOAD, BD, CAD, T1D, T2D, HT, IBD and RA (see Section Data Availability for links to the databases). The LOAD dataset had controls included. All other datasets had no controls included, and we used the British Cohort 1958 provided by the Wellcome Trust Case Control Consortium (WTCCC) as controls. In [Supplementary Tables S2–S10](#), we show the detailed sample and SNP numbers before and after the epiJSON tool for the datasets discussed in this manuscript. The LOAD dataset was provided by the TGen consortium (23,24), the other datasets were provided by the WTCCC consortium (25) (see Data Availability).

A replication study for SNP sets discovered in the disease LOAD was conducted using data from the UK Biobank database (www.ukbiobank.ac.uk) (26). Individuals were selected using ICD-10 coding (27) (field 41 270), see [Supplementary Table S11](#). We constructed two replication data sets. One containing individuals with mixed ancestry and a sub-data set containing only individuals with British ancestry. Individuals in controls matched the individuals in cases in age (field 34) and gender (field 31) proportionally. Replication was performed on the results of the discovery study using the tools NeEDL, higher-order baseline, second-order baseline, MACOED, and LinDen. The Pearson correlation coefficient was calculated (28) to determine the correlation of statistical scores between the discovery and replication dataset for the candidate SNP sets for LOAD of the NeEDL.

Epistasis models

For statistical modeling, we use the framework introduced by Blumenthal *et al.* (8). Let $G = (g_{i,s}) \in \{0, 1, 2\}^{I \times S}$ be a genotype matrix, where I and S denote the sets of all individuals (patients) and SNPs, respectively, and the entry $g_{i,s}$ encodes the number of minor alleles of individual i at SNP s . Moreover, let $y \in \mathbb{P}^I$ be a phenotype vector, where $\mathbb{P}$ is the set of all possible phenotypes (e.g., $\mathbb{P} = \{0, 1\}$ for case-versus-control data). We call a tuple $M = (f_{G,y}, \sigma)$ an *epistasis model* if $\sigma \in \{MIN, MAX\}$ is the model sense and $f_{G,y} : \mathcal{P}(S) \rightarrow \mathbb{R}$ is an objective function that assigns objective values $f_{G,y}(S)$ to SNP sets $S \subseteq S$ which quantify the statistical evidence for the SNPs $s \in S$ being involved in epistatic interaction.

A wide variety of epistasis models have been proposed in the literature, including Bayesian network scores (such as the K2-score (29–34)), variants of multifactor dimensionality reduction (35–39), variants of polynomial regression (40–42), the P -value of the χ^2 -test, which is arguably the most widely used epistasis model (29–34), or the maximum likelihood model (MLM) introduced by Blumenthal *et al.* (8).

Equipped with these definitions (and assuming $\sigma = MIN$; if $\sigma = MAX$ we instead have to maximize in (Equation (1)), the problem of detecting SNPs involved in epistatic interaction can, in a first attempt, be formalized as the problem of finding an SNP set S^* that solves the optimization problem

$$S^* \in \arg \min \{f_{G,y}(S) \mid S \subseteq S \wedge LB \leq |S| \leq UB\}, \quad (1)$$

where $LB \geq 2$ and $UB \geq LB$ are user-specified lower and upper bounds on the size of the solution SNP set. The problem with this naïve formulation is that the search space is huge, and solving the optimization problem is hence computationally very expensive. In NeEDL, we mitigate this shortcoming by restricting the search space to SNP sets S , which induce a connected subgraph $\mathcal{N}[S]$ in an SSI network $\mathcal{N} = (\mathcal{S}, \mathcal{E})$ where two SNPs $s_1, s_2 \in \mathcal{S}$ are connected by an edge $s_1 s_2 \in \mathcal{E}$ if there exists prior evidence that they might be involved in a biologically meaningful interaction (see the following subsection for details on the construction of $\mathcal{N}$). That is, NeEDL uses the following computational model:

$$S^* \in \arg \min \{ f_{G,y}(S) \mid S \subseteq \mathcal{S} \wedge LB \leq |S| \leq UB \wedge \mathcal{N}[S] \text{ is connected} \} \quad (2)$$

For our benchmark, we use four epistasis models: the P -value of the χ^2 -test, the K2-score, the MLM score, and the NLL gain of a quadratic regression model in comparison to a linear regression model. The former three models are all based on the penetrance table $T_{G,y,S}(\mathbf{g}) = \{ \{y_i \mid i \in \mathcal{I} : g_{i,s} = g_s \forall s \in S\} \}$ of the scored SNP set S . For each possible genotype $\mathbf{g} \in \{0, 1, 2\}^S$ at the scored SNP set S , the cell $T_{G,y,S}(\mathbf{g})$ of the penetrance table contains the multiset of phenotypes of individuals whose genotype at S matches $\mathbf{g}$. For all three models, the score $f_{G,y}(S)$ is small if the phenotypes are unevenly distributed across the cells of $T_{G,y,S}$ (see Blumenthal *et al.* (8) for details). For the P -value of the χ^2 -test, the K2-score, and the MLM score, we hence have $\sigma = MIN$ and small scores $f_{G,y}(S)$ indicate that the genotype at S is predictive of phenotypic variation.

The NLL gain of quadratic versus linear regression captures another dimension of the concept of epistasis. Here, we first fit two linear or logistic (depending on whether phenotypes are quantitative or categorical) regression models as follows ($\mathbf{G}_{\bullet,s}$ is the column for the SNP s in $\mathbf{G}$ and $\odot$ denotes element-wise multiplication):

$$\mathbf{y} \sim \theta_0 + \sum_{s \in S} \theta_s \mathbf{G}_{\bullet,s} + \sum_{\{s,s'\} \in \binom{S}{2}} \theta_{s,s'} \mathbf{G}_{\bullet,s} \odot \mathbf{G}_{\bullet,s'} \quad (3)$$

$$\mathbf{y} \sim \theta'_0 + \sum_{s \in S} \theta'_s \mathbf{G}_{\bullet,s} \quad (4)$$

Subsequently, we compute NLLs for the fit models θ and θ' and define our score as $f_{G,y}(S) = \text{NLL}(\theta') - \text{NLL}(\theta)$. For the NLL gain, large scores hence indicate that we can better predict the phenotypes when considering multiplicative interactions between all pairs of SNPs contained in the scored SNP set S than when considering only marginal effects (i.e. we have $\sigma = MAX$). Since the NLL gain can result in negative values, we excluded such values from the analysis.

Construction of the SNP-SNP interaction network

To construct $\mathcal{N} = (\mathcal{S}, \mathcal{E})$, we require a many-to-many SNP-to-gene mapping $\pi \subseteq (\mathcal{S} \times \mathcal{G})$ ($\mathcal{G}$ is the set of all genes), as well as a gene-gene network $\mathcal{N}' = (\mathcal{G}, \mathcal{E}')$ (Figure 1A). In NeEDL, we obtain π from dbSNP (7) and use the BioGRID (43) PPI network to construct $\mathcal{N}'$. That is, two genes $g_1, g_2 \in \mathcal{G}$ are connected by an edge $g_1 g_2 \in \mathcal{E}'$ if their encoded proteins are connected by an edge in BioGRID. Note that in BioGRID, two proteins are connected by an edge if and only if a physical interaction between them is supported by at least two publi-

cations or two experimental systems (e.g. affinity purification-mass spectrometry and yeast-2-hybrid). However, NeEDL can be easily extended to use other SNP-to-gene mappings and/or gene-gene networks. For each SNP $s \in \mathcal{S}$, let $\pi[s] = \{g \in \mathcal{G} \mid (s, g) \in \pi\}$ be the image of s under the mapping π . Equipped with π and $\mathcal{N}'$, we construct $\mathcal{N}$ by connecting to SNPs $s_1, s_2 \in \mathcal{S}$ by an edge $s_1 s_2 \in \mathcal{E}$ if and only if at least one of the following conditions holds:

- The SNPs s_1 and s_2 are mapped to at least one common gene, i.e. $\pi[s_1] \cap \pi[s_2] \neq \emptyset$.
- There is a SNP s_3 and genes $g_1, g_2 \in \pi[s_3]$ such that $g_1 \in \pi[s_1]$ and $g_2 \in \pi[s_2]$.
- The SNPs s_1 and s_2 are mapped to genes that are connected in the gene-gene network $\mathcal{N}'$, i.e. there is an edge $g_1 g_2 \in \mathcal{E}'$ such that $g_1 \in \pi[s_1]$ and $g_2 \in \pi[s_2]$.

Note that, with this construction, SNPs that are left unmapped by π correspond to isolated nodes in $\mathcal{N}$ and are hence not contained in any feasible solution of the optimization problem specified in Equation (2). In NeEDL, we hence remove all unmapped SNPs before (heuristically) solving the optimization problem, which further reduces the size of the search space.

Besides the dbSNP mapping procedure described above, we also assessed a mapping based on expression quantitative trait loci (eQTL) information. We downloaded all available studies with gene expression data and condition label 'naive' from the eQTL Catalogue (44). The studies in the catalog contain data from various different tissues. We converted the Ensembl gene identifiers to gene symbols with the help of biomaRt (45) to be compatible with the BioGRID dataset. The retrieved data was condensed into a single tabular file containing the dbSNP rsID, the gene name, the tissue, and the reported P -value. At runtime, NeEDL allows the user to select all tissues or a subset of those of interest. All mapping entries are retrieved from the mentioned condensed eQTL file and filtered for the given criteria. Afterward, NeEDL filters the remaining interactions based on their Benjamini-Hochberg-corrected P -value (46). For comparison with the dbSNP-based mapping, we selected all mapping information from all available tissues and applied a significance level of 0.05 after multiple testing correction. We then conducted a NeEDL run with the same parameters as our dbSNP-based runs but exchanged the dbSNP mapping with the eQTL mapping.

Local search with multi-start and simulated annealing

We use the local search with multi-start and simulated annealing to heuristically solve the optimization problem specified in Equation (2). NeEDL supports all epistasis models benchmarked by Blumenthal *et al.* (8). Simulated annealing is a general meta-heuristic to escape local optima in local search by accepting deteriorations from local optima with probabilities that decrease as the algorithm runs (47). For NeEDL, we adapted a simulating annealing algorithm for graph edit distance computation proposed by Riesen *et al.* (48), using its implementation available in GEDLIB (49,50).

We decided to use local search to solve our optimization problem instead of other meta-heuristics such as ant colony optimization (51) or genetic programming (52) because it can be easily parallelized and yields anytime algorithms that can be interrupted before convergence. Both of these properties

are important in the context of epistasis detection where run-time efficiency is crucial. Similarly, there are various other strategies besides simulated annealing which can be used to escape local optima in local search, e.g. variable neighborhood descent (53) or randomized postprocessing (54). Again, run-time considerations were the main reason for our choice of simulated annealing as a strategy to escape local optima.

Algorithm 1 provides a high-level description of our algorithm. It computes a set $\mathcal{S}^* \subseteq \mathcal{F}_{LB,UB}$ of up to n locally optimal SNP sets, where n is a hyper-parameter that can be set by the user and $\mathcal{F}_{LB,UB} = \{S \subseteq \mathcal{S} \mid LB \leq |S| \leq UB \wedge \mathcal{N}[S] \text{ is connected}\}$ is the set of all feasible solutions. Computation of the n locally optimal SNP sets is parallelized, which is the main reason for NeEDL's excellent runtime performance (see 'Implementation Details' for details).

Algorithm 1: Overview of local search and simulated annealing.

Input: Genotype matrix $\mathbf{G}$, phenotype vector $\mathbf{y}$, SSI network $\mathcal{N} = (\mathcal{S}, \mathcal{E})$, number of initial SNP sets n , lower bound on SNP set size LB , upper bound on SNP set size UB , time limit L , maximal number of iterations I , start and end probabilities $1 > p_1 > p_I > 0$.

Output: Set $\mathcal{S}^* \subseteq \mathcal{F}_{LB,UB}$ of up to n locally optimal SNP sets.

```

1  $\mathcal{S}^* \leftarrow \emptyset$ ;  $\alpha \leftarrow (\log(p_1)/\log(p_I))^{1/(I-1)}$ ;
2 parallel for  $j \in \{1, \dots, n\}$  do
3    $\mathcal{S}' \leftarrow \text{connected-subgraph}(\text{size} = LB)$ ;
4    $f^* \leftarrow f_{\mathbf{G},\mathbf{y}}(\mathcal{S}')$ ;  $\mathcal{S}'' \leftarrow \mathcal{S}'; f' \leftarrow f_{\mathbf{G},\mathbf{y}}(\mathcal{S}'')$ ;
5   converged  $\leftarrow \text{False}$ ;  $\Delta \leftarrow 0$ ;  $i \leftarrow 1$ ;
6   while  $i \leq I$  and not time limit reached and not converged do
7      $\mathcal{S}'' \leftarrow \arg\min\{f_{\mathbf{G},\mathbf{y}}(S) \mid S \in \mathcal{F}_{LB,UB}(\mathcal{S}')\}$ ;  $f'' \leftarrow f_{\mathbf{G},\mathbf{y}}(\mathcal{S}'')$ ;
8     if  $f'' < f^*$  then
9        $\mathcal{S}' \leftarrow \mathcal{S}''$ ;  $f' \leftarrow f''$ ;
10      if  $f' < f^*$  then
11         $\mathcal{S}^* \leftarrow \mathcal{S}'; f^* \leftarrow f'$ ;
12      else
13         $\delta \leftarrow (f'' - f')$ ;  $\Delta \leftarrow \Delta + \delta$ ;  $\bar{\delta} \leftarrow \Delta/i$ ;  $p_i \leftarrow p_1^{\alpha^{i-1}}$ ;
14         $r \leftarrow \text{random-number}(\text{min} = 0, \text{max} = 1)$ ;
15        if  $r < p_i^{\delta/\bar{\delta}}$  then
16           $\mathcal{S}' \leftarrow \mathcal{S}''$ ;  $f' \leftarrow f''$ ;
17        else
18          converged  $\leftarrow \text{True}$ ;
19    $\mathcal{S}^* \leftarrow \mathcal{S}^* \cup \{\mathcal{S}^*\}$ ;
20 return  $\mathcal{S}^*$ ;

```

To construct the j^{th} locally optimal SNP set, we initialize it as a randomly drawn connected subgraph of size LB (again, our objective to keep NeEDL's runtime low was the reason why we decided against more elaborate seeding strategies). We maintain the best encountered SNP set $\mathcal{S}^*$, as well as the currently processed SNP set $\mathcal{S}'$, together with their objective values f^* and f' . As long as a given time limit or a maximal number of iterations has not been reached and our descent in the j^{th} region of the search space has not converged, we compute a candidate SNP set $\mathcal{S}''$ as the best SNP set among the set $\mathcal{F}_{LB,UB}(\mathcal{S}')$ of all locally reachable SNP sets from $\mathcal{S}'$. We define $\mathcal{F}_{LB,UB}(\mathcal{S}') = \{S \in \mathcal{F} \mid |S \Delta \mathcal{S}'| = 1\}$ as the set of all feasible SNP sets that differ from $\mathcal{S}'$ in exactly one SNP (details on this step are provided below). We then check whether $\mathcal{S}''$ improves the currently processed SNP set $\mathcal{S}'$. If so, we update $\mathcal{S}'$ and also $\mathcal{S}^*$ if $\mathcal{S}'$ is now the best encountered SNP set.

On the other hand, if the objective value $\mathcal{S}''$ exceeds the objective value of $\mathcal{S}'$ by $\delta > 0$, we nonetheless accept it as our new currently processed SNP set with probability $p_i^{\delta/\bar{\delta}}$, where $\bar{\delta}$ is the mean deterioration from the objective value of the currently processed SNP set up to the current iteration i . That is, the larger δ , the lower the probability of accepting $\mathcal{S}''$ as our new currently processed SNP set. The baseline acceptance probability p_1 for the first iteration is a hyper-parameter that can be provided by the user (defaulted to $p_1 = 0.8$ in NeEDL). As the algorithm runs, it is updated using a suitably defined cooling factor α such that, when the maximum number of iterations I has been reached, it equals a user-specified end probability $p_I < p_1$ (defaulted to $p_I = 0.01$ in NeEDL).

To compute $\mathcal{S}''$, we exhaustively enumerate the set $\mathcal{F}_{LB,UB}(\mathcal{S}')$ of all locally reachable SNP sets from $\mathcal{S}'$. This can

be done by constructing modified SNP sets using the following three types of edit operations:

- SNP insertions (only if $|S'| < UB$): Insert an SNP $s \in \mathcal{S} \setminus S'$ that is connected to one of the SNPs already contained in S' .
- SNP removals (only if $|S'| > LB$): Remove a SNP $s \in S'$ such that the induced subgraph $\mathcal{N}[S' \setminus \{s\}]$ remains connected (i.e., s must not be an articulation point).
- SNP substitutions: Substitute a SNP $s_1 \in S'$ by a SNP $s_2 \in \mathcal{S} \setminus S'$ such that the induced subgraph $\mathcal{N}[(S' \setminus \{s_1\}) \cup \{s_2\}]$ remains connected (i.e., if s_1 is an articulation point, s_2 must bridge the connected components resulting from removal of s_1).

Statistical methods

We generate random SNP sets for empirical estimation of the false-positive rate. For epistasis detection tools other than NeEDL, we sample 1000 random SNP sets of size two as a baseline, whereas for NeEDL, we estimate the distribution of the number of SNPs in candidate sets based on the findings of our results and then randomly sample 1000 candidate SNP sets following this size distribution.

We generate two random SSI networks to evaluate the information that can be gained by transforming a PPI network into an SSI network. We either shuffle the labels to obtain a topology-preserving SSI network or shuffle the edges by a probability function to obtain an expected degree-preserving network as discussed in Lazareva *et al.* (55). Either way, this leads to a network where each SNP has other neighbors in the SSI network that we expect to result in the loss of biological information (Supplementary Figure S2A).

All P -values obtained by NeEDL and all competitors are corrected with the Benjamini-Hochberg procedure (46). In addition, we provide a baseline of randomly chosen candidate SNP sets, whose SNPs do not need to be connected via the SSI network. The baseline for the competitors contains 1000 randomly sampled sets with the size $|\text{sl}| = 2$, as all competitors (except PoCos, which return hundreds of SNPs, and since the statistical scores cannot be calculated with such a high number of SNPs, PoCos was excluded) only return candidate SNP sets with the size of two. NeEDL finds candidate SNP sets with variable sizes $|\text{sl}| \in [2, \infty]$.

Phenotype shuffling

We compared our original results with runs on datasets where we shuffled the phenotype vector while keeping the genotype matrix unchanged. These newly created datasets thus do not have any connection between the individuals' genotypes and phenotypes. We repeated this procedure 100 times for each dataset to reduce the influence of random effects on our analysis. This procedure led to 100 shuffled datasets for each disease. Due to the computational complexity of this analysis, we restricted the experiments to the three diseases BD, T2D and RA. We conducted one NeEDL run, and one Linden run for each shuffled dataset. We excluded MACOED from this analysis as the tool's extremely high runtime and memory requirements made it impossible to run it 300 times with our computational resources. Further, we replicated our analysis of the influence of the SSI network on these shuffled datasets as well. For that, we ran both of our network randomization methods ten times on each of the 100 phenotype-shuffled datasets,

resulting in 2,000 NeEDL runs per disease. In total we conducted 6300 NeEDL runs and 300 Linden runs on the shuffled datasets (Supplementary File 1).

Gene set enrichment analysis

Affected genes present in the highest ranks of penetrance SNP combinations for LOAD, IBD, and T1D were subjected to GSEA (<http://www.gsea-msigdb.org/gsea/index.jsp>, date of last access 8 March 2023) to determine if statistically significant enrichment in specific Gene Ontology (GO) gene sets could be identified (56,57). The query gene sets for GSEA are contained in Supplementary File 2.

Quantum computing

While large, fault-tolerant quantum computers are expected to outperform classical devices, complexity theory suggests that quantum computing may not be able to solve combinatorial optimization problems in polynomial time, similar to classical computing. Nevertheless, Grover's algorithm (58) can provide a quadratic speedup in terms of query complexity for quantum circuit-based quantum hardware (Supplementary Materials 1 and 2), and quantum annealers (Supplementary Materials 1 and 2) can provide the same speedup (59), under ideal, noiseless conditions. For particular instances of NP-hard optimization problems, super-polynomial speedups are possible (60), and certain problems can achieve even larger speedups by exploiting their specific structural features (61). Currently, only prototypes of quantum hardware are available, and heuristics must be run on them, with satisfactory results depending on the problem (62). The relation between Grover's algorithm and the heuristics used in our work is detailed in Supplementary Material 4.

We propose a novel quantum-enhanced method to address the initial seeding problem in NeEDL. Specifically, instead of selecting the initial seed randomly, our method employs an optimization technique to identify the most promising candidates. The underlying idea is that a better initial seed can significantly improve the quality of local search and reduce the time required to obtain high-quality solutions, assuming the optimization technique adds negligible overhead. To achieve this, we utilize a variant of the maximum clique problem to analyze the matrix of pairwise SNP correlations and identify the candidate set of SNPs that maximizes the sum of pairwise correlations. This problem is formulated as a quadratic unconstrained binary optimization (QUBO) whose objective function to minimize is defined as a quadratic polynomial over binary variables and consists of a linear combination of the solution cost and the associated constraints. By adjusting the weights of the linear combination, we can enforce or relax different constraints, such as the size of the candidate set. We formulate the QUBO problem as follows:

$$Q = \sum_{\ell=1}^N \left[\lambda_0 \left(\sum_i x_{i\ell} - K \right)^2 - \lambda_1 \sum_{(i,j) \notin E} w_{i\ell,j\ell} x_{i\ell} x_{j\ell} \right] + \lambda_2 \sum_{\ell=1}^N \sum_{m=\ell+1}^N \sum_{i=1}^{|V|} x_{i\ell} x_{im} \quad (5)$$

The formulation returns, saved in the matrix of binary variables x , N candidate sets of SNPs, each consisting of K SNPs. The parameter λ_0 penalizes candidate sets that are not of size K , λ_1 rewards sets with a higher sum of pairwise interactions

w , and λ_2 penalizes solutions with similar sets (63). For additional details, see Supplementary Materials 3. We solve the optimization problem by applying quantum annealing on a D-Wave quantum annealer, QAOA (quantum approximate optimization algorithm) on IBM Perth and IBM Lagos quantum processing units, thermal annealing (64), parallel tempering (65), and Gurobi (66) on classical hardware as a classical baseline. These algorithms are heuristics, and their potential speedup must be determined empirically on a case-by-case basis. The functioning of these quantum algorithms is detailed in Supplementary Materials 4, Supplementary Figures S3 and S4.

To address the limitations in quantum hardware resources, we utilize the community detection method Leiden (67) to identify densely connected sections of arbitrary size in the SNPs network. This approach allows us to divide the problem into smaller sub-instances, enabling efficient computation on any quantum hardware. Specifically, quantum annealing algorithms can handle problem sizes on the order of 100 SNPs, while IBM quantum processing units can manage up to 10 SNPs. Our quantum software module called *quepistasis*, implements the seeding procedure using the optimization technique. It is responsible for creating the QUBO formulation based on the pairwise correlation of SNPs and some configuration, interfacing with the platforms of D-Wave and IBM (and also supporting quantum hardware available on Microsoft Azure), and returning the set of SNPs. Its structure is shown in Supplementary Materials 4 and Supplementary Figure S5.

We test NeEDL with the quantum computing seeding procedure over three different datasets: BD, RA, and T2D. The purpose of our experiments is to verify how much quantum computing helps in finding better seeds and thus speeds up the local search process. To accommodate the limited amount of resources (namely the number of qubits and the number of gates) of the quantum computer, we subsample each of the three datasets into three different subsets, having 100, 500 and 1000 SNPs. Minimizing the quantum resources is essential for keeping the *fidelity*, which is the measure of the accuracy of the actual quantum operation compared to the ideal one, above a certain threshold, guaranteeing reliable results. The timing information in Supplementary Figure S5 is obtained from the logs generated by NeEDL. For the quantum procedure run on IBM devices, such timing is spoiled by the queue we must wait to execute each task, and such time is normalized. The scalability discussion in Figure 5 is obtained by linear interpolation of the seeding time using each technique for different sample sizes of each dataset. We show the experimental setup and provide some comments about the results obtained in Supplementary Materials 6, Supplementary Figure S6, Supplementary Table S12.

Results

NeEDL outperforms existing epistasis detection tools

We compared NeEDL to the existing epistasis detection tools LinDen (10) and MACOED (12) on the same input data as NeEDL. All tools were run with default hyper-parameters (see Material and Methods, for NeEDL, see Supplementary Tables S13–S15). PoCos (11), Potpourri (9) and BiologicalEpistasis (13) were excluded due to unresolvable implementation problems and/or excessive runtime or memory requirements

(Supplementary Table S16). To ensure fair competition, we compared the SNP sets using four widely used scores in this research area (8), including the P -value of the χ^2 -test (optimized for by LinDen), the Bayesian network score K2 (optimized for by MACOED), the negative log-likelihood score of the maximum likelihood model (MLM) (optimized for by NeEDL per default) and the differences between the negative log-likelihoods of fitted linear and quadratic regression models (NLL gain) (8). For the P -value of the χ^2 -test, the Bayesian network score K2, and the MLM score, low scores indicate promising SNP sets. Conversely, the higher the NLL gain, the more promising the scored SNP set (see Material and Methods). To show the robustness of the results and to control if higher-order candidate SNP sets achieve more significant associations just because of the number of SNPs, we generated baselines by randomly sampling 1000 size-matched SNP sets (size 2 for LinDen and MACOED, higher-order for NeEDL) to estimate the false positive rate (see Material and Methods). All P -values (both for the background distributions and the results) were adjusted using the Benjamini-Hochberg procedure (46).

Since the local search algorithm returns a result for each start seed, we obtain thousands of results. To find a suitable cut-off, we show the MLM and K2 scores for the 100 top-ranked SNP sets computed by NeEDL (Figure 2A, Supplementary Figures S7–S9). Interestingly, we observe phase transitions (i.e., an abrupt change to a lower significance of the score over linear increasing ranks) for both scores, which typically occur after 50 SNP sets (except for the LOAD and IBD datasets, with transitions between the ranks 20 and 30, and the RA dataset, with transitions at 60 SNP sets). We hypothesize that this score gap represents the barrier between promising and not promising epistasis candidate sets and thus focused on the top 50 candidates (i.e. where most transitions occur, Supplementary Figure S8) selected based on the respective metric each tool is optimizing for (NeEDL: MLM score, LinDen: P -value of χ^2 -test, MACOED: K2 score).

MACOED was executed 100 times to account for randomized subroutines, leading to the union of the top 50 candidate SNP sets for each run for MACOED. The data shown in Figure 2B, Supplementary Figure S7B, and Supplementary File 1 demonstrates that NeEDL outperforms LinDen, MACOED, and the baselines across all metrics and datasets. Even in instances where it performs comparably on some metrics on certain datasets, NeEDL still surpasses the competitor on at least one other metric across all datasets in terms of statistical association. LinDen, as well as the second-order- and higher-order baseline, are outperformed on all metrics over all datasets. On all datasets, the χ^2 -test P -values obtained for the SNP sets computed by NeEDL approach the maximum number of decimal places available in standard computation. With the more complex MLM and K2 scores, differences become more visible, and we see that NeEDL markedly outperforms MACOED in all datasets (Figure 2B, Supplementary Figure S7B). It is noteworthy that when the cut-off of the rank is set at 25 for LOAD, and IBD and 60 for RA, the advantage of NeEDL over MACOED becomes more distinct (Supplementary Figure S9). NeEDL also yields significantly lower K2 scores than MACOED on all datasets, even though NeEDL did not optimize for this score. Since NeEDL optimizes for the MLM score, we can see that NeEDL outperforms all competitors and baselines on this score. In terms of NLL gain, NeEDL again outperforms its competitors on all datasets except IBD, where

the NLL gain is similar to MACOED. Also, for LOAD, very low values of the NLL gain in MACOED indicate that, unlike NeEDL, MACOED mainly finds SNP sets with strong additive main effects. In terms of required computational resources, NeEDL and LinDen (both executable on desktop computers) are much more efficient than MACOED, which requires a super-computer (Supplementary Table S17).

Figure 2C and Supplementary Figures S7C and 9 show that, in most datasets, the best scores are achieved with three to seven interacting SNPs. Since we ran NeEDL with a maximum SNP set size of 10, NeEDL likely does not exploit combinatorial statistical artifacts but prioritizes SNPs that may be involved in protein-protein interactions and thus be more likely to be biologically meaningful. Additionally, we investigated if eQTL-based SNP-to-gene mappings could provide more insights into SNP-SNP relationships than the mostly positional mappings provided by dbSNP. For this, we constructed an SSI network based on eQTLs from the eQTL Catalogue (44) (see Material and Methods for details) and then ran NeEDL on this network. Supplementary Figure S10C shows the obtained MLM scores in comparison to the scores obtained with the dbSNP-based network. The eQTL-based approach leads to worse scores on all of our datasets. A possible explanation for this is that the eQTL-based network is substantially smaller than the dbSNP-based network (Supplementary Figure S10A and B), suggesting that information on eQTLs is still too sparse to be useful for epistasis detection.

To further examined whether the use of the biological priors encoded in the dbSNP-based SSI network indeed guides NeEDL toward more promising SNP sets, we generated 100 randomized networks with preserved topology (by shuffling the node labels) and 100 randomized networks where the SNPs' node degrees are preserved in expectation (Supplementary Figure S2A). For this analysis, we randomly selected three of the eight datasets (BD, T2D and RA) to avoid excessive runtime. MLM scores on the original SSI networks are, on average, better than those on the randomized networks (Figure 2D).

NeEDL identifies epistatic associations beyond random effects

To investigate whether the scores obtained by NeEDL may be explainable by random effects alone, we ran NeEDL and its competitors (except for MACOED due to its extremely high resource requirements) on the BD, T2D and RA datasets with randomly shuffled phenotypes (Results in Supplementary File 1). We observe that the score gaps we obtained with the original phenotypes after around 50 SNP sets disappear (Figure 3A). Moreover, w.r.t. all four scoring metrics, NeEDL yields substantially better results for the original dataset than for the phenotype-shuffled datasets, and the gap between NeEDL and LinDen disappears for the shuffled phenotypes (Figure 3B). Further, the SNP sets returned by NeEDL for the shuffled datasets contain more SNPs (Figure 3C) than those returned for the original datasets (Figure 2C), potentially due to the fact that, in the absence of a true signal in the data, a higher number of variables leads to an improvement of the statistical scores because of combinatorial statistical artifacts. Additionally, we observe that, for shuffled phenotypes, running NeEDL with a PPI-based SSI network does not lead to improved scores compared to random networks (Figure 3D)—in contrast to the results obtained for the orig-

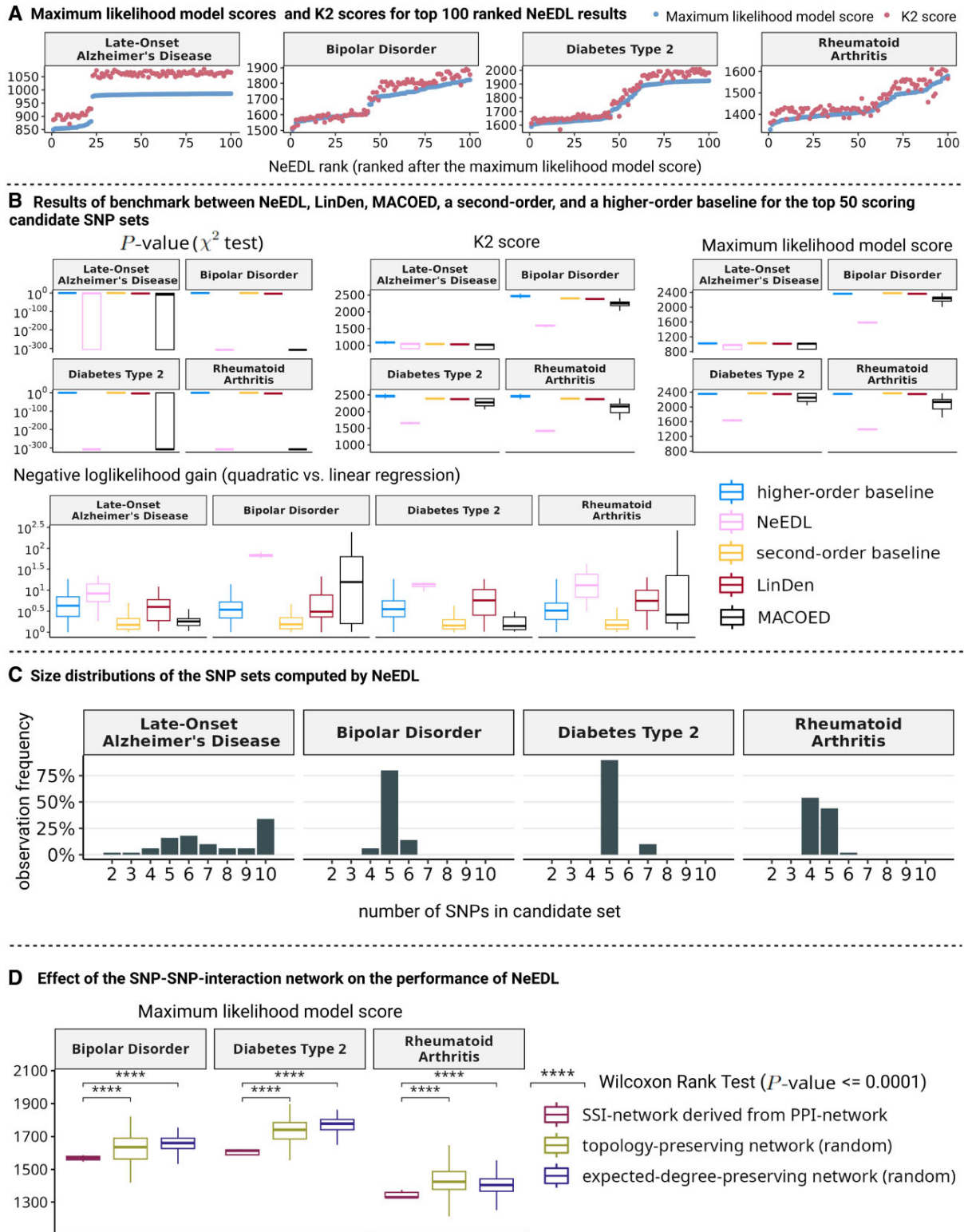
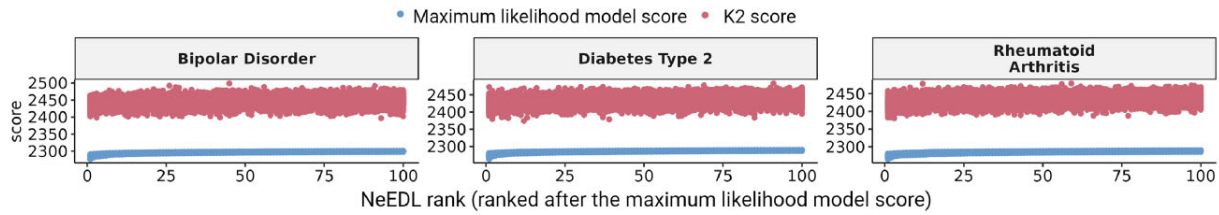
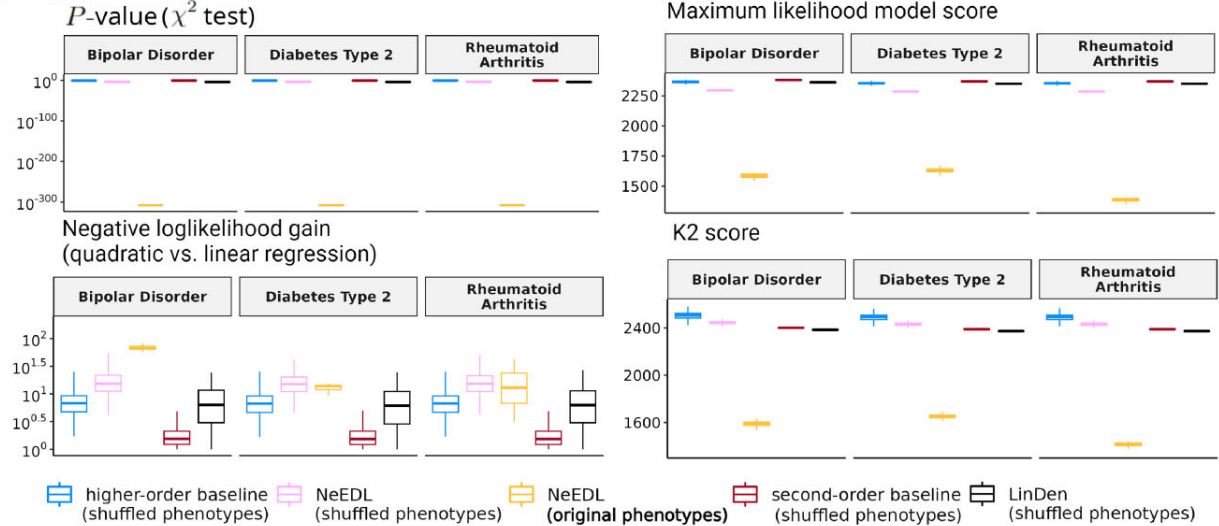


Figure 2. Quantitative evaluation of the SNP sets computed by NeEDL for Late-Onset Alzheimer's Disease, Bipolar Disorder, Diabetes Type 2, and Rheumatoid Arthritis. Results for Coronary Artery Disease, Diabetes Type 1, Hypertension, and Inflammatory Bowel Disease can be found in [Supplementary Figure S7](#). (A) Visualization of the maximum likelihood model score and the K2 score of the top 100 candidate SNP sets ranked by NeEDL. (B) A benchmark study between NeEDL, LinDen, MACOED, a second-order baseline (i.e. 1000 random sampled pairs of SNPs), and a higher-order baseline (i.e. 1000 randomly sampled sets consisting of multiple SNPs) shows that NeEDL outperforms in statistical significance existing epistasis detection tools w.r.t. four different evaluation metrics. (C) Analyzing the number of SNPs included in NeEDL's output SNP sets reveals that the most promising SNP sets typically contain between three and seven SNPs. (D) Comparing maximum likelihood model scores of NeEDL results against those obtained using randomized networks demonstrates that the use of the SSI network indeed leads to the discovery of more promising SNP sets.

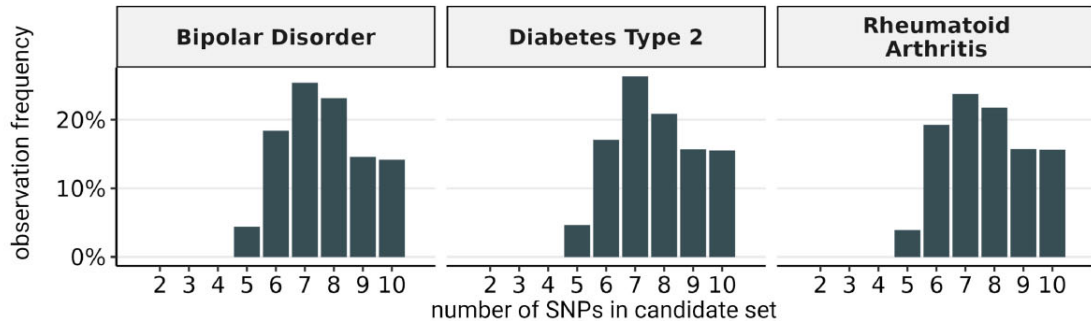
A Maximum likelihood model scores and K2 scores for top 100 ranked NeEDL results (shuffled phenotypes)



B Phenotype shuffling benchmark between NeEDL, LinDen, a second-order, and a higher-order baseline for the top 50 scoring candidate SNP sets



C Size distributions of the SNP sets computed by NeEDL (shuffled phenotypes)



D Effect of the SNP-SNP-interaction network on the performance of NeEDL (shuffled phenotypes)

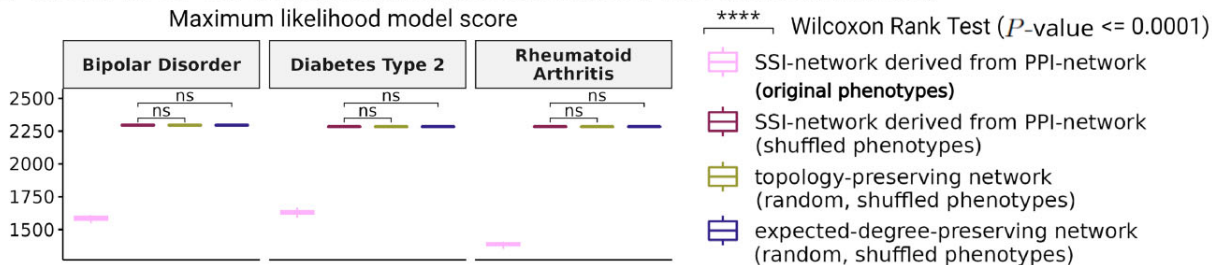


Figure 3. Evaluation of the SNP sets computed by NeEDL and LinDen with shuffled phenotypes for Bipolar Disorder, Diabetes Type 2, and Rheumatoid Arthritis. **(A)** Maximum likelihood model and K2 scores for the top 100 SNP sets reported by NeEDL (ranked by the maximum likelihood model score). **(B)** Comparison of the scores of the top 50 results of NeEDL and LinDen and the second- and higher-order baseline. As a reference, the scores of the original NeEDL runs without phenotype shuffling are also shown (yellow box plots). **(C)** SNP set size distribution of NeEDL's results. **(D)** Maximum likelihood model scores obtained by NeEDL with and without network rewiring and phenotype shuffling.

inal phenotypes where we did observe statistically significant improvements (Figure 2D). Overall, these results are strong evidence that NeEDL indeed picks up on true biological signals and that its usage of a PPI-based SSI network is crucial for this.

Consistent results across independent cohorts

We conducted a replication study to validate the findings of NeEDL and to demonstrate the consistency of the results across independent datasets for LOAD. For this, we retrieved two independent replication datasets (independent from the discovery dataset used in the benchmark) from the UK Biobank (see Material and methods), one with individuals of British ancestry (similar to the LOAD dataset) and one with individuals of mixed ancestry (i.e. 10% of the samples are of mixed ancestries like Latino, African, or Asian) to estimate population-specific effects. From the SNP sets computed by NeEDL on the LOAD dataset (i.e. the dataset used for discovery), we selected the 50 SNP sets with the most significant MLM scores (to be consistent with the benchmark) and re-computed their χ^2 -test *P*-values, MLM scores, K2 scores and NLL gains.

Figure 4 shows that we could replicate the phase transition observed in the LOAD dataset used for the benchmark (Figure 2A): The top 22 SNP sets and the top 23 SNP sets of the replication data set of British ancestry and mixed ancestry, respectively, replicate with a significant χ^2 -test *P*-values < 0.05 (Supplementary Figure S11). The MLM scores of replication correlate (Pearson correlation of 0.995) highly with the MLM scores of the benchmark study (Figure 4B). We can further observe a drop in the Pearson correlation to 0.993 in the mixed ancestry dataset, suggesting that the population-specific genetic architecture might play a role in epistasis detection. Similar to the LOAD dataset, a phase transition occurs in *P*-values and MLM scores after the top 22 SNP sets (Supplementary Figure S11). We also re-computed χ^2 -test *P*-values, MLM scores and K2 scores for the 50 best SNP sets computed by MACOED and LinDen on the LOAD benchmark dataset. Consistent with the discovery study, NeEDL with its top 50 candidate sets outperforms LinDen in both replication datasets (Figure 4C). MACOED is outperformed in the *P*-values of the χ^2 -test, the K2 score, and the MLM score. However, MACOED yielded similar results or slightly less significant scores with the NLL-gain. Overall, the replication results are consistent with the results of the benchmark in the discovery dataset.

Epistasis candidates uncovered by NeEDL are biologically plausible

By utilizing NeEDL, we identified potential epistatic candidate SNP sets that have been selected based on their maximum likelihood model scores. We selected LOAD, T1D and IBD (the latter two in Supplementary Materials 7) as they are well-represented in the literature. In each of these diseases, various SNPs of a single gene in different combinations with SNPs of several other genes are present in the majority of the statistically most significant candidate SNP sets (Figure 4B, Supplementary Figure S11). We evaluated the top genes before the gap that were affected by the predicted SNP interactions by performing gene set enrichment analysis (GSEA) (56,57) to identify statistically significant enrichment in Gene On-

tology (56) and KEGG (68) (see Materials and methods for details).

APOE was present in the 25 highest penetrance SNP combinations for the disease LOAD. APOE and the lipoprotein metabolism pathway have been previously recognized as the most critical gene and pathway for Alzheimer's disease risk (24,69). Since APOE mutation status alone is not a reliable predictor of Alzheimer's disease, a better understanding of epistatic interactions may be instrumental in developing better diagnostics. NeEDL identified 47 additional genes exhibiting potential epistatic SNP combinations with APOE (Supplementary File 2). GSEA determined that 15 genes (APOE, SORL1, SMAD3, LRP2, YBX3, TP63, PRKCB, DAB1, LRP1, IFI16, UBQLN1, A2M, DYRK1A, HMGXB4 and TRAF3IP1), including APOE, were enriched in a gene set that includes the process to reduce the response to the stimulus (GO:0048585; $k/K = 0.0088$; *P*-value = 6.25×10^{-11} ; FDR *q*-value = 9.82×10^{-7}). Of note, one of the epistatic SNP combinations identified by NeEDL was rs429358 and rs7412 in the APOE gene, which has been previously identified as a potential epistatic SNP combination (70) (note that the good results of NeEDL's competitor MACOED on the LOAD dataset can be explained by the fact that this combination of SNPs was found also by MACOED).

In the Human Protein Atlas (71), we can observe that the identified genes are significantly expressed in tissues that are reported to be affected by LOAD, T1D or IBD (Supplementary Figure S12) (72–84).

Quantum computing improves seeding of local search

The GWAS datasets employed in this study cover up to 140 000 SNPs due to necessary cleaning and filtering steps (Supplementary Tables S2–S10), a small fraction of the 84.7 million SNPs (85) in the human genome. It is hence likely that, in the future, we will see a sharp increase in the number of covered and mappable SNPs. In the NeEDL workflow, this would result in an SSI network containing millions of SNPs. To explore such a huge network via randomly seeded local search with reasonable coverage, we would have to dramatically increase the number of initial solutions—certainly, beyond the capacities of classical high-performance computing clusters. We thus explored the use of quantum computing in NeEDL (see Materials and methods for details) to find a promising set of initial solutions for the local search instead of starting with random seeds as in the classical implementation (Supplementary Figure S2B). To be able to use a quantum computer, we modeled the problem of finding promising initial solutions as a variant of the maximum clique problem, transformed it into an instance of the QUBO problem (86), and then solved it heuristically using several quantum computing machines and simulators, including the quantum annealer D-Wave Advantage 6.1, the superconducting-based IBM Perth and IBM Lagos devices.

To test this approach, we subsampled our BD, RA, and T2D datasets to 100, 500 and 1000 SNPs and ran NeEDL with random and quantum-computing-based seeding. Running quantum-computing-based seeding on the entire datasets is still infeasible due to limitations of current quantum computing hardware. Supplementary Figure S5 shows that when using quantum computing, fewer but higher-quality candidate SNP sets are returned, and that the speed up these high-

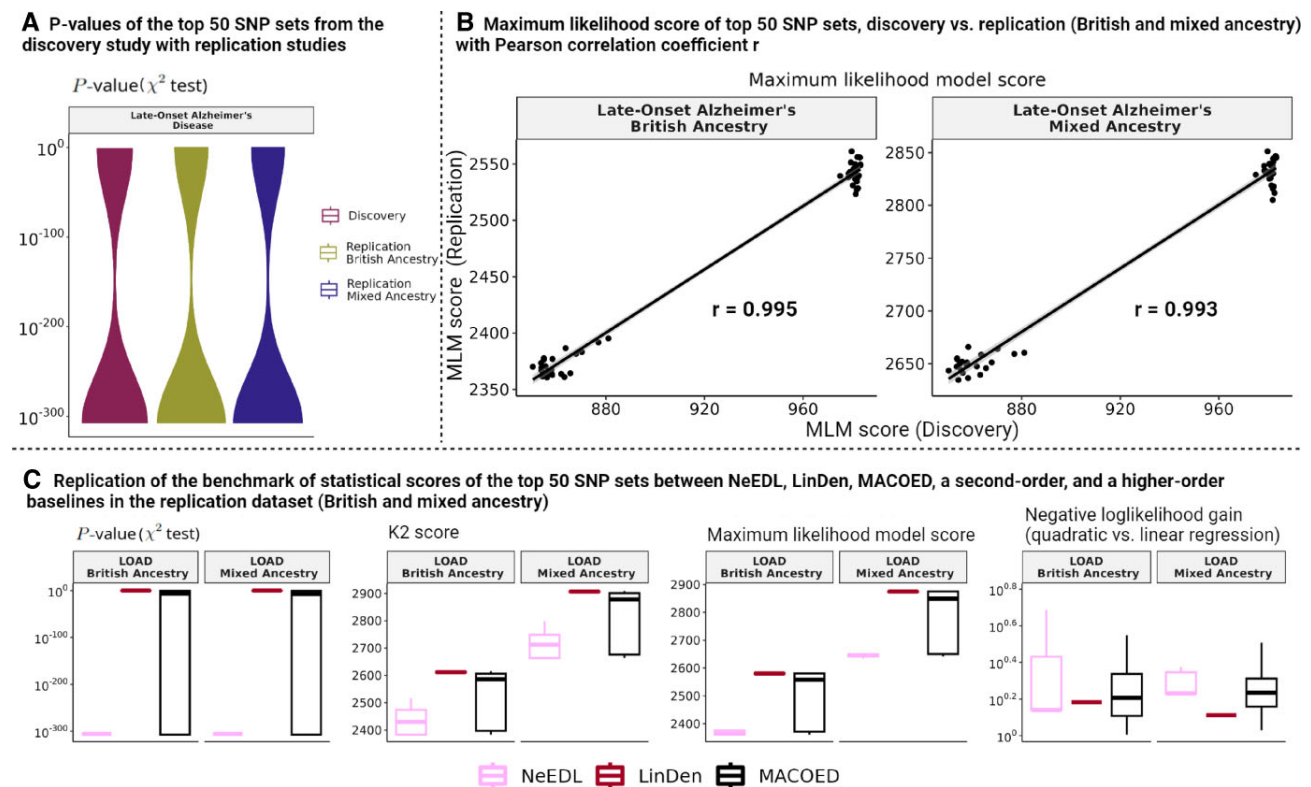


Figure 4. Replication studies in two independent data sets from UK Biobank with British and mixed ancestry. **(A)** P -values across discovery and the two replication studies. **(B)** Correlation of the MLM score between the discovery study and the two replication studies. **(C)** Replication of the benchmarking between NeEDL, LinDen, MACOED, a second-order baseline and a higher-order baseline in two replication data sets.

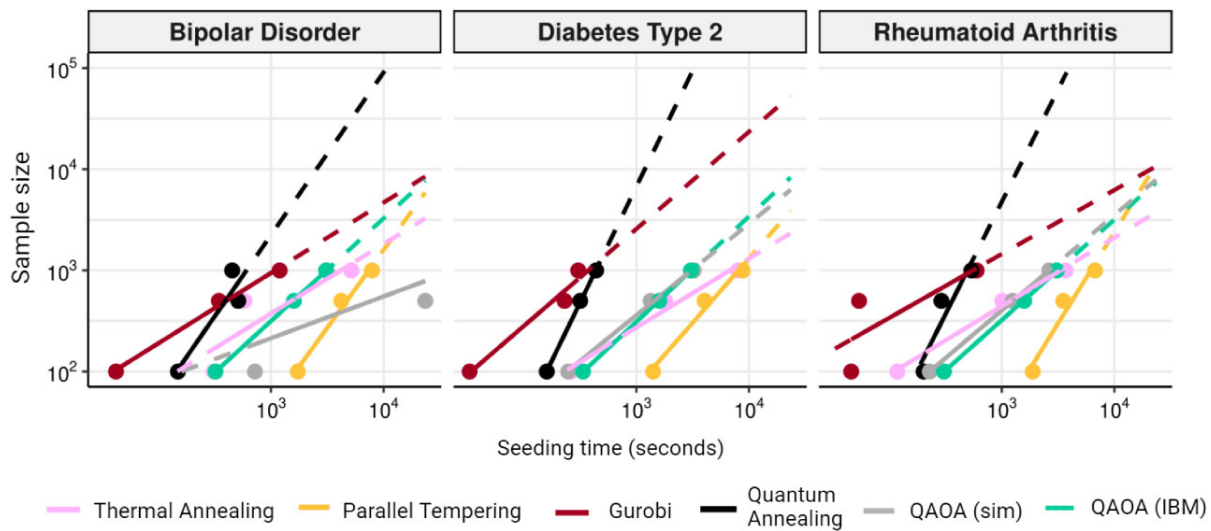


Figure 5. Experiments on the quantum computers: Expected scaling of the quantum hardware. The different quantum and classical devices used to perform the optimization process show different scaling, with the quantum devices (Quantum Annealing, QAOA) having a higher slope than classical devices (thermal annealing, parallel tempering, and Gurobi), suggesting a possible scaling advantage.

quality initial solutions induce in the downstream local search outweighs the additional runtime needed for the quantum-computing-based seeding. Since quantum computing hardware cannot yet process big data, we estimate the speed up on realistically sized GWAS data across different devices (Figure 5). Notably, any quantum device, whether simulated or executed on actual hardware, scales better than the baseline clas-

sical algorithm. We may hence anticipate quantum computation to surpass the classical baseline in the future, showcasing a potential quantum advantage once more powerful quantum computing hardware will become available, even though currently only small subproblems can be addressed. To conduct our quantum computing experiments, we used more than 4.1 million seconds of computing time on quantum resources.

Exploring epistatic interactions in the Epistasis Disease Atlas

Since more than two million CPU hours were necessary to calculate the results reported here, we make them widely and easily accessible through the web-based Epistasis Disease Atlas (<https://epistasis-disease-atlas.com>). The Epistasis Disease Atlas provides results of eight heritable diseases (see Section Datasets). It enables users to browse, visualize, interpret, and link candidate SNP sets and their interactions within the SSI network to published literature and a variety of external databases and tools (see Materials and methods for details, (87)). The Epistasis Disease Atlas can further be queried through an application programming interface (API) where we offer both R and Python packages (Supplementary Materials 8, including Supplementary Figures S12–S15).

Discussion

The initially high expectations towards genome-wide association studies have thus far not completely been fulfilled due to the missing heritability problem (4,5), which suggests that individual genetic variants cannot explain the genetic architecture of complex and polygenic heritable diseases. However, the missing heritability is likely overestimated since epistatic interactions are implicitly ignored in these estimates (88). While statistical models and computational tools for identifying epistatic interactions have been proposed, the space of possible interactions to be tested is vast and could thus far mostly tackle pairwise interactions. With NeEDL, we overcame this problem and suggest promising higher-order SNP sets. NeEDL outperforms the existing tools LinDen and MA-COED, which are limited to pairwise interactions, across all commonly used statistical models and offers new insights into the genetic architecture of complex diseases. We show evidence that candidate SNP sets that are strongly associated with disease phenotypes are in the order of three to seven, are biologically plausible, and can be verified in independent cohorts.

The candidate SNP sets identified by NeEDL represent locally optimal solutions within specific areas of an SSI network and are thus limited by the quality of the priors in this network. While the current version of NeEDL generally outperforms random priors, we see room for further improvement in designing the SSI networks. For instance, PPI networks such as BioGRID may contain a substantial amount of missing interactions such that a globally optimal SNP set cannot be found by NeEDL, i.e., due to yet unknown or tissue-specific PPIs. Similarly, existing PPI networks neglect the consequences of alternative splicing on functional interactions (89,90). Furthermore, SNPs that alter codon efficiency during the translation process can reduce or inhibit protein activity (91), which ultimately can alter a PPI network. More disease-associated variants are found in non-coding compared to coding regions of the DNA, suggesting a major role of gene regulation in human diseases (92). This motivates further research in SSI networks that consider enhancer-gene interaction and chromatin remodeling. Including gene-regulatory interactions will also pose a significant challenge for interpretability, as the current strategy of GSEA and KEGG enrichment is limited to established interactions between genes and proteins. Moreover, for the dbSNP- and eQTL based SSI networks used in this study, we consider only SNPs that have previously been associated

with genes in, respectively, dbSNP and the eQTL Catalogue. Other associations are neglected and possible misannotation in the external databases used to generate our SSI network are inherited. In summary, the higher the quality of the prior knowledge NeEDL can use to reduce the search space, the more meaningful the results will be.

We should further consider the limitations of the statistical models employed in NeEDL. Recent work by Blumenthal *et al.* (8) have demonstrated significant differences in the detection power of various statistical models, concluding that the MLM score offers the best tradeoff between quality and performance, outperforming the quadratic regression model, Bayesian model, and variance model. The *P*-values obtained through the χ^2 -test are extremely small (approaching the minimum C++ double value) and preclude further differentiation between candidate SNP sets. Hence, we recommend using the MLM score. Covariates, such as age, comorbidities, environmental factors, lifestyle factors, and population structure, could negatively impact the discovery of epistatic interactions. While regression models can, in principle, control for such effects (which is also supported by NeEDL), this further increases the computational demand beyond currently available resources. Further algorithmic improvements in mining the SSI network, maybe through graph neural networks, are conceivable and will be a promising research direction (90). As a potential solution to further speed up the local search algorithm of NeEDL, we considered various quantum algorithms and devices, showing that epistasis detection poses a real-world scenario in which quantum advantage could be achieved and exploited. In the near future, quantum algorithms (potentially enhanced via more sophisticated approaches such as warm-started QAOA (93)) may hence allow NeEDL to handle increasingly complex datasets with millions of SNPs.

Replication of epistasis candidate SNP sets in independent cohorts, as shown here, can provide valuable confirmation of statistical associations and offer a testbed for assessing the clinical potential of epistatic interaction scores similar to polygenic risk scores. To date, multi-SNP epistasis hypotheses are difficult to verify experimentally. Despite these challenges, future gene editing studies will be needed for their verification and to unravel the functional consequences of epistatic interactions in disease pathophysiology. To optimally support this process, we developed the Epistasis Disease Atlas, a user-friendly and interactive web resource that will allow the biomedical community to extract testable hypotheses directly and work towards an improved understanding of disease mechanisms as the basis for developing targeted therapeutic interventions. We note that the application of NeEDL is wider than just human disease research and see the potential for applying NeEDL in model organisms such as *Arabidopsis thaliana* or *Mus musculus*, where gene editing and functional studies are more readily achievable.

In conclusion, NeEDL is an epistasis detection tool to leverage prior biological knowledge for the systematic discovery of higher-order candidate SNP sets and quantum computing. NeEDL thus lays the foundation for charting the complex genetic architecture underlying heritable diseases. NeEDL lifts epistasis detection to a systems-oriented network biology level and shows that seamless integration of quantum computing has the potential to transform computational genomics, highlighting its role in biomedical research.

Data availability

The LOAD data set is restricted and can be found at <https://www.tgen.org>. The users need to apply for the dataset. The BD (EGAD00000000003), CAD (EGAD00000000004), T1D (EGAD00000000008), T2D (EGAD00000000009), HT (EGAD00000000006), IBD (EGAD00000000005), RA (EGAD00000000007) and the British 1958 British Birth Cohort (EGAD00000000001) datasets are restricted and can be found at <https://www.sanger.ac.uk/legal/DAA/MasterController> and <https://edam.sanger.ac.uk/#/>. The user needs to apply for the datasets. The results of NeEDL stored in the Epistasis Disease Atlas are freely available under the CC BY-NC 4.0 license.

In the future, we plan to include the following datasets in the Epistasis Disease Atlas: AS (EGAD00000000010), ATD (EGAD00000000011), MS (EGAD00000000012), BRCA (EGAD00000000013), TB (EGAD00000000016), UC (EGAD00000000025), PD (EGAD00000000057), AK (EGAD00010000150), SP (EGAD00010000262), IS (EGAD00010000264), CD (EGAD00010000246); these data sets are restricted and can be found at <https://www.sanger.ac.uk/legal/DAA/MasterController>. The user needs to apply for the data sets. The datasets for replication are restricted and can be found at the UK Biobank database (www.ukbiobank.ac.uk), under project ID 54273.

A graphical overview of the whole project is provided in [Supplementary Figure S16](#). [Supplementary File 1](#) (all results and scripts to plot them) is available on figshare: <https://doi.org/10.6084/m9.figshare.25780350.v1>.

Code availability

The source code of NeEDL, the quantum computing module, and the R Shiny App is freely available under the GPLv3 license at GitHub: <https://github.com/biomedbigdata/NeEDL>. All features of NeEDL are dockerized and available at Dockerhub: <https://hub.docker.com/r/bigdatainbiomedicine/needl>.

Supplementary data

[Supplementary Data](#) are available at NAR Online.

Acknowledgements

We want to thank Christina Trummer for the design of the NeEDL logo and the Epistasis Disease Atlas logo, as well as the design of the webpage. We further want to thank Simona Dedurkeviciute for her advice on webpage design and architecture. Images were created with <https://www.biorender.com>. Parts of the images were designed using resources from Flaticon.com under a paid license. ChatGPT version 4, under a paid license, was used to rephrase parts of this manuscript. The database diagrams were created with the support of the LucidApp, which gratefully provided us with a free premium account. Grammarly, under a paid license, was used for spell and grammar checks. Overleaf, under a paid license, was used to draft the manuscript in LaTeX. Paperpile, under a paid license, was used to collect references in the right format. We would like to thank Matt Huentelman from The Translational Genomics Research Institute in Phoenix, Arizona, for sharing the TGen II cohort data (LOAD). This study makes use of data generated by the Wellcome TrustCase-ControlConsortium. A

full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under awards 076113, 085475 and 090355. The authors gratefully acknowledge the Leibniz Supercomputing Centre for funding this project by providing computing time and support on its Linux Cluster. The authors gratefully acknowledge the Gauss Centre for Supercomputing e.V. (www.gauss-centre.eu) for funding this project by providing computing time on the GCS Supercomputer SuperMUC-NG at Leibniz Supercomputing Centre (www.lrz.de). We acknowledge the CINECA award under the ISCRA initiative, for the availability of high-performance computing resources and support (QPU D-Wave). Access to the IBM Quantum Services was obtained through the IBM Quantum Hub at CERN. The views expressed are those of the authors and do not reflect the official policy or position of IBM, the IBM Q team. We acknowledge support from Microsoft's Azure Quantum for providing credits and access to the quantum hardware used in this paper.

Author contributions: M.H.O. designed and developed most parts of the software architectural concepts, database design, and features of epiJSON, NeEDL, calculate_scores, the quantum-computing module, the Epistasis Disease Atlas, and the R Shiny App of the Epistasis Disease Atlas. MHO implemented the initial idea, conceptualized and conducted most analyses, wrote the manuscript, and held the role of the project manager and project leader. JP implemented large parts of NeEDL, benchmarked NeEDL against the competitors, and implemented the concept of the real-time calculation of the scores for the Epistasis Disease Atlas. M.I. implemented the quantum computing module of NeEDL and conducted several analyses on various quantum-computing hardware. S.B. conducted literature research, applied for most of the data sets, and implemented the R Shiny App of the Epistasis Disease Atlas. MHO and SB produced the overview video on the EDA. A.F., A.G.P. and K.B. conducted the replication studies of the potential candidate SNP sets in independent cohorts. AM and CH helped with the technical conceptualization and implementation of the backend (database, API and file server) of the Epistasis Disease Atlas. M.H.A. implemented and designed the front end of the Epistasis Disease Atlas. C.H. implemented the automatic pipeline to integrate new data sets into the database of the Epistasis Disease Atlas. N.T. implemented the features of the Integrated Genome Browser. K.A. implemented the R and Python packages of the Epistasis Disease Atlas. M.P. and M.W. conceptualized and executed the big memory runs and long-time runs of NeEDL and MACOED on the LRZ supercluster. E.S. implemented the epiJSON tool and the linkage-disequilibrium options of NeEDL. ES further reformatted the datasets of the eight discussed diseases into NeEDL readable format. P.W., H.F. and S.K. helped with analyses. M.V.H. implemented the co-variate and ancestry options. I.L. tested the software of NeEDL and the Epistasis Disease Atlas for bugs. LSCHW and LHAF conducted further research on epistatic SNPs in regulatory elements. G.G. and A.A. processed the raw Illumina data for the additional ten datasets to NeEDL readable format that are about to be integrated into the Epistasis Disease Atlas but not further discussed in this manuscript. L.H.A.C. and O.T. conducted further research on epistatic SNPs in alternative splicing site regions. N.W. designed the architecture of the Figures and revised the manuscript. SGL and GC conducted further research on epistatic SNPs in structural biology and binding sites. MC conducted further research on

epistatic SNPs that change translational speed. J.J., H.K.L., P.F. and L.H.E.N. conducted biological plausibility studies and drafted parts of the manuscript. A.Man. performed the theoretical analysis for the quantum computing module of NeEDL. F.M., H.C.G. and K.V.S. provided expertise in epistasis detection throughout the project, suggested algorithmic changes, and revised the manuscript. L.S.C.H.U., L.I. and M.G. provided quantum-computing expertise and quantum-computing hardware and helped calibrate the systems. L.W. and D.E. provided expertise on the math for the co-variant and mixed-ancestry options of NeEDL. A.D.P., A.Man. and M.G. supervised the quantum-computing part of NeEDL. P.F. and L.H.E.N. supervised the biomedical part of the project. T.K., J.B., M.L. and D.B. jointly supervised the project. J.B. and T.K. had the initial idea for network-enhanced higher-level epistasis detection. DB had crucial ideas regarding algorithmic optimization solutions. J.B. and D.B. first suggested the use of quantum computing. D.B. implemented the statistical models and, together with MHO, wrote most parts of the manuscript. This project took more than five and a half years to implement, execute, and evaluate—which is the reason for eight first authors and five senior authors. All authors revised the manuscript and gave their approval.

Funding

Technical University Munich—Institute for Advanced Study, funded by the German Excellence Initiative (in part); Intramural Research Program (IRP) of the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK, in part); M.G. is supported by CERN through the CERN Quantum Technology Initiative; A.Man. is supported by the Foundation for Polish Science (FNP), IRAP project ICTQT [2018/MAB/5], co-financed by the EU Smart Growth Operational Programme; German Federal Ministry of Education and Research (BMBF) within the framework of the *e:Med* research and funding concept [*ZX1908A/01ZX2208A*, *01ZX1910D/01ZX2210D*]; European Union's Horizon 2020 research and innovation programme [777111]; this publication reflects only the authors' view and the European Commission is not responsible for any use that may be made of the information it contains; Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) [422216132, 516188180]; the research of LI and LSCHU is partially funded by the Bavarian State Ministry of Science and the Arts as part of the Munich Quantum Valley; European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie [813533 to M.L.F.P.M., 860895 to TRANSYS]; FNRS convention PDR T.0294.24 'Expanded PRS embracing pathways and interactions for increased clinical utility'. Funding for open access charge: TUM (one of the corresponding authors markus.list@tum.de, has the right to publish at OUP journals). M.P. was supported by the German Federal Ministry of Education and Research (BMBF; Grant No. 031L0305A).

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the author(s) used ChatGPT 4 under a paid premium license to rephrase the text for more clarity, conciseness, and grammatical correctness. After using this tool/service, the author(s) reviewed and edited the

content as needed and take(s) full responsibility for the content of the publication.

Conflict of interest statement

During the course of the project, H.C.G. became a full-time employee of Novo Nordisk Ltd. M.W. is a founder and shareholder of MSAID GmbH and OmicScouts GmbH. He has no operational role in either company.

References

1. Heap,G.A., Trynka,G., Jansen,R.C., Bruinenberg,M., Swertz,M.A., Dinesen,L.C., Hunt,K.A., Wijmenga,C., Vanheel,D.A. and Franke,L. (2009) Complex nature of SNP genotype effects on gene expression in primary human leucocytes. *BMC Med. Genom.*, **2**, 1.
2. Bush,W.S. and Moore,J.H. (2012) Chapter 11: Genome-wide association studies. *PLoS Comput. Biol.*, **8**, e1002822.
3. MacArthur,J., Bowler,E., Cerezo,M., Gil,L., Hall,P., Hastings,E., Junkins,H., McMahon,A., Milano,A., Morales,J., *et al.* (2017) The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.*, **45**, D896–D901.
4. Gibson,G. (2012) Rare and common variants: twenty arguments. *Nat. Rev. Genet.*, **13**, 135–145.
5. Lippert,C., Listgarten,J., Davidson,R.I., Baxter,S., Poon,H., Kadie,C.M. and Heckerman,D. (2013) An exhaustive epistatic SNP association analysis on expanded Wellcome Trust data. *Sci. Rep.*, **3**, 1099.
6. Manolio,T.A., Collins,F.S., Cox,N.J., Goldstein,D.B., Hindorf,L.A., Hunter,D.J., McCarthy,M.I., Ramos,E.M., Cardon,L.R., Chakravarti,A., *et al.* (2009) Finding the missing heritability of complex diseases. *Nature*, **461**, 747–753.
7. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
8. Blumenthal,D.B., Baumbach,J., Hoffmann,M., Kacprowski,T. and List,M. (2020) A framework for modeling epistatic interaction. *Bioinformatics*, **37**, 1708–1716.
9. Caylak,G., Tastan,O. and Cicek,A.E. (2020) Potpourri: an epistasis test prioritization algorithm via diverse SNP selection. *J. Comput. Biol.*, **28**, 365–377.
10. Cowman,T. and Koyutürk,M. (2017) Prioritizing tests of epistasis through hierarchical representation of genomic redundancies. *Nucleic Acids Res.*, **45**, e131.
11. Ayati,M. and Koyutürk,M. (2014) Prioritization of genomic locus pairs for testing epistasis. In: *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics, BCB '14*. Association for Computing Machinery, NY, pp. 240–248.
12. Jing,P.-J. and Shen,H.-B. (2015) MACOED: a multi-objective ant colony optimization algorithm for SNP epistasis detection in genome-wide association studies. *Bioinformatics*, **31**, 634–641.
13. Duroux,D., Climente-González,H., Azencott,C.-A. and Van Steen,K. (2022) Interpretable network-guided epistasis detection. *Gigascience*, **11**, giab093.
14. Banchi,L., Fingerhuth,M., Babej,T., Ing,C. and Arrazola,J.M. (2020) Molecular docking with Gaussian boson sampling. *Sci. Adv.*, **6**, eaax1950.
15. Boev,A., Rakitko,A., Usmanov,S., Kobzeva,A., Popov,I., Ilinsky,V., Kiktenko,E. and Fedorov,A. (2021) Genome assembly using quantum and quantum-inspired annealing. *Sci. Rep.*, **11**, 13183.
16. Nałęcz-Charkiewicz,K. and Nowak,R.M. (2022) Algorithm for DNA sequence assembly by quantum annealing. *BMC Bioinformatics*, **23**, 122.
17. Sarkar,A., Al-Ars,Z. and Bertels,K. (2021) QuASer: Quantum Accelerated de novo DNA sequence reconstruction. *PLoS One*, **16**, e0249850.

18. Vakili, M.G., Gorgulla, C., Nigam, A., Bezrukov, D., Varoli, D., Aliper, A., Polykovsky, D., Padmanabha Das, K.M., Snider, J., Lyakisheva, A., *et al.* (2024) Quantum computing-enhanced algorithm unveils novel inhibitors for KRAS. arXiv doi: <https://arxiv.org/abs/2402.08210>, 13 february 2024, preprint: not peer reviewed.
19. Siek, J., Lumsdaine, A. and Lee, L.-Q. (2002) *The Boost Graph Library: User Guide and Reference Manual*. Addison-Wesley.
20. Csardi, G., Nepusz, T. and Others (2006) The igraph software package for complex network research. *InterJournal, Complex Syst.*, **1695**, 1–9.
21. Liu, F. and Chaudhary, V. (2003) A practical OpenMP compiler for system on chips. In: *OpenMP Shared Memory Parallel Programming*. Springer, Berlin, Heidelberg, pp. 54–68.
22. Marees, A.T., de Kluiver, H., Stringer, S., Vorspan, F., Curis, E., Marie-Claire, C. and Derks, E.M. (2018) A tutorial on conducting genome-wide association studies: quality control and statistical analysis. *Int. J. Methods Psychiatr. Res.*, **27**, e1608.
23. Kunkle, B.W., Grenier-Boley, B., Sims, R., Bis, J.C., Damotte, V., Naj, A.C., Boland, A., Vronskaya, M., van der Lee, S.J., Amli-Wolf, A., *et al.* (2019) Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates A β , tau, immunity and lipid processing. *Nat. Genet.*, **51**, 414–430.
24. Reiman, E.M., Webster, J.A., Myers, A.J., Hardy, J., Dunckley, T., Zismann, V.L., Joshupura, K.D., Pearson, J.V., Hu-Lince, D., Huentelman, M.J., *et al.* (2007) GAB2 alleles modify Alzheimer's risk in APOE epsilon4 carriers. *Neuron*, **54**, 713–720.
25. Wellcome Trust Case Control Consortium, Craddock, N., Hurles, M.E., Cardin, N., Pearson, R.D., Plagnol, V., Robson, S., Vukcevic, D., Barnes, C., Conrad, D.F., *et al.* (2010) Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature*, **464**, 713–720.
26. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., *et al.* (2018) The UK Biobank resource with deep phenotyping and genomic data. *Nature*, **562**, 203–209.
27. World Health Organization (1993) In: *The ICD-10 classification of mental and behavioural disorders: diagnostic criteria for research*, World Health Organization.
28. Best, D. and Roberts, D. (1975) Algorithm AS 89: the upper tail probabilities of Spearman's rho. *J. Roy. Stat. Soc. Ser. C (Appl. Stat.)*, **24**, 377–379.
29. Caylak, G., Tastan, O. and Cicek, A.E. (2021) Potpourri: an epistasis test prioritization algorithm via diverse SNP Selection. *J. Comput. Biol.*, **28**, 365–377.
30. Cowman, T. and Koyutürk, M. (2017) Prioritizing tests of epistasis through hierarchical representation of genomic redundancies. *Nucleic Acids Res.*, **45**, e131.
31. Guan, B., Zhao, Y. and Sun, W. (2018) Ant colony optimization with an automatic adjustment mechanism for detecting epistatic interactions. *Comput. Biol. Chem.*, **77**, 354–362.
32. Guan, B. and Zhao, Y. (2019) Self-adjusting ant colony optimization based on information entropy for detecting epistatic interactions. *Genes*, **10**, 114.
33. Schüpbach, T., Xenarios, I., Bergmann, S. and Kapur, K. (2010) FastEpistasis: a high performance computing solution for quantitative trait epistasis. *Bioinformatics*, **26**, 1468–1469.
34. Wang, Y., Liu, X., Robbins, K. and Rekaya, R. (2010) AntEpiSeeker: detecting epistatic interactions for case-control studies using a two-stage ant colony optimization algorithm. *BMC Research Notes*, **3**, 117.
35. Cao, X., Yu, G., Ren, W., Guo, M. and Wang, J. (2019) DualWMDR: Detecting epistatic interaction with dual screening and multifactor dimensionality reduction. *Hum. Mutat.*, **41**, 719–734.
36. Gola, D., John, J. M.M., van Steen, K. and König, I.R. (2015) A roadmap to multifactor dimensionality reduction methods. *Brief. Bioinform.*, **17**, 293–308.
37. Yu, W., Lee, S. and Park, T. (2016) A unified model based multifactor dimensionality reduction framework for detecting gene–gene interactions. *Bioinformatics*, **32**, i605–i610.
38. Ritchie, M.D., Hahn, L.W., Roodi, N., Bailey, L.R., Dupont, W.D., Parl, F.F. and Moore, J.H. (2001) Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am. J. Hum. Genet.*, **69**, 138–147.
39. Sinnott-Armstrong, N.A., Greene, C.S. and Moore, J.H. (2010) Fast genome-wide epistasis analysis using ant colony optimization for multifactor dimensionality reduction analysis on graphics processing units. In: *Proceedings of the 12th annual conference on Genetic and evolutionary computation - GECCO '10*. ACM Press.
40. Ansarifard, J. and Wang, L. (2019) New algorithms for detecting multi-effect and multi-way epistatic interactions. *Bioinformatics*, **35**, 5078–5085.
41. Wu, T.T., Chen, Y.F., Hastie, T., Sobel, E. and Lange, K. (2009) Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, **25**, 714–721.
42. North, B.V., Curtis, D. and Sham, P.C. (2005) Application of logistic regression to case-control association studies involving two causative loci. *Hum. Hered.*, **59**, 79–87.
43. Oughtred, R., Stark, C., Breitkreutz, B.-J., Rust, J., Boucher, L., Chang, C., Kolas, N., O'Donnell, L., Leung, G., McAdam, R., *et al.* (2019) The BioGRID interaction database: 2019 update. *Nucleic Acids Res.*, **47**, D529–D541.
44. Kerimov, N., Hayhurst, J.D., Peikova, K., Manning, J.R., Walter, P., Kolberg, L., Samoviča, M., Sakthivel, M.P., Kuzmin, I., Trevanion, S.J., *et al.* (2021) A compendium of uniformly processed human gene expression and splicing quantitative trait loci. *Nat. Genet.*, **53**, 1290–1299.
45. Durinck, S., Spellman, P.T., Birney, E. and Huber, W. (2009) Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.*, **4**, 1184–1191.
46. Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.*, **125**, 289–300.
47. Kirkpatrick, S., Gelatt, C.D. and Vecchi, M.P. (1983) Optimization by simulated annealing. *Science*, **220**, 671–680.
48. Riesen, K., Fischer, A. and Bunke, H. (2017) Improved graph edit distance approximation with simulated annealing. In: Foggia, P., Liu, C.-L. and Vento, M. (eds.) *GbRPR 2017, Vol. 10310 of LNCS*. Springer, Cham, pp. 222–231.
49. Blumenthal, D.B., Bougleux, S., Gamper, J. and Brun, L. (2019) GEDLIB: A C++ library for graph edit distance computation. In: Conte, D., Ramel, J.-Y. and Foggia, P. (eds.) *GbRPR 2019, Vol. 11510 of LNCS*. Springer, Cham, pp. 14–24.
50. Blumenthal, D.B., Boria, N., Gamper, J., Bougleux, S. and Brun, L. (2020) Comparing heuristics for graph edit distance computation. *VLDB J.*, **29**, 419–458.
51. Dorigo, M., Birattari, M. and Stutzle, T. (2006) Ant colony optimization. *IEEE Comput. Intell. Mag.*, **1**, 28–39.
52. Koza, J.R. and Poli, R. (2005) *Genetic Programming*. Springer, US, Boston, MA, pp. 127–164.
53. Duarte, A., Sánchez-Oro, J., Mladenović, N. and Todosijević, R. (2018) *Variable Neighborhood Descent*. Springer International Publishing, Cham, pp. 341–367.
54. Boria, N., Blumenthal, D.B., Bougleux, S. and Brun, L. (2019) Improved local search for graph edit distance. *Pattern Recognit. Lett.*, **129**, 19–25.
55. Lazareva, O., Baumbach, J., List, M. and Blumenthal, D.B. (2021) On the limits of active module identification. *Brief. Bioinform.*, **22**, bbab066.
56. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., *et al.* (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 15545–15550.

57. Mootha,V.K., Lindgren,C.M., Eriksson,K.-F., Subramanian,A., Sihag,S., Lehar,J., Puigserver,P., Carlsson,E., Ridderstråle,M., Laurila,E., *et al.* (2003) PGC-1 α -responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.*, **34**, 267–273.
58. Grover,L.K. (1997) Quantum mechanics helps in searching for a needle in a haystack. *Phys. Rev. Lett.*, **79**, 325.
59. Roland,J. and Cerf,N.J. (2002) Quantum search by local adiabatic evolution. *Phys. Rev. A*, **65**, 042308.
60. Pirnay,N., Ulitzsch,V., Wilde,F., Eisert,J. and Seifert,J.-P. (2024) An in-principle super-polynomial quantum advantage for approximating combinatorial optimization problems via computational learning theory. *Sci. Adv.*, **10**, eadj5170.
61. Aaronson,S. (2022) How Much structure is needed for huge quantum speedups? arXiv doi: <https://arxiv.org/abs/2209.06930>, 14 September 2022, preprint: not peer reviewed.
62. King,A.D., Raymond,J., Lanting,T., Harris,R., Zucca,A., Altomare,F., Berkley,A.J., Boothby,K., Ejtemaee,S., Enderud,C., *et al.* (2023) Quantum critical dynamics in a 5,000-qubit programmable spin glass. *Nature*, **617**, 16–66.
63. Incudini,M., Tarocco,F., Mengoni,R., Di Pierro,A. and Mandarino,A. (2022) Computing graph edit distance on quantum devices. *Quant. Mach. Intell.*, **4**, 24.
64. Kirkpatrick,S., Gelatt,C.D. and Vecchi,M.P. (1983) Optimization by simulated annealing. *science*, **220**, 671–680.
65. Earl,D.J. and Deem,M.W. (2005) Parallel tempering: theory, applications, and new perspectives. *Phys. Chem. Chem. Phys.*, **7**, 3910–3916.
66. Gurobi Optimization, LLC. (2024) Gurobi optimizer reference manual. <https://www.gurobi.com/documentation/current/refman/index.html>. version 11.
67. Traag,V.A., Waltman,L. and van Eck,N.J. (2019) From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.*, **9**, 5233.
68. Kanehisa,M. (2002) The KEGG database. *Novartis Found. Symp.*, **247**, 91–101.
69. Guo,X., Qiu,W., Garcia-Milian,R., Lin,X., Zhang,Y., Cao,Y., Tan,Y., Wang,Z., Shi,J., Wang,J., *et al.* (2017) Genome-wide significant, replicated and functional risk variants for Alzheimer's disease. *J. Neural. Transm.*, **124**, 1455–1471.
70. Kulminski,A.M., Shu,L., Loika,Y., He,L., Nazarian,A., Arbeeve,K., Ukraintseva,S., Yashin,A. and Culminskaya,I. (2020) Genetic and regulatory architecture of Alzheimer's disease in the APOE region. *Alzheimers. Dement.*, **12**, e12008.
71. Thul,P.J. and Lindskog,C. (2018) The human protein atlas: a spatial map of the human proteome. *Protein Sci.*, **27**, 233–244.
72. Braak,H. and Braak,E. (1991) Neuropathological staging of Alzheimer-related changes. *Acta Neuropathol.*, **82**, 239–259.
73. Braak,H., Thal,D.R., Ghebremedhin,E. and Del Tredici,K. (2011) Stages of the pathologic process in Alzheimer disease: age categories from 1 to 100 years. *J. Neuropathol. Exp. Neurol.*, **70**, 960–969.
74. Thal,D.R., Rüb,U., Orantes,M. and Braak,H. (2002) Phases of A beta-deposition in the human brain and its relevance for the development of AD. *Neurology*, **58**, 1791–1800.
75. Filip,P., Canna,A., Moheet,A., Bednarik,P., Grohn,H., Li,X., Kumar,A.F., Olawsky,E., Eberly,L.E., Seaquist,E.R., *et al.* (2020) Structural alterations in deep brain structures in type 1 diabetes. *Diabetes*, **69**, 2458–2466.
76. Knapp,M., Tu,X. and Wu,R. (2019) Vascular endothelial dysfunction, a major mediator in diabetic cardiomyopathy. *Acta Pharmacol. Sin.*, **40**, 1–8.
77. Schuster,D.P. and Duvvuri,V. (2002) Diabetes mellitus. *Clin. Podiatr. Med. Surg.*, **19**, 79–107.
78. Eizirik,D.L., Pasquali,L. and Cnop,M. (2020) Pancreatic β -cells in type 1 and type 2 diabetes mellitus: different pathways to failure. *Nat. Rev. Endocrinol.*, **16**, 349–362.
79. Gillespie,K.M. (2006) Type 1 diabetes: pathogenesis and prevention. *CMAJ*, **175**, 165–170.
80. Granlund,L., Hedin,A., Korsgren,O., Skog,O. and Lundberg,M. (2022) Altered microvasculature in pancreatic islets from subjects with type 1 diabetes. *PLoS One*, **17**, e0276942.
81. Stefański,A., Wolf,J., Harazny,J.M., Miskowska-Nagórna,E., Wolnik,B., Murawska,J., Narkiewicz,K. and Schmieder,R.E. (2023) Impact of type 1 diabetes and its duration on wall-to-lumen ratio and blood flow in retinal arterioles. *Microvasc. Res.*, **147**, 104499.
82. Kiseleva,E., Ryabkov,M., Baleev,M., Bederina,E., Shilyagin,P., Moiseev,A., Beschastnov,V., Romanov,I., Gelikonov,G. and Gladkova,N. (2021) Prospects of intraoperative multimodal OCT application in patients with acute mesenteric ischemia. *Diagnostics (Basel)*, **11**, 705.
83. Ahmed,M. (2021) Ischemic bowel disease in 2021. *World J. Gastroenterol.*, **27**, 4746–4762.
84. Green,B.T. and Tendler,D.A. (2005) Ischemic colitis: a clinical review. *South. Med. J.*, **98**, 217–222.
85. 1000 Genomes Project Consortium, Auton,A., Brooks,L.D., Durbin,R.M., Garrison,E.P., Kang,H.M., Korbel,J.O., Marchini,J.L., McCarthy,S., McVean,G.A., *et al.* (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.
86. Chapuis,G., Djidjev,H., Hahn,G. and Rizk,G. (2019) Finding maximum cliques on the D-wave quantum annealer. *J. Signal Process. Syst.*, **91**, 363–377.
87. Maier,A., Hartung,M., Abovsky,M., Adamowicz,K., Bader,G.D., Baier,S., Blumenthal,D.B., Chen,J., Elkjaer,M.L., Garcia-Hernandez,C., *et al.* (2024) Drugst.One — a plug-and-play solution for online systems medicine and network-based drug repurposing. *Nucleic Acids Res.*, **52**, W481–W488.
88. Zuk,O., Hechter,E., Sunyaev,S.R. and Lander,E.S. (2012) The mystery of missing heritability: genetic interactions create phantom heritability. *Proc. Natl. Acad. Sci.*, **109**, 1193–1198.
89. Louadi,Z., Yuan,K., Gress,A., Tsoy,O., Kalinina,O.V., Baumbach,J., Kacprowski,T. and List,M. (2021) DIGGER: exploring the functional role of alternative splicing in protein interactions. *Nucleic Acids Res.*, **49**, D309–D318.
90. Hernández-Lorenzo,L., Hoffmann,M., Scheibling,E., List,M., Matías-Guiu,J.A. and Ayala,J.L. (2022) On the limits of graph neural networks for the early diagnosis of Alzheimer's disease. *Sci. Rep.*, **12**, 17632.
91. Cortazzo,P., Cerveñansky,C., Marín,M., Reiss,C., Ehrlich,R. and Deana,A. (2002) Silent mutations affect in vivo protein folding in *Escherichia coli*. *Biochem. Biophys. Res. Commun.*, **293**, 537–541.
92. Watanabe,K., Stringer,S., Frei,O., Umičević Mirkov,M., de Leeuw,C., Polderman,T.J., van der Sluis,S., Andreassen,O.A., Neale,B.M. and Posthuma,D. (2019) A global overview of pleiotropy and genetic architecture in complex traits. *Nat. Genet.*, **51**, 1339–1348.
93. Tate,R., Moondra,J., Gard,B., Mohler,G. and Gupta,S. (2023) Warm-started QAOA with custom mixers provably converges and computationally beats Goemans-Williamson's Max-Cut at low circuit depths. *Quantum*, **7**, 1121.

Received: November 22, 2023. Revised: July 12, 2024. Editorial Decision: July 15, 2024. Accepted: August 1, 2024

© The Author(s) 2024. Published by Oxford University Press on behalf of Nucleic Acids Research.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

DISEASE MECHANISM VALIDATION

4.1 INTRODUCTION

In silico approaches for disease mechanism identification and DR candidate prediction are supposed to help limit the search space when looking for new remedies for underserved diseases. High certainty results cannot be guaranteed so far, as often underlying data appears to be too noisy and methods too sensitive. This means that resulting hypotheses about disease modules or DR candidates require additional steps of filtering in order to provide more confident results. For this, some kind of relevancy assessment is necessary. Especially for module detection, where purpose-built algorithms exist but a gold standard of existing and curated disease modules is lacking, complementary evaluation methods are needed.

Lacking dedicated solutions, current options include enrichment-based methods, to assess if the functional annotation or enriched functions within a module align with the disease or condition in question. Often, tools like DAVID¹ [270], g:Profiler² [271], and WebGestalt³ [272] are used to carry out standard enrichment analysis and functional annotation. Their design and goal does not include means of comparative quantification out of the box, e.g. to assess disease module quality. For that, expert-driven curation and interpretation of the coherence of enrichment vectors is necessary.

Another option is the statistical evaluation of these sets, assessing how far from a random selection they are. This is known as empirical p-value calculation and is implemented, for example, in NeDRex (used in NeDRex-Web) to easily assess the statistical significance of identified disease modules and drug rankings. While having a quantifying property by nature, this approach lacks biological context and only considers a random background, mostly similar network topology to the predicted result.

4.2 RESULTS

With **DIGEST (Publication D)** [273], we present a new *in silico* validation approach that combines functional coherence comparison and statistical probability evaluation to establish a quantifiable quality measure. Statistical significance is assessed given the network background, while using functional annotation and enrichment to evaluate the coherence of individual sets (Figure 4.1). This

¹ Unavailable by now under published URL

² <https://biit.cs.ut.ee/gprofiler/gost>

³ <https://www.webgestalt.org/>

general approach follows the assumption that a connected disease module should be relatively homogeneous in the molecular functions it is associated with. With increasing heterogeneity of function, chances rise that the set contains entries that do not fit the same functional context.

Naturally, some disease module identification methods like DOMINO produce results that are not connected in the interactome and they should not be. Each connected component can rather be regarded as a separate functional component of the disease module, as for example expected in cancer disorders, where several gain- or loss-of-function mutations have to accumulate to cause the phenotype [236]. For this, clustering-based evaluation is possible in DIGEST, allowing to assess the intra-cluster coherence for each connected component. Further, for individual connected components representing disease modules the option exists to compare against a reference set of genes, to evaluate if the functional annotation of the result module is, for example, still similar to the initial seed set of genes. Alternatively, a reference disease can be set, specifically relevant in the case of disease module prediction, to assess coherence with that original disease. In any process, the individual contribution of each gene to the final significance score can be computed and used to refine disease modules and hypotheses.

Addressing the touched upon issue of disease definitions, DIGEST provides additional functions for assessing the genetic coherence of a list of disorders instead of genes. We regarded this function highly interesting for assessing certain hypotheses concerning shared genetic mechanisms underlying certain disorders.

To make all of the discussed functions available, usable, and integrable on several levels, firstly a Python package for local use or integration into other projects was developed. To increase the usability and accessibility for non-computer scientists, a web application was added. It allows manual input of data or the configuration of analyses through its API allowing the integration into other web-focused applications, like Drugst.One, to offer an impactful tool for *in silico* disease module validation, addressing *Bottleneck 3: Lack of in silico validation strategies*.

4.3 DIGEST

4.3.1 Summary

DIGEST (*in silico* validation of **disease** and **gene** sets, clusterings or subnetworks) combines functional annotation and empirical p-value-based evaluation techniques to quantify the relevance of gene and disease sets. Specifically in the field of network-based DR this novel approach closes the gap between disease module prediction and experimental research. By improving over existing by-proxy methods using functional annotation to assess module relevancy and adherence with the disorder of interest, DIGEST can help to narrow down lists of potential disease module hypotheses to relevant module candidates that are worth further looking into. The Python-based tool is made available to clinical and pharmacological experts through an easy to use web interface at digest-validation.net⁴. A standalone REST API further allows programmatic integration into existing non-Python-based tools, whereas the core Python package allows full integration in other Python projects. To achieve this, an automatically updating database with disease and gene ID

⁴ <https://digest-validation.net>

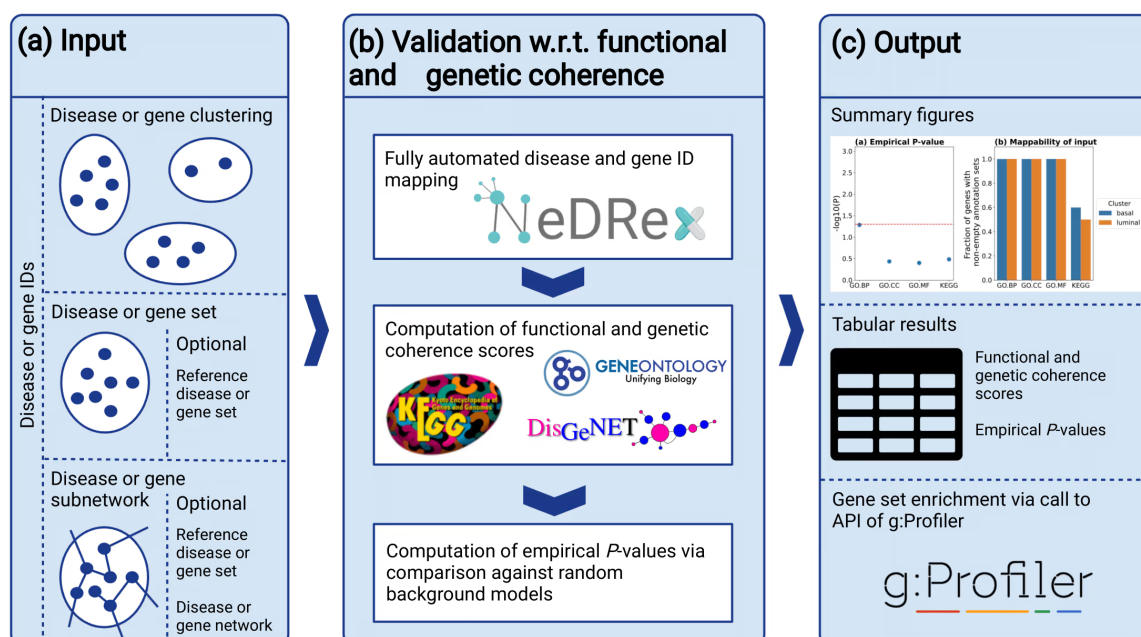
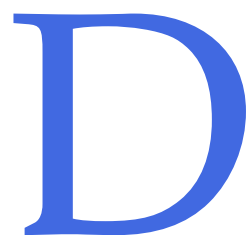


Figure 4.1: DIGEST (Reprinted from publication D [273] with permission from Oxford University Press): Overview of *in silico* validation functions provided. **(a)** Gene or disease sets, clusterings, or subnetworks can **(b)** be evaluated with or without a reference to compare to regarding their functional and genetic coherence, respectively. For this, information from NeDRex, KEGG, Gene Ontology, and others is integrated allowing the support of multiple ID spaces, the functional and genetic annotation, and the establishment of a background signal for empirical p -value calculation. **(c)** Results are persistently reachable through a stable URL and present options to deeply analyze the composition of the input selection of diseases or genes based on their functional annotation to assess mechanistic hypotheses and refine or expand the search for additional evaluation.

mapping information and modules for conducting enrichment analyses as well as the inclusion of several similarity measures to quantify genetic and genomic similarities were implemented.

DIGEST's ability to assess the coherence of gene sets, clusterings, and subnetworks enables the evaluation of disease module detection algorithms. For the results of some other published work DIGEST was able to show that the extracted modules are classified as non-significant whereas in another case DIGEST provided additional evidence about the hypothesis by Sadegh *et al.* [129] that Alzheimer's disease and hyperlipidemia are sharing some mechanistic background.



DIGEST

Klaudia Adamowicz, Andreas Maier, Jan Baumbach, and David B. Blumenthal

“Online *in silico* validation of disease and gene sets, clusterings or subnetworks with DIGEST”

Published in: *Briefings in Bioinformatics*, 2022, Vol. 23, No. 4, bbac247.

Published July 2022.

DOI: [10.1093/bib/bbac247](https://doi.org/10.1093/bib/bbac247)

Complete Reference:

Adamowicz, K., Maier, A., Baumbach, J., & Blumenthal, D. B. (2022). Online *in silico* validation of disease and gene sets, clusterings or subnetworks with DIGEST. *Briefings in Bioinformatics*, 23(4), bbac247. <https://doi.org/10.1093/bib/bbac247>

Authors' Contribution: *I am not a first author of the DIGEST publication, yet significantly contributed to the accessibility and usability of the published tool. My main contribution was the conceptualization and implementation of the DIGEST web application, including the UI, backend, data updating system, and RESTful API. Respective contributions to the publication were made by me.*

Code Availability: The source code for the DIGEST Python package can be found at github.com/bionetslab/digest. The website is available at github.com/bionetslab/digest-web and the backend at github.com/bionetslab/digest-api.

Online *in silico* validation of disease and gene sets, clusterings or subnetworks with DIGEST

Klaudia Adamowicz, Andreas Maier, Jan Baumbach and David B. Blumenthal

Corresponding author: David B. Blumenthal, Department Artificial Intelligence in Biomedical Engineering, Friedrich-Alexander-Universität Erlangen-Nürnberg, Werner-von-Siemens-Str. 61, 91052 Erlangen, Germany. Tel.: +49 9131 8570676; E-mail: david.b.blumenthal@fau.de

Abstract

As the development of new drugs reaches its physical and financial limits, drug repurposing has become more important than ever. For mechanistically grounded drug repurposing, it is crucial to uncover the disease mechanisms and to detect clusters of mechanistically related diseases. Various methods for computing candidate disease mechanisms and disease clusters exist. However, in the absence of ground truth, *in silico* validation is challenging. This constitutes a major hurdle toward the adoption of *in silico* prediction tools by experimentalists who are often hesitant to carry out wet-lab validations for predicted candidate mechanisms without clearly quantified initial plausibility. To address this problem, we present DIGEST (*in silico* validation of disease and gene sets, clusterings or subnetworks), a Python-based validation tool available as a web interface (<https://digest-validation.net>), as a stand-alone package or over a REST API. DIGEST greatly facilitates *in silico* validation of gene and disease sets, clusterings or subnetworks via fully automated pipelines comprising disease and gene ID mapping, enrichment analysis, comparisons of shared genes and variants and background distribution estimation. Moreover, functionality is provided to automatically update the external databases used by the pipelines. DIGEST hence allows the user to assess the statistical significance of candidate mechanisms with regard to functional and genetic coherence and enables the computation of empirical *P*-values with just a few mouse clicks.

Keywords: Systems medicine, *in silico* validation, Functional and genetic coherence

Introduction

The main objective of systems medicine is to uncover molecular mechanisms driving complex diseases and thereby pave the way for causally effective treatment. A plethora of computational approaches designed to support this overall aim exist, ranging from classical differential gene expression analysis tools [18] over comorbidity pattern analysis frameworks [11] to network-based disease mechanism mining methods [4, 10, 20, 21]. Independently of the concrete algorithmic model, computational systems medicine approaches often return sets, clusterings or subnetworks of genes or diseases as output. For instance, the typical output of differential gene expression analyses are sets of differentially expressed genes (DEGs) for case versus control studies or clusterings of DEGs for different disease subtypes. Comorbidity analyses, on the other hand, often return sets or clusterings of diseases that are predicted to share common molecular mechanisms. Disease mechanism mining approaches typically return small induced subnetworks in protein–protein interaction (PPI) or gene regulatory networks.

In order to have translational potential, pre-clinical studies are necessary to test the hypotheses generated by computational systems medicine approaches in cell lines or animal models. However, such studies are typically resource-intensive, which implies that the hypotheses derived via computational means must have solid prior plausibility in order to convince pre-clinical researchers to invest time and money in follow-up hypothesis testing. Consequently, tools are needed which allow to *in silico* quantify the hypotheses' initial plausibility in a user-friendly way.

A widely used approach for *in silico* validation of hypotheses generated by computational systems medicine approaches is to employ functional annotations available in data sources such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) [14] or the Gene Ontology (GO) [2, 9]. Various tools exist which allow enrichment analyses against these data sources, including DAVID [13], PANTHER [26], g:Profiler [32] and WebGestalt [22]. However, to the best of our knowledge, all existing tools have two limitations: Firstly, only gene sets are accepted as input and neither disease sets nor gene or disease

Klaudia Adamowicz is doctoral researcher at the Institute for Computational Systems Biology at the University of Hamburg.

Andreas Maier is doctoral researcher at the Institute for Computational Systems Biology at the University of Hamburg.

Jan Baumbach is professor and head of the Institute for Computational Systems Biology at the University of Hamburg. He obtained his PhD in Computer Science from Bielefeld University.

David B. Blumenthal is professor and head of the Biomedical Network Science Lab at the Department Artificial Intelligence in Biomedical Engineering of the Friedrich-Alexander-Universität Erlangen-Nürnberg. He obtained his PhD in Computer Science from the Free University of Bozen-Bolzano.

Received: February 28, 2022. Revised: May 25, 2022. Accepted: May 26, 2022

© The Author(s) 2022. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

Table 1. Features of DIGEST and existing online validation tools. Check marks in parentheses indicate that DIGEST provides the respective features via calls to the API of g:Profiler.

	DAVID	PANTHER	g:Profiler	WebGestalt	DIGEST
Tabular results					
Empirical P-values	X	X	X	X	✓
Enriched annotations	✓	✓	✓	✓	(✓)
Most frequent annotations	X	X	X	X	✓
Summary figures					
Empirical P-values	X	X	X	X	✓
Enriched annotations	X	✓	X	✓	X
Most frequent annotations	X	✓	X	✓	✓
Annotation distribution	X	X	X	X	✓
Annotation graph	X	X	X	✓	X
Publication-ready figures	X	X	X	✓	✓
Supported input types					
Gene set	✓	✓	✓	✓	✓
Gene clustering	X	X	X	X	✓
Gene subnetwork	X	X	X	X	✓
Disease set	X	X	X	X	✓
Disease clustering	X	X	X	X	✓
Disease subnetwork	X	X	X	X	✓
Supported data sources					
Gene ontology	✓	✓	✓	✓	✓
Biological pathways	✓	✓	✓	✓	✓
Regulatory motifs in DNA	✓	X	✓	✓	X
Protein databases	✓	✓	✓	✓	X
Phenotype ontology	✓	X	✓	✓	X
Gene-disease associations	✓	X	X	✓	✓
Variant-disease associations	X	X	X	X	✓
Automated mapping					
Automated gene ID mapping	✓	✓	✓	✓	✓
Gene ID conversion	✓	X	✓	✓	X
Automated disease ID mapping	X	X	X	X	✓
Support beyond <i>homo sapiens</i>	✓	✓	✓	✓	X
Orthology-based gene mapping	X	X	✓	X	X
Supported analyses					
Gene set enrichment analysis	✓	✓	✓	✓	(✓)
Reference-free functional coherence	✓	✓	✓	✓	✓
Reference-based functional coherence	X	X	X	✓	✓
Reference-free genetic coherence	X	X	X	✓	✓
Reference-based genetic coherence	X	X	X	X	✓

clusterings or subnetworks are supported. Secondly, functionality to assess the statistical significance of the obtained enrichment scores in comparison to random background models is missing.

In order to fill these gaps, we present DIGEST (short for ‘in silico validation of **d**isease and **g**ene **s**ets or **cl**usterings’) — an *in silico* validation tool for computational systems medicine approaches that allows the validation of different input types in a fully automated and user-friendly way. DIGEST is available as a web service (<https://digest-validation.net>), as a Python package (<https://pypi.org/project/biodigest>) and as a REST API (<https://api.digest-validation.net>). Table 1 provides an overview of its main features in comparison to the existing tools DAVID, PANTHER, g:Profiler and WebGestalt. Note that DIGEST is intended to complement rather than replace these tools. For instance, DIGEST focuses on human diseases and so does not offer support beyond *homo sapiens*. Moreover, DIGEST is not primarily designed for classical gene set enrichment analysis [39], although

this feature is indirectly supported via calls to the API of g:Profiler.

Results

Overview of DIGEST and supported input types

The main result of this paper is the DIGEST validation tool itself. Figure 1 provides an overview of its functionality and the supported validation routes. Similar to existing approaches, DIGEST can perform functional analysis on gene sets using GO and KEGG. In addition to the gene set enrichment analysis, gene sets can also be validated for functional coherence via pairwise comparisons of the contained genes’ functional annotations. Like WebGestalt, DIGEST also offers the possibility to provide a reference gene set. Alternatively, a reference disease ID can be provided by the user. In addition to analyzing gene sets, DIGEST allows validation of gene clusterings based on clustering quality measures such as the Dunn index [8], the silhouette

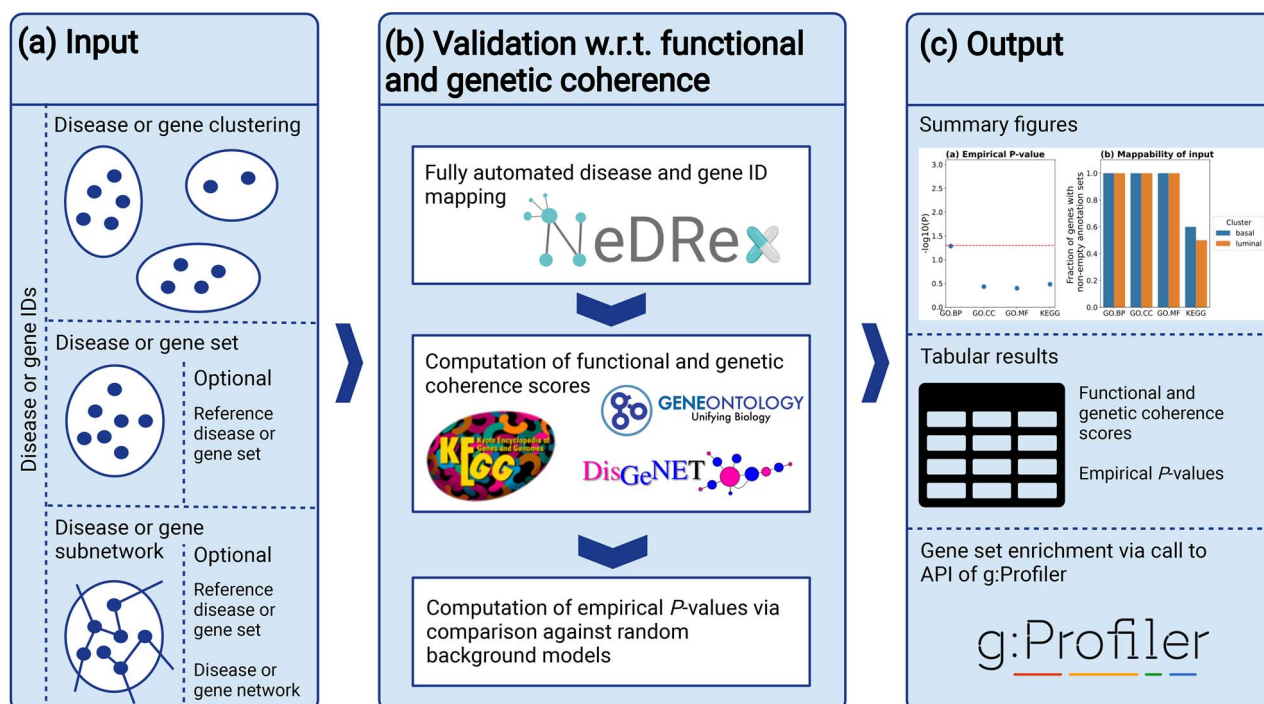


Fig. 1. Overview of validation routes provided by DIGEST. **(A)** As input, DIGEST supports sets, clusterings (i.e. sets of sets) and subnetworks of genes or diseases. For disease and gene sets or subnetworks, a second set or a disease ID can optionally be provided as reference. **(B)** DIGEST queries NeDRex [36] to automatically map the gene or disease IDs provided as input to the required namespaces. Depending on the provided input, functional or genetic coherence scores are computed using GO [2, 9] or KEGG [14] enrichment or genetic associations obtained from DisGeNET [30]. Subsequently, empirical P-values are computed by comparing the obtained scores against two random background models. **(C)** The results of the validations are provided in the form of summary figures, as well as in tabular format. Moreover, gene set enrichment analysis is supported via calls to the API of g:Profiler.

score [34] and the Davies–Bouldin index [6]. As in the gene set analysis, functional annotations from GO and KEGG are used for computing the gene distances underlying these measures. Finally, DIGEST can be used to *in silico* validate induced subnetworks in gene–gene networks.

Unlike all existing approaches we are aware of, DIGEST also supports disease IDs as input. As for genes, sets and clusterings of disease IDs are supported, as well as induced subnetworks in disease–disease networks. The input can be validated w. r. t. functional coherence based on annotations obtained from KEGG. Moreover, DIGEST supports a genetic coherence analysis for diseases based on shared genes or variants extracted from DisGeNET [30].

A further novelty of DIGEST w. r. t. existing approaches is that it allows to evaluate whether the obtained scores of functional or genetic coherence are statistically significant. This is important because scores such as the Dunn index or the silhouette score are often difficult to interpret. In order to be able to judge whether an obtained score is ‘good’ or ‘bad’, a comparison against a random background is required. Therefore, DIGEST provides random background models for all input types and uses them to compute empirical P-values quantifying the significance of the obtained scores.

In the sequel, we showcase how to use DIGEST for validation of gene and disease sets, clusterings and

subnetworks. Technical and algorithmic details can be found in the ‘Methods’ section.

In silico validation of gene sets

Since prior studies have linked lipid and cholesterol metabolism to Alzheimer’s disease (AD) [7], Sadegh *et al.* [36] hypothesized that hyperlipidemia-associated genes obtained from OMIM [1] and DisGeNET [30] constitute a promising point of departure for network-based AD drug repurposing. We ran DIGEST’s reference-based gene set validation route to assess this hypothesis, using the AD-associated genes as query and the hyperlipidemia-associated genes as reference set (the genes contained in the sets are listed in Supplementary Table 1). To compute empirical P-values, we used a random background model that preserves the sizes of the contained genes’ annotation sets.

The results are shown in Figure 2. While significant P-values were obtained for all three types of GO terms (BP: biological process; CC: cellular component; MF: molecular function), results were not significant for KEGG pathways. This can be explained by the fact that the Jaccard index between KEGG pathways containing, respectively, AD- and hyperlipidemia-associated genes is 0.0 (see the red vertical line in Figure 2b). Nonetheless, the results lend evidence to the conjecture that AD is mechanistically linked to hyperlipidemia and that

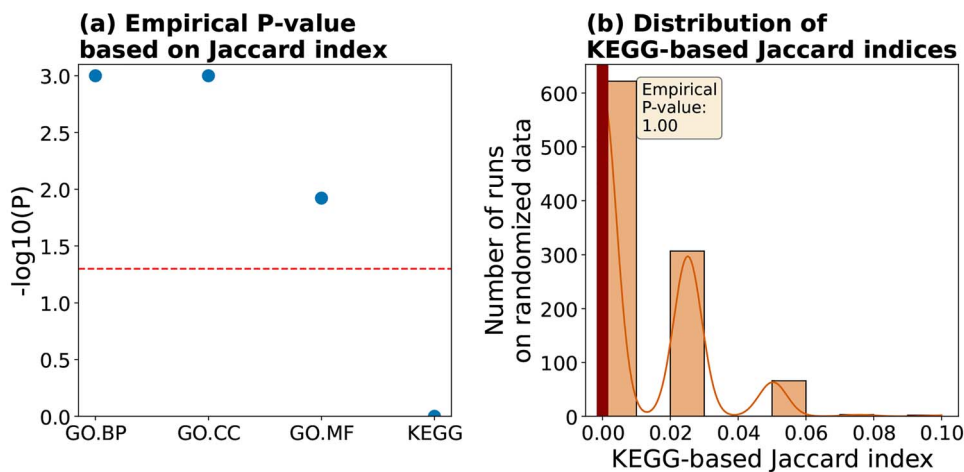


Fig. 2. Results of reference-based gene set validation route for hyperlipidemia-based AD drug repurposing approach suggested by Sadegh et al. [36]. **(A)** Empirical P-values obtained by comparing the mean Jaccard index of pairwise annotation sets against random background. **(B)** KEGG-based Jaccard indices for input gene set (red vertical line) and gene sets generated by background model.

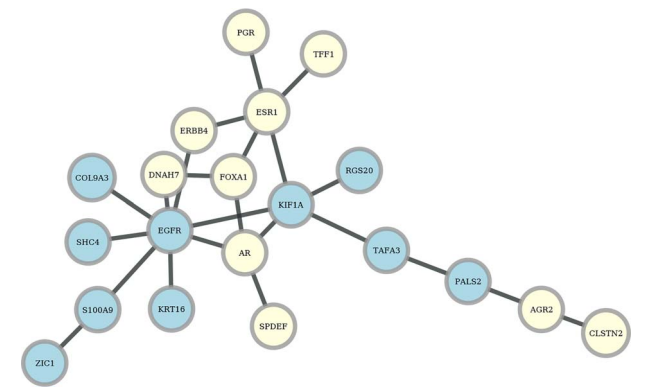


Fig. 3. Candidate pathomechanisms for luminal (blue) and basal (yellow) breast cancer reported in [20]. Genes are connected by an edge if the BioGRID [29] database contains a PPI for their encoded proteins.

hyperlipidemia-associated genes are a promising starting point for *in silico* AD drug repurposing.

In silico validation of gene clusterings

BiCoN [20] is a network medicine tool which allows binary patient stratification based on gene expression data. At the same time, it extracts two subgraphs from PPI networks that explain the patient stratification and hence constitute candidate pathomechanisms. In the original publication, the authors used BiCoN on gene expression data for breast cancer and obtained a patient clustering that almost perfectly matches the luminal versus basal breast cancer sub-types. Figure 3 shows the two candidate pathomechanisms. Here, we interpreted these predicted pathomechanisms as a gene clustering with two blocks, and used DIGEST’s gene clustering validation route to assess its functional coherence.

Figure 4 shows the results of the analysis. We did not obtain significant P-values for any of the four gene annotation options supported by DIGEST. The results hence provide little evidence for the hypothesis that the two sub-networks extracted by BiCoN indeed correspond to distinct pathomechanisms.

Table 2. List of diseases linked to cGMP signaling by Langhauser et al. [17] and their respective MeSH IDs.

Disease	MeSH ID
Stroke	D020521
Alzheimer disease	D000544
Dementia	D003704
Atherosclerosis	D050197
Asthma	D001249
Diabetes mellitus type 2	D003924
Parkinson disease	D010300
Heart failure	D006333
Migraine disorders	D008881
Myocardial infarction	D009203
Hypertension	D006973
Obesity	D009765

In silico validation of disease sets

Using a network-based systems medicine approach, Langhauser et al. [17] identified a set of 12 diseases hypothesized to share a common mechanism linked to cyclic guanosine monophosphate (cGMP) signaling. The set of these presumably mechanistically related diseases, specified in the MeSH namespace as in the original publication, is shown in Table 2.

We used DIGEST’s reference-free disease set validation route with random background models that preserve annotation set size to assess the genetic and functional coherence of this set of diseases. The results are shown in Figure 5. For all three annotation types, the disease set exhibits a significant internal coherence. Using DIGEST, we could hence provide further evidence for Langhauser et al.’s hypothesis that the diseases listed in Table 2 are mechanistically related.

In silico validation of disease clusterings

To exemplify the use of DIGESTS’s disease clustering validation route, we analyzed the clustering of the ICD-10 three-character codes induced by chapter IX ‘Diseases of the circulatory system’ (see Figure 6). This analysis helps

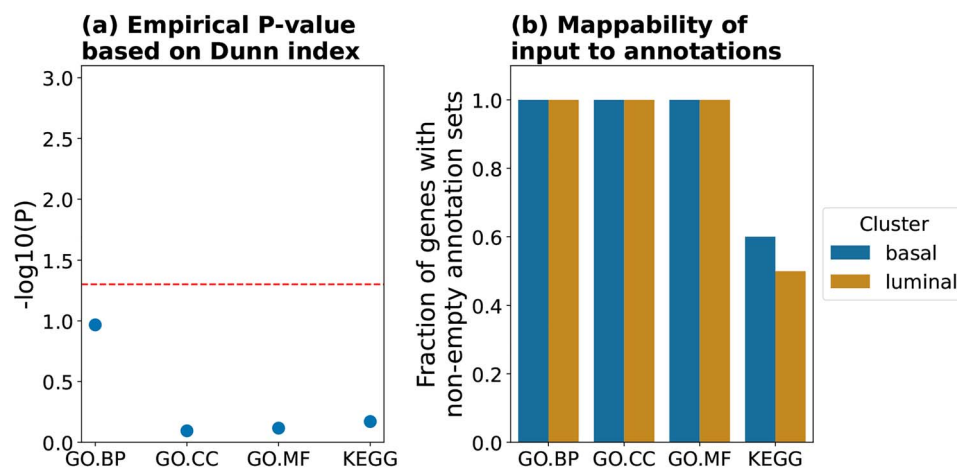


Fig. 4. Results of gene clustering validation route for luminal and basal breast cancer pathomechanisms predicted by Lazareva et al. [20]. **(A)** Empirical P-values obtained by comparing the clustering's Dunn index against random background. **(B)** Frequency of genes with non-empty sets of assigned functional annotations.

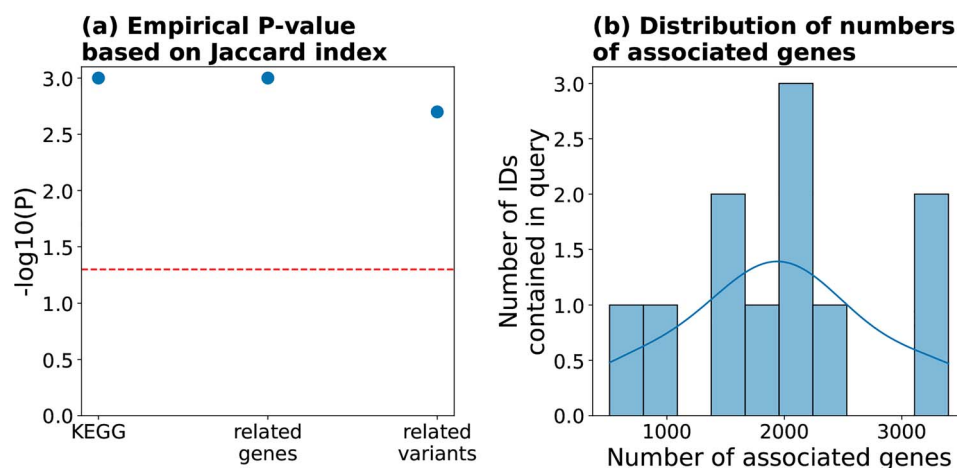


Fig. 5. Results of reference-free disease set validation route for diseases linked to cGMP signaling by Langhauser et al. [17]. **(A)** Empirical P-values obtained by comparing the mean Jaccard index of pairwise annotation sets against random background. **(B)** Distribution of number of related genes associated with each ID.

IX Diseases of the circulatory system

- I00–I02 Acute rheumatic fever
- I05–I09 Chronic rheumatic heart diseases
- I10–I15 Hypertensive diseases
- I20–I25 Ischaemic heart diseases
- I26–I28 Pulmonary heart disease and diseases of pulmonary circulation
- I30–I52 Other forms of heart disease
- I60–I69 Cerebrovascular diseases
- I70–I79 Diseases of arteries, arterioles and capillaries
- I80–I89 Diseases of veins, lymphatic vessels and lymph nodes, not elsewhere classified
- I95–I99 Other and unspecified disorders of the circulatory system

Fig. 6. Blocks of chapter IX of the ICD-10 disease ontology obtained from <https://icd.who.int/browse10/2019/en/>.

to assess to which extent chapter IX of the ICD-10 disease ontology is mechanistically grounded, i.e. whether the contained sub-blocks exhibit a stronger functional and genetic intra- than inter-block coherence.

Figure 7 shows the results of our analysis. Interestingly, we did not observe statistically significant results for any of the three annotation types. Moreover, all of the top 10 most frequently occurring associated genes are present in most of the clusters, indicating that clusters' genetic coherence is low. These findings confirm the increasing awareness in the field that currently used disease ontologies are problematic because they do not mirror the underlying pathomechanisms [28].

In silico validation of gene subnetworks

KeyPathwayMineR [25] is a *de novo* network enrichment tool to predict candidate pathomechanisms starting with a list of DEGs. For the original publication, the authors ran their tool on gene expression data from COVID-19 patients and healthy controls and on a human PPI network obtained from BioGRID to derive a candidate host mechanism involved in COVID-19.

We used DIGEST's reference-free gene subnetwork validation route to assess the internal functional coherence of this candidate mechanism. As shown in Figure 8, we

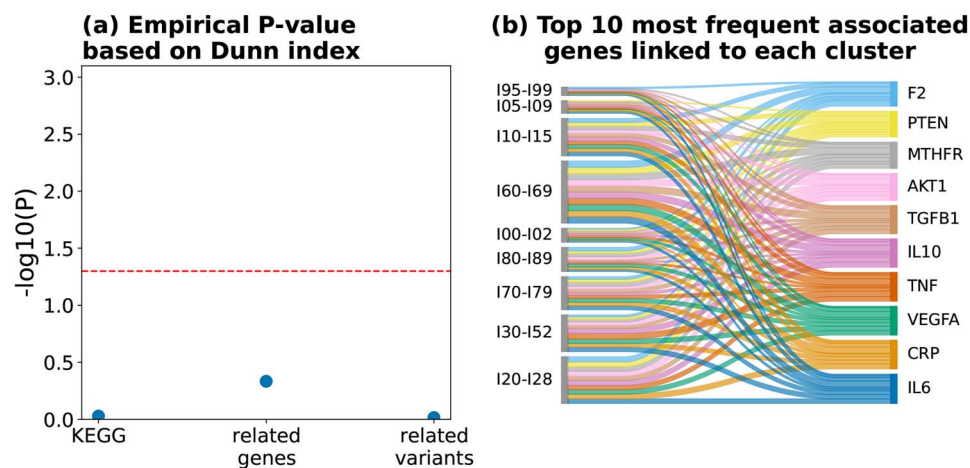


Fig. 7. Results of disease clustering validation route for clustering induced by the blocks of ICD-10 chapter IX. **(A)** Empirical P-values obtained by comparing the clustering's Dunn index against random background. **(B)** Sankey plot linking each cluster to the top 10 most frequently occurring associated genes.

did not obtain any significant P-values. In fact, even for the GO.BP annotations which led to the smallest P-value, the mean Jaccard index obtained for the predicted host mechanism is < 0.05 (Figure 8b). This renders unlikely that the subnetwork identified by Mechteridis *et al.* [25] indeed represents *one* host pathway that is hijacked in SARS-CoV-2 infection (it might still represent several, mechanistically disjoint pathways).

To elucidate possible reasons that might explain the non-significant results, we computed gene-level significance contributions (see 'Methods' for details and Figure 8c for the results obtained for the GO.MF-based P-values). That is, we estimated to which extent the individual genes contained in the COVID-19 subnetwork discovered by KeyPathwayMineR positively or negatively affect functional coherence. We then carried out gene set enrichment analysis with g:Profiler on, respectively, the 10 genes with largest positive and largest negative effects on coherence w. r. t. GO.MF annotations (supported in DIGEST via direct links to g:Profiler from the results view).

Interestingly, the 10 genes with largest positive effects are enriched with annotations related to virus response (see Supplementary Figure 1), while the 10 genes with largest negative effects are enriched with more generic annotations (see Supplementary Figure 2). A possible explanation for the non-significant results is hence that KeyPathwayMineR returns a large number of genes with rather generic functions—possibly, because such genes tend to constitute hubs in PPI networks. The fact that the genes with large positive significance contributions assume rather peripheral positions in the subnetwork computed by KeyPathwayMineR provides further evidence for this conjecture (see red nodes in Figure 8c).

Scalability tests

To evaluate the scalability of DIGEST, we compared the API query execution times for the reference-free gene

set validation route against functional gene set enrichment queries against the APIs of PANTHER and g:Profiler (results shown in Table 3; for details on the test setup, see 'Methods'). Since DIGEST is the only tool that runs permutation tests, it is not surprising that it is clearly the slowest of the three tools, with a mean query time of around 13 min on gene sets of size 100 if run with $n = 1000$ randomizations. To further improve usability of the web interface, we provide a stable URL to the results immediately after starting the *in silico* validation. By saving this URL, the user does not have to wait for the analysis to finish but can return to the results page at any later point. Note that if run with $n = 1$ randomizations, DIGEST's query is around twice the query time of PANTHER. Setting the number of randomizations to $n = 1$ of course does not make sense for validation purposes, but yields a fairer comparison against PANTHER and g:Profiler, which do not run permutation tests.

Discussion

With DIGEST, we introduce a user-friendly tool for the *in silico* validation of sets, clusterings or subnetworks of genes and diseases. DIGEST supports the fully automated computation of empirical P-values for hypotheses generated by computational systems medicine approaches and hence helps to increase the likelihood that promising hypotheses are further carried toward translation in pre-clinical studies.

It is important to emphasize that the P-values computed by DIGEST only quantify initial plausibility of hypotheses generated by computational systems medicine approaches. Whether these hypotheses are ultimately valid and indeed correspond to disease mechanisms can only be established via follow-up pre-clinical and clinical studies. DIGEST is hence not intended as a tool to substitute downstream wet-lab validation, but as a tool that allows to further narrow

Table 3. Mean execution times of API queries against PANTHER, g:Profiler and DIGEST for gene set queries with varying numbers of genes (in seconds). DIGEST was run with $n = 1000$ and $n = 1$ randomizations and without estimation of significance contributions.

# Genes	PANTHER	g:Profiler	DIGEST ($n = 1000$)	DIGEST ($n = 1$)
10	79.51	33.89	701.06	149.77
50	80.91	33.67	762.03	148.41
100	88.68	33.98	806.83	148.13

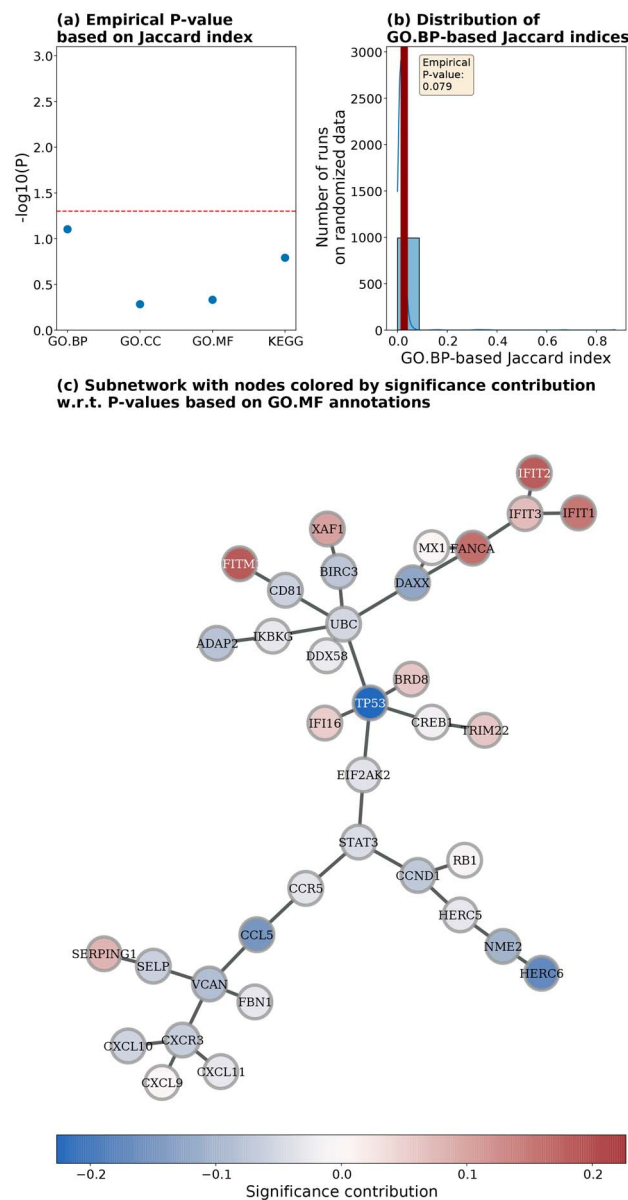


Fig. 8. Results of reference-free gene subnetwork validation route for candidate COVID-19 host mechanism predicted by Mechteridis *et al.* [25]. **(A)** Empirical P-values obtained by comparing the mean Jaccard index of pairwise annotation sets against random background. **(B)** Mean GO.BP-based Jaccard indices for input subnetwork (red vertical line) and gene subnetworks generated by random background model. **(C)** Significance contributions toward empirical P-values based on GO.MF annotations of individual genes contained in the predicted COVID-19 host mechanism.

down the search space and focus on the most promising hypotheses. Moreover, the P-values returned by DIGEST have to be interpreted carefully, since they are based on the

biological knowledge encoded in the KEGG, GO and DisGeNET databases. The scrutinized hypotheses are hence always validated against current knowledge, although they typically aim at generating new insights into pathomechanisms. Moreover, biological databases are known to be heavily affected by study bias [19, 37], i.e. few well-studied genes and diseases are overrepresented w. r. t. the number of associations. In DIGEST, we address this shortcoming using random background models that preserve annotation set sizes. However, this only partially solves the problem, because we cannot distinguish between genes that are ‘true biological hubs’ and genes that appear to be hubs due to study bias.

Finally, we would like to point out that it is possible that DIGEST’s empirical P-values are inconsistent (e.g. see results for the drug repurposing use case shown in Figure 2). In such a situation, it is important that the user carefully assesses which of the disease or gene annotation databases used by DIGEST are indeed relevant for their use case. For instance, in the above-mentioned drug repurposing use case, one might decide to disregard the non-significant KEGG-based P-value, since it has been argued that KEGG pathways poorly reflect pathomechanisms [28].

Methods

Supported disease and gene namespaces

DIGEST validates sets, clusterings and subnetworks of genes and diseases by comparing suitably defined annotations sets obtained via queries against KEGG, GO and DisGeNET (see subsequent sections for details). Since KEGG, GO and DisGeNET require genes and diseases to be given using specific namespaces (see Supplementary Table 2), DIGEST queries NeDRex [36] to automatically map the input provided by the user to the required namespaces. The following input formats are supported:

- Supported gene namespaces: Entrez [23], Ensembl [12], gene symbols [40], UniProt IDs of encoded proteins [3].
- Supported disease namespaces: Mondo [27], OMIM [1], SNOMED CT [35], UMLS [5], ORPHA [31], MeSH [33], DOID [38], ICD-10 (<https://icd.who.int/browse10/2010/en>).

For all input types and validation routes, identifiers that cannot be mapped to the required namespace are ignored. Moreover, DIGEST ignores diseases and genes for

which the annotation sets obtained from KEGG, GO or DisGeNET are empty.

Statistical validation

Let X be any input of the supported input types specified above. Moreover, let $S(X)$ be a functional relevance score for X and $\mathcal{M}$ be a background model which allows to generate randomized input $\mathcal{M}(X)$, preserving some of the properties of X . Assume that, w. r. t. S , larger means better. We compute empirical P -values as

$$P := \frac{1 + \sum_{i=1}^n [S(X) \leq S(\mathcal{M}(X)_i)]}{n + 1}, \quad (1)$$

where n is a user-specified parameter determining the P -value resolution (defaulted to $n := 1000$ in DIGEST) and $[\cdot]$ is the Iverson bracket (i.e. $[\text{true}] := 1$ and $[\text{false}] := 0$). In the following text, we detail how the scores S and the background models $\mathcal{M}$ are defined for the supported input types. Note that separate P -values are computed for each annotation type. For instance, for reference-free gene set validation (see details below), DIGEST computes empirical P -values based on KEGG pathways, GO cellular components, GO molecular functions and GO biological processes.

Scores and random background for gene sets

DIGEST supports the following three routes for *in silico* validating a set of genes X :

- (1) Validation against a reference disease set X_R : Assesses whether the functional relevance of X w. r. t. diseases contained in X_R is statistically significant.
- (2) Validation against a reference gene set X_R : Assesses whether the functional similarity between genes contained in X and X_R is statistically significant.
- (3) Reference-free validation: Assesses whether the internal functional coherence of the genes contained in X is statistically significant.

For the first two routes, sets $\mathcal{A}_X$ and $\mathcal{A}_R$ of functional annotations are obtained for, respectively, the gene set X and the reference X_R or d_R (see the next paragraph for details). Subsequently, the functional relevance score $S(X)$ is computed either as the Jaccard index $S_{JI}(X) := JI(\mathcal{A}_X, \mathcal{A}_R) := |\mathcal{A}_X \cap \mathcal{A}_R| / |\mathcal{A}_X \cup \mathcal{A}_R|$ or as the overlap coefficient $S_{OC}(X) := OC(\mathcal{A}_X, \mathcal{A}_R) := |\mathcal{A}_X \cap \mathcal{A}_R| / \min\{|\mathcal{A}_X|, |\mathcal{A}_R|\}$, depending on the choice of the user.

For the first validation route, $\mathcal{A}_R$ is defined as union of the set of KEGG pathways associated with diseases $d \in X_R$ and $\mathcal{A}_X$ is defined as the set of KEGG pathways associated with genes $g \in X$ (note that singleton disease sets $X_R = \{d\}$ are allowed). For the second validation route, KEGG and GO are used to obtain the set of functional annotations. In either case, $\mathcal{A}_X$ is defined as the set of annotations associated with genes $g \in X$ or, optionally, as the set of annotations enriched with X as a whole. $\mathcal{A}_R$ is defined either as the set of annotations associated with genes

$g \in X_R$ or as the set of annotations enriched with X_R as a whole.

For the third, reference-free validation route, sets $\mathcal{A}_g$ of functional annotations are obtained from GO and KEGG for each gene $g \in X$ independently. Depending on the choice of the user, $S(X)$ is then computed either as the mean Jaccard index or as the mean overlap coefficient over all pairs $(\mathcal{A}_g, \mathcal{A}_{g'})$ of annotation sets for $g, g' \in X$ with $g \neq g'$.

Two random background models are supported: The first model draws genes uniformly without replacement to compute fully randomized gene sets $\mathcal{M}(X)$ of size $|X|$. The second model maintains some information from X and constructs randomized gene sets $\mathcal{M}(X)$ where the distribution of the contained genes' annotation set sizes (approximately) matches the distribution of the annotation set sizes of the genes contained in X .

Scores and random background for gene clusterings

DIGEST supports the following three scoring measures for *in silico* validating gene clusterings $X = C_1, \dots, C_m$: The Dunn index $S_{DI}(X) := \min_{i \neq j} \delta(C_i, C_j) / \max_i \Delta(C_i)$, the Davies–Bouldin index $S_{DBI}(X) := \sum_{i=1}^m [\max_{j \neq i} \Delta(C_i) + \Delta(C_j)] / \delta(C_i, C_j) / m$ and the silhouette score $S_{SS}(X) := \sum_{g \in \bigcup C_i} [a(g) - b(g)] / \max\{a(g), b(g)\} / n$. $\Delta(C_i)$ denotes the mean intra-cluster distance (i.e. the mean distance between genes $g, g' \in C_i$ with $g \neq g'$), $\delta(C_i, C_j)$ denotes the mean inter-cluster distance (i.e. the mean distance between genes $g \in C_i$ and $g' \in C_j$), $a(g)$ denotes the mean distance between a fixed gene $g \in C_i$ and all other genes in C_i , and $b(g)$ is defined as $b(g) := \min_{j \neq i} b(g, C_j)$, where $b(g, C_j)$ denotes the mean distance between a fixed gene $g \in C_i$ and the genes contained in C_j .

The gene distances underlying the Dunn index, the Davies–Bouldin index, and the silhouette score are defined as $1 - JI(\mathcal{A}_g, \mathcal{A}_{g'})$ or as $1 - OC(\mathcal{A}_g, \mathcal{A}_{g'})$, depending on whether the user wants to use the Jaccard index or the overlap coefficient as underlying set similarity measure. $\mathcal{A}_g$ and $\mathcal{A}_{g'}$ are functional annotations for the genes g and g' obtained from GO and KEGG. The background model randomly shuffles the assigned clusters between the genes $g \in \bigcup_{i=1}^m C_i$, preserving the number of clusters m and the cluster size distribution.

Scores and random background for disease sets

DIGEST supports the following two routes for *in silico* validating a set of diseases X :

- (1) Validation against a reference gene set X_R : Assesses whether the functional coherence between the diseases contained in X and the genes contained X_R is statistically significant.
- (2) Validation against a reference disease set X_R : Assesses whether the genetic similarity between diseases contained in X and X_R is statistically significant.

- (3) Reference-free validation: Assesses whether the internal genetic coherence of the diseases contained in X is statistically significant.

Functional relevance scores for the validation routes are defined in almost exactly the same way as for the validation routes for gene sets explained above. The only difference is that, in addition to KEGG pathways, disease–gene and disease–variant associations also obtained from DisGeNET are used to construct the annotation sets underlying the relevance scores, while no GO annotations are employed.

As for gene sets, two random background models are supported: (1) A fully randomized model which uniformly samples genes sets $\mathcal{M}(X)$ of size $|X|$. (2) A partially randomized model which samples genes sets $\mathcal{M}(X)$ such that the distribution of the sizes of the contained diseases' annotation sets (approximately) matches the distribution of annotation set sizes for X .

Scores and random background for disease clusterings

DIGEST can also be used to *in silico* validate disease clusterings $X = C_1, \dots, C_m$. Like for gene clusterings, the Dunn index, the Davies–Bouldin index or the silhouette score can be used as clustering quality score $S(X)$. The underlying disease distances are defined in terms of the Jaccard index or the overlap coefficient of disease annotation sets $\mathcal{A}_d$ obtained from DisGeNET (genes or variants) or KEGG. The background model is analogous to the one for gene clusterings, i.e. diseases $d \in \bigcup_{i=1}^m C_i$ are randomly shuffled across the m clusters under the constraint that the cluster sizes are preserved.

Scores and random background for induced gene or disease subnetworks

Finally, DIGEST can be used to *in silico* validate subnetworks induced by a gene or disease set provided by the user. More precisely, let $G = (V, E)$ be a gene–gene or disease–disease network and $X \subseteq V$ be the user-provided gene or disease set. Also the network G can be provided by the user. If no network is provided, DIGEST obtains G via queries to NeDRex as follows: If X is a set of genes, DIGEST constructs G as a gene–gene network based on experimentally validated PPIs obtained from IID [15] (i.e. two genes are connected by an edge if their encoded proteins have been shown to physically interact). If X is a set of diseases, G is constructed as a disease–disease network based on shared disease–gene associations obtained from OMIM and DisGeNET (i.e. two diseases are connected by an edge if they have at least one associated gene in common).

To validate the subgraph $G[X]$ induced by X , we compute functional relevance scores $S(X)$ in exactly the same way as for gene and disease sets (see details above). As background model, we compute gene or disease sets $\mathcal{M}(X)$ such that the induced subgraph $G[\mathcal{M}(X)]$ matches $G[X]$ in terms of numbers and sizes of connected compo-

nents. For this, we first decompose $G[X]$ into its connected components $(X_i)_{i=1}^k$. Subsequently, we construct $\mathcal{M}(X)$ via k random network expansions along the edges of G from k randomly selected seed nodes $v_i \in V$. We stop the random expansion from node v_i as soon as the constructed connected component has reached the size $|X_i|$.

Significance contributions of individual genes or diseases

DIGEST also allows to compute the contributions of individual genes or diseases toward (non-)significance (implemented as an optional feature since it increases DIGEST's runtime linearly w. r. t. the size of the input). Let X be any input supported by DIGEST, x be a disease or gene contained in X and $X - x$ be a slightly modified input where x has been removed. We quantify x 's significance contribution as

$$\Delta_P(x) := P(X - x) - P(X), \quad (2)$$

where $P(X)$ and $P(X - x)$ are the empirical P -values obtained for the original and modified inputs, respectively. If $\Delta_P(x) > 0$, x has a positive effect on X 's genetic or functional coherence. If $\Delta_P(x) < 0$, the opposite is the case.

Note that the gene- or disease-level significance contributions are especially interesting for mixed-message results. If the P -values are extreme ($P(X) \approx 0$ or $P(X) \approx 1$), it may well happen that they are insensitive to removal of individual genes or diseases from the input, which then leads to $\Delta_P(x) = 0$ for many or even all genes or diseases x contained in the input (see Supplementary Figure 3 for such a situation).

Summary figures

DIGEST offers fully automated creation of plots to visualize the results. Let T be all annotation types used for computation of the relevance scores $S(X)$. Moreover, for each annotation type $t \in T$ and each element $x \in X$, let $\mathcal{A}_x^t$ be the set of annotations from t for x and $\Delta_P^t(x)$ be the significance contribution of a gene or disease x on the w. r. t. the P -value based on annotation type t . We provide plots to visualize the calculated empirical P -values and plots to show the mappability of the input. The following plots are generated to visualize the empirical P -value visualization (see Supplementary Figure 3 for an overview):

- A plot showing all calculated empirical P -values for all annotation types $t \in T$ (see Figure 2a, Figure 4a, Figure 5a, Figure 7a and Figure 8a).
- A heatmap showing the significance contributions $\Delta_P^t(x)$ for all annotation types $t \in T$ and the top 15 genes or diseases x contained in the input X with the largest absolute significance contributions across all annotation types ($\max_x \max_t |\Delta_P^t(x)|$, see Supplementary Figure 3b).

- For every annotation type $t \in T$: Heatmaps showing the significance contributions $\Delta_p^t(x)$ for all annotation types $t' \in T$ for (i) the top 10 genes or diseases with the largest positive significance contributions w. r. t. t ($\max_x \Delta_p^t(x)$, see Supplementary Figure 3c), and (ii) the top 10 genes or diseases with the largest negative significance contributions w. r. t. t ($\min_x \Delta_p^t(x)$, see Supplementary Figure 3d).
- For every annotation type $t \in T$ if X is the node set of an induced subnetwork: A visualization of the subnetwork $G[X]$ where the individual genes' or diseases' significance contributions $\Delta_p^t(x)$ are encoded as node colors (see Figure 8c).
- For every annotation type $t \in T$: A plot showing the relevance score $S(X)$, the distribution of all relevance scores $S(\mathcal{M}(X))$ for the background model, as well as the resulting empirical P-value (see Figure 2b and Figure 8b).

The following plots are generated to visualize mappability (see Supplementary Figure 4 for an overview):

- A plot showing the frequency of elements $x \in X$ with non-empty annotation sets ($\mathcal{A}_x^t \neq \emptyset$) for all annotation types $t \in T$ (see Figure 4b).
- For every annotation type $t \in T$: A plot showing the distribution of annotation set sizes $|\mathcal{A}_x^t|$ for the elements $x \in X$ (see Figure 5b).
- For every annotation type $t \in T$: A Sankey plot linking the elements $x \in X$ to the top 10 most frequently occurring annotation terms of type t . For clusterings, term occurrences are normalized by the sizes of the clusters they are appearing in to avoid overrepresentation of larger clusters (see Figure 7b).

Setup of scalability tests

To evaluate the scalability of DIGEST in comparison to existing tools, we ran queries with identical gene sets against the APIs of DIGEST, PANTHER and g:Profiler (DAVID and WebGestalt were excluded due to lack of APIs that allow query execution time measurements). For this, we generated 3×10 random gene sets of sizes 10, 50 and 100, respectively. We then ran PANTHER's 'Functional enrichment test', g:Profiler's 'g:GOST' analysis (performs functional enrichment analysis) and DIGEST's reference-free gene set validation route and measured the query execution times. The number of randomizations used for DIGEST's permutation tests was set to $n = 1000$ (default in web interface) and $n = 1$ (yields fairer comparison against tools that do not run permutation tests).

Implementation of Python package

The backend is implemented in Python 3 and pulls the disease ID mappings from the API of NeDRex, the disease attribute mappings (disease-associated genes and variants) from DisGeNET and the shared KEGG pathways from the API of KEGG. Gene set enrichment is computed via GSEapy's Enrichr module [16, 39]. The remaining

mappings are obtained via the biothings_client package (https://github.com/biothings/biothings_client.py). A YAML file is provided to easily set up an environment with all dependencies.

Implementation of web interface and REST API

The web interface and REST API components are deployed using Docker and separated into four containers: (1) A backend service running a Django REST framework for request handling and execution of DIGEST's Python package. (2) A Redis service for task queuing and execution of incoming validation requests. (3) A PostgreSQL database to store and manage validation results. (4) A frontend implemented in vue.js, providing a graphical configuration option for execution of DIGEST, as well as the API documentation.

The REST API provides programmatic access to DIGEST's validation routes. Requests receive a unique task ID, used to check the execution status and request results. Moreover, set-up files for DIGEST's Python package can be obtained via the API, which improve the initialization time by loading the latest pre-built scoring matrices and mapping files from the server. Requests submitted through the web interface use the unique task ID to generate a permanently accessible URL for the result page, allowing users to save it and return to the page at a later point. Besides the obtained empirical P-values, the result page presents the input data, the selected parameter configuration, as well as summary figures. If the user desires to compute gene- or disease-level significance contributions, the necessary computations are scheduled during idle times of the server in order not to block faster jobs. The current status of the computations can be checked at any time using the unique task URL. Moreover, if the user provides an email address, DIGEST notifies them once all computations have been carried out.

Key Points

- Solid prior evidence is necessary to carry hypotheses generated by computational systems medicine approaches toward translation.
- Existing validation tools only support limited input types and have limited statistical analysis capacities.
- DIGEST overcomes this limitation by computing empirical P-values quantifying functional and genetic coherence.

Availability

DIGEST's web interface and REST API are available at <https://digest-validation.net> and <https://api.digest-validation.net>, respectively. The Python package is available at <https://pypi.org/project/biodigest>. The source codes of the Python package, the web interface and

the REST API are freely available at <https://github.com/bionetslab/digest>, <https://github.com/bionetslab/digest-web> and <https://github.com/bionetslab/digest-api>, under the terms of the GNU General Public License, Version 3. A Jupyter notebook for getting familiar with DIGEST's Python package and reproducing the results reported in this paper is available at <https://github.com/bionetslab/digest-tutorial>. To further increase reproducibility, we generated an AIME report [24], which is available at <https://aime.report/ms7V2J>.

Supplementary data

Supplementary data are available online at *Briefings in Bioinformatics*.

Author contributions statement

K. A. and D. B. B. conceived and designed the platform. K. A. developed the Python backend. A. M. implemented the web interface. J. B. and D. B. B. supervised the project. All authors provided critical feedback and discussion and assisted in interpreting the results, writing the manuscript and improving the web service.

Acknowledgements

Figure 1 was created with BioRender.com.

Funding

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements No. 777111 (A.M., J.B.). This publication reflects only the authors' view and the European Commission is not responsible for any use that may be made of the information it contains. This work was supported by the German Federal Ministry of Education and Research (BMBF) within the framework of the e:Med research and funding concept (grant 01ZX1908A and grant 01ZX1910D) (J.B.). J.B. was partially funded by his VILLUM Young Investigator Grant No. 13154.

References

- Amberger JS, Bocchini CA, Scott AF, et al. OMIM.org: leveraging knowledge across phenotype-gene relationships. *Nucleic Acids Res* 2019;**47**(D1):D1038–43.
- Ashburner M, Ball CA, Blake JA, et al. Gene Ontology: tool for the unification of biology. *Nat Genet* 2000;**25**(1):25–9.
- Bateman A, Martin MJ, Orchard S, et al. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res* 2021;**49**:D480–9.
- Bernett J, Krupke D, Sadegh S, et al. Robust disease module mining via enumeration of diverse prize-collecting Steiner trees. *Bioinformatics* 2022.
- Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004;**32**(D267):1.
- Davies DL, Bouldin DW. A cluster separation measure. *IEEE Trans Pattern Anal Mach Intell*, PAMI-1:224–227 1979.
- Di Paolo G, Kim T-W. Linking lipids to alzheimer's disease: cholesterol and beyond. *Nat Rev Neurosci* 2011;**12**(5):284–96.
- Dunn JC. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *J Cybern* 2008;**3**(32–57):1.
- Gene Ontology Consortium. The gene ontology resource: enriching a GOLD mine. *Nucleic Acids Res* 2021;**49**(D1):D325–34.
- Ghiassian SD, Menche J, Barabási AL. A DIseAse MOdule Detection (DIAMOND) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. *PLoS Comput Biol* 2015;**11**(1004120):4.
- Guo M, Yanan Y, Wen T, et al. Analysis of disease comorbidity patterns in a large-scale china population. *BMC Med Genomics* 2019;**12**(Suppl 12):177.
- Howe KL, Achuthan P, Allen J, et al. Ensembl 2021. *Nucleic Acids Res* 2020;**49**(D1):D884–91.
- Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 2009;**4**:44–57.
- Kanehisa M, Sato Y, Kawashima M, et al. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res* 2016;**44**(D1):D457–62.
- Kotlyar M, Pastrello C, Malik Z, et al. IID 2018 update: context-specific physical protein-protein interactions in human, model organisms and domesticated species. *Nucleic Acids Res* 2019;**47**(D1):D581–9.
- Kuleshov MV, Jones MR, Rouillard AD, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res* 2016;**44**(W1):W90–7.
- Langhauser F, Casas AI, Dao V-T-V, et al. A diseasome cluster-based drug repurposing of soluble guanylate cyclase activators from smooth muscle relaxation to direct neuroprotection. *NPJ Syst Biol Appl* 2018;**4**:8.
- Law CW, Chen Y, Shi W, et al. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* 2014;**15**(2):R29.
- Lazareva O, Baumbach J, List M, et al. On the limits of active module identification. *Brief Bioinform* 2021;**22**(5):bbab066.
- Lazareva O, Canzar S, Yuan K, et al. BiCoN: network-constrained biclustering of patients and omics data. *Bioinformatics* 2021;**37**:2398–404.
- Levi H, Elkon R, Shamir R. DOMINO: a network-based active module identification algorithm with reduced rate of false calls. *Mol Syst Biol* 2021;**17**(e9593).
- Liao Y, Wang J, Jaehnig EJ, et al. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res* 2019;**47**:W199–205.
- Maglott D, Ostell J, Pruitt KD, et al. Entrez gene: gene-centered information at NCBI. *Nucleic Acids Res* 2011;**39**(Database issue):D52–7.
- Matschinske J, Alcaraz N, Benis A, et al. The AIME registry for artificial intelligence in biomedical research. *Nat Methods* 2021;**18**:1128–31.
- Mechteridis K, Lauber M, Baumbach J, et al. KeyPathwayMineR: De novo pathway enrichment in the R ecosystem. *Front Genet* 2021;**12**:812853.
- Mi H, Ebert D, Muruganujan A, et al. PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. *Nucleic Acids Res* 2021;**49**:D394–403.

27. Mungall CJ, McMurtry JA, Köhler S, et al. The monarch initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Res* 2017;**45**(D1):D712–22.
28. Nogales C, Mamdouh ZM, List M, et al. Network pharmacology: curing causal mechanisms instead of treating symptoms. *Trends Pharmacol Sci* 2022;**43**(2):136–50.
29. Oughtred R, Stark C, Breitkreutz B-J, et al. The BioGRID interaction database: 2019 update. *Nucleic Acids Res* 2019;**47**(D1):D529–41.
30. Piñero J, Ramírez-Anguila JM, Saüch-Pitarch J, et al. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res* 2020;**48**(D1):D845–55.
31. Rath A, Olry A, Dhombres F, et al. Representation of rare diseases in health information systems: the orphanet approach to serve a wide range of end users. *Hum Mutat* 2012;**33**(5):803–8.
32. UKU Raudvere, Liis Kolberg, IVAN Kuzmin, TAMBET Arak, PRIIT Adler, HEDI Peterson, and JAAK Vilo. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.*, **47**:W191–8, 2019.
33. Rogers FB. Medical subject headings. *Bull Med Libr Assoc* 1963;**51**:114–6.
34. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math* 1987;**20**:53–65.
35. Ruch P, Gobeill J, Lovis C, et al. Automatic medical encoding with SNOMED categories. *BMC Med Inform Decis Mak* 2008;**8** Suppl (1):S6.
36. Sadegh S, Skelton J, Anastasi E, et al. Network medicine for disease module identification and drug repurposing with the NeDRex platform. *Nat Commun* 2021;**12**:6848.
37. Schaefer MH, Serrano L, Andrade-Navarro MA. Correcting for the study bias associated with protein-protein interaction measurements reveals differences between protein degree distributions from different cancer types. *Front Genet* 2015;**6**:260.
38. Schriml LM, Mittra E, Munro J, et al. Human disease ontology 2018 update: classification, content and workflow expansion. *Nucleic Acids Res* 2019;**47**:D955–62.
39. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 2005;**102**(43):15545–50.
40. Tweedie S, Braschi B, Gray K, et al. Genenames.org: the HGNC and VGNC resources in 2021. *Nucleic Acids Res* 2021;**49**:D939–46.

ANALYSIS OF TOOL INTEGRATION AND COMMUNITY REACH

While the preceding sections presented the solutions to addressing gaps in the network-based DR workflow, a primary goal has been the successful implementation of accessible resources and the distribution of these tools as either self-sufficient, intrinsically feature rich platforms, or synergistic elements. Rather than functioning as isolated software packages, some tools were engineered with cross-platform interoperability in mind to cover the full *in silico* DR value chain. For example, technical protocols were established to allow for the seamless transfer of gene and protein sets from identification tools, such as the Epistasis Disease Atlas (EDA), directly into the Drugst.One Standalone environment for prioritization, and subsequently into DIGEST for statistical validation. This synergistic use of the presented tools can especially be seen in [Figure 1.3](#).

To quantify the success of the democratization effort, usage analytics were collected from July 2023 to April 2026. Metrics were derived from Apache server logs and package download statistics gathered. The logs were filtered through a multi-layered pipeline to approximate human interaction. Still as it will be discussed later, the derived numbers can and should only be regarded as this: an approximation. [Table 5.1](#) collects specific and relevant metrics for the presented resources with the number of uniquely recorded Internet Protocol (IP) addresses as a proxy for reach, and the number of requests as a proxy for accessibility. The code and output log of all scripts used to create the metrics and exclusion list is publicly available¹.

Table 5.1: Reach and usage metrics for the developed network medicine tools (as of end of April 2026). Data excludes known web crawlers and a list of 45,398 exclusion IP addresses. The Metric / Component column describes the specific component of a tool that is being assessed and how; “Total” in that context refers to a collective evaluation of components of the current tool or if specifically mentioned, a specific version of the tool. As unique IP addresses are used as a metric, result sets of individual components of the same tool can and often do record the same IP addresses, making it impossible to simply add up final counts of unique IP addresses over all components to receive a total.

Tool	Interaction Category	Metric / Component	Value
A - NeDRex-Web	Web Reach	Unique Website Visitors	4,517 IPs
		Total Unique Recorded Interactions	7,838 IPs

Continued on next page...

¹ https://github.com/AndiMajore/dissertation_tool_usage_scripts

Table 5.1 continued from previous page

Tool	Interaction Category	Metric / Component	Value
B - Drugst.One	Web Reach	Unique Website Visitors (Direct)	26,877 IPs
		Plugin-Only Reach (External Websites)	39,275 IPs
		Website + Plugin (Total Reach)	66,152 IPs
		Total Unique Stable Version Visitors (stable.drugst.one)	4,114 IPs
		Total Unique Main Version Visitors (drugst.one)	113,701 IPs
		Total Unique Recorded Interactions	115,704 IPs
	Referral Traffic	Unique Sites referring to Standalone	8 (including drugst.one)
		Standalone Unique Referrers	2,081 IPs
		Standalone Page Total Referrals	6,980 requests
	Development	Unique Plugin Integrators: reachable (all recorded)	59 (78)
		Reach through Plugin Integrations (Drugst.One + External)	47,132 IPs
		Python Package Downloads	~27,000 ²
C - NeEDL & EDA	Web Reach	Unique Website Visitors	12,708 IPs
	Development	NeEDL docker image download	966 ³
		Total Unique Recorded Interactions	16,020 IPs
D - DIGEST	Web Reach	Website Visitors	11,344 IPs
		Total Unique Recorded Interactions	13,082 requests
	API / Referral	API-Only Reach (External Referral)	1,738 IPs

Continued on next page...

Table 5.1 continued from previous page

Tool	Interaction Category	Metric / Component	Value
	Development	Total API Interactions	2,730 IPs
		Python Package Downloads	~34,000
Total Suite	Cumulative Reach	All Unique Device Interactions	136,302 IPs
	Cumulative Dev	Combined Tool Downloads	~62,000⁴

Figure 5.1 provides a more general statistical overview of these adoption patterns analyzing the unique device interactions recorded. Figure 5.1(A) depicts the cumulative growth of unique IP interactions aggregated per tool proxying the outreach and showing a consistent upward trajectory across the evaluation period for most tools. When compared against a baseline of other tools hosted at the same institute and server (pseudonymized as “app-X” in gray), three of the four presented solutions consistently occupy the top quarter of reach. Figure 5.1(B) shows the same metric, but split up for each of the tools’ individual components, when relevant and available, instead of aggregating all information. “drugstone-cdn” (90,687 uniques) and “drugstone-website” (26,877 uniques) representing the endpoints reached most devices. “drugstone-cdn” stands for the Drugst.One Content Delivery Network (CDN). It is responsible to serve and distribute the drugstone.js and styles.css files to anyone loading a website in which the Drugst.One plugin is integrated. It also serves a specific Drugst.One button style to simplify referring to Drugst.One’s Standalone page from other resources.

Figure 5.2 highlights total service usage in requests in Figure 5.2(A) again the tool-wide aggregates, and Figure 5.2(B) the component specific counts. Figure 5.2(A) reveals that Drugst.One generated more than 10 million requests (10,592,044 hits), the next tool following with less than a quarter of that. In Figure 5.2(B), the origin of this high number of requests can be dissected and linked mostly to the Drugst.One backend generating close to 6 million requests (5,984,935 hits), marking it as the most requested endpoint in the institute’s ecosystem. Again three of the presented solutions are found at the top quarter of the list of the in total 27 web applications, yet NeDRex-Web being now in there, whereas in Figure 5.1(A) EDA was in the top quarter.

5.1 METHODOLOGICAL CHALLENGES IN WEB-BASED SUCCESS QUANTIFICATION

A critical finding of this analysis is the inherent difficulty in accurately quantifying the reach of scientific software via web analytics. It is essential to stress once more that recorded metrics represent an approximation rather than an absolute count of unique human users due to several technical constraints:

1. **Automated Traffic and Noise:** A significant portion of web server logs consists of non-human requests from automated crawlers, indexers, and malicious scanners [274, 275]. Research

2 <https://clickpy.clickhouse.com/dashboard/drugstone>

3 <https://hub.docker.com/r/bigdatainbiomedicine/needl>

4 <https://clickpy.clickhouse.com/dashboard/biodigest> and <https://clickpy.clickhouse.com/dashboard/biodigest-light>

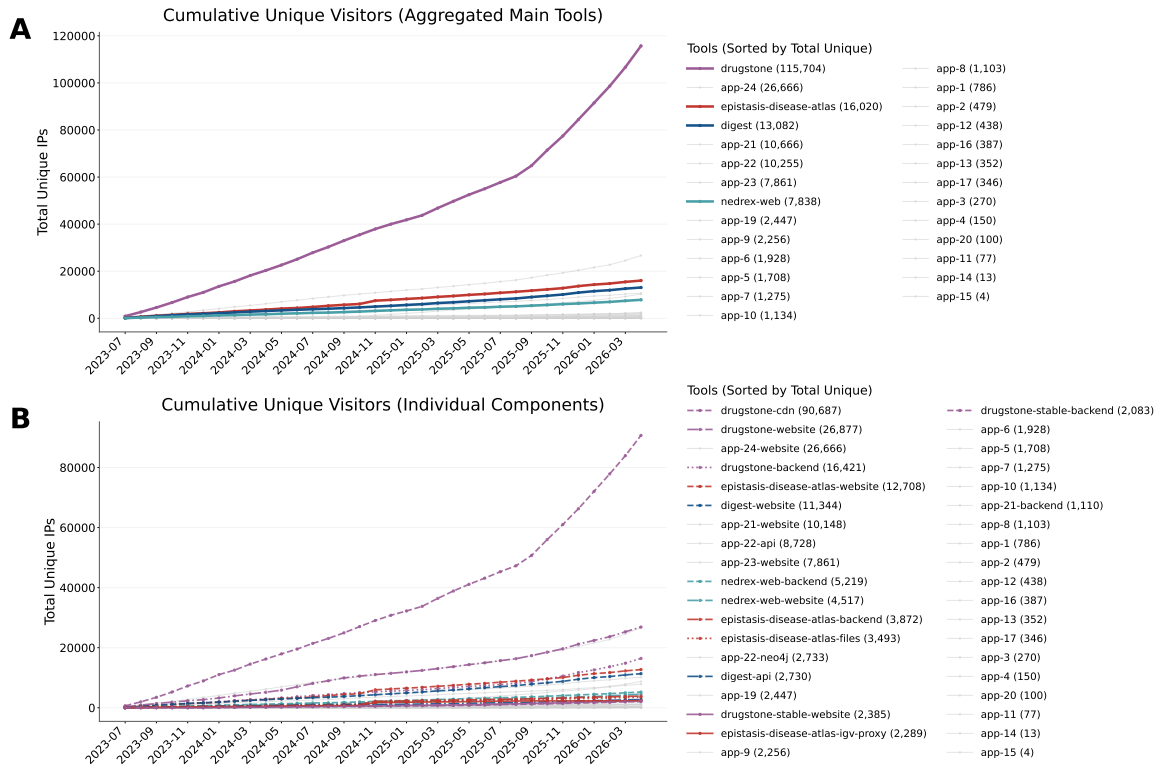


Figure 5.1: Webtool Reach: Cumulative Unique Interactions: Apache vhost logs are processed by splitting all tools into individual logs (where necessary), applying some filters to remove non-genuine requests as far as possible, and then plotting unique IP interactions aggregated over all components of each tool (A) and split into the individual components statistics, where possible, (B) over time. The colors are in line with Figure 1.3, where purple corresponds to Drugst.One related services, turquoise to NeDRex-Web, red to the EDA and blue to DIGEST. The tools in gray are 23 other publicly hosted web tools on the same web server for comparison. In July 2023 the server was first set up, hence no older logs exist, even though some tools predate this timepoint.

suggests up to 67% of bot traffic can mimic browser headers, yet many fail to execute client-side scripts [276].

2. **IP-to-User Mapping Inconsistencies:** Standard metrics like unique IPs do not perfectly reflect unique individuals even if they only create genuine traffic. A single person using multiple devices or browsers is counted multiple times, while multiple researchers sharing a single network in a laboratory environment might run over the same IP address and thus may be recorded as one.
3. **Data Filtering Strategy:** To mitigate these issues, a conservative filtering pipeline was applied. For Drugst.One, a unique architectural detail was leveraged: the loading of the Drugst.One package from its CDN was used as a “human signal”. Requests from IPs that never loaded this plugin were flagged as non-genuine when requests from the Standalone page were recorded, as that contains an integrated Drugst.One plugin. While this approach likely excludes a small number of legitimate users with dynamic IPs, it prevents the overestimation of reach. This method identified a list of 45,398 IPs to be excluded. This list was used in filtering all application logs even though the Drugst.One logs are obviously impacted most significantly, as can be seen in the unfiltered versions of Figure 5.1 and especially Figure 5.2 in, respectively, Figure A.1 and Figure A.2.

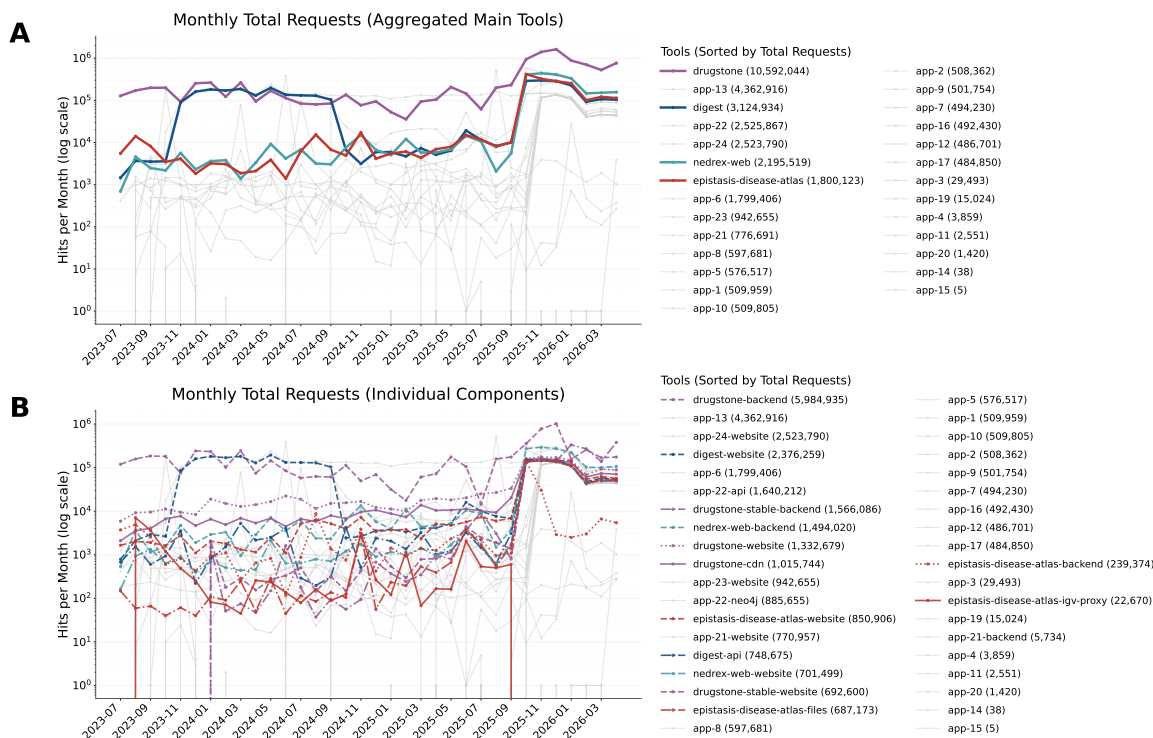


Figure 5.2: Webtool Usage: Monthly Total Requests (Hits): Logged requests serve as a proxy for general continued interest and recurring usage. This figure highlights the absolute monthly usage statistics (in number of requests logged) as a proxy for general continued interest in and usage of the tool, regardless of it being new or recurrent users: **(A)** Aggregated monthly hits per tool, showing Drugst.One’s steadily high request volume. **(B)** Component-specific monthly hits, revealing the high traffic levels sustained by backends and APIs compared to front-facing websites.

One phenomenon that can be observed most clearly in Figure 5.2 is an irrational increase in traffic across all tools beginning in September 2025. Even with more than 45,000 IPs filtered out, this surge is most likely linked to a heavy increase in the capabilities and existence of agentic web crawling and knowledge distillation [277]. This indicates that the filtering mechanisms were only partially successful and highlights why exact numbers should be viewed as orders of magnitude rather than precise headcounts.

To ensure scientific transparency all scripts for filtering, processing, and analysis are provided in a dedicated GitHub repository⁵. This includes the “run_all.sh.log” file, which records the console output of the final execution of the pipeline that produced Figure 5.1 and Figure 5.2 and the comprehensive values presented in Table 5.1, included in the general README.md file. However, full reproducibility by third parties is not possible, as adhering to data privacy regulations (GDPR) means that the raw Apache logs cannot be provided publicly, as recorded IP addresses are classified as personal data.

The implementation of Drugst.One and DIGEST as integrable resources enabled the expansion of the feature set of other existing tools. The referral system documented 39,275 users who transitioned from third-party websites into the repurposing ecosystem via the Drugst.One plugin, and 1,738 purely API-based interactions with the DIGEST validation service.

5 https://github.com/AndiMajore/dissertation_tool_usage_scripts

DISCUSSION

The following section serves as the discussion of the impact of the works presented in this thesis, first individually how they achieve the initially set objectives, and then jointly assessing their overall and combined contribution to the field of network-driven disease mechanism and DR research. The overarching goal of this dissertation was to address the bottlenecks in the network medicine value chain through an integrated and accessible software ecosystem. The usage metrics presented in the previous chapter provide evidence of the success and limitations of this democratization strategy.

6.1 DISEASE MODULE AND DRUG REPURPOSING CANDIDATE IDENTIFICATION

With the development of NeDRex-Web and Drugst.One I aimed to address the previously defined *Bottleneck 1: A lack of democratization* through *Goal 1: Integrated and accessible end-to-end solutions for network-based drug repurposing*. Both projects achieve this goal more than any previous publications did but not without their individual limitations.

6.1.1 NeDRex-Web

NeDRex-Web was designed as a robust, standalone, fully-functional web application facilitating access to all functionalities the NeDRex platform has to offer. It even expands functionalities to facilitate the use of more nuanced information and tools, like the integration of comorbidity data for shared mechanism research, highlighting identified drug disease combinations registered in ClinicalTrials.gov¹, tissue-specific PPI networks to limit the noise of interactions for disorders with a clear primary tissue where the phenotype is presenting itself, and adding promising module detection algorithms to address existing limitations [192, 227].

NeDRex-Web recorded 7,838 total unique visitors, reflecting its specialized role in the ecosystem (Table 5.1). The usage metrics in Figure 5.1 and Figure 5.2 suggests its role as the “deep-dive” and more expert focused platform. As described in Chapter 2, NeDRex-Web facilitates complex exploratory research tasks with in-depth configuration options and history management. Tasks that require significantly more time investment from the user than the other presented solutions. Additionally, from all the presented tools, it was the last published. The small but steady, linear growth shown in Figure 5.1(A) indicates a stable and loyal user base, while the lower IP count compared to distributed tools confirms that “deep-dive” platforms cater to high-intent specialists

¹ <https://clinicaltrials.gov>

rather than casual visitors. This is verified through a rather steady 1,000 to 10,000 requests per month (Figure 5.2(B)). When compared to all other recorded tools, as a baseline, the same pattern is visible, as NeDRex-Web is under the top 25% of most recorded requests (Figure 5.1(A)), yet not under the top 25% of most reach (Figure 5.2(A)). The recent increase in traffic to more than 100,000 requests per month is most likely non-genuine as discussed later (Figure 5.2).

While it provides the lowest-hurdle UI in the NeDRex ecosystem so far for non-computationally trained experts, its interface is still not perfectly intuitive. The juxtaposition of the attempt to provide extensive functionalities that do not limit users to specific research paths and an easy-to-use UI was a challenge that I –without any structured or professional training in web-development and user experience design– was initially not able to fully overcome. This was improved upon by hands-on-sessions with clinical experts to gather their feedback. Further, extensive documentation about all available functions was added to address issues where it was unclear to users why some functions would sometimes stay hidden to them. All these learnings were further applied in the following projects, most prominently Drugst.One.

One obvious option for improvement is the expansion of the offer of disease module and drug ranking prediction algorithms. While ROBUST is part of the available algorithms for module detection, the newer and possibly better ROBUST-bias-aware version is not, so far. Similarly, for drug prioritization there are only general network algorithms available; the purposely developed network proximity algorithm, presented in the introduction (Section 1.4.3.3), is so far not integrated.

Additionally, as the NeDRex database expanded its content past the initially published edge and node types, NeDRex-Web could be updated to use or provide more than it currently does. This includes information like drug side effect information and disease phenotypes, which, as the Firoozbakht *et al.* [231] showed, potentially contains valuable information.

Even though it could be improved and expanded in some ways, NeDRex-Web is already a valuable tool for direct, on-demand access to NeDRex and with that an end-to-end platform for network-based disease module detection and DR, making hypothesis generation and exploration easier than ever before.

6.1.2 *Drugst.One*

Drugst.One was a complementary attempt to NeDRex-Web. All functions and the interface are kept as simple as possible and Drugst.One provides a plugin for any website and is not just simply one web tool available on one page. This way, we aimed to upgrade web applications and other tools that previously ended with the identification of a list of genes or proteins, which are usually long and need follow-up analysis to draw any conclusion, to drug-repurposing ready websites with minimal effort. We engaged with a broad set of collaborators to understand their needs, identified integration challenges, and gathered additional feedback, incorporating and addressing as many of their concerns as possible. The general reception appears to have been great, as more than 25 tools integrated Drugst.One one way or another, to increase the potential impact of the findings they provide through their tools. While some of these tools integrating Drugst.One are DR focused projects developed by us or other members of our lab, like SCANet [278] for the

extraction of key differential regulatory networks from single cell data or ProHarMed [279] for the harmonization and exploration of proteomic meta-study information, many external projects did so as well. Very recently, an update to the Integrated Interactions Database² was published, adding docking functionally and within that the visualization of results through an integrated Drugst.One component [280]. A list of tools integrating or otherwise interacting with Drugst.One at the time of publishing this dissertation can be found in Table A.1. This includes additional seven tools that, I identified through our server logs and verified, are using Drugst.One omitting others that I could not verify manually and/or are not published yet.

The simplicity of Drugst.One’s interface and functions comes with a trade-off. Deep exploratory research, and network-connectivity searches such as those provided by NeDRex-Web are not possible out of the box. Iterative exploration of networks has been a feature from the beginning, but a limited number of tools were available to help filtering or selecting relevant subnetworks. Some of these network-refinement focused shortcomings were since addressed by a separate publication updating Drugst.One and releasing plugin version 2.0.0 [281].

While Drugst.One, significantly lowers the barrier to generating DR hypotheses, it is not intended to replace the expert-driven evaluation necessary to confirm a candidate’s clinical relevance. Instead, the platform facilitates the identification of mechanistically plausible candidates, which must then be refined through expert domain knowledge and supplementary analyses before proceeding to experimental validation.

With Drugst.One we were not only able to develop another web interface that allows the execution of some specific pipeline, but to provide researchers within and beyond the DR community with a building block they can build and extend upon. The usage metrics presented in Chapter Chapter 5 provide empirical support for this strategy. 20 external tools were verified to integrate the plugin, but the number could be as high as 59 according to the list generated from the log entries after filtering all 78 recorded referrers for domains that have been taken offline since the requests were made. Those do contain duplications, due to slight changes in domain names, websites that are not in use anymore, and some development or unpublished pages, but this list is not extensively bloated. Seemingly, seven different pages refer users to Drugst.One’s Standalone page (Table 5.1). Close to 7,000 referrals from all those sources were recorded. In total, Drugst.One amassed traffic from over 115,000 unique IPs (~113,700 through drugst.one; ~4,100 through stable.drugst.one). The 39,275 plugin-only IP interactions, which is an order of magnitude higher than the reach of standalone deep-dive platforms, confirm a successful contribution in addressing *Bottleneck 1: A lack of democratization*. Drugst.One is not requiring experts to migrate to a new platform, but embedding specialized tools within their existing workflows. The comparably large request volume shown in Figure 5.2(A) (peaking at more than 10 million hits), specifically the Drugst.One Backend shown in Figure 5.2(B) (peaking at nearly 6 million hits) goes hand in hand with the steady increase in unique IPs requesting resources from the Drugst.One CDN (the distribution server for the plugin) seen in Figure 5.1(B), further suggesting that Drugst.One has achieved a distribution reach far beyond the home institute’s primary domain, despite the limitations of those metrics.

As implicitly communicated through the existence of the Drugst.One CDN component, so far the distribution of Drugst.One is handled by an in-house controlled component. While this was

² <https://iid25.ophid.utoronto.ca/>

ideal for early development and internal testing stages, the distribution of the bundled resources should be equally possible through e.g. npm³, the node package manager. This would also allow the integration of the Drugst.One to change from through two lines of code in the html document head section to just installing the package as a dependency of a website project. This could further increase the reach generated through adoption and distribution, as it would allow the use of standard channels instead of unconventional integration strategies.

While the presented metrics are not directly translatable into the same number of unique users, the generated reach is significantly higher than in the other projects, clearly helping with the broader distribution of DR research tools and thus increasing the odds of reaching researchers who can drive impactful discoveries.

6.2 SEED GENE IDENTIFICATION

Addressing the issue of disease gene identification for DR research, we introduced the Epistasis Disease Atlas (EDA), which –to our knowledge– is the first resource for human epistasis disease set candidates. It is coupled with an explorer supporting research and refinement of epistatic interaction hypotheses for currently eight different disease datasets. The platform is powered by a novel method for epistasis detection, called NeEDL, introducing a significant speed up in evaluation due to basing it on network medicine ideas, resulting in evaluation depths no prior existing tool was able to reach with the same statistical significance. It is further scalable due to its quantum computer-ready implementation. Through the integration of Drugst.One, EDA provides direct access to DR tools.

While serving as a great additional resource for data-driven disease genes, using those epistatic interaction sets as hints for seed genes of a DR search, might introduce less novel insights than initially hoped. This is because NeEDL uses the interactome to filter down potential SNP interaction sets and thus overlaps heavily with module detection approaches, where interactomes are used to expand the selection of initially known key-player genes of a disease. So far, no systematic evaluation on the relevance of the epistatic interaction sets as DR starting point has been done, which also is partly due to the sets only having a size of five SNPs on average. This means that on average a maximum of five genes are present in each set which, for complex disorders like Alzheimer's or diabetes, is highly unlikely to cover all relevant genes that play a role. The size is of course a limitation due to current computational power even though for this project a significant amount of computational hours on super computers were used. Given NeEDL, the increasing availability of higher-qubit QC machines, epistasis research might receive a large boost in the future. Nevertheless, evaluating the currently identified sets could already reveal isolated mechanisms perturbed in the multi-factorial disorders analyzed and provide new insights into potential treatment options.

Further, reducing the number of potential SNP interactions may significantly accelerate the process, but would also lead to the exclusion of interactions that could lead to candidate sets of impact. As the field of epistasis detection is yet to determine higher-order examples of epistatic

³ <https://www.npmjs.com/>

effects, it is also not clear which statistical metric provides the best measure.

The community's need for novel genetic starting points is evidenced by the engagement with the EDA. With a total of 16,020 unique recorded IPs, the EDA demonstrates a robust interest in exploring higher-order epistatic interactions. This suggests that researchers are actively looking beyond traditional GWAS results to advance in disease-cause identification or to find more robust seed genes for their drug repurposing pipelines. As visualized in Figure 5.1(B), the EDA website (top red line) is a highly visited resource, indicating that researchers are actively looking beyond traditional GWAS results to find more robust seed genes for their drug repurposing pipelines. What is apparent is that the whole of EDA experienced a rather high number of unique visitors (Figure 5.1(A)), compared to a rather high number of requests, when comparing positions among other tools (Figure 5.1(B)). This could point to an initial high interest in a resource like this, but a lack of continued interest. Deeper and more page specific user behavior analysis would need to be conducted to confirm this assumption, though. The containerized Docker image for running analyses with NeEDL was downloaded 966 times so far suggesting the interest in exploring other GWAS datasets for epistatic interactions not presented or integrated in the publication.

All in all, this means that the initially raised *Bottleneck 2: The seed gene knowledge gap* has been mainly addressed, yet the presented work will only be one part in the larger puzzle of identifying causative genes for all disorders with genetic components and enabling the research of meaningful precision disease modules. *Goal 2: Integrating network medicine and epistasis research to facilitate identification of novel data-based disease genes for a new, robust drug repurposing angle*, has been fulfilled development of NeEDL, the systematic application thereof on eight large scale GWAS dataset and with publishing the results in the EDA, the first resource for publicly published epistatic interaction candidate sets for human diseases.

6.3 *in silico* VALIDATION

DIGEST is an attempt to provide quantifiable and biology-backed *in silico* validation for disease mechanism hypotheses. It allows for evaluation of sets, clusterings, or networks of diseases or genes based on shared genetic or functional coherence, respectively, with and without a reference to compare against. With this it provides a powerful resource in the field of network medicine, to evaluate hypotheses generated herein but also to assess the quality of tools in absence of gold standard annotations for disease module identification. Through offering DIGEST as an easy-to-use web application, a Python package, and a REST API, we enable seamless integration of its methods into existing resources and research pipelines, significantly lowering adoption barriers and maximizing its impact for developers and end users. With this, *Goal 3: Development of novel, and integrable purpose built validation algorithms for predicted disease modules* is achieved taking one step further towards eliminating *Bottleneck 3: Lack of in silico validation strategies*.

It is clear that DIGEST is explicitly focused on the evaluation of mechanistic disease background, but novel methods for the evaluation of drug target sets would be equally relevant. While building on different functional annotation resources like KEGG and GO for gene-level evaluation, these resources are driven by humanly crafted pathways and potentially miss annotations which can lead to a skewed final coherence score. This further means that study bias is inadvertently propagated

to our evaluations this way potentially impacting the resulting p-values, even though random backgrounds are sampled adequately.

Also, as previously described in Section 1.3.3, disease modules do not necessarily need to be completely coherent with regard to their enriched function. In complex disorders multiple key functions may be affected to present a certain phenotype or the pleiotropic nature of genes involved in several, potentially dissimilar functions. Still, it can be sufficient to gain critical insights into complex disorders. Specifically, its clustering evaluation option allows for the assessment of modules with disconnected functional components, such as those produced by DOMINO. A notable caveat, however, is that high interactome connectivity can cause separate functional modules to link together, potentially obscuring DIGEST's assessment. Due to this and the possibility of situation-dependent p-value inconsistencies, as reported in the original publication, the tool is best suited for cases where gene-set coherence is intrinsically valuable to the research question. P-values should therefore be regarded as metrics for functional property quantification rather than binary indicators of translatability. To support the necessary expert curation, DIGEST provides means of transparently looking at annotation information, like the mappability plots, evaluation of individual significance contribution, and plots about the generally top enriched functions to allow critical assessment and interpretation of results.

Interestingly, the usage metrics for DIGEST reveal a distinct split in user personas. While 11,344 unique visitors were recorded accessing the web interface (Figure 5.1(B)), the ~34,000 downloads of the DIGEST Python package (Table 5.1) indicate that *Bottleneck 3: Lack of in silico validation strategies* is a significant gap within automated bioinformatics pipelines. The additionally recorded 1,738 requests directly to the API, suggests that also this offer is being used. This high rate of programmatic adoption shows that this statistical framework is being integrated into other, possibly unrelated discovery workflows, thereby maximizing the impact of the method beyond the scope of a single web tool and hopefully driving translational research. Furthermore, Figure 5.2(B) shows that the DIGEST API has consistent one to tens of thousands of requests per month, suggesting it is being integrated as a persistent quality control step in larger discovery workflows, thereby maximizing its impact beyond the scope of a single web tool.

6.4 SUMMARY

The overall goal of this thesis was the development of accessible and usable web tools for network-based DR research, that have the potential of reducing the translation gap of *in silico* hypothesis generation and experimental validation. Here I want to assess the joint impact of the presented work towards this goal and network-based DR research in general. As one quantifiable proxy, the number of unique IP addresses logged accessing relevant parts of the developed systems are comparatively presented in Table 5.1 and Figure 5.1.

Based on the network-based DR workflow, as shown in Figure 1.3, I addressed shortcomings with the individual projects, but also leveraged synergies between them. We identified, based on previous research, that the usual starting point of disease gene selection can benefit from data-grounded approaches and saw potential in genetics data, especially epistasis detection, as this field is all about interactions, fitting well into the context of network, and generally also systems

medicine. To this end, we designed NeEDL, a novel epistasis detection method, and built EDA to provide an open and public resource to analyze and explore the results. Enabling tools like the EDA to easily add DR functionalities, we developed Drugst.One to close the gap between candidate gene sets and a complete, but simple suite of DR functionalities. For more advanced and complex exploratory research and more detailed analyses, I additionally developed NeDRex-Web, an online interface for the NeDRex platform, providing not only existing features, but also extending them to allow additional filtering and refinement at crucial steps. Seeing the final step of *in silico* validation as a crucial one, we published DIGEST, a novel method to quantify coherence of gene and disease sets. It is integrated in Drugst.One, allowing for the analysis of module detection results. These presented publications highlight my dedication to improve on limitations in the existing pipeline of network-based DR research tools. As shown by integrating our own tools in not only the presented ones, but also e.g. a novel module detection tool ROBUST-Web [199] or single cell analysis software SCANet [278], we heavily focused on the integrateability and otherwise interoperability, to not attempt to replace existing workflows and tools researchers might use, but to allow the expansion of them.

All individually presented works represent first-in-their-class tools that revolutionize accessibility and democratization of the field of DR. Certainly, there are further unaddressed gaps in the network-based acdr pipeline and potential to improve the presented list of tools, to jointly increase their impact even further.

Firstly, a fundamental question is, how much information can actually be drawn from a heavily reductionistic view on medicine and biology, as network medicine attempts. We often see, e.g. in interactomes and disease definitions, that aggregated information introduces noise that is propagated through the individual steps of the pipeline, and if not properly handled, will overshadow or destroy the true signal present. Additionally, information, like the PPI layer, might hold true for the general healthy case, but mutations in the diseased might influence the interactions protein can form alongside their general availability in the cell. Data-grounded approaches, like EDA, are a step towards combatting this in the first step already, yet epistasis detection is still a field of uncertainty itself, making it hard to fully trust in this new starting point for DR. In module detection and drug ranking steps in NeDRex-Web, we added the option to use expression-based filtered versions of the general interactome in an attempt to provide a version with way less noise in cases where a disease phenotype can be linked to a clear tissue. Similarly, in Drugst.One we allow the overlay of tissue-level expression information, providing a similar, yet more manual way of removing links, that biologically might not make sense. With DIGEST, we generally attempted to have a quantifiable approach to address the topic of *in silico* hypothesis evaluation, in an attempt to be able to flag insignificant disease modules that might not be worth pursuing further. Thus, offering DIGEST within the NeDRex ecosystem and also in would be an important and logical addition. Still, so far, DIGEST only covers disease and gene set evaluation, as the validation of DR candidates would require different information and potentially entirely different approaches. This leaves the simple empirical p-value based evaluation introduced in NeDRex and in NeDRex-Web as the currently only available option in the presented systems. Despite all these concerns, all presented tools are open source and, except EDA, implemented with the capability of updating to the latest information available in their sources. Consequently, the open-source nature of these platforms facilitates the development of community-driven derivatives. Additionally, “passive” refinement of the tools is ensured as the underlying biological data, like disease-gene, drug-target,

and protein-protein links, evolve toward higher levels of fidelity and lower noise thresholds by more data being generated and analyzed every day.

While EDA provides a new starting point for disease mechanism and DR analyses through the link to Drugst.One, a systematic evaluation of a potential overlap between epistasis and disease-module hypotheses has not been carried out so far. This idea was raised in the introductory section and EDA represents the first resource that enables such an analysis, as before it, no public resource for researching epistatic interaction candidate sets was available. Through the nature of NeEDL, building on information from interactome networks, the candidate sets probably are biased towards a list of variants, where linked genes interact with each other within the interactome. This by itself just serves as a logical filter to reduce the search space, yet might introduce circular reasoning when systematically analyzing result sets with regard to their adherence or overlap with disease modules. Although this perspective was not explored extensively yet, it could offer interesting and valuable potential for future research.

While one main focus of all projects was usability, there are still stark differences in the level of accessibility and intuitiveness of the presented tools. Initially, NeDRex-Web was partially not as intuitive to use as intended, mainly because of its complex feature set and analytical freedom, but after some rounds of targeted feedback and improvements, not only the interface of this project improved, but the learnings made their way into the development and improvement of the other, later developed, projects.

Jointly assessing the results across the four tools reveals an interconnected system of synergistic resources that is greater than the sum of its parts. The collective footprint of the suite, spanning over 136,000 recorded unique device visits provides empirical proof that focusing on usability and integrability is effective. The longitudinal data in [Figure 5.1](#) underscores this success, showing that the individual components of the developed tools show consistently higher usage patterns than most of the 23 other active bioinformatics applications on the same server throughout the recording period. The data further identifies two primary adoption pathways: a "distribution-" or "integration-centric" path via Drugst.One and partially DIGEST and a "method-centric" path via NeDRex-Web and the Epistasis Disease Atlas. By offering both plug-and-play modules and deep-dive platforms, this dissertation has successfully provided different entry points into network medicine, catering to both the experts in need of usable but powerful software and developers in search of reusable solutions.

As touched upon earlier, [Figure 5.1\(B\)](#) most likely recorded the heavy increase in the capabilities and the existence of agentic web crawling and knowledge distillation [277]. This is one reason why this increase was not discussed as a natural increase of interest in the tools. It shows that the filtering mechanisms applied to reduce non-genuine requests were far from fully successful. Unfiltered versions of [Figure 5.1](#) and are available as [Figure A.1](#) and [Figure A.2](#), respectively, for the interested reader. It shows that without filtering the trends do not look too dissimilar, yet the number of requests is sometimes more than four times as large. This again should visualize to the reader that the exact numbers are not to be taken at face value, but simply provide an order of magnitude of reached users. Most prominently, the resources used to identify the list of IPs, so the Drugst.One CDN, is impacted.

One large issue that I could identify with multiple systems was that research requires both up-to-date and reproducible information. As all of our resources are attempting to provide the most up-to-date information, thus updating their underlying databases on a weekly basis, reproducibility of the exact results cannot be guaranteed. With `stable.drugst.one`⁴, we therefore specifically deployed a version of Drugst.One, that updates only once a year only and all versions of Drugst.One provide downloads for the datasets used at that point in time through the documentation page. DIGEST comes with the option to run the Python package locally, which ensures that the dataset can be controlled and does not update automatically and in NeDRex-Web we provide transparent information about the build version and date of NeDRex and all underlying databases. While we are aware that reproducibility is one of the most important factors in research, we realize, that providing options for up to date and historic data alike, e.g. by allowing the user to select a past timepoint with the according state of information, requires work and especially storage space, that is not feasible to offer besides for big projects with dedicated funding. But at least monthly or yearly dumps and versions to choose from could be a feasible and valuable addition.

⁴ <https://stable.drugst.one>

CONCLUSION AND OUTLOOK

In this thesis, I presented new tools for network-driven DR research with the goal of making the field more accessible to the broader community and especially clinical and pharmacological experts. I attempted to address the most pressing bottlenecks in the current space of network medicine, module detection, and DR leading to four individually presented publications.

7.1 SUMMARY OF PRESENTED RESULTS

One major challenge of computational tools and with that a goal of this thesis was the development of useful and usable resources to bring network-based DR to a broader user base. The result are four standalone web tools, not only published as papers but as web applications, that are highly integrable into research processes and easily accessible in various ways. The use patterns over time in [Figure 5.1](#) suggest that these tools have transitioned from experimental proof-of-concept into resources with recurring usage. This is complemented by the 136,302 unique IP interactions and ~62,000 programmatic downloads recorded during the evaluation period.

NeDRex-Web, is an extensive web application for the exploration and analysis of large heterogeneous networks focusing on DR. It builds on data available through NeDRexDB, but is a fully standalone tool with the sole purpose to provide fast and easy access to the KG stored in NeDRexDB without the need of installing additional software. It integrates additional module detection algorithms and allows the use of tissue-specific interactomes for module detection. Exploratory freedom was one large goal, providing solutions that allow the near unlimited refinement of networks to assist any step necessary for relevant hypothesis generation.

Drugst.One was developed as a solution for any online (and even on-premise) operating resource in need of an interactive network explorer for proteins or genes. Its main result was the development of the Drugst.One plugin, a JavaScript bundle, that any web tool can integrate, use, and configure with only three lines of code. The plugin is highly customizable allowing it to be used as a simple interactive network visualization tool, or as a fully functional DR explorer offering module detection, drug ranking, and *in silico* validation tools in addition to handling all the data integration and updating procedures on a separate server.

NeEDL, a novel algorithm published in the same project, was used to predict the most likely SNP-sets for eight different large GWAS studies. Epistasis Disease Atlas (EDA) is the first online resource providing results of epistasis detection tools. It stores and publishes the results of NeEDL, which enables exploration on a SNP-SNP or Gene-Gene level, links the results to external resources

like g:Profiler, Interactive Genomics Viewer [282], or Drugst.One, and even allows the modification of a selected SNP set and the recomputation of epistasis scoring algorithms.

DIGEST is an *in silico* validation tool accessible through a Python package, a well-documented API, or the online UI. It allows for the assessment of functional coherence of gene sets and clusterings, as well as the assessment of genetic coherence of disease sets and clusterings, with and without a disease or gene reference given statistical rigor in mechanistic disease definition hypotheses.

The Power of Synergistic Software

The implementation of specifically Drugst.One and DIGEST as integrable resources enabled us to easily expand the functionality of some of the presented work and complement the existing feature set, as highlighted throughout the previous sections and in Figure 1.3(B). This way, Drugst.One integrates DIGEST such that users can select genes or a sub-network to validate, the processing is automatically started through the REST API of DIGEST, and the user is simply redirected to the DIGEST result page. Drugst.One, in turn, was used to bring DR research capabilities to EDA. Based on the currently selected SNP-SNP and therefore derived Gene-Gene network, users can choose to use the EDA selection as a starting point for subsequent DR research and are redirected to the Drugst.One standalone page, with the exact Gene-Gene network loaded.

The synergistic potential of this work is best evidenced by the referral metrics. Tens of thousands of users entered the drug repurposing ecosystem through third-party platforms via Drugst.One. This demonstrates that providing a cohesive workflow—from seed gene identification to statistical validation—successfully facilitates the transition from *in silico* hypothesis to real-world medical research.

7.2 CONCLUSION

The overarching goal of this dissertation was to democratize network medicine for drug repurposing. The analysis of usage metrics demonstrates that this goal has been substantially met. By addressing the democratization, knowledge, and validation bottlenecks through an integrated suite of accessible web tools and synergistic APIs, this work has meaningfully lowered the threshold for domain experts to contribute to computational drug discovery.

With NeDRex-Web, the first goal of an integrated and accessible end-to-end solution for the field was reached, providing an application that can be simply accessed from any modern browser and provides a toolkit for complex network medicine analyses drawing from a large heterogeneous KG, facilitating the access to in depth research. Network-medicine experts can dive deep into relationships between diseases, genes, drugs, and more from anywhere at any time, with as few limitations as possible, while having many options for filtering the sometimes overwhelming amount of information.

In a complementary way, Drugst.One addresses the same goal by bringing DR tools from dedicated web applications to any website or tool. It provides an option to have a simple but versatile end-to-end solution to where information about disease-associated genes or proteins is generated, with very limited work. This way, Drugst.One contributes to the availability of DR hypothesis generation by having not a single web presence but being adopted in an ever growing number of resources.

Following a novel approach for starting DR research in a data-driven way by integrating network medicine and epistasis research was a second goal reached by implementing and publishing the EDA. It is the first resource for human epistatic interaction candidate sets, powered by a novel network medicine-driven and quantum computing-ready epistasis detection algorithm called NeEDL. EDA integrates directly into network medicine by allowing the refinement of the candidate sets based on SNP-interaction network refinements, the recomputation of these candidate sets scores, and the direct start from there with a DR search through Drugst.One.

Finally, with goal three and the development of DIGEST, we provided a new way of quantifying the relevancy of gene and disease sets, clusterings, or networks, targeting specifically network medicine and therein disease mechanism research. This way the tool can help assess the quality of disease modules for certain hypotheses or even refine them based on the fundamental biological information provided.

By moving from algorithmic benchmarks to real-world usage, I have demonstrated that democratizing network medicine is not merely a technical challenge, but a distribution challenge. The comparably large reach of these tools confirms that providing accessible, integrated, and especially integrable workflows is a viable and invaluable path to enable the accessibility to useful research tools and thus hopefully the acceleration of research into drug repurposing hypotheses.

All in all, I think that with EDA as a novel way to detect disease genes, NeDRex-Web as an all-in-one platform for complex, and Drugst.One as a plugin-based, more generalistic solution to network-based DR, as well as DIGEST for final *in silico* disease mechanism validation, I heavily contributed to the availability, usability, and democratization of DR research. With that, I sincerely hope that the work included as part of this thesis can help generate, explore, or validate at least one hypothesis about a new remedy that translates to real world patient benefits.

7.3 OUTLOOK

The software ecosystem presented in this thesis establishes a foundational pipeline for accessible, network-driven DR. However, several methodological and conceptual frontiers remain to be explored to fully realize the potential of systems medicine.

To address the inherent heterogeneity of the disease module identification landscape, an ongoing research project of mine is focusing on consensus-based solutions. CAMI [283] aims to provide unified “ensemble” module definitions, lowering the hurdle for users unsure of which algorithmic output to prioritize.

Furthermore, the integration of LLMs represents a paradigm shift in accessibility. By shifting from traditional UIs to conversational interfaces, as seen in ChatDRex [284] and DrugRepoChatter [285], we can empower domain experts to execute complex graph queries and mechanistic explorations using natural language.

Future works on LLMs and other foundation models can go further in taking a more active role in generating and validating DR hypotheses. Harnessing this potential within network medicine, NetMedGPT is another project we recently preprinted that shows promising results for using AI models to extract novel insights about drug indications and contraindications, drug targets, adverse drug reactions and more [190]. Again a publicly available chat-interface gives the research community access to this powerful tool. Here, less reductionistic networks could be employed, as human “simplicity” could limit signals foundation models could pick up on.

A particularly promising avenue for accelerating the transition from *in silico* tools to clinical trials lies in the integration of EHRs. While current network medicine workflows typically rely on population-level knowledge, LLMs offer the unique capability to extract high-detail patient information from large-scale, often unstructured clinical notes. This data can be condensed into patient-specific KGs, which represent an individual’s unique health trajectory and molecular profile. By applying the network medical approaches presented in this thesis such as module identification and drug prioritization directly onto these patient-specific KGs, we can move toward a truly personalized DR paradigm closing in on the goals of predictive, preventive, personalized, and participatory (P4) medicine. This allows for the identification of treatments tailored to a patient’s specific mechanistic perturbations and enables the use of real-world evidence to retrospectively validate computational hypotheses before commencing expensive prospective trials.

Beyond interface and EHR integration, the foundational data structures of network medicine require moving beyond purely reductionist representations. Future generation repurposing workflows could prioritize gene regulatory networks and weighted, multi-layered “omics” graphs. Incorporating edge directionality and tissue-specific regulatory dynamics will allow for more biologically faithful modeling of disease perturbations. Additionally, the challenge of inconsistent disease definitions remains a significant hurdle for interoperability. Future research should explore latent space representations of diseases, embedding phenotypic and molecular data into high-dimensional vectors that are agnostic of symptom-based nomenclature. While our initial attempts at such embeddings have yielded mixed results, training specialized biomedical models remains a priority for achieving cleaner use of data from disparate ID spaces.

Ultimately, continued democratization of network medicine ensures that the collective expert knowledge of the clinical and pharmacological communities is no longer sidelined by technical complexity. By providing the tools to bridge the gap between high-dimensional molecular interplay and real-world clinical data, we can transform medicine from a symptom-based art into a mechanism-driven science. This work serves as a step toward fulfilling the original promise of the postgenomic era: a future where every patient can be treated and a deep molecular understanding of disease serves as the definitive map to their treatment recovery or even cure.

- [1] C. D. Mathers, T. Boerma, and D. Ma Fat, 'Global and regional causes of death', en, *Br. Med. Bull.*, vol. 92, no. 1, pp. 7–32, 2009 (cit. on p. 1).
- [2] C. Mathers, G. Stevens, D. Hogan, W. R. Mahanani, and J. Ho, 'Global and regional causes of death: Patterns and trends, 2000–15', en, in *Disease Control Priorities, Third Edition (Volume 9): Improving Health and Reducing Poverty*, The World Bank, Dec. 2017, pp. 69–104 (cit. on p. 1).
- [3] F. S. Collins, 'Shattuck lecture—medical and societal consequences of the human genome project', en, *N. Engl. J. Med.*, vol. 341, no. 1, pp. 28–37, Jul. 1999 (cit. on p. 1).
- [4] F. S. Collins and V. A. McKusick, 'Implications of the human genome project for medical science', en, *JAMA*, vol. 285, no. 5, pp. 540–544, Feb. 2001 (cit. on p. 1).
- [5] J. N. C. Bailey, M. A. Pericak-Vance, and J. L. Haines, 'The impact of the human genome project on complex disease', en, *Genes (Basel)*, vol. 5, no. 3, pp. 518–535, Jul. 2014 (cit. on p. 1).
- [6] D. C. Koboldt, K. M. Steinberg, D. E. Larson, R. K. Wilson, and E. R. Mardis, 'The next-generation sequencing revolution and its impact on genomics', en, *Cell*, vol. 155, no. 1, pp. 27–38, Sep. 2013 (cit. on p. 1).
- [7] Y. Zhang, H. Chen, R. Li, K. Sterling, and W. Song, 'Amyloid β -based therapy for alzheimer's disease: Challenges, successes and future', en, *Signal Transduct. Target. Ther.*, vol. 8, no. 1, p. 248, Jun. 2023 (cit. on p. 1).
- [8] G. B. Frisoni *et al.*, 'The probabilistic model of alzheimer disease: The amyloid hypothesis revised', en, *Nat. Rev. Neurosci.*, vol. 23, no. 1, pp. 53–66, Jan. 2022 (cit. on p. 1).
- [9] E. Margolis, *Why drug development is slowing down – and what to do about it*, en, <https://ideas.newsrx.com/blog/why-drug-development-is-slowing-down-and-what-to-do-about-it>, Accessed: 2025-7-18, Feb. 2024 (cit. on p. 1).
- [10] S. Leucht, B. Helfer, G. Gartlehner, and J. M. Davis, 'How effective are common medications: A perspective based on meta-analyses of major drugs', en, *BMC Med.*, vol. 13, no. 1, p. 253, Oct. 2015 (cit. on pp. 1, 2, 4).
- [11] R. Thomas and R. S. Parikh, 'Benefit versus risk', in *Glaucoma*, Elsevier, 2015, 537–541.e1 (cit. on p. 1).
- [12] N. J. Schork, 'Personalized medicine: Time for one-person trials', en, *Nature*, vol. 520, no. 7549, pp. 609–611, Apr. 2015 (cit. on pp. 1, 4).
- [13] M. R. Nelson, T. Johnson, L. Warren, A. R. Hughes, S. L. Chissoe, C.-F. Xu, and D. M. Waterworth, 'The genetics of drug efficacy: Opportunities and challenges', en, *Nat. Rev. Genet.*, vol. 17, no. 4, pp. 197–206, Apr. 2016 (cit. on pp. 1, 2, 3).
- [14] S. Simoons and I. Huys, 'R&D costs of new medicines: A landscape analysis', en, *Front. Med. (Lausanne)*, vol. 8, p. 760762, Oct. 2021 (cit. on p. 2).
- [15] N. Berdigaliyev and M. Aljofan, 'An overview of drug discovery and development', en, *Future Med. Chem.*, vol. 12, no. 10, pp. 939–947, May 2020 (cit. on p. 2).
- [16] Office of the Commissioner, *The drug development process*, en, <https://www.fda.gov/patients/learn-about-drug-and-device-approvals/drug-development-process>, Accessed: 2026-2-13, Feb. 2020 (cit. on p. 2).
- [17] U.S. Food and Drug Administration. 'Step 3: Clinical research', U.S. Food and Drug Administration. (Jan. 4, 2018), [Online]. Available: <https://www.fda.gov/patients/drug-development-process/step-3-clinical-research> (visited on 05/12/2026) (cit. on p. 2).
- [18] D. Sun, W. Gao, H. Hu, and S. Zhou, 'Why 90% of clinical drug development fails and how to improve it?', en, *Acta Pharm. Sin. B.*, vol. 12, no. 7, pp. 3049–3062, Jul. 2022 (cit. on p. 2).
- [19] A. D. Hingorani *et al.*, 'Improving the odds of drug development success through human genomics: Modelling study', en, *Sci. Rep.*, vol. 9, no. 1, p. 18911, Dec. 2019 (cit. on p. 2).
- [20] H.-G. Eichler, E. Abadie, A. Breckenridge, B. Flamion, L. L. Gustafsson, H. Leufkens, M. Rowland, C. K. Schneider, and B. Bloechl-Daum, 'Bridging the efficacy-effectiveness gap: A regulator's perspective on addressing variability of drug response', en, *Nat. Rev. Drug Discov.*, vol. 10, no. 7, pp. 495–506, Jul. 2011 (cit. on p. 2).
- [21] S. W. Shin, J. S. Lee, S. Abdi, S. J. Lee, and K. H. Kim, 'Antipsychotics for patients with pain', en, *Korean J. Pain*, vol. 32, no. 1, pp. 3–11, Jan. 2019 (cit. on p. 2).
- [22] A. V. Morant, V. Jagalski, and H. T. Vestergaard, 'Labeling of disease-modifying therapies for neurodegenerative disorders', en, *Front. Med. (Lausanne)*, vol. 6, p. 223, Oct. 2019 (cit. on pp. 2, 3).

- [23] G. Hasler, 'Pathophysiology of depression: Do we have any solid evidence of interest to clinicians?', en, *World Psychiatry*, vol. 9, no. 3, pp. 155–161, Oct. 2010 (cit. on p. 2).
- [24] A. M. Abeles, M. H. Pillinger, B. M. Solitar, and M. Abeles, 'Narrative review: The pathophysiology of fibromyalgia', en, *Ann. Intern. Med.*, vol. 146, no. 10, pp. 726–734, May 2007 (cit. on p. 2).
- [25] H. I. Roessler, N. V. A. M. Knoers, M. M. van Haelst, and G. van Haaften, 'Drug repurposing for rare diseases', en, *Trends Pharmacol. Sci.*, vol. 42, no. 4, pp. 255–267, Apr. 2021 (cit. on pp. 2, 3, 5).
- [26] S. Wang, Y. Liu, Y. Feng, J. Zhang, J. Swinnen, Y. Li, and Y. Ni, 'A review on curability of cancers: More efforts for novel therapeutic options are needed', en, *Cancers (Basel)*, vol. 11, no. 11, p. 1782, Nov. 2019 (cit. on p. 2).
- [27] National Institutes of Health (US) and Biological Sciences Curriculum Study, *Understanding emerging and re-emerging infectious diseases*, en. National Institutes of Health, 2007 (cit. on p. 2).
- [28] H. Nishio, E. T. E. Niba, T. Saito, K. Okamoto, Y. Takeshima, and H. Awano, 'Spinal muscular atrophy: The past, present, and future of diagnosis and treatment', en, *Int. J. Mol. Sci.*, vol. 24, no. 15, p. 11939, Jul. 2023 (cit. on p. 2).
- [29] M. C. Bierlaagh, D. Muilwijk, J. M. Beekman, and C. K. van der Ent, 'A new era for people with cystic fibrosis', en, *Eur. J. Pediatr.*, vol. 180, no. 9, pp. 2731–2739, Sep. 2021 (cit. on p. 2).
- [30] C. Nogales, Z. M. Mamdouh, M. List, C. Kiel, A. I. Casas, and H. H. H. W. Schmidt, 'Network pharmacology: Curing causal mechanisms instead of treating symptoms', en, *Trends Pharmacol. Sci.*, vol. 43, no. 2, pp. 136–150, Feb. 2022 (cit. on pp. 2, 5).
- [31] Å. Johansson, O. A. Andreassen, S. Brunak, P. W. Franks, H. Hedman, R. J. F. Loos, B. Meder, E. Melén, C. E. Wheelock, and B. Jacobsson, 'Precision medicine in complex diseases-molecular subgrouping for improved prediction and treatment stratification', en, *J. Intern. Med.*, vol. 294, no. 4, pp. 378–396, Oct. 2023 (cit. on pp. 3, 4, 7, 17).
- [32] T. J. Hwang, D. Carpenter, J. C. Lauffenburger, B. Wang, J. M. Franklin, and A. S. Kesselheim, 'Failure of investigational drugs in late-stage clinical development and publication of trial results', en, *JAMA Intern. Med.*, vol. 176, no. 12, pp. 1826–1833, Dec. 2016 (cit. on p. 3).
- [33] I. W. Mak, N. Evaniew, and M. Ghert, 'Lost in translation: Animal models and clinical trials in cancer treatment', en, *Am. J. Transl. Res.*, vol. 6, no. 2, pp. 114–118, Jan. 2014 (cit. on p. 3).
- [34] P. McGonigle and B. Ruggeri, 'Animal models of human disease: Challenges in enabling translation', en, *Biochem. Pharmacol.*, vol. 87, no. 1, pp. 162–171, Jan. 2014 (cit. on p. 3).
- [35] L. B. Kinter, R. DeHaven, D. K. Johnson, and J. J. DeGeorge, 'A brief history of use of animals in biomedical research and perspective on non-animal alternatives', en, *ILAR J.*, vol. 62, no. 1-2, pp. 7–16, Dec. 2021 (cit. on p. 3).
- [36] I. Zucker and B. J. Prendergast, 'Sex differences in pharmacokinetics predict adverse drug reactions in women', en, *Biol. Sex Differ.*, vol. 11, no. 1, p. 32, Jun. 2020 (cit. on p. 3).
- [37] *Fast track, breakthrough therapy, accelerated approval, priority review*, en, <https://www.fda.gov/patients/learn-about-drug-and-device-approvals/fast-track-breakthrough-therapy-accelerated-approval-priority-review>, Accessed: 2025-7-23, Jun. 2023 (cit. on p. 3).
- [38] A. E. Muller, U. Theuretzbacher, and J. W. Mouton, 'Use of old antibiotics now and in the future from a pharmacokinetic/pharmacodynamic perspective', en, *Clin. Microbiol. Infect.*, vol. 21, no. 10, pp. 881–885, Oct. 2015 (cit. on p. 3).
- [39] A. C. Fanaroff and M. T. Roe, 'Contemporary reflections on the safety of long-term aspirin treatment for the secondary prevention of cardiovascular disease', en, *Drug Saf.*, vol. 39, no. 8, pp. 715–727, Aug. 2016 (cit. on p. 3).
- [40] B. T. Saragiotto, G. C. Machado, M. L. Ferreira, M. B. Pinheiro, C. Abdel Shaheed, and C. G. Maher, 'Paracetamol for low back pain', en, *Cochrane Database Syst. Rev.*, vol. 2019, no. 6, p. CD012230, Jun. 2016 (cit. on p. 3).
- [41] M. Pincus, 'Management of digoxin toxicity', en, *Aust. Prescr.*, vol. 39, no. 1, pp. 18–20, Feb. 2016 (cit. on p. 3).
- [42] L. Aguirre Dávila, K. Weber, U. Bavendiek, J. Bauersachs, J. Wittes, S. Yusuf, and A. Koch, 'Digoxin-mortality: Randomized vs. observational comparison in the DIG trial', en, *Eur. Heart J.*, vol. 40, no. 40, pp. 3336–3341, Oct. 2019 (cit. on p. 3).
- [43] H. P. Selker and H. Pardes, 'Changing the incentives for development of low-cost high-impact treatments', en, *J. Clin. Transl. Sci.*, vol. 1, no. 5, pp. 267–268, Oct. 2017 (cit. on p. 3).

- [44] C. I. E. Smith, P. Bergman, and D. W. Hagey, 'Estimating the number of diseases - the concept of rare, ultra-rare, and hyper-rare', en, *iScience*, vol. 25, no. 8, p. 104 698, Aug. 2022 (cit. on p. 3).
- [45] S. Nguengang Wakap, D. M. Lambert, A. Olry, C. Rodwell, C. Gueydan, V. Lanneau, D. Murphy, Y. Le Cam, and A. Rath, 'Estimating cumulative point prevalence of rare diseases: Analysis of the orphanet database', en, *Eur. J. Hum. Genet.*, vol. 28, no. 2, pp. 165–173, Feb. 2020 (cit. on pp. 3, 5).
- [46] E. Tambuyzer *et al.*, 'Therapies for rare diseases: Therapeutic modalities, progress and challenges ahead', en, *Nat. Rev. Drug Discov.*, vol. 19, no. 2, pp. 93–111, Feb. 2020 (cit. on p. 3).
- [47] S. C. Groft, M. Posada, and D. Taruscio, 'Progress, challenges and global approaches to rare diseases', en, *Acta Paediatr.*, vol. 110, no. 10, pp. 2711–2716, Oct. 2021 (cit. on p. 3).
- [48] M. Gabay, 'The orphan drug act: An appropriate approval pathway for treatments of rare diseases?', en, *Hosp. Pharm.*, vol. 54, no. 5, pp. 283–284, Oct. 2019 (cit. on p. 3).
- [49] N. Bagley, B. Berger, A. Chandra, C. Garthwaite, and A. D. Stern, 'The orphan drug act at 35: Observations and an outlook for the twenty-first century', en, *Innov. Policy Econ.*, vol. 19, pp. 97–137, Jan. 2019 (cit. on p. 3).
- [50] R. Santos *et al.*, 'A comprehensive map of molecular drug targets', en, *Nat. Rev. Drug Discov.*, vol. 16, no. 1, pp. 19–34, Jan. 2017 (cit. on p. 3).
- [51] I. Ratnani, *Gap between practices of evidence-based medicine and personalized medicine: A barrier in learning*, en, <https://hdl.handle.net/10657/10723>, Accessed: 2025-12-30, Aug. 2021 (cit. on p. 4).
- [52] S. E. Jackson and J. D. Chester, 'Personalised cancer medicine: Personalised cancer medicine', en, *Int. J. Cancer*, vol. 137, no. 2, pp. 262–266, Jul. 2015 (cit. on p. 4).
- [53] W. Sadee, D. Wang, K. Hartmann, and A. E. Toland, 'Pharmacogenomics: Driving personalized medicine', en, *Pharmacol. Rev.*, vol. 75, no. 4, pp. 789–814, Jul. 2023 (cit. on p. 4).
- [54] H. Vellekoop *et al.*, 'The net benefit of personalized medicine: A systematic literature review and regression analysis', en, *Value Health*, vol. 25, no. 8, pp. 1428–1438, Aug. 2022 (cit. on p. 4).
- [55] S. Akodad, X. Niu, B. Secades, and H. Stevens, 'Impact of drug repurposing between 1985 and 2024 on pharmaceutical innovation', en, *Commun. Med. (Lond.)*, vol. 6, no. 1, p. 84, Jan. 2026 (cit. on p. 4).
- [56] J. M. Pulley *et al.*, 'Accelerating precision drug development and drug repurposing by leveraging human genetics', en, *Assay Drug Dev. Technol.*, vol. 15, no. 3, pp. 113–119, Apr. 2017 (cit. on p. 4).
- [57] S. Shah, M. M. Dooms, S. Amaral-Garcia, and M. Igoillo-Esteve, 'Current drug repurposing strategies for rare neurodegenerative disorders', en, *Front. Pharmacol.*, vol. 12, p. 768 023, Dec. 2021 (cit. on p. 4).
- [58] A. P. Challa, R. R. Lavieri, J. T. Lewis, N. M. Zaleski, J. K. Shirey-Rice, P. A. Harris, D. M. Aronoff, and J. M. Pulley, 'Systematically prioritizing candidates in genome-based drug repurposing', en, *Assay Drug Dev. Technol.*, vol. 17, no. 8, pp. 352–363, Nov. 2019 (cit. on p. 4).
- [59] T. I. Oprea and J. Mestres, 'Drug repurposing: Far beyond new targets for old drugs', en, *AAPS J.*, vol. 14, no. 4, pp. 759–763, Dec. 2012 (cit. on p. 4).
- [60] A. E. Clout, O. Della Pasqua, M. G. Hanna, M. Orlu, and R. D. S. Pitceathly, 'Drug repurposing in neurological diseases: An integrated approach to reduce trial and error', en, *J. Neurol. Neurosurg. Psychiatry*, vol. 90, no. 11, pp. 1270–1275, Nov. 2019 (cit. on p. 4).
- [61] H. Xu, J. Li, X. Jiang, and Q. Chen, 'Electronic health records for drug repurposing: Current status, challenges, and future directions', en, *Clin. Pharmacol. Ther.*, vol. 107, no. 4, pp. 712–714, Apr. 2020 (cit. on p. 4).
- [62] H. A. Mucke, 'The case of galantamine: Repurposing and late blooming of a cholinergic drug', en, *Future Sci. OA*, vol. 1, no. 4, FSO73, Nov. 2015 (cit. on p. 4).
- [63] A. Palumbo *et al.*, 'Thalidomide for treatment of multiple myeloma: 10 years later', en, *Blood*, vol. 111, no. 8, pp. 3968–3977, Apr. 2008 (cit. on p. 4).
- [64] S. K. Teo, K. E. Resztak, M. A. Scheffler, K. A. Kook, J. B. Zeldis, D. I. Stirling, and S. D. Thomas, 'Thalidomide in the treatment of leprosy', en, *Microbes Infect.*, vol. 4, no. 11, pp. 1193–1202, Sep. 2002 (cit. on p. 4).
- [65] H. A. Ghofrani, I. H. Osterloh, and F. Grimminger, 'Sildenafil: From angina to erectile dysfunction to pulmonary hypertension and beyond', en, *Nat. Rev. Drug Discov.*, vol. 5, no. 8, pp. 689–702, Aug. 2006 (cit. on p. 5).
- [66] D. A. Sica, 'Minoxidil: An underused vasodilator for resistant or severe hypertension', en, *J. Clin. Hypertens. (Greenwich)*, vol. 6, no. 5, pp. 283–287, May 2004 (cit. on p. 5).
- [67] P. Suchonwanit, S. Thammarucha, and K. Leerunyakul, 'Minoxidil and its use in hair disorders: A review', en, *Drug Des. Devel. Ther.*, vol. 13, pp. 2777–2786, Aug. 2019 (cit. on p. 5).

- [68] V. Altabas, Z. Orlović, and M. Baretić, 'Addressing the shortage of GLP-1 RA and dual GIP/GLP-1 RA-based therapies—a systematic review', en, *Diabetology (Basel)*, vol. 6, no. 6, p. 52, Jun. 2025 (cit. on p. 5).
- [69] W. Qin, J. Yang, C. Deng, Q. Ruan, and K. Duan, 'Efficacy and safety of semaglutide 2.4 mg for weight loss in overweight or obese adults without diabetes: An updated systematic review and meta-analysis including the 2-year STEP 5 trial', en, *Diabetes Obes. Metab.*, vol. 26, no. 3, pp. 911–923, Mar. 2024 (cit. on p. 5).
- [70] R. Siqueira, V. Fernandes, L. Ferreira, V. De, T. Glaser, H. Ulrich, and C. Lameu, 'Complex diseases demand novel treatment strategies: Understanding drug combination', en, *Drug Comb. Ther.*, vol. 4, no. 2, p. 6, 2022 (cit. on p. 5).
- [71] B. He, C. Lu, G. Zheng, X. He, M. Wang, G. Chen, G. Zhang, and A. Lu, 'Combination therapeutics in complex diseases', en, *J. Cell. Mol. Med.*, vol. 20, no. 12, pp. 2231–2240, Dec. 2016 (cit. on p. 5).
- [72] Y. Cha *et al.*, 'Drug repurposing from the perspective of pharmaceutical companies: Drug repurposing in pharmaceutical companies', en, *Br. J. Pharmacol.*, vol. 175, no. 2, pp. 168–180, Jan. 2018 (cit. on p. 5).
- [73] R. Malla, S. Viswanathan, S. Makena, S. Kapoor, D. Verma, A. A. Raju, M. Dunna, and N. Muniraj, 'Revitalizing cancer treatment: Exploring the role of drug repurposing', en, *Cancers (Basel)*, vol. 16, no. 8, Apr. 2024 (cit. on p. 5).
- [74] A. H. Jonker, D. O'Connor, M. Cavaller-Bellaubi, C. Fetro, M. Gogou, P. A. C. 't Hoen, M. de Kort, H. Stone, N. Valentine, and A. M. G. Pasmooij, 'Drug repurposing for rare: Progress and opportunities for the rare disease community', en, *Front. Med. (Lausanne)*, vol. 11, p. 1352803, Jan. 2024 (cit. on p. 5).
- [75] F. R. Weth, G. B. Hoggarth, A. F. Weth, E. Paterson, M. P. J. White, S. T. Tan, L. Peng, and C. Gray, 'Unlocking hidden potential: Advancements, approaches, and obstacles in repurposing drugs for cancer therapy', en, *Br. J. Cancer*, vol. 130, no. 5, pp. 703–715, Mar. 2024 (cit. on p. 5).
- [76] F. Bertolini, V. P. Sukhatme, and G. Bouche, 'Drug repurposing in oncology—patient and health systems opportunities', en, *Nat. Rev. Clin. Oncol.*, vol. 12, no. 12, pp. 732–742, Dec. 2015 (cit. on p. 5).
- [77] R. Feldman, 'May your drug price be evergreen', en, *J. Law Biosci.*, vol. 5, no. 3, pp. 590–647, Dec. 2018 (cit. on p. 5).
- [78] M. Flores, G. Glusman, K. Brogaard, N. D. Price, and L. Hood, 'P4 medicine: How systems medicine will transform the healthcare sector and society', en, *Per. Med.*, vol. 10, no. 6, pp. 565–576, 2013 (cit. on p. 6).
- [79] C. P. Hopper, P. N. Zambrana, U. Goebel, and J. Wollborn, 'A brief history of carbon monoxide and its therapeutic origins', en, *Nitric Oxide*, vol. 111–112, pp. 45–63, Jun. 2021 (cit. on p. 6).
- [80] GTEx Consortium *et al.*, 'Genetic effects on gene expression across human tissues', en, *Nature*, vol. 550, no. 7675, pp. 204–213, Oct. 2017 (cit. on p. 6).
- [81] K. Lage, N. T. Hansen, E. O. Karlberg, A. C. Eklund, F. S. Roque, P. K. Donahoe, Z. Szallasi, T. S. Jensen, and S. Brunak, 'A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes', en, *Proc. Natl. Acad. Sci. U. S. A.*, vol. 105, no. 52, pp. 20870–20875, Dec. 2008 (cit. on p. 6).
- [82] B. J. Strober, R. Elorbany, K. Rhodes, N. Krishnan, K. Tayeb, A. Battle, and Y. Gilad, 'Dynamic genetic regulation of gene expression during cellular differentiation', en, *Science*, vol. 364, no. 6447, pp. 1287–1290, Jun. 2019 (cit. on p. 6).
- [83] R. Froissart, M. Piraud, A. M. Boudjemline, C. Vianey-Saban, F. Petit, A. Hubert-Buron, P. T. Eberschweiler, V. Gajdos, and P. Labrune, 'Glucose-6-phosphatase deficiency', en, *Orphanet J. Rare Dis.*, vol. 6, no. 1, p. 27, May 2011 (cit. on p. 6).
- [84] O. Omarjee *et al.*, 'Monogenic lupus: Dissecting heterogeneity', en, *Autoimmun. Rev.*, vol. 18, no. 10, p. 102361, Oct. 2019 (cit. on p. 6).
- [85] J. Krainer, S. Siebenhandl, and A. Weinhäusel, 'Systemic autoinflammatory diseases', en, *J. Autoimmun.*, vol. 109, no. 102421, p. 102421, May 2020 (cit. on p. 6).
- [86] B. Zhang *et al.*, 'Cross-tissue molecular responses in the liver and blood after toxicant exposures', en, Oct. 2025 (cit. on p. 6).
- [87] S. Oparil, M. A. Zaman, and D. A. Calhoun, 'Pathogenesis of hypertension', en, *Ann. Intern. Med.*, vol. 139, no. 9, pp. 761–776, Nov. 2003 (cit. on p. 6).
- [88] H. Chen, L. Zhang, M. Liu, Y. Li, and Y. Chi, 'Multi-omics research on angina pectoris: A novel perspective', en, *Aging Dis.*, Dec. 2024 (cit. on p. 6).

- [89] M. Martínez-García and E. Hernández-Lemus, 'Data integration challenges for machine learning in precision medicine', en, *Front. Med. (Lausanne)*, vol. 8, p. 784455, 2021 (cit. on p. 7).
- [90] S. Sadegh *et al.*, 'Lacking mechanistic disease definitions and corresponding association data hamper progress in network medicine and beyond', en, *Nat. Commun.*, vol. 14, no. 1, p. 1662, Mar. 2023 (cit. on pp. 7, 8, 11, 17, 23, 24, 26, 34, 61, 62).
- [91] A. J. Espay, P. Brundin, and A. E. Lang, 'Precision medicine for disease modification in parkinson disease', en, *Nat. Rev. Neurol.*, vol. 13, no. 2, pp. 119–126, Feb. 2017 (cit. on p. 7).
- [92] M. Benson, 'Clinical implications of omics and systems medicine: Focus on predictive and individualized treatment', en, *J. Intern. Med.*, vol. 279, no. 3, pp. 229–240, Mar. 2016 (cit. on p. 7).
- [93] H. Pang, J. Lin, S. Luo, G. Huang, X. Li, Z. Xie, and Z. Zhou, 'The missing heritability in type 1 diabetes', en, *Diabetes Obes. Metab.*, vol. 24, no. 10, pp. 1901–1911, Oct. 2022 (cit. on p. 7).
- [94] C. Queitsch, K. D. Carlson, and S. Girirajan, 'Lessons from model organisms: Phenotypic robustness and missing heritability in complex disease', en, *PLoS Genet.*, vol. 8, no. 11, e1003041, Nov. 2012 (cit. on p. 7).
- [95] L. Powers, 'An expansion of jones's theory for the explanation of heterosis', *Am. Nat.*, vol. 78, no. 776, pp. 275–280, May 1944 (cit. on p. 7).
- [96] P. C. Phillips, 'Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems', en, *Nat. Rev. Genet.*, vol. 9, no. 11, pp. 855–867, Nov. 2008 (cit. on p. 7).
- [97] W. M. Pardridge, 'Drug transport across the blood-brain barrier', en, *J. Cereb. Blood Flow Metab.*, vol. 32, no. 11, pp. 1959–1972, Nov. 2012 (cit. on p. 7).
- [98] K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Barabási, 'The human disease network', en, *Proc. Natl. Acad. Sci. U. S. A.*, vol. 104, no. 21, pp. 8685–8690, May 2007 (cit. on p. 8).
- [99] K.-I. Goh and I.-G. Choi, 'Exploring the human diseasome: The human disease network', en, *Brief. Funct. Genomics*, vol. 11, no. 6, pp. 533–542, Nov. 2012 (cit. on p. 8).
- [100] T. Pawson and R. Linding, 'Network medicine', en, *FEBS Lett.*, vol. 582, no. 8, pp. 1266–1270, Apr. 2008 (cit. on p. 8).
- [101] M. Zitnik *et al.*, 'Current and future directions in network biology', en, *Bioinform. Adv.*, vol. 4, no. 1, vbae099, Aug. 2024 (cit. on p. 9).
- [102] C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, 'BioGRID: A general repository for interaction datasets', en, *Nucleic Acids Res.*, vol. 34, no. Database issue, pp. D535–9, Jan. 2006 (cit. on pp. 11, 13).
- [103] D. Szklarczyk *et al.*, 'The STRING database in 2025: Protein networks with directionality of regulation', en, *Nucleic Acids Res.*, vol. 53, no. D1, pp. D730–D737, Jan. 2025 (cit. on pp. 11, 15, 39).
- [104] B. Zdrazil *et al.*, 'The ChEMBL database in 2023: A drug discovery platform spanning multiple bioactivity data types and time periods', en, *Nucleic Acids Res.*, vol. 52, no. D1, pp. D1180–D1192, Jan. 2024 (cit. on pp. 11, 13).
- [105] L. M. Schriml *et al.*, 'Human disease ontology 2018 update: Classification, content and workflow expansion', en, *Nucleic Acids Res.*, vol. 47, no. D1, pp. D955–D962, Jan. 2019 (cit. on p. 11).
- [106] D. S. Wishart *et al.*, 'DrugBank 5.0: A major update to the DrugBank database for 2018', en, *Nucleic Acids Res.*, vol. 46, no. D1, pp. D1074–D1082, Jan. 2018 (cit. on pp. 11, 14).
- [107] S. Avram *et al.*, 'DrugCentral 2021 supports drug discovery and repositioning', en, *Nucleic Acids Res.*, vol. 49, no. D1, pp. D1160–D1169, Jan. 2021 (cit. on pp. 12, 14).
- [108] S. C. Dyer *et al.*, 'Ensembl 2025', en, *Nucleic Acids Res.*, vol. 53, no. D1, pp. D948–D957, Jan. 2025 (cit. on p. 12).
- [109] World Health Organization, *The international statistical classification of diseases and related health problems, ICD-10*, 10th ed. Genève, Switzerland: World Health Organization, Oct. 2015 (cit. on p. 12).
- [110] C. J. Mungall *et al.*, 'The monarch initiative: An integrative data and analytic platform connecting phenotypes to genotypes across species', en, *Nucleic Acids Res.*, vol. 45, no. D1, pp. D712–D722, Jan. 2017 (cit. on p. 12).
- [111] E. W. Sayers *et al.*, 'Database resources of the national center for biotechnology information in 2025', en, *Nucleic Acids Res.*, vol. 53, no. D1, pp. D20–D29, Jan. 2025 (cit. on p. 12).
- [112] J. S. Amberger, C. A. Bocchini, A. F. Scott, and A. Hamosh, 'OMIM.org: Leveraging knowledge across phenotype-gene relationships', en, *Nucleic Acids Res.*, vol. 47, no. D1, pp. D1038–D1043, Jan. 2019 (cit. on pp. 12, 15).

- [113] M. Q. Stearns, C. Price, K. A. Spackman, and A. Y. Wang, 'SNOMED clinical terms: Overview of the development process and project status', en, *Proc. AMIA Symp.*, pp. 662–666, 2001 (cit. on p. 12).
- [114] O. Bodenreider, 'The unified medical language system (UMLS): Integrating biomedical terminology', en, *Nucleic Acids Res.*, vol. 32, no. Database issue, pp. D267–70, Jan. 2004 (cit. on p. 12).
- [115] UniProt Consortium, 'UniProt: The universal protein knowledgebase in 2025', en, *Nucleic Acids Res.*, vol. 53, no. D1, pp. D609–D617, Jan. 2025 (cit. on pp. 12, 15).
- [116] D. Alonso-López, F. J. Campos-Laborie, M. A. Gutiérrez, L. Lambourne, M. A. Calderwood, M. Vidal, and J. De Las Rivas, 'APIID database: Redefining protein–protein interaction experimental evidences and binary interactomes', en, *Database*, vol. 2019, baz005, Jan. 2019 (cit. on p. 13).
- [117] A. P. Davis, C. J. Grondin, R. J. Johnson, D. Sciaky, J. Wiegiers, T. C. Wiegiers, and C. J. Mattingly, 'Comparative toxicogenomics database (CTD): Update 2021', en, *Nucleic Acids Res.*, vol. 49, no. D1, pp. D1138–D1143, Jan. 2021 (cit. on p. 14).
- [118] M. Cannon *et al.*, 'DGIdb 5.0: Rebuilding the drug-gene interaction database for precision medicine and drug discovery platforms', en, *Nucleic Acids Res.*, vol. 52, no. D1, pp. D1227–D1235, Jan. 2024 (cit. on p. 14).
- [119] D. Grissa, A. Junge, T. I. Oprea, and L. J. Jensen, 'Diseases 2.0: A weekly updated database of disease-gene associations from text mining and data integration', en, *Database (Oxford)*, vol. 2022, baac019, Mar. 2022 (cit. on p. 14).
- [120] J. Pinero, N. Queralt-Rosinach, A. Bravo, J. Deu-Pons, A. Bauer-Mehren, M. Baron, F. Sanz, and L. I. Furlong, *DisGeNET: A discovery platform for the dynamical exploration of human diseases and their genes*, 2015 (cit. on p. 14).
- [121] M. Kotlyar, C. Pastrello, Z. Ahmed, J. Chee, Z. Varyova, and I. Jurisica, 'IID 2021: Towards context-specific protein interaction analyses by increased coverage, enhanced annotation and enrichment analysis', en, *Nucleic Acids Res.*, vol. 50, no. D1, pp. D640–D647, Jan. 2022 (cit. on pp. 14, 39).
- [122] S. Orchard *et al.*, 'The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases', en, *Nucleic Acids Res.*, vol. 42, no. Database issue, pp. D358–63, Jan. 2014 (cit. on p. 14).
- [123] S. Lobentanzer *et al.*, 'Democratizing knowledge representation with BioCypher', en, *Nat. Biotechnol.*, Jun. 2023 (cit. on p. 15).
- [124] D. Türei *et al.*, 'Integrated intra- and intercellular signaling knowledge for multicellular omics analysis', en, *Mol. Syst. Biol.*, vol. 17, no. 3, e9923, Mar. 2021 (cit. on p. 15).
- [125] D. S. Himmelstein, A. Lizée, C. Hessler, L. Brueggeman, S. L. Chen, D. Hadley, A. Green, P. Khankhanian, and S. E. Baranzini, 'Systematic integration of biomedical knowledge prioritizes drugs for repurposing', en, *Elife*, vol. 6, e26726, Sep. 2017 (cit. on pp. 15, 16, 21, 22, 33, 39, 61).
- [126] A. Buniello *et al.*, 'Open targets platform: Facilitating therapeutic hypotheses building in drug discovery', en, *Nucleic Acids Res.*, vol. 53, no. D1, pp. D1467–D1475, Jan. 2025 (cit. on pp. 16, 21, 33, 61).
- [127] European Commission. 'A european platform for drug repurposing', Community Research and Development Information Service (CORDIS). (Apr. 22, 2026), [Online]. Available: <https://cordis.europa.eu/article/id/464696-a-european-platform-for-drug-repurposing> (visited on 05/12/2026) (cit. on p. 16).
- [128] European Commission. 'Precision drug repurposing for europe and the world', Community Research and Development Information Service (CORDIS). (2022), [Online]. Available: <https://cordis.europa.eu/project/id/101057619> (visited on 05/12/2026) (cit. on pp. 16, 136).
- [129] S. Sadegh *et al.*, 'Network medicine for disease module identification and drug repurposing with the NeDRex platform', en, *Nat. Commun.*, vol. 12, no. 1, p. 6848, Nov. 2021 (cit. on pp. 16, 20, 21, 22, 23, 31, 33, 61, 87).
- [130] Q. Wang and R. Xu, 'Disease comorbidity-guided drug repositioning: A case study in schizophrenia', en, *AMIA Annu. Symp. Proc.*, vol. 2018, pp. 1300–1309, Dec. 2018 (cit. on p. 17).
- [131] K. Shameer, G. Dow, B. S. Glicksberg, K. W. Johnson, Y. Ze, M. S. Tomlinson, B. Readhead, J. T. Dudley, and I. J. Kullo, 'A network-biology informed computational drug repositioning strategy to target disease risk trajectories and comorbidities of peripheral artery disease', en, *AMIA Summits Transl. Sci. Proc.*, vol. 2017, pp. 108–117, May 2018 (cit. on p. 17).
- [132] A. B. Das, 'Lung disease network reveals impact of comorbidity on SARS-CoV-2 infection and opportunities of drug repurposing', en, *BMC Med. Genomics*, vol. 14, no. 1, p. 226, Sep. 2021 (cit. on p. 17).

- [133] E. Dervić, K. Ledebur, S. Thurner, and P. Klimek, 'Comorbidity networks from population-wide health data: Aggregated data of 8.9m hospital patients (1997-2014)', en, *Sci. Data*, vol. 12, no. 1, p. 215, Feb. 2025 (cit. on p. 17).
- [134] T. T. T. Truong, B. Panizzutti, J. H. Kim, and K. Walder, 'Repurposing drugs via network analysis: Opportunities for psychiatric disorders', en, *Pharmaceutics*, vol. 14, no. 7, p. 1464, Jul. 2022 (cit. on p. 17).
- [135] W. Kuswanto and M. C. Baker, 'Repurposing drugs for the treatment of osteoarthritis', en, *Osteoarthritis Cartilage*, vol. 32, no. 8, pp. 886–895, Aug. 2024 (cit. on p. 17).
- [136] J. Menche, A. Sharma, M. Kitsak, S. D. Ghiassian, M. Vidal, J. Loscalzo, and A.-L. Barabási, 'Disease networks. uncovering disease-disease relationships through the incomplete interactome', en, *Science*, vol. 347, no. 6224, p. 1 257 601, Feb. 2015 (cit. on p. 17).
- [137] A. Reyes-Palomares, R. Rodríguez-López, J. A. G. Ranea, F. Sánchez-Jiménez, and M. A. Medina, 'Global analysis of the human pathophenotypic similarity gene network merges disease module components', en, *PLoS One*, vol. 8, no. 2, e56653, Feb. 2013 (cit. on p. 17).
- [138] D.-Y. Cho, Y.-A. Kim, and T. M. Przytycka, 'Chapter 5: Network biology approach to complex diseases', en, *PLoS Comput. Biol.*, vol. 8, no. 12, e1002820, Dec. 2012 (cit. on p. 17).
- [139] K. G. Becker, 'The common variants/multiple disease hypothesis of common complex genetic disorders', en, *Med. Hypotheses*, vol. 62, no. 2, pp. 309–317, 2004 (cit. on p. 17).
- [140] D. S. Gardner and E. S. Tai, 'Clinical features and treatment of maturity onset diabetes of the young (MODY)', en, *Diabetes Metab. Syndr. Obes.*, vol. 5, pp. 101–108, May 2012 (cit. on pp. 17, 18).
- [141] C. Dravet, 'Dravet syndrome history: Review', en, *Dev. Med. Child Neurol.*, vol. 53 Suppl 2, no. s2, pp. 1–6, Apr. 2011 (cit. on p. 18).
- [142] C. Dravet, 'The core dravet syndrome phenotype: Core dravet syndrome', en, *Epilepsia*, vol. 52 Suppl 2, pp. 3–9, Apr. 2011 (cit. on p. 18).
- [143] A. Sharma *et al.*, 'A disease module in the interactome explains disease heterogeneity, drug response and captures novel pathways and genes in asthma', en, *Hum Mol Genet*, vol. 24, no. 11, pp. 3005–3020, Jun. 2015 (cit. on p. 18).
- [144] J. Eder, R. Sedrani, and C. Wiesmann, 'The discovery of first-in-class drugs: Origins and evolution', en, *Nat. Rev. Drug Discov.*, vol. 13, no. 8, pp. 577–587, Aug. 2014 (cit. on p. 18).
- [145] M. L. Bolognesi and A. Cavalli, 'Multitarget drug discovery and polypharmacology', en, *ChemMedChem*, vol. 11, no. 12, pp. 1190–1192, Jun. 2016 (cit. on p. 18).
- [146] W. Sun, P. E. Sanderson, and W. Zheng, 'Drug combination therapy increases successful drug repositioning', en, *Drug Discov. Today*, vol. 21, no. 7, pp. 1189–1195, Jul. 2016 (cit. on p. 18).
- [147] Z. Tanoli *et al.*, 'Computational drug repurposing: Approaches, evaluation of in silico resources and case studies', en, *Nat. Rev. Drug Discov.*, vol. 24, no. 7, pp. 521–542, Jul. 2025 (cit. on pp. 18, 19).
- [148] J. Li, S. Zheng, B. Chen, A. J. Butte, S. J. Swamidass, and Z. Lu, 'A survey of current trends in computational drug repositioning', en, *Brief. Bioinform.*, vol. 17, no. 1, pp. 2–12, Jan. 2016 (cit. on pp. 18, 19, 29).
- [149] E. Sam and P. Athri, 'Web-based drug repurposing tools: A survey', en, *Brief. Bioinform.*, vol. 20, no. 1, pp. 299–316, Jan. 2019 (cit. on p. 19).
- [150] I. Aqeel, A. Majid, A. Albanyan, and H. Wasfi, 'Drug repurposing targeting COVID-19 3CL protease using molecular docking and machine learning regression approaches', en, *Sci. Rep.*, vol. 15, no. 1, p. 18 722, May 2025 (cit. on pp. 19, 20).
- [151] A. W. Thorman *et al.*, 'Accelerating drug discovery and repurposing by combining transcriptional signature connectivity with docking', en, *Sci. Adv.*, vol. 10, no. 35, eadj3010, Aug. 2024 (cit. on p. 19).
- [152] B. Arsić, B. S. Ilić, A. Maier, M. Hartung, J. Janjić, J. Milićević, and J. Baumbach, 'Drug repurposing for Kala-Azar', en, *Pharmaceutics*, vol. 17, no. 8, p. 1021, Aug. 2025 (cit. on p. 20).
- [153] K.-L. Tsai, S.-Y. Chang, and L.-W. Yang, 'DRDOCK: A drug repurposing platform integrating automated docking, simulations and a log-odds-based drug ranking scheme', Feb. 2021 (cit. on p. 20).
- [154] F. Wang, F.-X. Wu, C.-Z. Li, C.-Y. Jia, S.-W. Su, G.-F. Hao, and G.-F. Yang, 'ACID: A free tool for drug repurposing using consensus inverse docking strategy', en, *J. Cheminform.*, vol. 11, no. 1, p. 73, Nov. 2019 (cit. on p. 20).
- [155] C. M. Labbé, J. Rey, D. Lagorce, M. Vavruša, J. Becot, O. Sperandio, B. O. Villoutreix, P. Tufféry, and M. A. Miteva, 'MTiOpenScreen: A web server for structure-based virtual screening', en, *Nucleic Acids Res.*, vol. 43, no. W1, W448–54, Jul. 2015 (cit. on pp. 20, 21).

- [156] J.-H. Gan, J.-X. Liu, Y. Liu, S.-W. Chen, W.-T. Dai, Z.-X. Xiao, and Y. Cao, 'DrugRep: An automatic virtual screening server for drug repurposing', en, *Acta Pharmacol. Sin.*, vol. 44, no. 4, pp. 888–896, Apr. 2023 (cit. on pp. 20, 21).
- [157] R. Shukla *et al.*, 'Signature-based approaches for informed drug repurposing: Targeting CNS disorders', en, *Neuropsychopharmacology*, vol. 46, no. 1, pp. 116–130, Jan. 2021 (cit. on p. 20).
- [158] C. Yang *et al.*, 'A survey of optimal strategy for signature-based drug repositioning and an application to liver cancer', en, *Elife*, vol. 11, e71880, Feb. 2022 (cit. on p. 20).
- [159] S. A. El-Gawad, Y. Mikhael, and R. K. Arafa, 'Signature-based drug repositioning: Tackling speeding up drug discovery of anticancer drugs employing recently developed machine learning tools', en, in *Interdisciplinary Cancer Research*, Cham: Springer Nature Switzerland, 2024, pp. 271–299 (cit. on p. 20).
- [160] Q. Duan *et al.*, 'L1000CDS2: LINCS L1000 characteristic direction signatures search engine', en, *NPJ Syst. Biol. Appl.*, vol. 2, no. 1, pp. 1–12, Aug. 2016 (cit. on pp. 20, 21).
- [161] C. Andronis, A. Sharma, V. Virvilis, S. Deftereios, and A. Persidis, 'Literature mining, ontologies and information visualization for drug repurposing', en, *Brief. Bioinform.*, vol. 12, no. 4, pp. 357–368, Jul. 2011 (cit. on p. 20).
- [162] B. Chen, Y. Ding, and D. J. Wild, 'Assessing drug target association using semantic linked data', en, *PLoS Comput. Biol.*, vol. 8, no. 7, e1002574, Jul. 2012 (cit. on p. 20).
- [163] L. B. Tari and J. H. Patel, 'Systematic drug repurposing through text mining', en, *Methods Mol. Biol.*, vol. 1159, pp. 253–267, 2014 (cit. on p. 20).
- [164] D. R. Swanson, 'Fish oil, raynaud's syndrome, and undiscovered public knowledge', en, *Perspect. Biol. Med.*, vol. 30, no. 1, pp. 7–18, 1986 (cit. on p. 20).
- [165] A. Junge and L. J. Jensen, 'CoCoScore: Context-aware co-occurrence scoring for text mining applications using distant supervision', en, *Bioinformatics*, vol. 36, no. 1, pp. 264–271, Jan. 2020 (cit. on p. 20).
- [166] J. Dagdelen, A. Dunn, S. Lee, N. Walker, A. S. Rosen, G. Ceder, K. A. Persson, and A. Jain, 'Structured information extraction from scientific text with large language models', en, *Nat. Commun.*, vol. 15, no. 1, p. 1418, Feb. 2024 (cit. on p. 20).
- [167] C. Yan *et al.*, 'Leveraging generative AI to prioritize drug repurposing candidates for alzheimer's disease with real-world clinical validation', en, *NPJ Digit. Med.*, vol. 7, no. 1, p. 46, Feb. 2024 (cit. on p. 20).
- [168] Z. Ji, T. Yu, Y. Xu, N. Lee, E. Ishii, and P. Fung, 'Towards mitigating LLM hallucination via self reflection', in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds., Singapore: Association for Computational Linguistics, 2023, pp. 1827–1843 (cit. on p. 20).
- [169] K. K. Y. Ng, I. Matsuba, and P. C. Zhang, 'RAG in health care: A novel framework for improving communication and decision-making by addressing LLM limitations', en, *NEJM AI*, vol. 2, no. 1, Jan. 2025 (cit. on p. 20).
- [170] J. Wei, L. Zhuo, X. Fu, X. Zeng, L. Wang, Q. Zou, and D. Cao, 'DrugReAlign: A multisource prompt framework for drug repurposing based on large language models', en, *BMC Biol.*, vol. 22, no. 1, p. 226, Oct. 2024 (cit. on pp. 20, 21).
- [171] Z. Gao, P. Ding, and R. Xu, 'KG-Predict: A knowledge graph computational framework for drug repurposing', en, *J. Biomed. Inform.*, vol. 132, no. 104133, p. 104 133, Aug. 2022 (cit. on p. 20).
- [172] Y. Zhu, C. Che, B. Jin, N. Zhang, C. Su, and F. Wang, 'Knowledge-driven drug repurposing using a comprehensive drug knowledge graph', en, *Health Informatics J.*, vol. 26, no. 4, pp. 2737–2750, Dec. 2020 (cit. on p. 20).
- [173] P. Perdomo-Quinteiro and A. Belmonte-Hernández, 'Knowledge graphs for drug repurposing: A review of databases and methods', en, *Brief. Bioinform.*, vol. 25, no. 6, Sep. 2024 (cit. on p. 20).
- [174] B. Abu-Salih, M. Al-Qurishi, M. Alweshah, M. Al-Smadi, R. Alfayez, and H. Saadeh, 'Healthcare knowledge graph construction: A systematic review of the state-of-the-art, open issues, and opportunities', en, *J. Big Data*, vol. 10, no. 1, p. 81, May 2023 (cit. on p. 20).
- [175] I. F. do Valle, H. G. Roweth, M. W. Malloy, S. Moco, D. Barron, E. Battinelli, J. Loscalzo, and A.-L. Barabási, 'Network medicine framework shows that proximity of polyphenol targets and disease proteins predicts therapeutic effects of polyphenols', en, *Nat. Food*, vol. 2, no. 3, pp. 143–155, Mar. 2021 (cit. on p. 21).
- [176] T. C. Waller, J. A. Berg, A. Lex, B. E. Chapman, and J. Rutter, 'Compartment and hub definitions tune metabolic networks for metabolomic interpretations', en, *Gigascience*, vol. 9, no. 1, giz137, Jan. 2020 (cit. on p. 21).

- [177] S. Horvath and J. Dong, 'Geometric interpretation of gene coexpression network analysis', en, *PLoS Comput. Biol.*, vol. 4, no. 8, e1000117, Aug. 2008 (cit. on p. 21).
- [178] A. Jiménez, M. J. Merino, J. Parras, and S. Zazo, 'Explainable drug repurposing via path based knowledge graph completion', en, *Sci. Rep.*, vol. 14, no. 1, p. 16 587, Jul. 2024 (cit. on p. 21).
- [179] M. C. Cobanoglu, Z. N. Oltvai, D. L. Taylor, and I. Bahar, 'BalestraWeb: Efficient online evaluation of drug-target interactions', en, *Bioinformatics*, vol. 31, no. 1, pp. 131–133, Jan. 2015 (cit. on pp. 21, 32).
- [180] Y.-W. Chen *et al.*, 'PharmOmics: A species- and tissue-specific drug signature database and gene-network-based drug repositioning tool', en, *iScience*, vol. 25, no. 4, p. 104 052, Apr. 2022 (cit. on p. 21).
- [181] A. Singh, 'Artificial intelligence for drug repurposing against infectious diseases', en, *Artif. Intell. Chem.*, vol. 2, no. 2, p. 100 071, Dec. 2024 (cit. on p. 21).
- [182] S. Ambreen, M. Umar, A. Noor, H. Jain, and R. Ali, 'Advanced AI and ML frameworks for transforming drug discovery and optimization: With innovative insights in polypharmacology, drug repurposing, combination therapy and nanomedicine', en, *Eur. J. Med. Chem.*, vol. 284, no. 117164, p. 117 164, Feb. 2025 (cit. on p. 21).
- [183] E. Anokian, J. Bernett, A. Freeman, M. List, L. P. Santamaría, Z. Tanoli, and S. Bonnin, 'Machine learning and artificial intelligence in drug repurposing—challenges and perspectives', en, *Drug Repurposing*, vol. 1, no. 1, Jul. 2024 (cit. on p. 21).
- [184] C. Choudhury, N. Arul Murugan, and U. D. Priyakumar, 'Structure-based drug repurposing: Traditional and advanced AI/ML-aided methods', en, *Drug Discov. Today*, vol. 27, no. 7, pp. 1847–1861, Jul. 2022 (cit. on p. 21).
- [185] K. Huang, P. Chandak, Q. Wang, S. Havaladar, A. Vaid, J. Leskovec, G. N. Nadkarni, B. S. Glicksberg, N. Gehlenborg, and M. Zitnik, 'A foundation model for clinician-centered drug repurposing', en, *Nat. Med.*, vol. 30, no. 12, pp. 3601–3613, Dec. 2024 (cit. on pp. 21, 22).
- [186] P. Pham, V. T. D. Nguyen, K. H. Cho, and T.-S. Hy, 'DrugPipe: Generative artificial intelligence-assisted virtual screening pipeline for generalizable and efficient drug repurposing', en, *Biol. Methods Protoc.*, vol. 10, no. 1, bpaf038, May 2025 (cit. on p. 21).
- [187] A. Bessadok and F. Grisoni, 'An explainable foundation model for drug repurposing', en, *Nat. Med.*, vol. 30, no. 12, pp. 3422–3423, Dec. 2024 (cit. on p. 21).
- [188] A. Ianevski, A. Kushnir, K. Nader, M. Miihkinen, H. Xhaard, T. Aittokallio, and Z. Tanoli, 'Repurpose-Drugs: An interactive web-portal and predictive platform for repurposing mono- and combination therapies', en, *Brief. Bioinform.*, vol. 25, no. 4, bbae328, May 2024 (cit. on p. 21).
- [189] W. Li, W. Ma, M. Yang, and X. Tang, 'Drug repurposing based on the DTD-GNN graph neural network: Revealing the relationships among drugs, targets and diseases', en, *BMC Genomics*, vol. 25, no. 1, p. 584, Jun. 2024 (cit. on p. 21).
- [190] F. Firoozbakht, S. Süwer, M. L. Elkjaer, D. E. Handy, A. Maier, J. Li, L. Lancashire, J. Loscalzo, and J. Baumbach, 'NetMedGPT - a network medicine foundation model for extensive disease mechanism mining and drug repurposing', Jan. 2026 (cit. on pp. 21, 119).
- [191] B. Wei, Y. Zhang, and X. Gong, 'DeepLPI: A novel deep learning-based model for protein-ligand interaction prediction for drug repurposing', en, *Sci. Rep.*, vol. 12, no. 1, p. 18 200, Oct. 2022 (cit. on p. 22).
- [192] O. Lazareva, J. Baumbach, M. List, and D. B. Blumenthal, 'On the limits of active module identification', en, *Brief. Bioinform.*, vol. 22, no. 5, Sep. 2021 (cit. on pp. 23, 25, 107).
- [193] Z. Piper, Z. M. Maamdouh, M. Pacheco Pachado, R. A. Reijnders, E. Pishva, and H. H. H. W. Schmidt, 'From genomic puzzles to network solutions: Charting parkinson's path using systematic review/ meta analysis as the compass', in *From Genomic Puzzles to Network Solutions: Charting Parkinson's Path using Systematic Review/ Meta Analysis as the Compass*, REPO4EU, 2023 (cit. on p. 24).
- [194] M. I. Love, W. Huber, and S. Anders, 'Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2', en, *Genome Biol.*, vol. 15, no. 12, p. 550, 2014 (cit. on pp. 24, 62).
- [195] S. D. Ghiassian, J. Menche, and A.-L. Barabási, 'A DisEASE MOdule detection (DIAMOnD) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome', en, *PLoS Comput. Biol.*, vol. 11, no. 4, e1004120, Apr. 2015 (cit. on pp. 25, 26, 32).
- [196] H. Levi, R. Elkon, and R. Shamir, 'DOMINO: A network-based active module identification algorithm with reduced rate of false calls', en, *Mol. Syst. Biol.*, vol. 17, no. 1, e9593, Jan. 2021 (cit. on pp. 25, 26).
- [197] J. Matschinske, M. Salgado-Albarrán, S. Sadegh, D. Bongiovanni, J. Baumbach, and D. B. Blumenthal, 'Individuating possibly repurposable drugs and drug targets for COVID-19 treatment through Hypothesis-Driven systems medicine using CoVex', en, *Assay Drug Dev. Technol.*, vol. 18, no. 8, pp. 348–355, Nov. 2020 (cit. on pp. 25, 26, 33, 38, 39).

- [198] J. Bennett, D. Krupke, S. Sadegh, J. Baumbach, S. P. Fekete, T. Kacprowski, M. List, and D. B. Blumenthal, 'Robust disease module mining via enumeration of diverse prize-collecting steiner trees', en, *Bioinformatics*, Jan. 2022 (cit. on pp. 25, 26, 32).
- [199] S. Sarkar, M. Lucchetta, A. Maier, M. M. Abdrabbou, J. Baumbach, M. List, M. H. Schaefer, and D. B. Blumenthal, 'Online bias-aware disease module mining with ROBUST-Web', en, *Bioinformatics*, May 2023 (cit. on pp. 25, 26, 33, 113).
- [200] D. Buzzao, M. Castresana-Aguirre, D. Guala, and E. L. L. Sonnhhammer, 'TOPAS, a network-based approach to detect disease modules in a top-down fashion', en, *NAR Genom. Bioinform.*, vol. 4, no. 4, lqac093, Dec. 2022 (cit. on pp. 25, 26).
- [201] Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen, 'Combating web spam with trustrank', in *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*, 2004, pp. 576–587 (cit. on pp. 25, 26).
- [202] L. Page, S. Brin, R. Motwani, and T. Winograd, *The PageRank citation ranking: Bringing order to the web*, <http://ilpubs.stanford.edu:8090/422/>, Accessed: 2025-9-3, Nov. 1999 (cit. on pp. 25, 27, 28).
- [203] 'Closeness centrality extended to unconnected graphs: The harmonic centrality index', *Closeness Centrality Extended to Unconnected Graphs: the Harmonic Centrality Index*, (cit. on pp. 25, 26, 27, 28).
- [204] G. Sabidussi, 'The centrality of a graph', en, *Psychometrika*, vol. 31, no. 4, pp. 581–603, Dec. 1966 (cit. on pp. 25, 26, 28).
- [205] L. C. Freeman, 'A set of measures of centrality based on betweenness', *Sociometry*, vol. 40, no. 1, p. 35, Mar. 1977 (cit. on pp. 25, 26).
- [206] K. Mechteridis, M. Lauber, J. Baumbach, and M. List, 'KeyPathwayMineR: De novo pathway enrichment in the R ecosystem', en, *Front. Genet.*, vol. 12, p. 812 853, 2021 (cit. on pp. 25, 26, 32).
- [207] O. Lazareva, S. Canzar, K. Yuan, J. Baumbach, D. B. Blumenthal, P. Tieri, T. Kacprowski, and M. List, 'BiCoN: Network-constrained biclustering of patients and omics data', en, *Bioinformatics*, Dec. 2020 (cit. on pp. 26, 27, 32, 39).
- [208] B. H. Junker, D. Koschützki, and F. Schreiber, 'Exploration of biological network centralities with CentiBiN', en, *BMC Bioinformatics*, vol. 7, no. 1, p. 219, Apr. 2006 (cit. on p. 27).
- [209] M. Girvan and M. E. J. Newman, 'Community structure in social and biological networks', en, *Proc. Natl. Acad. Sci. U. S. A.*, vol. 99, no. 12, pp. 7821–7826, Jun. 2002 (cit. on pp. 27, 28).
- [210] L. C. Freeman, 'Centrality in social networks conceptual clarification', en, *Soc. Networks*, vol. 1, no. 3, pp. 215–239, Jan. 1978 (cit. on pp. 27, 28).
- [211] X. Tang, J. Wang, J. Zhong, and Y. Pan, 'Predicting essential proteins based on weighted degree centrality', en, *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 11, no. 2, pp. 407–418, Mar. 2014 (cit. on p. 27).
- [212] A. Valdeolivas, L. Tichit, C. Navarro, S. Perrin, G. Odelin, N. Levy, P. Cau, E. Remy, and A. Baudot, 'Random walk with restart on multiplex and heterogeneous biological networks', en, *Bioinformatics*, vol. 35, no. 3, pp. 497–505, Feb. 2019 (cit. on pp. 27, 28, 63).
- [213] L. Sun, Z. Yin, and L. Lu, 'ISLRWR: A network diffusion algorithm for drug-target interactions prediction', en, *PLoS One*, vol. 20, no. 1, e0302281, Jan. 2025 (cit. on p. 27).
- [214] Y. Wang, M. Guo, Y. Ren, L. Jia, and G. Yu, 'Drug repositioning based on individual bi-random walks on a heterogeneous network', en, *BMC Bioinformatics*, vol. 20, no. Suppl 15, p. 547, Dec. 2019 (cit. on p. 27).
- [215] E. Guney, J. Menche, M. Vidal, and A.-L. Barábasi, 'Network-based in silico drug efficacy screening', en, *Nat. Commun.*, vol. 7, no. 1, p. 10331, Feb. 2016 (cit. on pp. 27, 28).
- [216] F. Cheng, R. J. Desai, D. E. Handy, R. Wang, S. Schneeweiss, A.-L. Barabási, and J. Loscalzo, 'Network-based approach to prediction and population-based validation of in silico drug repurposing', en, *Nat. Commun.*, vol. 9, no. 1, p. 2691, Jul. 2018 (cit. on p. 27).
- [217] G. Fiscon, F. Conte, L. Farina, and P. Paci, 'SAveRUNNER: A network-based algorithm for drug repurposing and its application to COVID-19', en, *PLoS Comput. Biol.*, vol. 17, no. 2, e1008686, Feb. 2021 (cit. on pp. 27, 28).
- [218] E. Ulgen, O. Ozisik, and O. U. Sezerman, 'PANACEA: Network-based methods for pharmacotherapy prioritization in personalized oncology', en, *Bioinformatics*, vol. 39, no. 1, btad022, Jan. 2023 (cit. on p. 28).
- [219] X. Sun, X. Jia, Z. Lu, J. Tang, and M. Li, 'Drug repositioning with adaptive graph convolutional networks', en, *Bioinformatics*, vol. 40, no. 1, btad748, Jan. 2024 (cit. on p. 28).

- [220] T. J. Rintala, A. Ghosh, and V. Fortino, 'Network approaches for modeling the effect of drugs and diseases', en, *Brief. Bioinform.*, vol. 23, no. 4, bbac229, Jul. 2022 (cit. on p. 28).
- [221] X. Shen, S. Pan, K.-S. Choi, and X. Zhou, 'Domain-adaptive message passing graph neural network', en, *Neural Netw.*, vol. 164, pp. 439–454, Jul. 2023 (cit. on p. 28).
- [222] W. Ju *et al.*, 'A comprehensive survey on deep graph representation learning', en, *Neural Netw.*, vol. 173, no. 106207, p. 106 207, May 2024 (cit. on p. 28).
- [223] A. S. Brown and C. J. Patel, 'A review of validation strategies for computational drug repositioning', en, *Brief. Bioinform.*, vol. 19, no. 1, pp. 174–177, Jan. 2018 (cit. on p. 29).
- [224] M. Pillai and D. Wu, 'Validation approaches for computational drug repurposing: A review', en, *AMIA Annu. Symp. Proc.*, vol. 2023, pp. 559–568, 2023 (cit. on p. 29).
- [225] A. G. Hoed, L. Prieto-Santamaría, and A. Rodríguez-González, 'Evaluating the influence of Disease-Gene associations in the significance of disease modules through the lens of network medicine', Jun. 2025 (cit. on p. 29).
- [226] H. Xu *et al.*, 'Validating drug repurposing signals using electronic health records: A case study of metformin associated with reduced cancer mortality', en, *J. Am. Med. Inform. Assoc.*, vol. 22, no. 1, pp. 179–191, Jan. 2015 (cit. on p. 29).
- [227] A.-L. Barabási, N. Gulbahce, and J. Loscalzo, 'Network medicine: A network-based approach to human disease', en, *Nat. Rev. Genet.*, vol. 12, no. 1, pp. 56–68, Jan. 2011 (cit. on pp. 30, 107).
- [228] J. Wang, G.-Y. Luo, T. Tian, Y.-Q. Zhao, S.-Y. Meng, J.-H. Wu, W.-X. Han, B. Deng, and J. Ni, 'Shared genetic basis and causality between schizophrenia and inflammatory bowel disease: Evidence from a comprehensive genetic analysis', en, *Psychol. Med.*, vol. 54, no. 10, pp. 2658–2668, Jul. 2024 (cit. on p. 30).
- [229] D. Ellinghaus *et al.*, 'Analysis of five chronic inflammatory diseases identifies 27 new associations and highlights disease-specific patterns at shared loci', en, *Nat. Genet.*, vol. 48, no. 5, pp. 510–518, May 2016 (cit. on p. 30).
- [230] P. Zhang, F. Wang, J. Hu, and R. Sorrentino, 'Exploring the relationship between drug side-effects and therapeutic indications', en, *AMIA Annu. Symp. Proc.*, vol. 2013, pp. 1568–1577, Nov. 2013 (cit. on p. 30).
- [231] F. Firoozbakht, M. L. Elkjaer, D. E. Handy, R.-S. Wang, Z. Chervontseva, M. Rarey, J. Loscalzo, J. Baumbach, and O. Tsoy, 'DREAMER: Exploring common mechanisms of adverse drug reactions and disease phenotypes through Network-Based analysis', Jul. 2024 (cit. on pp. 30, 108).
- [232] M. Lucchetta and M. Pellegrini, 'Finding disease modules for cancer and COVID-19 in gene co-expression networks with the Core&Peel method', en, *Sci. Rep.*, vol. 10, no. 1, p. 17 628, Oct. 2020 (cit. on p. 30).
- [233] M. Ashburner *et al.*, 'Gene ontology: Tool for the unification of biology. the gene ontology consortium', en, *Nat. Genet.*, vol. 25, no. 1, pp. 25–29, May 2000 (cit. on p. 30).
- [234] U. Raudvere, L. Kolberg, I. Kuzmin, T. Arak, P. Adler, H. Peterson, and J. Vilo, 'G:profiler: A web server for functional enrichment analysis and conversions of gene lists (2019 update)', en, *Nucleic Acids Res.*, vol. 47, no. W1, W191–W198, Jul. 2019 (cit. on p. 30).
- [235] D. W. Huang, B. T. Sherman, and R. A. Lempicki, 'Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources', en, *Nat. Protoc.*, vol. 4, no. 1, pp. 44–57, 2009 (cit. on p. 30).
- [236] D. Hanahan and R. A. Weinberg, 'Hallmarks of cancer: The next generation', en, *Cell*, vol. 144, no. 5, pp. 646–674, Mar. 2011 (cit. on pp. 30, 86).
- [237] A. Subramanian *et al.*, 'Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles', en, *Proc. Natl. Acad. Sci. U. S. A.*, vol. 102, no. 43, pp. 15 545–15 550, Oct. 2005 (cit. on p. 30).
- [238] M. Gustafsson *et al.*, 'Modules, networks and systems medicine for understanding disease and aiding diagnosis', en, *Genome Med.*, vol. 6, no. 10, p. 82, Oct. 2014 (cit. on p. 30).
- [239] A. R. Sonawane, J. Platig, M. Fagny, C.-Y. Chen, J. N. Paulson, C. M. Lopes-Ramos, D. L. DeMeo, J. Quackenbush, K. Glass, and M. L. Kuijjer, 'Understanding tissue-specific gene regulation', Feb. 2017 (cit. on p. 30).
- [240] N. Sahni *et al.*, 'Widespread macromolecular interaction perturbations in human genetic disorders', en, *Cell*, vol. 161, no. 3, pp. 647–660, Apr. 2015 (cit. on p. 30).
- [241] P. Sibilio *et al.*, 'In silico drug repurposing in COVID-19: A network-based analysis', en, *Biomed. Pharmacother.*, vol. 142, no. 111954, p. 111 954, Oct. 2021 (cit. on p. 31).
- [242] H. Levi, N. Rahmanian, R. Elkon, and R. Shamir, 'The DOMINO web-server for active module identification analysis', en, *Bioinformatics*, vol. 38, no. 8, pp. 2364–2366, Apr. 2022 (cit. on pp. 32, 39).

- [243] J. von Eichborn, M. S. Murgueitio, M. Dunkel, S. Koerner, P. E. Bourne, and R. Preissner, 'PROMISCUOUS: A database for network-based drug-repositioning', en, *Nucleic Acids Res.*, vol. 39, no. Database issue, pp. D1060–6, Jan. 2011 (cit. on p. 32).
- [244] K. Gallo, A. Goede, A. Eckert, B. Moahamed, R. Preissner, and B.-O. Gohlke, 'PROMISCUOUS 2.0: A resource for drug-repositioning', en, *Nucleic Acids Res.*, vol. 49, no. D1, pp. D1373–D1380, Jan. 2021 (cit. on pp. 32, 38).
- [245] S. Alaimo, V. Bonnici, D. Cancemi, A. Ferro, R. Giugno, and A. Pulvirenti, 'DT-Web: A web-based application for drug-target interaction and drug combination prediction through domain-tuned network-based inference', en, *BMC Syst. Biol.*, vol. 9 Suppl 3, no. S3, S4, Jun. 2015 (cit. on p. 33).
- [246] W. Ba-Alawi, O. Soufan, M. Essack, P. Kalnis, and V. B. Bajic, 'DASPfind: New efficient method to predict drug-target interactions', en, *J. Cheminform.*, vol. 8, no. 1, p. 15, Mar. 2016 (cit. on p. 33).
- [247] M. Hartung, E. Anastasi, Z. M. Mamdouh, C. Nogales, H. H. H. W. Schmidt, J. Baumbach, O. Zolotareva, and M. List, 'Cancer driver drug interaction explorer', en, *Nucleic Acids Res.*, vol. 50, no. W1, W138–W144, Jul. 2022 (cit. on pp. 33, 39).
- [248] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, 'Cytoscape: A software environment for integrated models of biomolecular interaction networks', en, *Genome Res.*, vol. 13, no. 11, pp. 2498–2504, Nov. 2003 (cit. on p. 33).
- [249] T. Doğan *et al.*, 'CROssBAR: Comprehensive resource of biomedical relations with knowledge graph representations', en, *Nucleic Acids Res.*, vol. 49, no. 16, e96, Sep. 2021 (cit. on p. 38).
- [250] A. Maier *et al.*, 'NeDRex-web: An interactive web tool for drug repurposing by exploring heterogeneous molecular networks', in *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Lisbon, Portugal: IEEE, Dec. 2024, pp. 429–434 (cit. on pp. 40, 41).
- [251] A. Maier *et al.*, 'Drugst.One - a plug-and-play solution for online systems medicine and network-based drug repurposing', en, *Nucleic Acids Res.*, vol. 52, no. W1, W481–W488, Jul. 2024 (cit. on pp. 40, 50).
- [252] S. A. Forbes *et al.*, 'COSMIC: Mining complete cancer genomes in the catalogue of somatic mutations in cancer', en, *Nucleic Acids Res.*, vol. 39, no. Database issue, pp. D945–50, Jan. 2011 (cit. on p. 61).
- [253] F. Martínez-Jiménez *et al.*, 'A compendium of mutational cancer driver genes', en, *Nat. Rev. Cancer*, vol. 20, no. 10, pp. 555–572, Oct. 2020 (cit. on p. 61).
- [254] L. Dressler *et al.*, 'Comparative assessment of genes driving cancer and somatic evolution in non-cancer tissues: An update of the network of cancer genes (NCG) resource', en, *Genome Biol.*, vol. 23, no. 1, p. 35, Jan. 2022 (cit. on p. 61).
- [255] P.-F. Chen *et al.*, 'Automatic international classification of diseases coding system: Deep contextualized language model with rule-based approaches', en, *JMIR Med. Inform.*, vol. 10, no. 6, e37557, Jun. 2022 (cit. on p. 61).
- [256] W. T. Garvey and J. I. Mechanick, 'Proposal for a scientifically correct and medically actionable disease classification system (ICD) for obesity', en, *Obesity (Silver Spring)*, vol. 28, no. 3, pp. 484–492, Mar. 2020 (cit. on p. 61).
- [257] M. A. Gargano *et al.*, 'The human phenotype ontology in 2024: Phenotypes around the world', en, *Nucleic Acids Res.*, vol. 52, no. D1, pp. D1333–D1346, Jan. 2024 (cit. on p. 61).
- [258] S. Anders and W. Huber, 'Differential expression analysis for sequence count data', en, *Genome Biol.*, vol. 11, no. 10, R106, Oct. 2010 (cit. on p. 62).
- [259] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth, 'Limma powers differential expression analyses for RNA-sequencing and microarray studies', en, *Nucleic Acids Res.*, vol. 43, no. 7, e47, Apr. 2015 (cit. on p. 62).
- [260] E. Masson *et al.*, 'Expanding ACMG variant classification guidelines into a general framework', en, *Hum. Genomics*, vol. 16, no. 1, p. 31, Aug. 2022 (cit. on p. 62).
- [261] K. M. T. H. Rahit and M. Tarailo-Graovac, 'Genetic modifiers and rare mendelian disease', en, *Genes (Basel)*, vol. 11, no. 3, E239, Feb. 2020 (cit. on p. 62).
- [262] L. Phan, H. Zhang, Q. Wang, R. Villamarin, T. Hefferon, A. Ramanathan, and B. Kattman, 'The evolution of dbSNP: 25 years of impact in genomic research', en, *Nucleic Acids Res.*, vol. 53, no. D1, pp. D925–D931, Jan. 2025 (cit. on p. 63).
- [263] 1000 Genomes Project Consortium *et al.*, 'A global reference for human genetic variation', en, *Nature*, vol. 526, no. 7571, pp. 68–74, Oct. 2015 (cit. on p. 63).
- [264] G. Caylak, O. Tastan, and A. E. Cicek, 'Potpourri: An epistasis test prioritization algorithm via diverse SNP selection', en, *J. Comput. Biol.*, vol. 28, no. 4, pp. 365–377, Apr. 2021 (cit. on p. 63).

- [265] T. Cowman and M. Koyutürk, 'Prioritizing tests of epistasis through hierarchical representation of genomic redundancies', en, *Nucleic Acids Res.*, vol. 45, no. 14, e131, Aug. 2017 (cit. on p. 63).
- [266] P.-J. Jing and H.-B. Shen, 'MACOED: A multi-objective ant colony optimization algorithm for SNP epistasis detection in genome-wide association studies', en, *Bioinformatics*, vol. 31, no. 5, pp. 634–641, Mar. 2015 (cit. on p. 63).
- [267] D. Duroux, H. Climente-González, C.-A. Azencott, and K. Van Steen, 'Interpretable network-guided epistasis detection', en, *Gigascience*, vol. 11, Feb. 2022 (cit. on p. 63).
- [268] M. Ayati and M. Koyutürk, 'Prioritization of genomic locus pairs for testing epistasis', in *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, Newport Beach California: ACM, Sep. 2014 (cit. on p. 63).
- [269] M. Graça, R. Nobre, L. Sousa, and A. Ilic, 'Distributed transformer for high order epistasis detection in large-scale datasets', en, *Sci. Rep.*, vol. 14, no. 1, p. 14579, Jun. 2024 (cit. on p. 63).
- [270] B. T. Sherman, M. Hao, J. Qiu, X. Jiao, M. W. Baseler, H. C. Lane, T. Imamichi, and W. Chang, 'DAVID: A web server for functional enrichment analysis and functional annotation of gene lists (2021 update)', en, *Nucleic Acids Res.*, vol. 50, no. W1, W216–W221, Jul. 2022 (cit. on p. 85).
- [271] J. Reimand, M. Kull, H. Peterson, J. Hansen, and J. Vilo, 'G:profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments', en, *Nucleic Acids Res.*, vol. 35, no. Web Server issue, W193–200, Jul. 2007 (cit. on p. 85).
- [272] Y. Liao, J. Wang, E. J. Jaehnig, Z. Shi, and B. Zhang, 'WebGestalt 2019: Gene set analysis toolkit with revamped UIs and APIs', en, *Nucleic Acids Res.*, vol. 47, no. W1, W199–W205, Jul. 2019 (cit. on p. 85).
- [273] K. Adamowicz, A. Maier, J. Baumbach, and D. B. Blumenthal, 'Online in silico validation of disease and gene sets, clusterings or subnetworks with DIGEST', en, *Brief. Bioinform.*, vol. 23, no. 4, Jul. 2022 (cit. on pp. 85, 87).
- [274] P. Marques, Z. Dabbabi, M.-M. Mironescu, O. Thonnard, F. Buontempo, I. Gashi, and A. Bessani, 'Using diverse detectors for detecting malicious web scraping activity', in *2018 48th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W)*, Luxembourg, Luxembourg: IEEE, Jun. 2018, pp. 67–68 (cit. on p. 104).
- [275] X. Li, B. A. Azad, A. Rahmati, and N. Nikiforakis, 'Good bot, bad bot: Characterizing automated browsing activity', in *2021 IEEE Symposium on Security and Privacy (SP)*, San Francisco, CA, USA: IEEE, May 2021, pp. 1589–1605 (cit. on p. 104).
- [276] R. Van Boxem, T. Barbette, C. Pelsser, and R. Sadre, 'Shy guys: A light-weight approach to detecting robots on websites', Mar. 2026. arXiv: 2603.28546 [cs.NI] (cit. on p. 105).
- [277] *AI-Driven bots surge in september 2025*, en, <https://www.equimedia.co.uk/resources/blog/ai-driven-bots-surge-september-2025>, Accessed: 2026-5-4 (cit. on pp. 106, 114).
- [278] M. Oubounyt, L. Adlung, F. Patroni, N. Wenke, A. Maier, M. Hartung, J. Baumbach, and M. Elkjaer, 'Inference of differential key regulatory networks and mechanistic drug repurposing candidates from scRNA-seq data with SCANet', *Bioinformatics*, vol. 39, no. 11, btad644, Oct. 2023 (cit. on pp. 108, 113).
- [279] K. Adamowicz, L. Arend, A. Maier, J. R. Schmidt, B. Kuster, O. Tsoy, O. Zolotareva, J. Baumbach, and T. Laske, 'Proteomic meta-study harmonization, mechanotyping and drug repurposing candidate prediction with ProHarMeD', en, *NPJ Syst Biol Appl*, vol. 9, no. 1, p. 49, Oct. 2023 (cit. on p. 109).
- [280] M. Kotlyar, C. Pastrello, M. Abovsky, A. Mizeranschi, A. Keating, L.-C. Cameron, V. Chandran, and I. Jurisica, 'IID 2025: Physical protein interaction data with detection types, co-purified protein sets, molecular docking, and immune cell networks', en, *Nucleic Acids Res.*, vol. 54, no. D1, pp. D593–D600, Jan. 2026 (cit. on p. 109).
- [281] L. M. Spindler, J. Kersting, Q. Manz, M. Hartung, A. Maier, Z. M. Mamdouh, A. I. Casas, J. Baumbach, and M. List, 'Drugst.One DREAM—Drug repurposing through expert annotation and modification', en, *Br. J. Pharmacol.*, no. bph.70495, May 2026 (cit. on p. 109).
- [282] H. Thorvaldsdóttir, J. T. Robinson, and J. P. Mesirov, 'Integrative genomics viewer (IGV): High-performance genomics data visualization and exploration', en, *Brief. Bioinform.*, vol. 14, no. 2, pp. 178–192, Mar. 2013 (cit. on p. 117).
- [283] A. Maier, K. Adamowicz, M. Le, M. Chow Castro, K. Kemnitz, O. Zolotareva, D. Blumenthal, and J. Baumbach, 'Integrated tools for enhancing network-based disease module identification, drug prioritization and repurposing', in *Integrated Tools for Enhancing Network-based Disease Module Identification, Drug Prioritization and Repurposing*, REPO4EU, Sep. 2023 (cit. on p. 118).
- [284] S. Süwer, K. Bagemihl, S. Baier, L. Dicunta, M. List, J. Baumbach, A. Maier, and F. M. Delgado-Chaves, 'Conversational no-code, multi-agentic disease module identification and drug repurposing prediction with ChatDRex', Feb. 2026. arXiv: 2511.21438 [cs.AI] (cit. on p. 119).

- [285] F. M. Delgado-Chaves *et al.*, 'DrugRepoChatter: A drug repurposing expert chatbot curated by the REPO4EU consortium', Jul. 2024 (cit. on p. [119](#)).

APPENDIX

A.1 THESIS SUPPLEMENTARY MATERIAL

Table A.1: Updated list of tools verifiably integrating Drugst.One (in alphabetical order). Excluded are four tools that are not publicly available or not published and one page that used it in the past, but removed it since.

Tool/Page	URL	Tool Description	Integration	Status
2D-molecular-visualization	https://2d-molecular-visualization.streamlit.app/	Collection of various online 2D visualization tools	Plugin	Done
BiCoN	https://exbio.wzw.tum.de/bicon/	Network-constrained patient stratification through biclustering	Plugin	Done
ChatDRex	https://apps.cosy.bio/chatdrex/	LLM-driven, multi-agent system interfacing with NeDRexAPI and other sources for network-based drug repurposing	Plugin	Done
DOMINO	http://domino.cs.tau.ac.il/	Active module identification with improved empirical validation	Link-out	Done
G-Browser	https://exbio.wzw.tum.de/genome-browser/	An enhanced genome browser plugin that seamlessly integrates data sources and functions for genetics research.	Plugin	Done
GraphFusion	https://github.com/CarlosJesusGH/GraphFusion	An intuitive web-based graph analytics, fusion, and visualization tool	Plugin	Done
GraphSimViz	https://graphsimeviz.net/	Visualization of diseasomes, drugomes, and drug-disease networks	Plugin	Done
HitSeekR	https://exbio.wzw.tum.de/hitseekr/	User-friendly tool for drug (target) identification in high-throughput screening	Plugin	Done
Interactive Enrichment Analysis	https://github.com/gladstone-institutes/Interactive-Enrichment-Analysis/	Enrichment analysis on multiple public datasets	Link-out	Done

Continued on next page...

Table A.1 continued from previous page

Tool/Page	URL	Tool Description	Integration	Status
IID - Molecular Docking	https://iid.ophid.utoronto.ca/	Large database integrating all sorts of protein interaction information and resources	Plugin	Done
mirDIP	https://ophid.utoronto.ca/mirDIP/	Integrated microRNA-target data integration portal	Plugin	Done
MetalinksDB	https://metalinks.omnipathdb.org/	Comprehensive database of intercellular metabolite-protein interactions	Plugin	Done
NAViGaTOR	https://ophid.utoronto.ca/navigator/	Network visualization and analysis software	Link-out	Done
NDEx IQuery	https://www.ndexbio.org/iquery/	Web tool for pathway and network-based gene set analysis	Plugin	Planned
NeEDL - Epistasis Disease Atlas	https://epistasis-disease-atlas.com	Web resource to visualize, investigate, and interpret higher-order genetic interactions of single nucleotide polymorphisms in 18 human heritable diseases.	Link-out	Done
NeEDL - R Shiny App	https://hub.docker.com/r/bigdatainbiomedicine/needl	R shiny app to visualize, investigate, and interpret higher-order genetic interactions of single nucleotide polymorphisms on locally computed datasets.	Plugin	Done
Nephro DIP	https://ophid.utoronto.ca/NephroDIP/	Database for nephrology-related tissue expression information	Plugin	Done
Olfaction	https://olfaction.gnext.gm.eurac.edu/	Exploration of GWAS on human olfactory identification summary statistics	Plugin	Done
Pan-UK-Biobank	https://panukbb-eur.gnext.gm.eurac.edu/	Exploration of GWAS Pan UK Biobank focusing on ancestry-enriched effects	Plugin	Done
openPIP	https://github.com/BaderLab/openPIP	Open platform to store and retrieve protein-protein interaction datasets.	Link-out	Done
pathDIP	https://ophid.utoronto.ca/pathDIP	Integrated pathway database and pathway enrichment analysis portal	Plugin	Done
Pathway Figure OCR	https://pfocr.wikipathways.org	Platform for browsing pathway information extracted from published figures.	Link-out	Done

Continued on next page...

Table A.1 continued from previous page

Tool/Page	URL	Tool Description	Integration	Status
ProHarMeD	https://apps.cosy.bio/proharmed	Closing the gap between (harmonized) proteomics results and mechanotyping / drug repurposing	Plugin	Done
ProteomicsDB	https://www.proteomicsdb.org/	Multi-omics and -organism explorer	Plugin	Done
REPO4EU	https://repo4.eu/	Official project page of the Horizon Europe project REPO4EU [128]	Plugin	Done
ROBUST-Web	https://robust-web.net/	ROBUST is a disease module identification tool.	Plugin	Done
SCANet	https://pypi.org/project/scanet/	SCANet is an all-in-one package for single-cell profiling covering the whole differential mechanotyping workflow, from inference of trait/cell-type-specific gene co-expression modules to mechanistic drug repurposing candidate prediction.	Python package	Done
Seed Connector Algorithm	https://github.com/bwh784/SCA	Identification of network modules by adding a minimal number of edges between the seed nodes.	Link-out	Done
UnPaSt	https://apps.cosy.bio/unpast	Visualizer and context explorer for unsupervised expression data bicluster results.	Plugin	Done
WikiPathways	https://wikipathways.org	Platform for browsing and visualizing pathways.	Link-out	Done

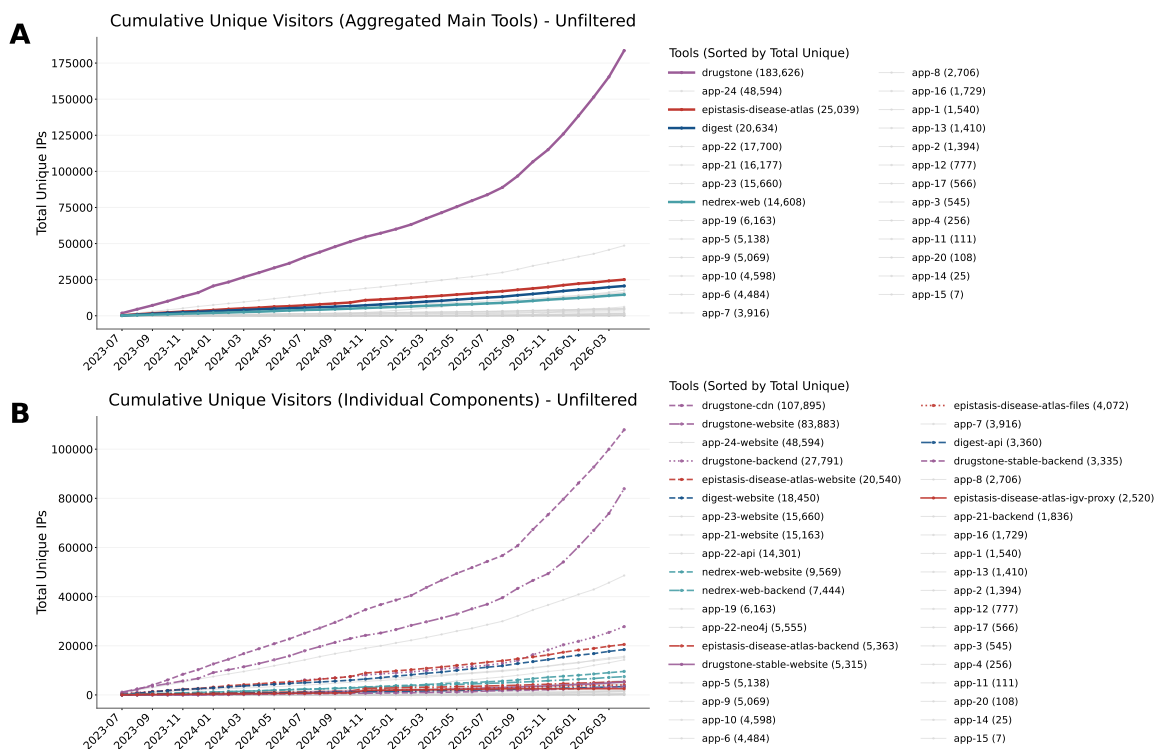


Figure A.1: Web Tool Reach: Cumulative Unique Interactions - Unfiltered: This figure was created equivalently to [Figure 5.1](#), but the completely unfiltered server logs were used as source of the statistics.

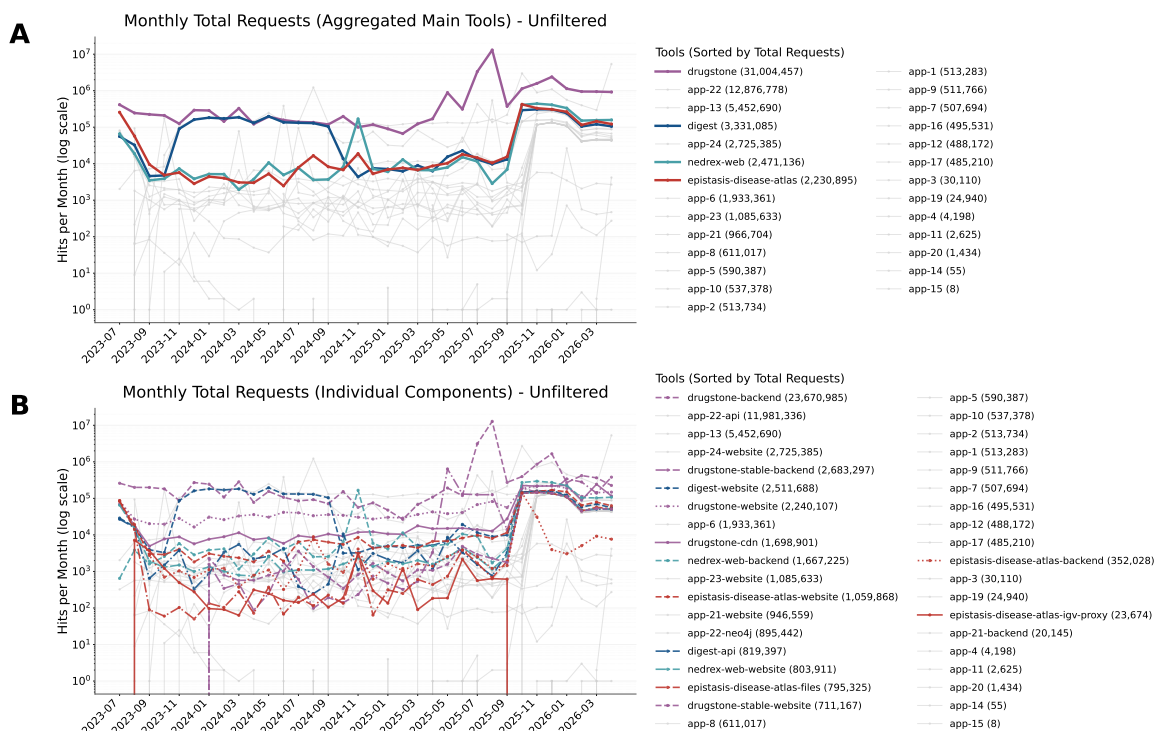


Figure A.2: Web Tool Usage: Monthly Total Requests (Hits) - Unfiltered: This figure was created equivalently to [Figure 5.2](#), but the completely unfiltered server logs were used as source of the statistics.

A.2 PUBLICATION RELATED MATERIAL

A.2.0.1 Publication Permissions

IEEE Republication Permission Statement

Publication A: NeDRex-Web

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

Requirements to be followed when using any portion of an IEEE copyrighted paper:

1. In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © [Year] IEEE.
2. In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
3. If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:

1. The following IEEE copyright/credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication].
2. Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
3. In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of the University of Hamburg's products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to [RightsLink](#) to learn how to obtain a License.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

Republishing License Information

Publication B: Drugst.One

This Agreement between Andreas Maier, Institute for Computational Systems Biomedicine ("You") and Oxford University Press consists of your license details and the terms and conditions provided by Oxford University Press and Copyright Clearance Center.

License Number	6266100193507
License Date	May 11, 2026
Publisher	Oxford University Press
Publication	<i>Nucleic Acids Research</i>
Content Title	Drugst.One — a plug-and-play solution for online systems medicine and network-based drug repurposing
Content Authors	Maier, Andreas; Hartung, Michael
Content Date	May 23, 2024
Type of Use	Thesis/Dissertation

Republication License Information**Publication C: NeEDL & EDA**

This Agreement between Andreas Maier ("You") and Oxford University Press consists of your license details and the terms and conditions provided by Oxford University Press and Copyright Clearance Center.

License Number	6266101039256
License Date	May 11, 2026
Publisher	Oxford University Press
Publication	<i>Nucleic Acids Research</i>
Content Title	Network medicine-based epistasis detection in complex diseases: ready for quantum computing
Content Authors	Hoffmann, Markus; Poschenrieder, Julian M
Content Date	Aug 23, 2024
Type of Use	Thesis/Dissertation

Republication License Information**Publication D: DIGEST**

This Agreement between Andreas Maier ("You") and Oxford University Press consists of your license details and the terms and conditions provided by Oxford University Press and Copyright Clearance Center.

License Number	6266111360550
License Date	May 11, 2026
Publisher	Oxford University Press
Publication	<i>Briefings in Bioinformatics</i>
Content Title	Online <i>in silico</i> validation of disease and gene sets, clusterings or subnetworks with DIGEST
Content Authors	Adamowicz, Klaudia; Maier, Andreas
Content Date	Jun 27, 2022
Type of Use	Thesis/Dissertation
Title of Your Work	Computational Democratization of Network Medicine for Drug Repurposing

A.2.0.2 Figure Attribution Statement**Figure Attribution Statement**

The schematic diagrams presented in [Figure 1.1](#), [Figure 1.2](#), and [Figure 1.3](#) of this dissertation were created using BioRender.com.

A.2.1 Authors' Contribution Statements for Supporting Publications

This section details my individual contributions to each co-authored work listed as a supporting publication in the beginning of this dissertation (section 1).

[a] Proteomic meta-study harmonization, mechanotyping and drug repurposing candidate prediction with ProHarMeD (2023)

For ProHarMeD, my primary contribution consisted of engineering and implementing the web application¹, thereby making the tool accessible via standard web browsers. I handled the deployment, continue to maintain the platform, and contributed to writing the manuscript.

[b] Inference of differential key regulatory networks and mechanistic drug repurposing candidates from scRNA-seq data with SCANet (2023)

Regarding SCANet, I supported the core developers by guiding the integration of Drugst.One into the SCANet platform. Additionally, I assisted with the network analysis and the interpretation of the resulting networks, and provided critical feedback to the manuscript.

[c] Democratizing knowledge representation with BioCypher (2023)

Within the BioCypher project, I contributed by leveraging the framework to generate a reproducible, archived version of NeDRex². In doing so, I provided extensive user-experience feedback to the core developer regarding usability, and contributed to writing and reviewing the manuscript.

[d] Online bias-aware disease module mining with ROBUST-Web (2023)

For ROBUST-Web, I guided the software development of the web interface³ for the bias-aware algorithmic update. My efforts included general web development feedback, provided information for integrating Drugst.One functionalities to implement the visualization runtime estimation feature, evaluating general usability, and managing tool deployment and maintenance. I also contributed directly to the manuscript text.

[e] Lacking mechanistic disease definitions and corresponding association data hamper progress in network medicine and beyond (2023)

I developed and implemented the GraphSimViz web application⁴ to grant public access to the novel methodologies and datasets generated within this project. I am responsible for the ongoing deployment and maintenance of the tool, and I authored the corresponding parts in the manuscript alongside providing general feedback.

¹ <https://apps.cosy.bio/proharmed/>

² https://github.com/repotrial/NeDRexDB_v1_Archive

³ <https://robust-web.net/>

⁴ <https://graphsimgviz.net/>

[f] DrugRepoChatter: A Drug Repurposing Expert Chatbot Curated by the REPO4EU Consortium (2024)

I contributed to DrugRepoChatter by assisting in curating the body of literature utilized throughout the project. Furthermore, I significantly facilitated the integration of open-source LLMs, the hosting thereof, and orchestrated the hosting infrastructure for the web application⁵. I also provided corresponding input during the manuscript preparation.

[g] Conversational no-code and multi-agentic disease module identification and drug repurposing prediction with ChatDRex (2025)

For ChatDRex, I served in a co-supervisory capacity to bring the project to a publication-ready state. I provided strategic guidance on software usability, detailed how to integrate Drugst.One, NeDRex, and broader network-based drug repurposing pipelines, and oversaw the manuscript draft through critical revisions.

[h] Drugst.One DREAM–Drug repurposing through expert annotation and modification (2026)

I supported the Drugst.One DREAM expansion (plugin version $\geq 2.1.0$) by continuously contributing architectural details and implementation insights. My responsibilities further included testing and publishing the proposed updates, as well as providing iterative feedback during the writing and revision of the manuscript.

[i] NetMedGPT – A network medicine foundation model for extensive disease mechanism mining and drug repurposing (2026)

In the NetMedGPT project, my contributions involved supporting the first author through conceptual discussions through my network-based drug repurposing knowledge and providing thorough review and feedback on the manuscript text.

⁵ <https://apps.cosy.bio/drugrepochatter/>

A.2.2 *Publication Supplements*

This section contains the original supplementary material of the four main publications this dissertation builds upon.



NEDREX-WEB
SUPPLEMENTARY MATERIAL

This section contains the original supplementary material published alongside the article:

“NeDRex-Web: An Interactive Web Tool for Drug Repurposing by Exploring Heterogeneous Molecular Networks”

Supplementary Material for NeDRex-Web

1 Methods

1.1 Implementation

The main technologies used to create NeDRex-Web are: i) vue.js (v2.5.11) as a primary component-based web-framework with vuetifyjs (v2.6.10) for component styling, fontawesome icons (v5.15.3) and vis-network (v9.1.9) of vis.js for interactive network visualization. ii) Spring Boot (v3.1.39) as a Java-based backend framework for the handling of GET and POST requests, WebSocket connections for request status updates and database access management. iii) A MySQL (v8.0.29) database to store the large heterogeneous network and user results. All three components are deployed using Docker, allowing to quickly change the host server if necessary, to distribute the applications on separate servers, or to deploy a locally running version. NeDRex-Web was tested on all common combinations of Chrome, Firefox, Safari, and Edge running on Linux, MacOS, or Windows installations (Table S1).

1.2 Interactive Network Visualization

A key part of any web-based network explorer is the interactive visualization. In NeDRex-Web, vis.js is used as a base framework to display up to 100,000 entities at once. A force-directed layout is initially prepared for each network on the backend server to save resources on the user's machine. Users have the option to select other network layout options, that are then computed on-demand. We use "portrait", "topographic", and "geodesic" layouts from the cartoGRAPHS package (1) allowing the efficient layouting of larger networks. As "portrait" and "topographic" layouts are provided for 3D layouting we provide all relevant binary combinations of the dimensions. All layouts are saved and reloaded in case the user decides to return to a network that has been visualized before. The network can be explored interactively by identifying first-degree neighbours and detailed information for individual nodes or by using a rectangular selection tool to select multiple nodes for follow-up analysis.

1.3 Input

Generally, users have three options to provide input in NeDRex-Web. These options are mostly standardized between the different exploration paths and assist in the selection of nodes or edges of different types from the knowledge graph. An input list can be defined using a combination of the available methods:

- **Manual ID input:** ID lists can either be pasted to the corresponding text area, where the delimiter will be automatically recognized, or uploaded as a file. The file upload supports single column lists or tsv tables as long as the IDs are located in the first column. The ID space for each node is the primary identifier type of source database (Table 1 rows 2, 3). The actual IDs can be prefixed by a lower case version of the ID space name. (Table 1 row 4).
- **Keyword search:** Most text-based node attributes are pre-processed and indexed to efficiently execute a substring search on all nodes of one type. This allows to provide suggestions that match the query in any attribute value responsively. Suggestions indicate the attribute and value that match the keyword as well as the main display name of the entry. There are different sorting styles for the list of suggestions and an option to add all suggestions at once.

Further, the keyword search is able to retrieve all entries connected to the search result and this way can add all genes associated with a specific disease or all pathways that include a specific protein, to name two examples.

- **Module identification based on gene expression data:** A special case of input is the expression file that the BiCoN algorithm needs to execute a module identification (MI) task. The matrix can be in tsv or csv format with IDs in the first column. Supported ID spaces are Symbols, EntrezIDs and Ensembl gene IDs.

1.4 Output

For the convenience of users, NeDRex-Web provides results in various ways and file formats:

- **On-page result summaries:** In NeDRex-Web result networks are visualized in the interactive graph browser and node or edge tables. The list of functions and degree of details provided at the primary result page differs between the exploration paths but any result can be loaded in the advanced view with the full list of functions available. Any entry of a table or visualized network can be inspected in detail, giving an overview over all relevant information stored in the NeDRexDB (2). IDs and other values are linked to corresponding external primary sources to enable access to the original entry.
- **Network export:** The resulting network can be downloaded in graphml format, an XML-based format supported by multiple network analysis libraries and Cytoscape (3). Also, extracted networks can be exported in form of node or edge tables in tsv format. Apart from the primary ID and the display name, users have the full control over attributes that should be included in the exported table.
- **List downloads:** In many cases users are able to save lists of entries. These can be manually created input lists, the identified disease modules, a ranked list of drugs, or a connector set in a connectivity search. In all cases, the lists can be uploaded as input to the tool in future queries.
- **Algorithms original outputs:** For all module identification and drug prioritization (DP) algorithm executions, the original result files are stored and provided on request. These may be provided as zip archive for approaches with multiple output files.

1.5 Drug Repurposing

NeDRex-Web implements a two-stage drug repurposing workflow consisting of disease module identification and drug prioritization steps and offers an easy-to-use interactive pipeline.

In the MI step, a set of user-defined seed genes or proteins is extended with the entities of the same type identified through the according interactome. This allows to identify additional drug targets and achieve better coverage of the genetic mechanism underlying the disease. NeDRex-Web provides users with the choice of six MI methods: BiCoN (4), DIAMOND (5), DOMINO (6), KeyPathwayMiner (7), ROBUST (8), and MuST (9). An explanation of the algorithms ideas and criteria to select an approach fitting the task can be found below (Section 1.9).

In the DP step, drugs that target nodes in the constructed modules are identified serving as potential drug candidates for the disease of interest. The candidate drugs are ranked using either TrustRank (10) or Closeness Centrality (11).

The drug repurposing result overview provides a table for seeds and module genes including algorithm-specific scores, assigned by the module identification method. A similarly structured drug table contains all identified candidates. All these entries are jointly presented in an interactive network visualisation. Drugs that are already in use for the queried disease are highlighted and enriched by information from ClinicalTrials.gov (12). To evaluate the statistical significance of the whole module set or drug ranking, empirical p-values can be computed (see Section 1.9.3).

1.6 Guided Connectivity Search

Guided Connectivity Search allows identifying direct or indirect connections between two sets of source and target entities. Both sets are defined in the same way as seed sets for drug repurposing queries, without the restriction of them being of type gene or protein only. The user specifies the desired type of connection, selecting from a list of available paths. Additionally, it is possible to filter edges and nodes by attributes and define an exclusion or inclusion list of connector nodes. The result overview provides the lists of source, target and on request the connector nodes with an interactive network representation of the extracted network.

1.7 Free Exploration

The Free Exploration provides a way for mostly unrestricted exploratory analysis of the large heterogeneous NeDRex-Web knowledge graph. A base network can be extracted by selecting node and edge types and setting filter options. Based on this initial selection, the sub-network is defined and provided in a visual network browser as well as in a tabular representation. Both views support the selection of nodes and edges for the creation of a standalone network. Further, all MI and DP algorithms are available to be applied on all or a selection of genes or proteins. Entries can be added by adding unused connection types and induced edge creation is possible. These options allow for the iterative analyses to answer each users specific questions without running into an exploratory dead-end after each query.

1.8 Result caching and History

One information that NeDRex-Web needs to save persistently in the user's browser is a cookie with a unique user ID. This ID is linked to all queries, algorithm executions and results of this specific user to provide an individual research history. Each result is also uniquely identifiable through a specific URL, allowing to share it and continue exploration at any time. The unique ID is further used to save hierarchical relationships between networks. Such a relationship is established when one "base" network is used to derive another one with the initial network being the "parent" and the derived one being "child" in the resulting hierarchy. These research hierarchies are the result of iterative explorations in NeDRex-Web which are presented in the user specific research history. The goal is to help users to trace their exploration steps and find old results quickly. This is further supported by allowing custom titles and descriptions of networks to save additional information about their context.

1.9 Algorithms

1.9.1 Disease module identification algorithms

NeDRex-Web provides the user with the choice of five seed-based MI approaches. DOMINO (6) and Key-PathwayMiner (KPM) (7) identify maximal connected subnetworks highly enriched with seed genes. KPM focuses on the whole seed input at once, where DOMINO iteratively refines the enriched regions resulting in potentially multiple disconnected modules. Together with ROBUST (8), which aims to identify only the most reproducible gene additions, these three approaches might be used for input genes, which describe a disease where the mechanism is not fully understood or multiple distinct disease mechanisms exist.

In contrast, MuST (9), the method the ROBUST algorithm was built on, and also DIAMOND (5) are better used to enrich the genetic background of an already well-defined disease by including first degree interaction partners as potentially druggable targets.

For unsupervised extraction of key genes that describe a disease mechanism, BiCoN (4) can be used. It biclusters expression data and uses the network to limit results to connected sets of genes.

BiCoN

BiCoN (Biclustering Constrained by Networks) (4) enriches general expression data-based biclustering of genes and patients/conditions by gene interaction information. It 'restricts biclusters to functionally related genes connected in molecular interaction networks' while 'maximizing the expression difference between two

subgroups of patients'. Both clusters of genes are seen as relevant for the analyzed condition and thus combined in the resulting module.

Parameters:

- **ID type:** Defines the expected ID type used in the gene expression matrix. Can be one of ['Symbol', 'Entrez', 'Ensembl'].
- **Expression matrix:** Defines the file to be uploaded containing the expression matrix.
- **Min & max:** Defines the minimal and maximal solution network size in genes that are used to bichuster patients or conditions.

DIAMOnD

DIAMOnD (DIseAse MOdule Detection) (5) is an algorithm that iteratively expands a disease module by identifying proteins that have significant connectivity patterns to known disease proteins (seeds). Rather than simply counting connections or focusing on dense communities, DIAMOnD evaluates the statistical significance of the connections between proteins and the current module. For each candidate protein, the algorithm calculates the probability that its observed or higher number of connections to seed proteins (k_s) could occur by chance, given its total number of connections (k) in the network. The connectivity significance is determined by a P -value defined as:

$$P\text{-value}(k, k_s) = \sum_{k_i=k_s}^k p(k, k_i) \quad (5)$$

where $p(k, k_i)$ is given by the hypergeometric distribution, representing the probability that a protein with k total connections has exactly k_i connections to seed proteins in a random configuration. At each iteration, the protein with the lowest P -value is added to the module, becoming a new seed for subsequent iterations. An extension of the algorithm introduces differential weighting, where connections to original seed proteins receive a weight $w = \alpha$ while connections to proteins added during the iterative process receive a weight $w = 1$. This modification allows for fine-tuning the relative importance of original disease proteins versus newly added module members.

Parameters:

- **N:** Defines the maximum number of iterations and thus the number of nodes to be added to the module.
- **Alpha:** Defines the seed node influence in the significance calculation. Connections to seed nodes are weighted according to α (default: 1) whereas non-seeds are weighted 1. An $\alpha > 1$ will prefer nodes with seed node connections over nodes with non-seed connections to the module.
- **P-cutoff:** Sets a threshold value for the module node p-value (default: 10^{-3}).

DOMINO

Discovery of Modules In Networks using Omic (6) is a quite recently developed active MI approach with the ability to resolve the issue of the over-representation of certain gene ontology terms in result modules. Further, the authors observed results to be too unspecific regarding the biological context. To combat this, DOMINO does not try to connect all active nodes by force but allows for completely independent clusters as long as active nodes are over-represented. NeDRex-Web uses DOMINO as MI method for user-defined sets of seed nodes.

Key Pathway Miner

Key Pathway Miner (KPM) (7) aims to identify functional modules through extracting connected subnetworks. This method relies on an indicator matrix based on which functionally active genes are extracted to identify key pathways within the gene-gene interactome, so network modules in which these genes are enriched. It will uncover all possible, maximal pathways that contain at most K non-active genes. Marking seed genes as active ones and adjusting the K parameter allows to provide a flexible method for MI that, depending on the configuration, is able to uncover various, unconnected disease mechanisms. In

NeDRex-Web the indicator matrix contains one column only which is defined by the set of input seeds.
Parameters:

- **K**: Defines how many non-seeds are allowed to be added in each identified pathway. Influences how freely KPM adds module nodes.

MuST

Minimal Steiner trees (9) are generated, meaning a minimal scoring tree is identified that is still only one connected component. This way, minimal cost paths within the network are extracted. The reasoning behind using multiple Steiner trees (MuST) is that PPIs are highly connected leading to non-unique solutions for the Steiner tree problem. Thus, selecting only one tree does not identify a disease mechanism but merely proposes one part of the overall module. Therefore, multiple trees are used to derive a prioritized list of target nodes the seeds are connected to, where the number of 'visits' of a non-seed node indicates the priority or significance of it.

Parameters:

- **Hub-penalty**: Defines a penalty for hubs and preferring lower network degree nodes (default: 0 $\rightarrow$ no penalty).
- **Multiple**: Enables the repeated execution and consensus creation mode. This helps to eliminate the random factor if multiple minimal scoring Steiner trees exist.
- **Trees**: Only shows if 'Multiple' enabled. Defines the number of equally parametrized trees to be constructed and ultimately combined into one.
- **Iterations**: The number of iterations of the Steiner tree construction (default: 10).

ROBUST

Robust disease module mining via enumeration of diverse prize-collecting Steiner trees (ROBUST) (8) combats the issue of robustness that MI methods face regarding input permutation, through the incorporation of non-deterministic steps. Like MuST, ROBUST generates various Steiner trees with the difference of using prize-collecting Steiner trees and seeds having higher priority value than non-seeds. This way the algorithm favours the addition of non-seed nodes only if they are far from being only random occurrences. This method is provided as an approach that is able to lead to the identification of significant and possibly distinct mechanisms.

Parameters:

- **Initial Fraction**: The parameter $\alpha \in (0, 1]$ defines how expensive the Steiner trees are allowed to become. A larger α will allow more diverse results and $\alpha = 0$ would for example only return the best Steiner trees.
- **Reduction Factor**: The parameter $\beta \in [0, 1)$ dictates how many non-seed nodes are allowed to be integrated into Steiner trees. Each non-seed value will decrease with the amount of Steiner trees it was captured with. $\beta = 0$ sets a non-seeds value to 0 once it was used the first time. β close to 1 can lead to limited exploration, because a non-seeds value will only slowly decrease over the integration in many Steiner trees.
- **Threshold**: The threshold $\tau \in (0, 1]$ determines the robustness/explorativeness of the result. The explorativeness increases (and robustness decreases) for lower τ , and *vice versa* for higher τ .
- **Trees**: Number of individual Steiner trees to be computed for the combination step.

1.9.2 Drug ranking algorithms

Closeness centrality

Closeness centrality (CC) (11) is one of several centrality measurements. These are used in network analysis to derive scores that translate to the importance of the nodes in the specific network. Closeness, in particular, is derived by calculating the sum over all shortest paths between the specific node n and any other o to the power of the negative one:

$$CC(n) = \frac{1}{\sum_o d(n,o)}$$

A low CC is then used as an indicator for possible candidates that target the module.

TrustRank

TrustRank(10) is derived from the well known PageRank algorithm (13). Its original purpose is to filter spam-sites that are not highly supported by other pages. This is achieved by assigning weights based on in- and outgoing edges to favor highly and punish weakly supported nodes. In theory, these weights are iteratively propagated through the network until the assigned scores or ranks stabilize. In a biological application, the ranked websites are drugs, and the most trusted ones are drugs that have high chances of targeting a large number of seeds.

Parameters:

- **Damping factor:** The damping factor for the network propagation step known from other PageRank implementations.

1.9.3 Validation algorithms

To measure significance of MI and DP results, NeDRex-Web provides permutation-based empirical p-value computation. Drugs known to be associated with the overall context of the search, i.e., mostly the queried disease that is under study, are needed for the validation, as those are used as reference to determine the empirical p-values. If the seed input was derived through the search for diseases by adding all associated genes using the NeDRex-Web suggestion function, this disease is saved and drugs are extracted using the drug-disorder indication edges. Through literature search or expert knowledge, this drug list may be manually extended or curated to provide a set representative for the context. For both algorithm groups, two p-values are derived. For MI results, these are the "empirical p-value" and "empirical (precision-based) p-value"; for DP results, the "empirical DCG-based p-value" and "empirical p-value without considering ranks". They are executed using the NeDRexAPI, and more information about the different validation methods is provided in the NeDRex publication (2).

Module validation

For the calculation of an empirical P-value, n random modules with the same number (c) and combined size (s) of connected components are generated. For this, first c random nodes are selected as starting points for each connected component, assuring that those are not first-degree neighbors to any other starting point or to their neighbors (i.e., not first- or second-degree neighbors). Then, for each connected component, random first-degree neighbors are picked iteratively and added to the connected component set, except if they would create connections to other connected components (either directly or through shared neighbors). This expansion continues until either the total module size reaches s or no more nodes can be added while maintaining the separation of components (with a limit of 20 unsuccessful iterations as a failsafe). This process is repeated n times to create n random modules. The lists of drugs targeting the result modules and each random module are determined and used to generate the following p-values:

- **Empirical p-value:** This value is determined by counting the occurrences of random modules that directly link to a higher number of reference drugs in comparison to the result module. This number of occurrences is then divided by the number of permutations (with a pseudo count of 1) to derive a p-value.

- **Empirical (precision-based) p-value:** This value is derived in the same way as the normal empirical p-value, but the definition of "better performing module" changes. Precision-based means that the module with the better precision with respect to connected drugs is determined as "better". This is calculated by dividing the number of reference drugs targeting the module by the number of all drugs targeting the module.

Drug Ranking Validation

For empirical p-value calculation for drug prioritization results, n randomly from the whole database selected and randomly ranked drug lists are derived matching the length of the result list. The random ones and the observed (result) drug lists are then scored using the reference list. It is evaluated in how many instances a random list performs better than the observed one. The performance measures are:

- **Empirical DCG-based p-value:** Here, the Discounted Cumulative Gain (14) is used to evaluate how good the ranking of the drug list is. With this function, drugs from the unranked reference list that occur at a smaller rank, thus high up in the drug list that is scored improve the score more than worse ranked occurrences. The cumulative score of the observed result and the random ones is then compared to derive the number of random results that perform better than the algorithmic result. Finally, the number of occurrences performing better than the observed result (with a pseudo count of 1) is divided by n to derive an empirical p-value.
- **Empirical p-value without considering ranks:** This measure only compares the number of reference drug occurrences in the observed ranking and the random rankings. The number of random drug lists with more reference drugs is counted to derive the p-value as before.

1.10 Runtime and Scaling

To evaluate runtime and scaling behavior differences between algorithms, we followed a standardized methodology using the API. To ensure realistic test cases, diseases from the database were categorized into bins (10, 25, 50, 100, and 250) based on their number of associated genes, allowing for a 15% deviation. Five diseases were randomly selected from each bin and their gene lists were saved as seed sets. These seed lists served as input for both module identification algorithms (DIAMOnD, MuST, KPM, DOMINO, ROBUST) and drug ranking algorithms (TrustRank, Closeness Centrality). Each algorithm combination was executed sequentially, with execution times measured and logged. The results are visualized in line graphs with error areas representing standard deviation across examples, shown in Figures S1 and S2, respectively. While KPM, DOMINO, and DIAMOnD demonstrate constant runtime independent of seed set size, ROBUST, Closeness Centrality, and TrustRank show slight input size dependency. In contrast, MuST exhibits apparent polynomial scaling with respect to input seed count. Code and data for reproducibility are available at: <https://github.com/repotrial/nedrex-web-runtime-scaling>.

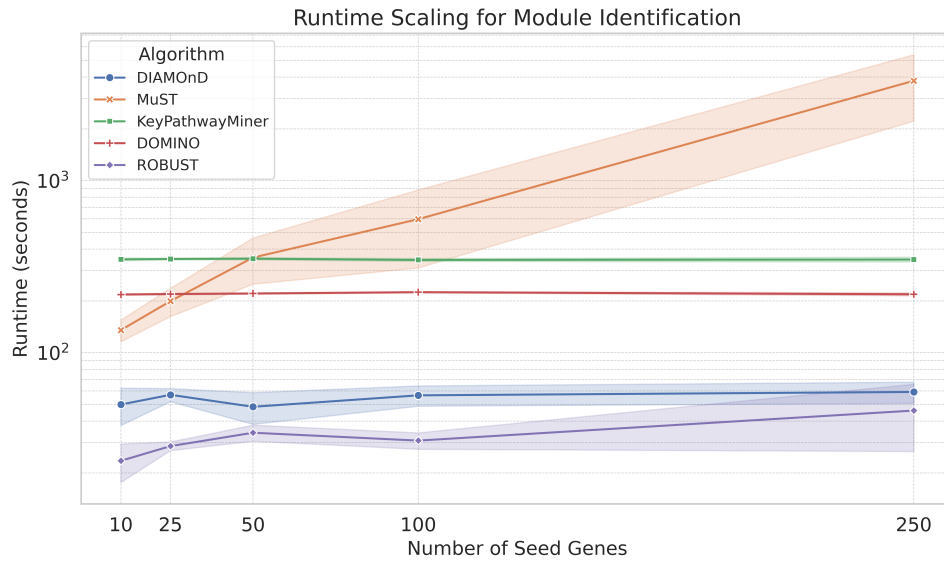


Figure S1: Runtime evaluation of module identification algorithms. KeyPathwayMiner and DOMINO demonstrate constant runtime regardless of seed count. While DIAMOnD hints to general seed count independence, some runtime variation is observed. ROBUST executes rapidly overall but exhibits linear scaling with input gene count. MuST, initially among the slower algorithms, displays polynomial runtime scaling with input size, becoming notably the slowest algorithm for seed sets exceeding 50 genes.

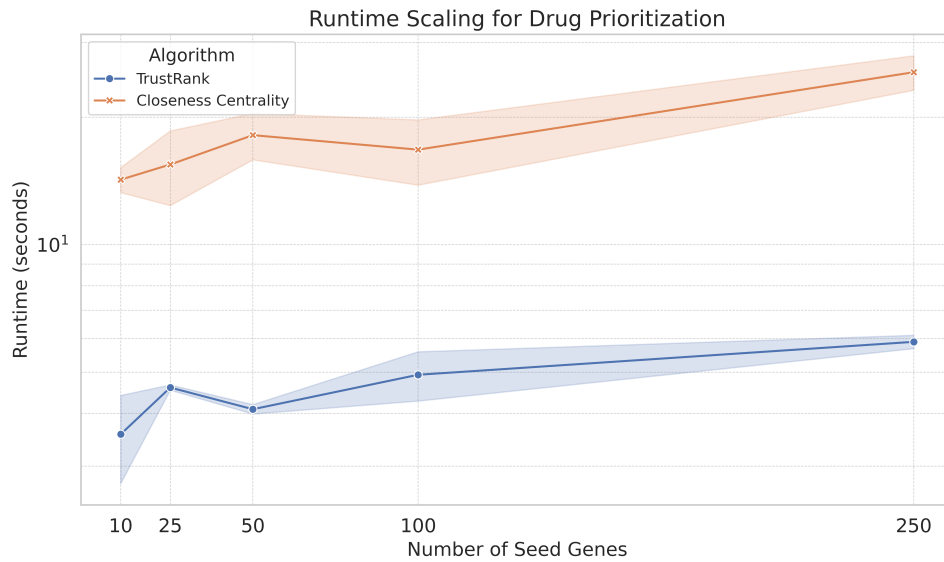


Figure S2: Runtime evaluation of drug prioritization algorithms. TrustRank and Closeness Centrality exhibit similar runtime scaling patterns with respect to input seed set size. TrustRank consistently performs approximately an order of magnitude faster than Closeness Centrality.

OS	Version	Chrome	Firefox	Safari	Edge
Linux	Mint 22 Cinnamon	127.0.6533.119	129.0	n/a	n/a
MacOS	Sonoma 14.5	127.0.6533.119	(123.0.1)	17.5	n/a
Windows	10/11	127.0.6533.120	192.0.1	n/a	127.0.2651.98

Table S1: NeDRex-Web tested browser compatibility table. Versions in brackets mean that the page is generally usable but minor unintended behaviour might be experience.

1.11 Additional Material

Additional information on NeDRex-Web is provided in form of a documentation¹ and tutorial videos², demonstrating the different exploration paths on simple example cases. NeDRex-Web is compatible with four most popular browsers and tested under three operating systems (Table S1).

2 Data & Attributes

In the following an overview about the five node and eleven edge types in the NeDRex-Web knowledge graph will be given. This includes the source databases and attributes of entries. The source database versions used to produce presented results are listed in Table S4

2.1 Nodes

An overview about the integrated node types and sources is given in Table S2.

Table S2: NeDRex-Web comprises 439,712 nodes of five different types.

Category	Node Types				
	<i>Disorder</i>	<i>Drug</i>	<i>Gene</i>	<i>Pathway</i>	<i>Protein</i>
Entries	24,254	15,235	193,456	2,679	204,088
Sources	MONDO	DrugBank	NCBI gene	Reactome	UniProt
ID space	MONDO	DrugBank	Entrez	Reactome	UniProt
ID example	mondo. 0009061	drugbank. DB09038	entrez.6524	reactome. R-HSA- 77111	uniprot. F2WIII
Attributes	'ID' 'display name' 'ICD-10' 'synonyms' 'description'	'ID' 'display name' 'category' 'drug group' 'inchi' 'cas number' 'mol. formula' 'IUPAC name' 'SMILES' 'sequences' 'synonyms' 'indication' 'description'	'ID' 'symbol' 'alt. symbols' 'chromosome' 'display name' 'gene type' 'genomic location' 'synonyms' 'description'	'ID' 'display name'	'ID' 'display name' 'gene name' 'Review Status' 'synonyms' 'sequence' 'comments'

¹https://docs.google.com/document/d/1BGp0wovJk_ERonojc6s9XQ7W2sV_Bgb-1jtjI7yj2bg/

²<https://www.youtube.com/watch?v=NDx2YVN7TpQ&list=PLY5U8UsdXZRnc0A05f9QjCMZSb065WKfV>

2.2 Edges

NeDRex-Web imports eight edge types from NeDRexDB. Three edge types are *inferred* from applying the protein to gene mapping provided through the protein - [encoded by] - gene edge on the according edge types. An overview over all edge types, the corresponding number of entries and their sources are combined in Table S3.

- **Disorder** - [*is comorbid with*] - **Disorder**: This edge provides information about co-occurrence of diseases in the same individual based of data provided by the Estonian Biobank, resulting in 4,414,177 comorbidity connections.
- **Disorder** - [*is subtype of*] - **Disorder**: This edge provides disease hierarchy information all integrated from MONDO (15). This directed acyclic ontology graph contains 36,529 connections.
- **Drug** - [*has contraindication*] - **Disorder**: DrugCentral (16) provides information about strong adverse side-effects of drugs applied in specific diseases. 13,809 contra-indications are imported into NeDRexDB.
- **Drug** - [*has indication*] - **Disorder**: The indication for drugs having a positive effect on a disease is encoded in 17,672 indication edges. Those are combined from DrugCentral (16) and the Comparative Toxicogenomics Database (CTD) (17).
- **Drug** - [*has target*] - **Gene**: This connection describes the genetic target of drugs. Normally this information is directed towards proteins, because drugs mostly bind to proteins to have some effect on patients system. Using the protein to drug mapping given by the protein - [encoded by] - gene edge, this edge is created on bases of the original drug - [has target] - protein data. This preprocessing step produces 29,866 unique edges and summarizes the actions of the protein targets.
- **Drug** - [*has target*] - **Protein**: The information about what is a drugs genetic target is essential in several steps of NeDRex-Web. The information is gathered from DrugCentral (16) and DrugBank (18) and combines to 30,088 unique entries. Information about the action of the drug on the protein is included.
- **Gene** - [*associated with*] - **Disorder**: The underlying genetic mechanism of a disease can be partially described by its known associations to genes. By integrating DisGeNET (19) and Online Mendelian Inheritance in Man (OMIM) (20) 33,304 association edges are established. A score provides the strength of the association.
- **Gene** - [*interacts with*] - **Gene**: The gene based interactome is a translated version of the normal protein-based interactome by using the protein to gene mapping edge protein - [encoded by] - gene. This process creates 1,382,888 unique edges and collects the entries for the attribute list 'evidence type' [predicted, orthologue, experimental] and conditions under which the interaction is expected including 'tissue', 'brain tissue', development stage', 'joint tissue' and 'subcellular location'.
- **Protein** - [*associated with*] - **Disorder**: This edge is created based on the gene - [associated with] - disorder edges using the protein - [encoded by] - gene edge to map between genes and proteins. Association scores are transferred to the 63,877 derived edges.
- **Protein** - [*encoded by*] - **Gene**: Information about the origin gene of a protein is extracted from UniProt (21) generating 30,389 mapping entries.
- **Protein** - [*in*] - **Pathway**: The pathways that proteins plays a role in are derived from Reactome (22). 126,564 edges are this way created in the NeDRexDB.
- **Protein** - [*interacts with*] - **Protein**: The interactome is essential for almost all analyses done on NeDRex-Web. 2,769,759 protein - protein interaction (PPI) edges are a combination of the integration of BioGRID (23), Inactive Ingredients Database (IID) (24) and IntAct (25). The source databases are stored as attributes alongside the annotated methods to determine the interaction (e.g biochemical) and the list of evidence types (predicted, experimental, orthologues). A fourth resource the Human Protein Atlas (HPA) (26) is integrated to allow the construction of condition specific interactomes.

The condition classes are 'tissues', 'brain tissues', 'joint tissues', development stages' and 'subcellular locations'.

Table S3: A summary table containing all 12 edge types in the NeDRex-Web knowledge graph.

Edge Type	Entries	Sources
Disorder - [is comorbid with] - Disorder	4,414,177	Estonian Biobank
Disorder - [is subtype of] - Disorder	36,529	MONDO
Drug - [has contraindication] - Disorder	13,809	DrugCentral
Drug - [has indication] - Disorder	17,672	DrugCentral, CTD
Drug - [has target] - Gene	29,866	<i>inferred</i>
Drug - [has target] - Protein	30,088	DrugCentral, DrugBank
Gene - [associated with] - Disorder	33,304	DisGeNET, OMIM
Gene - [interacts with] - Gene	1,545,159	<i>inferred</i>
Protein - [associated with] - Disorder	63,877	<i>inferred</i>
Protein - [encoded by] - Gene	30,389	UniProt
Protein - [in] - Pathway	126,564	Reactome
Protein - [interacts with] - Protein	2,769,759	BioGRID, IID, IntAct, (HPA)
Total	9,111,193	-

3 Use cases

3.1 Drug Repurposing for Cystic Fibrosis

For the first show case, cystic fibrosis (OMIM 219700, MONDO 0009061) was investigated for potential drug repurposing candidates. The seed list was generated through NeDRex-Web using all cystic fibrosis associated genes. A visualization of the result network can be found in Figure S3.

For module identification, DIAMOnD was employed with the following parameters:

- P -value threshold: 10^{-35}
- Rest: default

TrustRank was selected for drug ranking with the following configuration:

- Approved drugs only: disabled
- Rest: default

Source	Version
BioGRID (23)	v4.4.236
CTD (17)	release July 2024
DisGeNET (19)	v7.0
DrugCentral (16)	2023-11-10
DrugBank (18)	v5.1
HPA (26)	v23
IID (24)	2021-09
IntAct (25)	2023-05
MONDO (15)	release 2024-08-06
NCBI gene (27)	state 2024-08-15
OMIM (20)	release 2024-08-14
Reactome (22)	v89
UniProt (21)	release 2024_07

Table S4: All source database integrated in NeDRexDB and used to construct the NeDRex-Web knowledge graph.

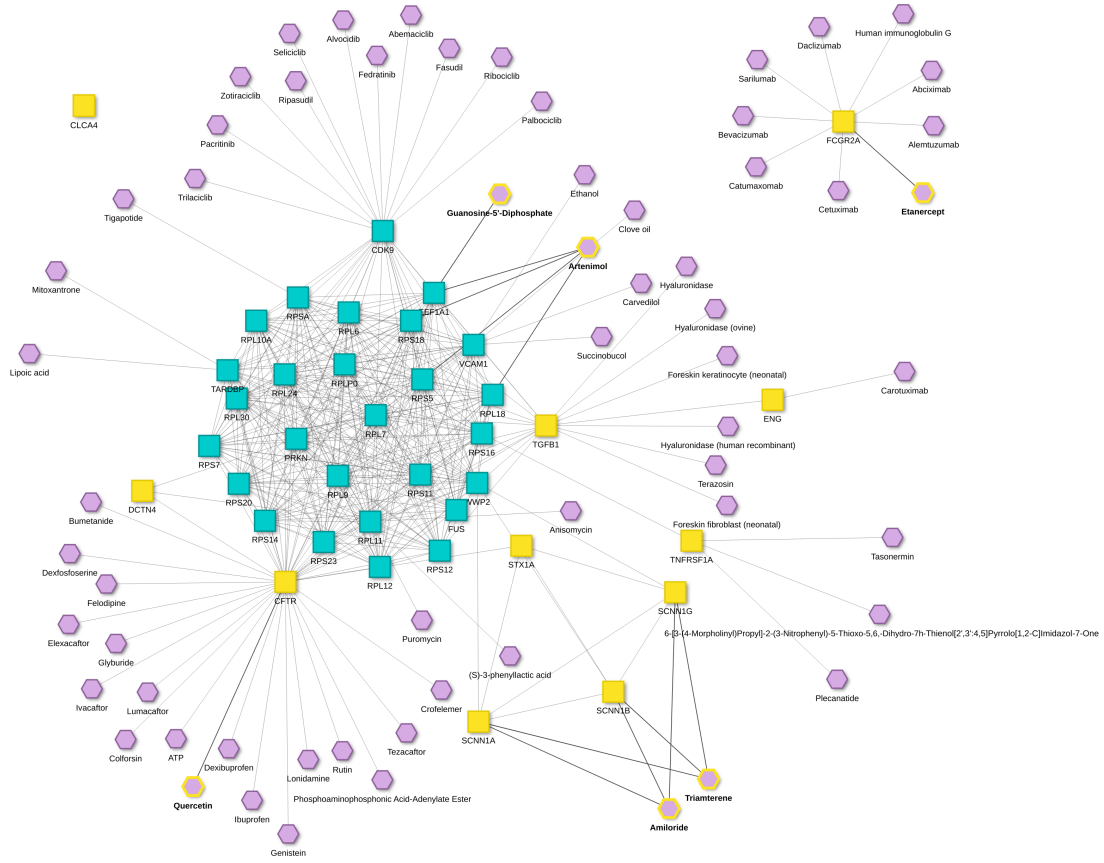


Figure S3: Full-sized version of Figure 3A - The resulting subnetwork from the disease module and drug repurposing pipeline performed for Cystic Fibrosis (CF). Yellow nodes depict the seeds associated with the CF disease in the NeDRexDB. Turquoise nodes are genes incorporated by the DIAMoND module (P -value threshold 10^{-35}), and purple nodes represent top drugs targeting all module genes after TrustRank drug prioritization. Drugs highlighted with a yellow border are drugs identified to have shown potential or might be candidates for use in CF treatment, namely amiloride, artemimol, etanercept, guanosine-5-diphosphate, quercetin, and triamterene.

3.2 Exploratory analysis of the positive outcomes of SGLT2 inhibitors in Heart Failure

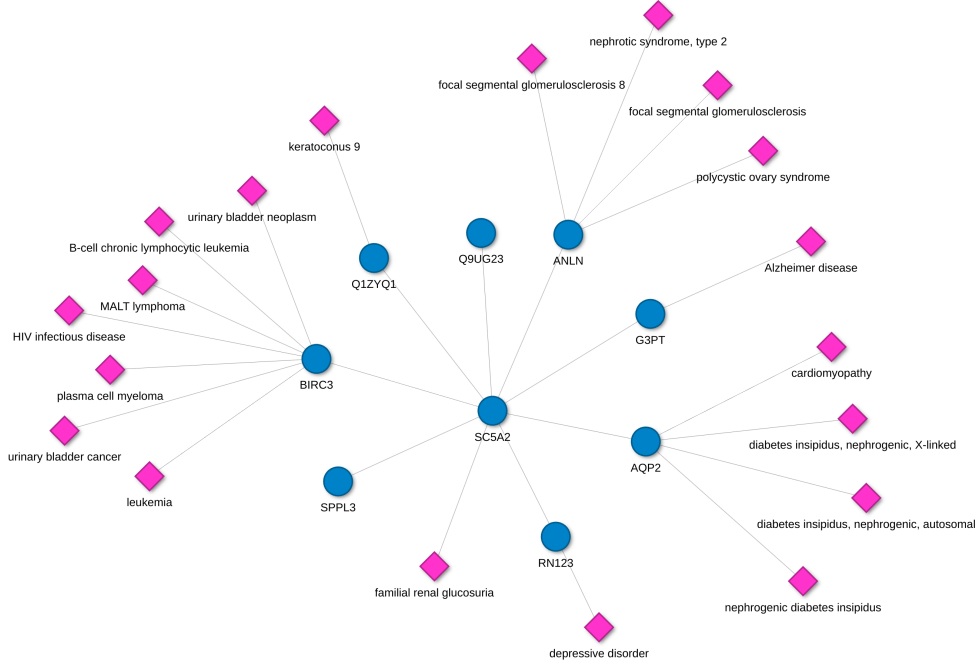


Figure S4: Full-sized version of Figure 3B - First neighbour PPI network of the *SGLT2* co-transporter (*SC5A2*, UniProt P31639) and associated disease phenotypes. On the right, Aquaporin 2 (*AQP2*) appeared directly connected to *SC5A2* and associated to cardiomyopathy. These results, supported also by previous research, suggest *AQP2* as a potential mechanism that could explain the positive outcomes of *SGLT2* inhibitors in heart failure.

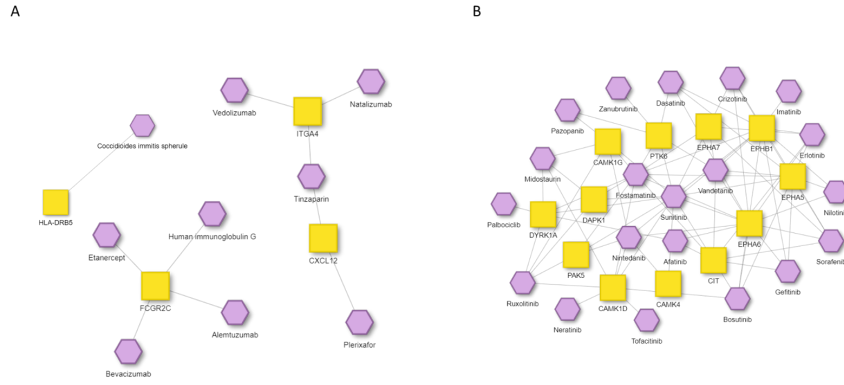
3.3 Drug prioritization for the heterogeneous brain lesions in patients with progressive multiple sclerosis

This exploratory study was conducted on a now updated version of the integrated interactome, thus the exact networks might differ in the latest version of NeDRex-Web.

Multiple sclerosis (MS) is a chronic inflammatory and neurodegenerative disease of the central nervous system (CNS) characterized by the formation of focal lesions (28). In the last decades, a rapid increase in treatment options has been developed (29). Despite the success of disease modifying therapies (DMTs) in the early disease phase, the compartmentalized chronic inflammation and neurodegeneration in progressive MS (PMS) are not prevented (30). In PMS, new lesions are rare, but there are still infiltrating immune cells inducing inflammatory demyelinating lesions (31), some of which can spontaneously repair (RL), become chronic active (CA) or subsequently become inactive (IL). We have established the PMS transcriptome of these different lesions (n=100) in the brain of ten patients with PMS versus five non-neurological diseased controls (32, 33). With NeDRex-Web, we used TrustRank (Figures S5.A, S6) to explore general MS treatment options based on expressed genes (FDR<0.05) shared among all the different lesion types. The set of these 264 expressed genes mapped to brain tissue specific PPIs is used as the starting point of the analysis. A gene cluster from the generated interactome contained two approved MS treatments, natalizumab and alemtuzumab, ranked 36 and 53 respectively (Figure S5A). Both lead to immune suppression e.g. natalizumab target the integrin (*ITGA4*) on lymphocytes to reduce infiltration into the brain tissue (Figure S5A). The MS susceptibility gene *HLA-DRB5* (34) and the key chemokine (*CXCL12*) in MS glia immunity (35) were also present, where plerixafor (rank 30) an inhibitor of the *CXCL12* receptor (*CXCR4*) and tinzaparin (rank

28) also targeting the *ITGA4* were suggested (Figure S5A).

Major differences exist between MS patients, especially those that stay in the early relapsing-remitting phase contra those that convert to the progressive phase. Adapting treatment to the progressive phase is critical for patient outcomes. In PMS the number of CA lesions are increased and associated with poor clinical prognosis (36). Therefore with NeDRex-Web, we wanted to identify candidates for PMS specific targets (n=294) unique for CA (FDR<0.001) mapped to brain tissue-specific PPIs (Figure S5B, S7). One of two CA-specific connected gene-drug components was dominated by Eph receptors (*EPHA6*, *EPHA5*, *EPHA7*, *EPHB1*), and the drug suggestions (e.g. bosutinib, sorafenib, erlotinib, pazopanib, ranked 7, 18, 19, 21) were also recommended in an astrocyte-microglia gene-drug module of the common MS lesion seeds (Figure S6B) indicating a synergic effect of same drugs on multiple MS pathological events (Figure S5B). Eph/ephrin bidirectional signaling regulate axon guidance through contact repulsion (37) and inducing axonal growth cone collapse in injured tissue (38). Besides their direct role in inhibition of regeneration by repulsive mechanisms, the Eph/ephrins can also regulate astrocyte function by supporting scar tissue formation (39). Thereby blocking of the Eph/ephrin signaling in MS lesions may play a central role in determining repairing outcomes and can be the direction for regeneration-based therapies of CA in PMS. To sum up, with NeDRex-Web we re-discovered already applied drugs in the MS clinic, but moreover the tool also revealed multiple new candidates for drug repurposing. The idea of using multiple drugs acting in a synergistic manner on the same and/or multiple genes expressed in all types of white matter lesions or even specifically for patients with CA lesions may increase treatment efficacy in PMS.



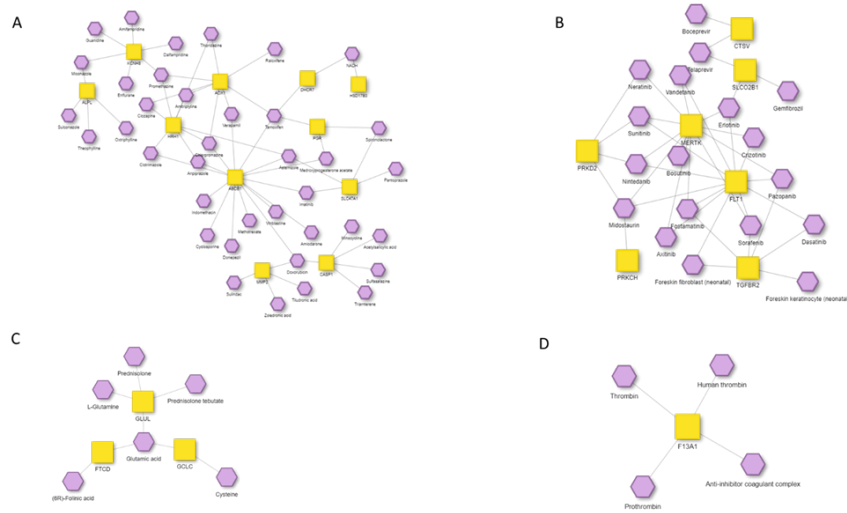


Figure S6: This depicts the connected components in the drug-target network for seeds shared among all white matter lesions types ($n=264$, $FDR < 0.05$ compared to non-neurological diseased controls). The drugs were identified and ranked with TrustRank. **(A)** The biggest component contained multiple seeds reflecting the pathological consequences of demyelination (*DHCR7*, *HSD17B3*, *MMP2*) and neuronal/axonal ion imbalance (*KCNH8*, *ABCB1*, *SLC47A1*). **(B)** A second component contained known activated receptors in MS representing the dynamic cross-talk between astrocytes and microglia: (i) the phagocytic tyrosine kinase receptor (*MERTK*) on microglia important for myelin clearance activated by transforming growth factor β (*TGF- β*) (40), (ii) the *TGF- β* receptor II (*TGFRBR2*), an important glia immune activator (41, 42) and (iii) the FMS-related tyrosine kinase 1 (*FLT-1/VEGFR-1*) important for proliferation and scar formation in astrocyte controlled by microglia (43). Several tyrosine kinase inhibitors are suggested (e.g. bosutinib, sorafenib, erlotinib, pazopanib). However, whether this MS activated microglia-astrocyte signaling should be activated or inhibited is unclear. **(C)** In another connected group, seeds are coding for enzymes involved in glutamatergic pathways (*FTCD*, *GCLC*, *GLUL*). Glutamate toxicity has long been suggested to contribute to MS disease process, and several compounds targeting the glutamatergic system in animal models of MS have been studied, however clinical trials have not been successful (44), and the drug suggestions are also the substrates of the dysregulated enzymes. **(D)** Lastly, four drug suggestions (rank 89, 90, 91, 97) targeting the coagulation factor XIII, A1 polypeptide (*F13A1*), is a well known CAM gene enriched in microglia in the MS brain (45).

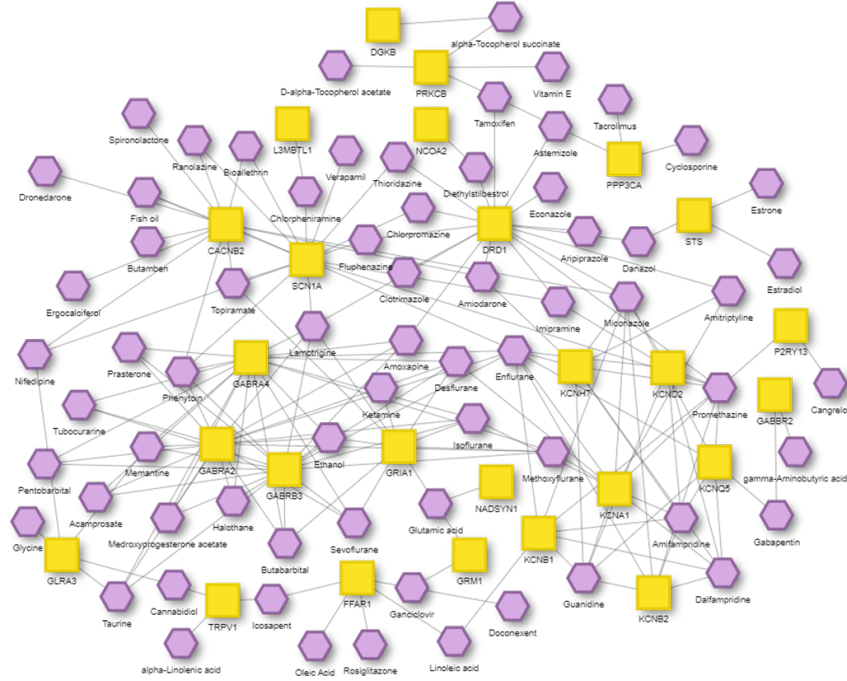


Figure S7: The biggest connected component in the gene-drug network of seeds uniquely expressed in the chronic active lesion type (CA) ($n=294$, $FDR < 0.001$ compared to non-neurological diseased controls) mapped to the brain tissue-specific experimentally validated PPI network. The drugs were ranked with TrustRank (find parameters below). The subnetwork is dominated by ionotropic glutamate receptors (*GRIA1*), metabotropic glutamate receptors (*GRM1*), modulators of glutamate (*STS*) with multiple suggestive drugs e.g. treatment of major depressive disorders (amoxapine, amitriptyline). Studies have also shown the role of glutamate excitotoxicity in the progression of MS (46, 47). Glutamate is the major excitatory neurotransmitter of the CNS, and it has a central role in a complex cellular interactome. However, an excess release of glutamate leads to the activation of ionotropic and metabotropic receptors, resulting in accumulation of toxic cytoplasmic Ca^{2+} and cell death. The presence of Na^+ channel (*SCN1A*) in the CA seeds may also reflect the imbalance of ion homeostasis. On contrary, this CA gene-drug component also contained seeds that may prevent excitotoxic neuronal damage as both *GABAA* receptors (*GABRA2*, *GABRA3*, *GABRB3*) are thought to inhibit depolarization-evoked Ca^{2+} influx, and increased expression of voltage-gated K^+ channels (*KCNA1*, *KCNB1*, *KCNB2*, *KCND2*, *KCNQ5*, *KCNH7*) are believed to limit neuronal excitability (48). Additionally, the microglia (*P2RY13*), fatty acid signaling (*FFAR1*) and candidates targets in Alzheimer Disease (*L3MBTL1* involved in proteasomal digestion) and the dopamine receptor (*DRD1*) are also present in the seeds. Therefore, understanding how the distinct channels and receptors contribute to the MS pathophysiology and excitotoxicity in the CA is still a challenge before therapeutic strategies. However using NeDrex-Web, the importance of excitotoxicity in the MS disease progression was emphasized and identification of functional gene-drug clusters for further investigation.

Parameters and Material

In both drug ranking cases, common shared genes (Table S5) and CA specific genes (Table S6), TrustRank was employed using the following configuration:

- TopX drugs: 100 (default)
- Only experimentally validated interactions: enabled (default)
- Add gene interactions: disabled
- Tissue-specific interactions: Brain
- Direct drugs only: enabled (default)
- Filter Element Drugs: enabled (default)
- damping factor: 0.85 (default)

2	4240	7048	1781	9332	9920	23143	22837	11098	55603	340485	728723
230	9124	7706	1119	8477	7168	84327	55503	23041	54502	221178	116159
834	5793	4864	9882	9987	8418	10457	51059	57664	58475	253827	196383
837	3269	3127	1601	6387	3676	25859	80704	10841	84899	127435	337873
316	5241	3992	3426	1522	9829	55244	57211	26146	54622	374887	143903
730	8832	8797	2752	8740	2035	51176	10018	23533	80833	115361	646755
463	4094	5763	3603	6693	3481	51365	25816	10971	51339	285668	153328
873	5937	8857	3487	3021	2294	10165	92610	10110	116362	283755	100505495
862	1832	4211	1193	6819	9452	84947	51735	80000	221395	197259	100463486
390	2683	1717	1806	8870	1356	91351	10602	63901	126669	153830	100130562
947	1107	1295	2313	4481	11309	25937	54510	51705	259236	116441	101928470
871	5243	3293	5420	4360	23231	55340	53827	28965	255488	131096	101928176
249	4053	9096	2729	2321	51363	84689	55303	84952	171586	338596	100463498
840	6319	3500	2162	2023	89857	27296	54855	58480	439949	201626	100288805
951	4162	9935	9411	3082	25793	23321	57091	10742	286467	284293	100288332
684	9371	3937	8365	3638	10491	64167	23266	54718	728430	400550	100271871
113	1281	5583	5239	7052	25865	10894	10461	26112	219855	162073	100420358
4313	9103	7253	1634	1803	23345	79745	64231	84285	219833	283316	100131320
2078	4773	5523	1278	2296	23208	84735	51703	23224	122953	387751	101928651
7705	7681	8829	2114	1515	51296	89846	26585	23180	171024	285735	100128890
8542	2635	5787	6697	3075	91607	90102	26157	80177	138639	170692	101234262
4121	4647	5151	6778	7439	55803	10788	93663	56895	155038	283349	104266957

Table S5: Table of 264 common MS genes (EntrezIDs)

67	2119	5577	9353	11197	27086	55342	78996	116931	219736	440482
412	2138	5579	9400	11234	27115	55384	79085	120114	222008	641467
670	2259	5688	9446	22854	27315	55531	79145	121260	253152	642236
783	2555	5693	9462	22866	27346	55729	79632	121456	255167	654429
814	2557	5753	9480	23057	27445	55755	79667	125228	261729	654790
1000	2562	5800	9516	23085	29100	55809	79698	127254	283392	677772
1004	2762	5936	9547	23092	29948	56479	79815	127700	283480	677775
1005	2774	6096	9568	23105	29951	56666	79968	134728	284215	729993
1007	2864	6323	9569	23132	50485	56704	80059	134829	284252	780813
1008	2890	6547	9644	23145	50939	57118	80063	138649	285220	100127950
1040	2911	6882	9840	23178	51334	57144	80195	139599	285484	100129696
1123	2967	7111	9851	23228	51361	57168	80742	140767	285905	100130621
1318	3131	7162	9961	23236	51399	57172	81491	145624	285958	100216544
1497	3208	7360	10009	23281	53335	57451	81669	145853	286046	100420155
1593	3736	7415	10129	23329	53829	57484	81849	146713	286826	100422558
1602	3745	7442	10311	23518	53905	57574	83746	148479	317772	100506244
1607	3751	7458	10412	23543	54212	57620	84301	149297	319085	100507043
1612	4017	7761	10499	23551	54477	57633	84451	150084	343990	100507381
1730	4166	7932	10552	23767	54566	57718	84679	151126	346157	100873986
1804	4208	8001	10721	24141	54682	57821	84867	151258	347730	100874113
1812	4338	8395	10777	25759	54715	57835	85377	151354	348487	101410534
1859	4628	8554	10793	25937	54952	60437	90134	151516	349152	101927193
1946	4850	8708	10824	26013	55061	63982	90417	152006	353116	101927972
2044	5121	9079	10891	26050	55154	64061	92715	157574	359822	105370424
2045	5210	9228	10936	26097	55191	64388	93986	157807	387978	
2047	5318	9312	10953	26586	55223	65258	113263	163223	401190	
2115	5530	9315	11113	27020	55257	65982	114880	171523	439921	

Table S6: Table of 294 CA lesion type specific genes (EntrezIDs)

References

1. Hütter, C. V., Sin, C., Müller, F., and Menche, J. (2, 2022) Network cartographs for interpretable visualizations. *Nature Computational Science* 2022 2:2, **2**, 84–89.
2. Sadegh, S., Skelton, J., Anastasi, E., Bernett, J., Blumenthal, D. B., Galindez, G., Salgado-Albarrán, M., Lazareva, O., Flanagan, K., Cockell, S., Nogales, C., Casas, A. I., Schmidt, H. H., Baumbach, J., Wipat, A., and Kacprowski, T. (11, 2021) Network medicine for disease module identification and drug repurposing with the NeDRex platform. *Nat. Commun.*, **12**, 1–12 [PubMed:34824199] [PubMed Central:PMC8617287] [doi:10.1038/s41467-021-27138-2].
3. Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (11, 2003) Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.*, **13**, 2498 [PubMed:14597658] [PubMed Central:PMC403769] [doi:10.1101/gr.1239303].
4. Lazareva, O., Canzar, S., Yuan, K., Baumbach, J., Blumenthal, D. B., Tieri, P., Kacprowski, T., and List, M. (8, 2021) BiCoN: network-constrained biclustering of patients and omics data. *Bioinformatics*, **37**, 2398–2404 [PubMed:33367514] [doi:10.1093/bioinformatics/btaa1076].
5. Ghiassian, S. D., Menche, J., and Barabási, A. L. (4, 2015) A DIseAse MOdule Detection (DIAMOnD) Algorithm Derived from a Systematic Analysis of Connectivity Patterns of Disease Proteins in the Human Interactome. *PLoS Comput. Biol.*, **11**, e1004120 [PubMed:25853560] [PubMed Central:PMC4390154] [doi:10.1371/journal.pcbi.1004120].
6. Levi, H., Elkon, R., and Shamir, R. (1, 2021) DOMINO: a network-based active module identification algorithm with reduced rate of false calls. *Mol. Sys. Biol.*, **17**, e9593 [PubMed:33471440] [PubMed Central:PMC7816759] [doi:10.15252/msb.20209593].
7. List, M., Alcaraz, N., Dissing-Hansen, M., Ditzel, H. J., Mollenhauer, J., and Baumbach, J. (7, 2016) KeyPathwayMinerWeb: online multi-omics network enrichment. *Nucleic Acids Res.*, **44**, W98–W104 [PubMed:27150809] [PubMed Central:PMC4987922] [doi:10.1093/nar/gkw373].
8. Bernett, J., Krupke, D., Sadegh, S., Baumbach, J., ndor Fekete, S. P., Kacprowski, T., List, M., and Blumenthal, D. B. (3, 2022) Robust disease module mining via enumeration of diverse prize-collecting Steiner trees. *Bioinformatics*, **38**, 1600–1606 [PubMed:34984440] [doi:10.1093/bioinformatics/btab876].
9. Sadegh, S., Matschinske, J., Blumenthal, D. B., Galindez, G., Kacprowski, T., List, M., Nasirigerdeh, R., Oubounyt, M., Pichlmair, A., Rose, T. D., Salgado-Albarrán, M., Späth, J., Stukalov, A., Wenke, N. K., Yuan, K., Pauling, J. K., and Baumbach, J. (7, 2020) Exploring the SARS-CoV-2 virus-host-drug interactome for drug repurposing. *Nat. Commun.*, **11**, 1–9 [PubMed:32665542] [PubMed Central:PMC7360763] [doi:10.1038/s41467-020-17189-2].
10. Gyöngyi, Z., Garcia-Molina, H., and Pedersen, J. O. (2004) Combating Web Spam with TrustRank. *Proc. 2004 VLDB Conf.*, pp. 576–587.
11. Bavelas, A. (6, 1950) Communication Patterns in Task-Oriented Groups. *J. Acoust. Soc.*, **22**, 725.
12. Zarin, D. A., Tse, T., Williams, R. J., and Carr, S. (11, 2016) Trial Reporting in ClinicalTrials.gov - The Final Rule. *N Engl J Med*, **375**, 1998–2004 [PubMed:27635471] [PubMed Central:PMC5225905] [doi:10.1056/NEJMSr1611785].
13. Page, L., Brin, S., Motwani, R., and Winograd, T. The PageRank Citation Ranking: Bringing Order to the Web. (1998).
14. Järvelin, K. and Kekäläinen, J. (October, 2002) Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.*, **20**(4), 422–446.

15. Mungall, C. J., McMurtry, J. A., Kohler, S., Balhoff, J. P., Borromeo, C., Brush, M., Carbon, S., Conlin, T., Dunn, N., Engelstad, M., Foster, E., Gouridine, J. P., Jacobsen, J. O., Keith, D., Laraway, B., Lewis, S. E., Xuan, J. N., Shefchek, K., Vasilevsky, N., Yuan, Z., Washington, N., Hochheiser, H., Groza, T., Smedley, D., Robinson, P. N., and Haendel, M. A. (1, 2017) The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Res.*, **45**, D712.
16. Avram, S., Bologa, C. G., Holmes, J., Bocci, G., Wilson, T. B., Nguyen, D. T., Curpan, R., Halip, L., Bora, A., Yang, J. J., Knockel, J., Sirimulla, S., Ursu, O., and Oprea, T. I. (1, 2021) DrugCentral 2021 supports drug discovery and repositioning. *Nucleic Acids Res.*, **49**, D1160–D1169 [PubMed:33151287] [PubMed Central:PMC7779058] [doi:10.1093/nar/gkaa997].
17. Davis, A. P., Grondin, C. J., Johnson, R. J., Sciaky, D., Wieggers, J., Wieggers, T. C., and Mattingly, C. J. (1, 2021) Comparative Toxicogenomics Database (CTD): update 2021. *Nucleic Acids Res.*, **49**, D1138–D1143 [PubMed:33068428] [PubMed Central:PMC7779006] [doi:10.1093/nar/gkaa891].
18. Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., Assempour, N., Iynkkaran, I., Liu, Y., Maclejewski, A., Gale, N., Wilson, A., Chin, L., Cummings, R., Le, D., Pon, A., Knox, C., and Wilson, M. (1, 2018) DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.*, **46**, D1074–D1082 [PubMed:29126136] [PubMed Central:PMC5753335] [doi:10.1093/nar/gkx1037].
19. Piñero, J., Ramírez-Anguita, J. M., Saüch-Pitarch, J., Ronzano, F., Centeno, E., Sanz, F., and Furlong, L. I. (1, 2020) The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res.*, **48**, D845–D855 [PubMed:31680165] [PubMed Central:PMC7145631] [doi:10.1093/nar/gkz1021].
20. Amberger, J. S., Bocchini, C. A., Scott, A. F., and Hamosh, A. (1, 2019) OMIM.org: leveraging knowledge across phenotype–gene relationships. *Nucleic Acids Res.*, **47**, D1038–D1043 [PubMed:30445645] [PubMed Central:PMC6323937] [doi:10.1093/nar/gky1151].
21. Bateman, A., Martin, M. J., O'Donovan, C., Magrane, M., Alpi, E., Antunes, R., Bely, B., Bingley, M., Bonilla, C., Britto, R., Bursteinas, B., Bye-Ajee, H., Cowley, A., Silva, A. D., Giorgi, M. D., Dogan, T., Fazzini, F., Castro, L. G., Figueira, L., Garmiri, P., Georghiou, G., Gonzalez, D., Hatton-Ellis, E., Li, W., Liu, W., Lopez, R., Luo, J., Lussi, Y., MacDougall, A., Nightingale, A., Palka, B., Pichler, K., Poggioli, D., Pundir, S., Pureza, L., Qi, G., Rosanoff, S., Saidi, R., Sawford, T., Shypitsyna, A., Speretta, E., Turner, E., Tyagi, N., Volynkin, V., Wardell, T., Warner, K., Watkins, X., Zaru, R., Zellner, H., Xenarios, I., Bougueleret, L., Bridge, A., Poux, S., Redaschi, N., Aimo, L., ArgoudPuy, G., Auchincloss, A., Axelsen, K., Bansal, P., Baratin, D., Blatter, M. C., Boeckmann, B., Bolleman, J., Boutet, E., Breuza, L., Casal-Casas, C., Castro, E. D., Coudert, E., Cuche, B., Doche, M., Dornevil, D., Duvaud, S., Estreicher, A., Famiglietti, L., Feuermann, M., Gasteiger, E., Gehant, S., Gerritsen, V., Gos, A., Gruaz-Gumowski, N., Hinz, U., Hulo, C., Jungo, F., Keller, G., Lara, V., Lemercier, P., Lieberherr, D., Lombardot, T., Martin, X., Masson, P., Morgat, A., Neto, T., Nospikel, N., Paesano, S., Pedruzzi, I., Pilbout, S., Pozzato, M., Pruess, M., Rivoire, C., Roechert, B., Schneider, M., Sigrist, C., Sonesson, K., Staehli, S., Stutz, A., Sundaram, S., Tognolli, M., Verbregue, L., Veuthey, A. L., Wu, C. H., Arighi, C. N., Arminski, L., Chen, C., Chen, Y., Garavelli, J. S., Huang, H., Laiho, K., McGarvey, P., Natale, D. A., Ross, K., Vinayaka, C. R., Wang, Q., Wang, Y., Yeh, L. S., and Zhang, J. (1, 2017) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **45**, D158–D169.
22. Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., Sidiropoulos, K., Cook, J., Gillespie, M., Haw, R., Loney, F., May, B., Milacic, M., Rothfels, K., Sevilla, C., Shamovsky, V., Shorser, S., Varusai, T., Weiser, J., Wu, G., Stein, L., Hermjakob, H., and D'Eustachio, P. (1, 2020) The reactome pathway knowledgebase. *Nucleic Acids Res.*, **48**, D498–D503.
23. Stark, C., Breitkreutz, B. J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (1, 2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.*, **34**, D535–D539 [PubMed:16381927] [PubMed Central:PMC1347471] [doi:10.1093/nar/gkj109].

24. Kotlyar, M., Pastrello, C., Malik, Z., and Jurisica, I. (1, 2019) IID 2018 update: context-specific physical protein-protein interactions in human, model organisms and domesticated species. *Nucleic Acids Res.*, **47**, D581–D589 [PubMed:30407591] [PubMed Central:PMC6323934] [doi:10.1093/nar/gky1037].
25. Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N. H., Chavali, G., Chen, C., Del-Toro, N., Duesbury, M., Dumousseau, M., Galeota, E., Hinz, U., Iannuccelli, M., Jagannathan, S., Jimenez, R., Khadake, J., Lagreid, A., Licata, L., Lovering, R. C., Meldal, B., Melidoni, A. N., Milagros, M., Peluso, D., Perfetto, L., Porras, P., Raghunath, A., Ricard-Blum, S., Roechert, B., Stutz, A., Tognolli, M., Roey, K. V., Cesareni, G., and Hermjakob, H. (1, 2014) The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.*, **42**, D358–D363 [PubMed:24234451] [PubMed Central:PMC3965093] [doi:10.1093/nar/gkt1115].
26. Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., Åsa Sivertsson, Kampf, C., Sjöstedt, E., Asplund, A., Olsson, I. M., Edlund, K., Lundberg, E., Navani, S., Szigartyo, C. A. K., Odeberg, J., Djureinovic, D., Takanen, J. O., Hober, S., Alm, T., Edqvist, P. H., Berling, H., Tegel, H., Mulder, J., Rockberg, J., Nilsson, P., Schwenk, J. M., Hamsten, M., Feilitzén, K. V., Forsberg, M., Persson, L., Johansson, F., Zwahlen, M., Heijne, G. V., Nielsen, J., and Pontén, F. (1, 2015) Tissue-based map of the human proteome. *Science*, **347**.
27. Maglott, D., Ostell, J., Pruitt, K. D., and Tatusova, T. (1, 2007) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **35**.
28. Compston, A. and Coles, A. (2008) Multiple sclerosis. *Lancet*, **372**, 1502–1517.
29. Hauser, S. L. and Cree, B. A. (12, 2020) Treatment of Multiple Sclerosis: A Review. *Am. J. Med.*, **133**, 1380–1390.e2.
30. 't Hart, B. A., Luchicchi, A., Schenk, G. J., Killestein, J., and Geurts, J. J. (6, 2021) Multiple sclerosis and drug discovery: A work of translation. *EBioMedicine*, **68** [PubMed:34044219] [PubMed Central:PMC8245896] [doi:10.1016/j.ebiom.2021.103392].
31. Absinta, M., Lassmann, H., and Trapp, B. D. (6, 2020) Mechanisms underlying progression in multiple sclerosis. *Curr. Opin. Neurol.*, **33**, 277–285.
32. Frisch, T., Elkjaer, M. L., Reynolds, R., Michel, T. M., Kacprowski, T., Burton, M., Kruse, T. A., Thomassen, M., Baumbach, J., and Illes, Z. (8, 2020) Multiple Sclerosis Atlas: A Molecular Map of Brain Lesion Stages in Progressive Multiple Sclerosis. *Netw Syst Med.*, **3**, 122 [PubMed:32954379] [PubMed Central:PMC7500075] [doi:10.1089/nsm.2020.0006].
33. Elkjaer, M. L., Frisch, T., Reynolds, R., Kacprowski, T., Burton, M., Kruse, T. A., Thomassen, M., Baumbach, J., and Illes, Z. (12, 2019) Molecular signature of different lesion types in the brain white matter of patients with progressive multiple sclerosis. *Acta Neuropathol. Commun.*, **7**, 1–17 [PubMed:31829262] [PubMed Central:PMC6907342] [doi:10.1186/s40478-019-0855-7].
34. Lang, H. L., Jacobsen, H., Ikemizu, S., Andersson, C., Harlos, K., Madsen, L., Hjorth, P., Sondergaard, L., Svejgaard, A., Wucherpfennig, K., Stuart, D. I., Bell, J. I., Jones, E. Y., and Fugger, L. (2002) A functional and structural basis for TCR cross-reactivity in multiple sclerosis. *Nat. Immunol.*, **3**, 940–943 [PubMed:12244309] [doi:10.1038/ni835].
35. Krumbholz, M., Theil, D., Cepok, S., Hemmer, B., Kivisäkk, P., Ransohoff, R. M., Hofbauer, M., Farina, C., Derfuss, T., Hartle, C., Newcombe, J., Hohlfeld, R., and Meinl, E. (1, 2006) Chemokines in multiple sclerosis: CXCL12 and CXCL13 up-regulation is differentially linked to CNS immune cell recruitment. *Brain*, **129**, 200–211 [PubMed:16280350] [doi:10.1093/brain/awh680].
36. Lassmann, H. (3, 2018) Multiple Sclerosis Pathology. *Cold Spring Harb. Perspect. Med.*, **8**.
37. Gallarda, B. W., Bonanomi, D., Müller, D., Brown, A., Alaynick, W. A., Andrews, S. E., Lemke, G., Pfaff, S. L., and Marquardt, T. (4, 2008) Segregation of axial motor and sensory pathways via heterotypic trans-axonal signaling. *Science*, **320**, 233–236 [PubMed:18403711] [PubMed Central:PMC3158657] [doi:10.1126/science.1153758].

38. Fujita, Y., Takashima, R., Endo, S., Takai, T., and Yamashita, T. (9, 2011) The p75 receptor mediates axon growth inhibition through an association with PIR-B. *Cell Death Dis.*, **2** [PubMed:21881600] [PubMed Central:PMC3186903] [doi:10.1038/cddis.2011.85].
39. Goldshmit, Y., McLenachan, S., and Turnley, A. (9, 2006) Roles of Eph receptors and ephrins in the normal and damaged adult CNS. *Brain Res. Rev.*, **52**, 327–345 [PubMed:16774788] [doi:10.1016/j.brainresrev.2006.04.006].
40. Shen, K., Reichelt, M., Kyauk, R. V., Ngu, H., Shen, Y. A. A., Foreman, O., Modrusan, Z., Friedman, B. A., Sheng, M., and Yuen, T. J. (3, 2021) Multiple sclerosis risk gene MERTK is required for microglial activation and subsequent remyelination. *Cell Rep.*, **34**, 108835 [PubMed:33691116] [doi:10.1016/j.celrep.2021.108835].
41. Zhang, S. Z., Wang, Q. Q., Yang, Q. Q., Gu, H. Y., Yin, Y. Q., Li, Y. D., Hou, J. C., Chen, R., Sun, Q. Q., Sun, Y. F., Hu, G., and Zhou, J. W. (11, 2019) NG2 glia regulate brain innate immunity via TGF- β 2/TGFBR2 axis. *BMC Med.*, **17**, 1–22.
42. Zöller, T., Schneider, A., Kleimeyer, C., Masuda, T., Potru, P. S., Pfeifer, D., Blank, T., Prinz, M., and Spittau, B. (10, 2018) Silencing of TGF β signalling in microglia results in impaired homeostasis. *Nat. Commun.*, **9**, 1–13.
43. Rothhammer, V., Borucki, D. M., Tjon, E. C., Takenaka, M. C., Chao, C. C., Ardura-Fabregat, A., Lima, K. A. D., Gutiérrez-Vázquez, C., Hewson, P., Staszewski, O., Blain, M., Healy, L., Neziraj, T., Borio, M., Wheeler, M., Dragin, L. L., Laplaud, D. A., Antel, J., Alvarez, J. I., Prinz, M., and Quintana, F. J. (5, 2018) Microglial control of astrocytes in response to microbial metabolites. *Nature*, **557**, 724–728 [PubMed:29769726] [PubMed Central:PMC6422159] [doi:10.1038/s41586-018-0119-x].
44. Macrez, R., Stys, P. K., Vivien, D., Lipton, S. A., and Docagne, F. (9, 2016) Mechanisms of glutamate toxicity in multiple sclerosis: biomarker and therapeutic opportunities. *Lancet Neurol.*, **15**, 1089–1102 [PubMed:27571160] [doi:10.1016/S1474-4422(16)30165-X].
45. Miedema, A., Gerrits, E., Brouwer, N., Jiang, Q., Kracht, L., Meijer, M., Nutma, E., Peferoen-Baert, R., Pijnacker, A. T., Wesseling, E. M., Wijering, M. H., Gabius, H. J., Amor, S., Eggen, B. J., and Kooistra, S. M. (12, 2022) Brain macrophages acquire distinct transcriptomes in multiple sclerosis lesions and normal appearing white matter. *Acta Neuropathol. Commun.*, **10**, 1–18 [PubMed:35090578] [PubMed Central:PMC8796391] [doi:10.1186/s40478-021-01306-3].
46. Srinivasan, R., Sailasuta, N., Hurd, R., Nelson, S., and Pelletier, D. (5, 2005) Evidence of elevated glutamate in multiple sclerosis using magnetic resonance spectroscopy at 3 T. *Brain*, **128**, 1016–1025 [PubMed:15758036] [doi:10.1093/brain/awh467].
47. Azevedo, C. J., Kornak, J., Chu, P., Sampat, M., Okuda, D. T., Cree, B. A., Nelson, S. J., Hauser, S. L., and Pelletier, D. (2014) In vivo evidence of glutamate toxicity in multiple sclerosis. *Ann. Neurol.*, **76**, 269–278 [PubMed:25043416] [PubMed Central:PMC4142752] [doi:10.1002/ana.24202].
48. Boscia, F., Elkjaer, M. L., Illes, Z., and Kukley, M. (6, 2021) Altered Expression of Ion Channels in White Matter Lesions of Progressive Multiple Sclerosis: What Do We Know About Their Function?. *Front. Cell. Neurosci.*, **15**, 214 [PubMed:34276310] [PubMed Central:PMC8282214] [doi:10.3389/fncel.2021.685703].



DRUGST.ONE
SUPPLEMENTARY MATERIAL

This section contains the original supplementary material published alongside the article:

**“Drugst.One — A plug-and-play solution for online systems
medicine and network-based drug repurposing”**

Drugst.One - A plug-and-play solution for online systems medicine and network-based drug repurposing

Supplementary information

1. Drugst.One ecosystem

The Drugst.One ecosystem consists of the Drugst.One plugin for developers and a content delivery system (CDS) to distribute it, a website, a backend server, and a Python package (Figure 2). Any communication between Drugst.One components is SSL encrypted using HTTPS. We use GitHub for versioning of plugin and backend as well as code example repositories, all of which can be found here: <https://github.com/drugst-one>. For the Drugst.One website, the CDN server, CI/CD pipelines and container registries we, for now, use a GitLab instance hosted by the RRZ, the computing center of the University of Hamburg.

1.1 Plugin

The Drugst.One web plugin visualizes and enriches protein-protein interaction (PPI) networks (Figure S1). Given a set of gene or protein identifiers, or small networks, interaction information can be completed automatically from the Drugst.One database and the users of the website can explore the nodes in their network context. With one click, transitive connections between otherwise unconnected proteins within the interactome can be identified or first-neighbor diseases and drugs can be added to the network. Further, tissue-specific expression of the proteins can be highlighted directly in the network.

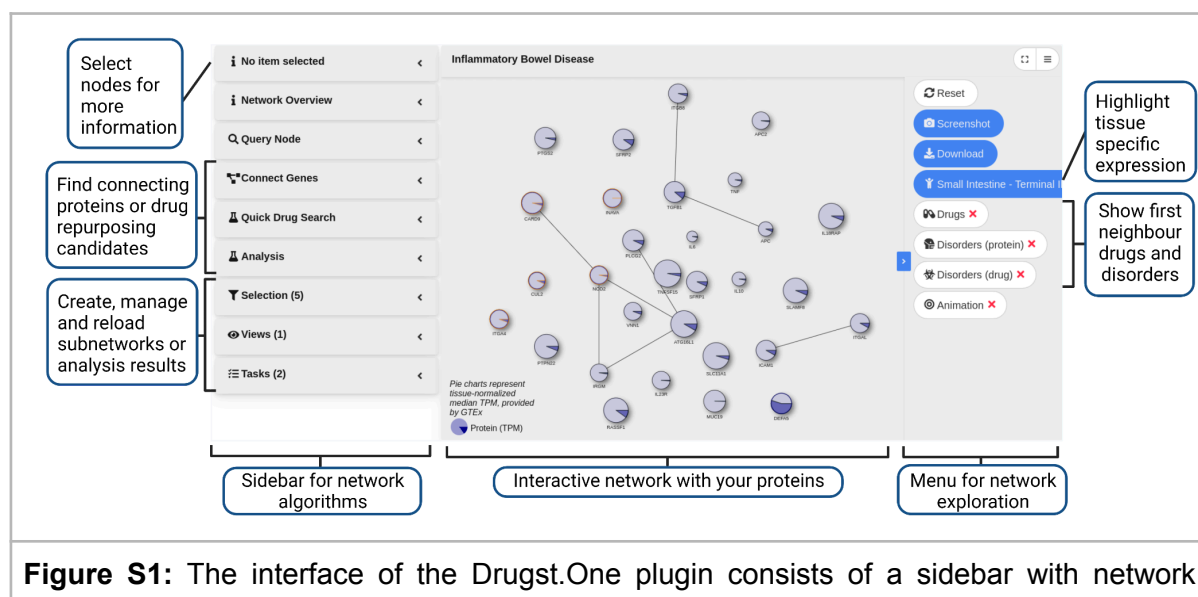


Figure S1: The interface of the Drugst.One plugin consists of a sidebar with network

information and options for drug repurposing algorithms (left), an interactive network visualization (center), and a menu with functions to modify the network (right). Individual components and positions of these elements can be configured. Visualized are the 30 genes from the inflammatory bowel disease use case (section 4 in Supplementary Notes).

Besides exploring the loaded genes or proteins, the plugin can be used to generate drug repurposing hypotheses. Using the explorative functionalities such as highlighting proteins based on their expression in tissues, input proteins for analysis tasks can be selected. Leveraging the power of specialized network algorithms, disease modules can be identified to find additional drug targets within the same genetic context and drug repurposing candidates can be prioritized (Figure S2). Analysis results are stored for future access and can be loaded directly into the interface where a network visualization helps to understand the relation of the result to the input nodes. New analyses can be started on the original network or based on previous analysis results, supporting an iterative approach to a transparent and user-driven drug repurposing.

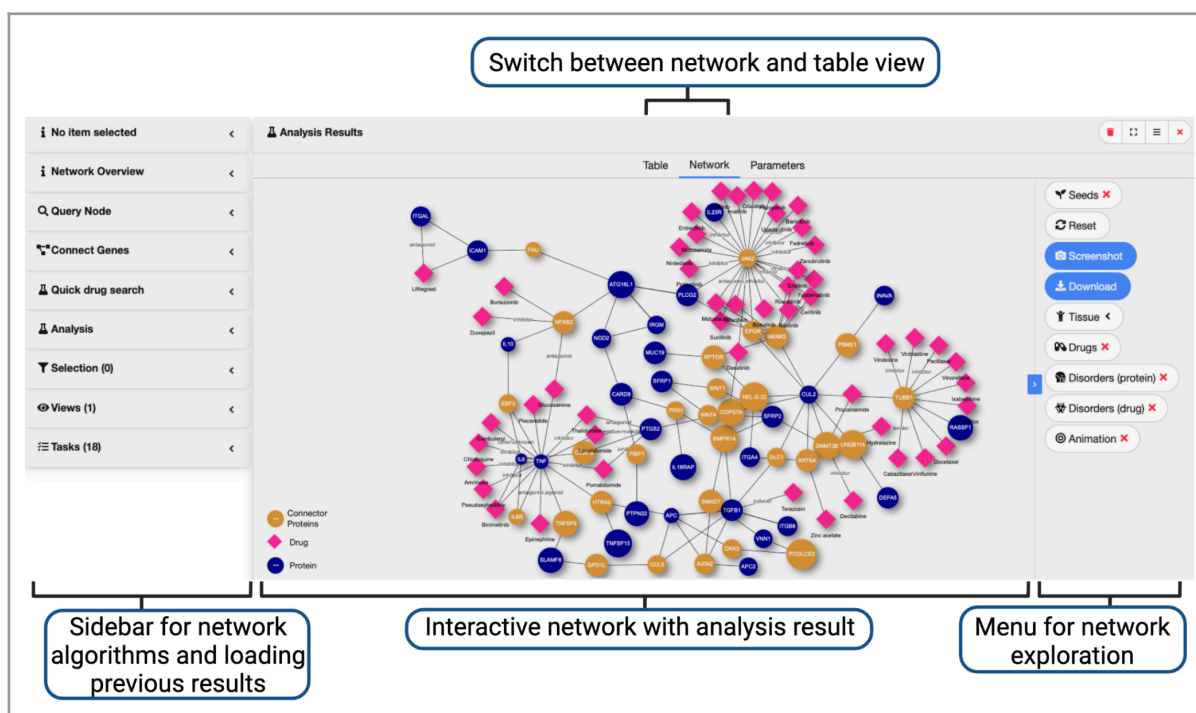


Figure S2: The results of drug target and drug repurposing candidates search for the inflammatory bowel disease use case (Supplements 4) are represented in the Drugst.One plugin as a network. A 'Quick Drug Search' was executed on all proteins; hence, a Multi-Steiner tree was used to connect the genes and a harmonic centrality algorithm to identify putative drug candidates.

Despite all these features the plugin remains completely serverless for hosting websites as all computations take place on the Drugst.One servers. The customizable plugin consists of a JavaScript-based web component, which can be fitted to the needs of the host website through JSON-formatted configuration strings. Functions, buttons, colors, and dimensions of the component can be adjusted to seamlessly blend in with the rest of the page. Style encapsulation guarantees no interference with styles from the host webpage. Lastly, the

Drugst.One plugin is lightweight, no heavy libraries will be added to the host's webpage to minimize loading times.

The latest Drugst.One plugin (at submission v1.1.20) is developed in Angular.js 14 with TypeScript 4.4.4 and compiled and packaged with npm 8.15.0 and node 16.17.1 into drugstone.js and drugstone.css files. For network visualization, vis.js (v9.1.6) is used and screenshot support is provided through dom-to-image-cross-origin (v.2.6.7).

1.2 Content delivery system

To enable the easy use and distribution of the plugin, we set up a simple content delivery system (CDS) to account for a large number of simultaneously loaded plugin instances on multiple hosting web pages. The CDS manages the plugin versions by maintaining all builds of previous plugin versions that can be accessed through fixed version identifiers. The latest stable release is tagged 'latest' and will be updated upon a new stable release. Unstable development versions can be tested using versions tagged as release candidates ('-rc'). The most current development version can be accessed through the tag 'nightly'. A list of all available plugin versions is generated after each build and is available at <https://cdn.drugst.one/>.

1.3 Website

The Drugst.One website (<https://drugst.one/>) serves as a documentation and exploration environment for the Drugst.One plugin. Many examples in the documentation and tutorial videos guide users through the individual features of the Drugst.One plugin as well as its integration.

The Drugst.One playground provides a graphical interface to conveniently generate custom "copy-and-paste" code snippets incorporating style and configuration choices made by the user. Further, a standalone version of the Drugst.One plugin is integrated into the website and can be accessed and modified by passing parameters using GET or POST HTTP requests, allowing developers to use the plugin without having to host a web service or incorporate the plugin in a webpage. Additionally, to facilitate and initiate the development of novel web tools presenting biomedical networks, it provides a website application template with the Drugst.One plugin.

1.4 Server

The Drugst.One backend server consists of a Django API (v3.2.19) connected to a PostgreSQL (v14) database in combination with a redis broker and worker (v7.0.11) and a Celery scheduler (v5.2.7). The backend interacts closely with the plugin, instantly returning information to the loaded proteins and adding the interactions between the proteins to create the PPI network. It further handles the asynchronous execution of the network algorithms in the analysis tasks by starting separate jobs on the redis server, allowing it to handle a large number of jobs in parallel or queue them accounting for the workload of Drugst.One instances on multiple websites. Most network algorithms (Supplement 3.2) are implemented

with graph-tool (version 2.55), a performant C++-based network library for Python, and pre-processed network files, resulting in fast execution times of usually less than 30 seconds.

1.5 Python package

Programmatic access is supported using Python 3.6 and newer. The Python package (<https://pypi.org/project/drugstone/>) can be installed using pip ('pip install drugstone') and supports the main functionalities of the plugin, including fetching protein information, PPIs, and conducting drug repurposing analyses. With the Drugst.One python package, a larger number of tasks can be executed in an automated fashion, then form the GUI of the plugin, empowering developers to integrate Dugst.One into custom workflows of their own programs.

2. Integration and customization

The Drugst.One ecosystem is a multi-component platform consisting of a website, the web plugin, a server, a content delivery system (CDS), and a Python package, as depicted in Figure 1 (for details, see Supplementary Notes 1).

The web plugin can be added to any webpage by importing one JavaScript and one stylesheet file from the <https://cdn.drugst.one> distribution server, and by adding the 'drugst-one' HTML tag to the source code of any system medicine tool's website (Figure S2). Features can be customized to a high degree through JSON configuration strings that are passed as attributes. This includes default states of on/off toggles, the network, and the node and edge groups that define the network style. The plugin is responsive to changes during runtime, allowing developers to add buttons or other controls to the host page, for example switching between networks. For seamless integration of the rendered plugin into any website, styling and coloring are controllable by adding specific CSS variables to the website stylesheet. To assist developers in the integration process, the Drugst.One website provides conclusive documentation of available parameters, features, and styles. It further offers an interactive configuration page at <https://drugst.one/playground> where configuration options are categorized, and the replication of a configured Drugst.One instance is achieved by simple copy-pasting of the generated code snippets to the developers' websites. This low-code approach allows bioinformatics researchers to provide the community with an interactive mechanism mining web tool within hours or even minutes instead of days. The lightweight Drugst.One JavaScript library connects to the Drugst.One data warehouse server, which handles all the computationally expensive work like data annotation, mapping, and asynchronous algorithm execution.

Alternatively, a standalone integration of Drugst.One is provided at <https://drugst.one/standalone>, which can be accessed and customized using URLs or POST-based requests. This way, results from any website or even a command line tool can be redirected to Drugst.One through a simple web service request (Supplementary Notes 2). Detailed documentation about all Drugst.One integration options can be found at <https://drugst.one/doc>.

2.1 Plugin

Developers can add the plugin to their own website using JavaScript. This may help to visualize any results containing a list of proteins or PPIs and to add the Drugst.One functionalities. Integration is done in three steps. Firstly, the Drugst.One libraries need to be loaded with the website by integrating the import statement into the head tag of your website. It is possible to define a specific version identifier to use a static version of the plugin or to use the tag 'latest' to automatically update to the latest stable version upon a new release.

```
<head>  
  <script src="https://cdn.drugst.one/latest/drugstone.js"></script>
```



```
<link rel="stylesheet"
href="https://cdn.drugst.one/latest/styles.css">
</head>
```

Secondly, the Drugst.One component needs to be placed on the website. After loading the Drugst.One libraries, the html tag 'drugst-one' becomes available. The position of the tag defines the position of the component on a webpage.

```
<drugst-one id='drugstone-component-id'></drugst-one>
```

Lastly, the component can be configured by passing options as JSON strings to the three attributes 'groups', 'config', and 'network'. The 'network' parameter can accept lists of nodes and optionally edges to construct the network. Each node and edge in the network is assigned to a node and edge group, respectively, which defines the styles of each member. Additionally, groups are assigned a group name that all its members inherit, e.g. 'protein' or 'drug'. Individual nodes and edges may receive individual styles to highlight them which will override the group styles. Because the network is based on vis.js, all node and edge attributes used in vis.js are applicable, including e.g. directed edge styles. Settings regarding the component itself e.g. to add or remove certain features, or which datasets to use may be passed to the 'config' parameter.

```
<drugst-one
groups='{ "nodeGroups":{ "Protein":{ "type": "Protein", "color": "#ff881f", "font":{ "color": "#ffffff", "groupName": "Protein", "shape": "ellipse" } }, "edgeGroups":{ "PPI":{ "color": "#111111", "groupName": "PPI" } } }'
config='{ "identifier": "symbol", "title": "Breast cancer example network" }'
network='{ "nodes": [ { "id": "BRCA1", "label": "BRCA1", "group": "Protein" }, { "id": "BRCA2", "label": "BRCA2", "group": "Protein" } ], "edges": [ { "from": "BRCA1", "to": "BRCA2" } ] }'>
</drugst-one>
```

The Drugst.One plugin adapts to any changes of the options immediately, allowing developers to add features like buttons, toggles, or selections to e.g. change or adjust the network at runtime.

Custom styling of the component can be achieved by setting global CSS parameters on the website. Through the prefix 'drgstn' it is ensured that CSS variables will not randomly collide with other style variables.

```
:root {
  --drgstn-primary:#347eee;
  --drgstn-secondary:#2e42f2;
  --drgstn-success:#48C774;
```

```
--drgstn-warning:#ffdd00;
--drgstn-danger:#ff2744;
--drgstn-background:#f8f9fa;
--drgstn-panel:#ffffff;
--drgstn-info:#61c43d;
--drgstn-text-primary:#151515;
--drgstn-text-secondary:#e0e0e0;
--drgstn-border:rgba(0, 0, 0, 0.2);
--drgstn-tooltip:rgba(74,74,74,0.9);
--drgstn-panel-secondary:#FFFFFF;
--drgstn-height:600px;
--drgstn-font-family:Helvetica Neue, sans-serif;
}
```

2.2 Standalone

In cases where it is not desired or possible to integrate the Drugst.One component to a website, e.g., because a tool has no website, the version hosted at the Drugst.One website is accessible using HTTP requests. When working with small networks and little customization of the plugin is required, it is sufficient to encode the parameters in GET requests.

```
https://drugst.one?nodes=PTEN,TP53&edges=PTEN%20TP53&autofillEdges=false
```

For the GET-based configuration, only selected parameters are available (https://drugst.one/doc#standalone_api).

Drugst.One buttons are provided to facilitate the integration.

```
<link rel="stylesheet"
href="https://cdn.drugst.one/libs/drugstone-buttons/0.0.1/drugstone-buttons.min.css">

<a class="drugstone-button drugstone-grey"
href="https://drugst.one/standalone?nodes=PTEN,TP53,BRCA2&autofillEdges=true&activateNetworkMenuButtonAdjacentDrugs=true&interactionDrugProtein=NeDRex&licensedDatasets=true" target="_blank">Drugst.One</a>
```

To pass more network data or extensive configuration parameters that are otherwise not available or exceed the URL limit of 2048 characters, POST requests can be used. An API endpoint expects the same options as the plugin and returns an identifier with which the network can be loaded using a GET request (i.e. an URL):

```
(Send options to API and GET identifier)
let networkID = post(
  'https://api.drugst.one/create_network',
  {
    network: {nodes: [...], edges: [...]},
    groups: {...}
    config: {...}
  }
)
```

```
(Load configuration in Drugst.One standalone with a network identifier)
https://drugst.one?id=<networkID>
```

Table S1. Systems medicine tools that integrate Drugst.One listed in alphabetical order. Options are ‘Link-out’, referring to a URL based redirect from the tool or website to the Drugst.One standalone page, ‘Plugin’, referring to the integration of the javascript-based plugin into the web tool, and programmatic access using the ‘Python package’. This table contains all collaboration partners that integrated Drugst.One into their projects.

Tool	URL	Tool Description	Integration	Integration status
BiCoN (1)	https://exbio.wzw.tum.de/bicon/	Network-constrained patient stratification through biclustering	Plugin	Done
DOMINO (2)	http://domino.cs.tau.ac.il/	Active module identification with improved empirical validation	Link-out	Done

Tool	URL	Tool Description	Integration	Integration status
G-Browser	https://exbio.wzw.tum.de/genome-browser/	An enhanced genome browser plugin that seamlessly integrates data sources and functions for genetics research.	Plugin	Done
GraphFusion	https://github.com/CarlosJesusGH/GraphFusion	An intuitive web-based graph analytics, fusion, and visualization tool	Plugin	Done
GraphSimViz (3)	https://graphsimviz.net/	Visualization of diseasesomes, drugomes, and drug-disease networks	Plugin	Done
HitSeekR (4)	https://exbio.wzw.tum.de/hitseekr/	User-friendly tool for drug (target) identification in high-throughput screening	Plugin	Done
Interactive Enrichment Analysis (5)	https://github.com/gladstone-institutes/Interactive-Enrichment-Analysis/	Enrichment analysis on multiple public datasets	Link-out	In progress
mirDIP (6)	https://ophid.utoronto.ca/mirDIP/	Integrated microRNA-target data integration portal	Plugin	Done
NAViGaTOR (7)	https://ophid.utoronto.ca/navigator/	Network visualization and analysis software	Link-out	In progress
NDEx IQuery (8)	https://www.ndexbio.org/iquery/	Web tool for pathway and network-based gene set analysis	Plugin	Planned
NeEDL - Epistasis Disease Atlas (9)	https://epistasis-disease-atlas.com	Web resource to visualize, investigate, and interpret higher-order genetic interactions of single nucleotide polymorphisms in 18 human heritable diseases.	Link-out	Done
NeEDL - R Shiny App (9)	https://hub.docker.com/r/bigdatainbio/medicine/ne	R shiny app to visualize, investigate, and interpret higher-order genetic interactions of single	Plugin	Done

Tool	URL	Tool Description	Integration	Integration status
	edl	nucleotide polymorphisms on locally computed datasets.		
openPIP (10)	https://github.com/BaderLab/openPIP	Open platform to store and retrieve protein-protein interaction datasets.	Link-out	Done
pathDIP (11)	https://ophid.utoronto.ca/pathDIP	Integrated pathway database and pathway enrichment analysis portal	Plugin	Done
Pathway Figure OCR (12)	https://pfocr.wikipathways.org	Platform for browsing pathway information extracted from published figures.	Link-out	Done
ProHarMeD (13)	https://apps.cosy.bio/proharmed	Closing the gap between (harmonized) proteomics results and mechanotyping / drug repurposing	Plugin	Done
ROBUST-Web (14)	https://robust-web.net/	ROBUST is a disease module identification tool.	Plugin	Done
SCANet (15)	https://pypi.org/project/scanet/	SCANet is an all-in-one package for single-cell profiling covering the whole differential mechanotyping workflow, from inference of trait/cell-type-specific gene co-expression modules to mechanistic drug repurposing candidate prediction.	Python package	Done
Seed Connector Algorithm (16)	https://github.com/bwh784/SCA	Identification of network modules by adding a minimal number of edges between the seed nodes.	Link-out	Done
UnPaSt	https://apps.cosy.bio/unpast	Visualizer and context explorer for unsupervised expression data bicluster results.	Plugin	Done
WikiPathways (17)	https://wikipathways.org	Platform for browsing and visualizing pathways.	Link-out	Done

3. Methods

3.1 Data integration

Table S2. List of association types and source databases stored in the Drugst.One data warehouse. The version refers to the latest state before submission. *Databases with an asterisk are integrated as a part of NeDRexDB.

Source	Version	Layers	Licensed
APID (18)	January 2019	Protein-Protein	no
BioGRID* (19)	2023-07-03	Protein-Protein	no
ChEMBL (20)	27	Protein-Drug	no
CTD* (21)	2023-07-03	Drug-Disorder	no
DGIdb (22)	4.2.0	Protein-Drug	no
DisGeNET*(23)	2023-07-03	Protein-Disorder	no
DrugBank (24)	5.1.8	Drug-Disorder	yes
DrugBank* (24)	2023-07-02	Protein-Drug	yes
DrugCentral* (25)	2023-07-03	Protein-Drug, Drug-Disorder	no
GTEx (26)	v8	Tissue Expression	no
IID* (27)	2023-07-03	Protein-Protein	no
IntAct* (28)	2023-07-03	Protein-Protein	no
NeDRex (29)	2.10.0	Protein-Protein, Protein-Drug, Protein-Disorder, Drug-Disorder	yes
NeDRex (29)	2.10.0	Protein-Protein, Protein-Drug, Protein-Disorder, Drug-Disorder	no
OMIM* (30)	27-12-2022	Protein-Disorder	yes
STRING (31)	11.0	Protein-Protein	no

3.2 Algorithms

Seven network mining algorithms are implemented in Drugst.One for module identification and/or drug prioritization. To keep the plugin lightweight and easy to use, algorithms were deliberately chosen due to their focus on basic network properties like degree centrality and network proximity. Depending on the type of network, e.g. sparsely or densely connected, users can try out different approaches to explore the search space of related drug targets and drug repurposing candidates. While all of the algorithms share some general parameters (Supplement 3.2.1), some of the algorithms may offer additional settings (Supplement 3.2.2 - 3.2.9).

3.2.1 General parameters

For each algorithm the following options exist:

- Result Size. The number of returned nodes (drug targets or drugs)
- Maximum Degree. Option to filter out hub genes. If set to > 0 , genes with a network degree in the complete gene-drug interaction network above this threshold will be excluded.
- Hub Penalty. Penalizes genes with a large degree in the network.
- Filter Edges. If set, only the shortest paths to the drug target or drug nodes will be displayed in the resulting network. Otherwise, all found pathways will be shown.

3.2.2 Betweenness centrality

Betweenness is obtained by finding the shortest paths for each pair of nodes in the network and assessing the number of shortest paths that pass through a particular node, such that a measure of the centrality of a node in a network global context is received. Betweenness centrality has been established as a common measurement in network biological applications (32) and is especially practical in finding communities in large networks (33). In Drugst.One, betweenness is based on the shortest paths between the seed nodes only and can be used to find drug targets with maximized connectivity to all seeds.

3.2.3 Harmonic centrality

Harmonic centrality (C_h) measurement can be described as the average shortest distance from each node to all other nodes in a network. This measurement is the equivalent of harmonic centrality for disconnected graphs. Formally speaking, it can be annotated as

$$C_h = \sum_{y \neq x}^y \left(\frac{1}{dist(x, y)} \right)$$

where x is a given node and $\frac{1}{dist(x, y)} = 0$ if $dist(x, y) = \infty$ (34). The closer a node is to other nodes, the higher the score. It has already been proven successful in a number of biological network problems for instance with metabolic or PPI networks (35, 36).

3.2.4 Degree centrality

Degree centrality (C_d) measurement is obtained by ranking the nodes in a network based on their degree, which is defined as the number of neighbors a node has divided by the total number of nodes in the network. It can be described as

$$C_d(x) = \frac{deg(x)}{n}$$

where x is a given node and $deg(x)$ is its degree. While it is a commonly used network analysis technique, it most importantly has been shown useful in the identification of essential proteins in PPI networks (37, 38). Thus, it is a simple approach for classifying the network-related importance of a particular protein. In Drugst.One, it can be used to discover valuable drug targets or drugs, based on the seed selection given by the user. However, ranking by node degree is prone to introduce research bias.

3.2.5 KeyPathwayMiner

KeyPathwayMiner (KPM) is an online tool developed by Alcaraz et al. for pathway enrichment analysis (39). Users can utilize KPM for their drug target search by selecting seed genes from the network and letting KPM find an interaction network of genes spanned by the seed genes. The resulting genes are functionally related to the seed nodes and therefore are suitable drug target candidates. Only one parameter k has to be set by the user, which defines the amount of permitted intermediate nodes that are neither part of the seed nodes nor the common pathway.

- Additional proteins k : Number of intermediate nodes allowed between the seed nodes.

3.2.6 Multi-Steiner tree

The Multi-Steiner tree algorithm (40) approximates the minimum spanning tree connecting the seed nodes in a reasonable time. The implementation is adopted from Ahmed et al. (41). It can be used to create a minimum spanning subnetwork between user-selected seed nodes, which happen to be central interaction partners between the seed nodes and thus represent favorable drug targets. The user can specify the number of Steiner trees computed to approximate a minimum spanning tree, and the tolerance indicating how much the subsequent trees may increase the number of edges a higher number of Steiner trees leads to more variations at the cost of a longer runtime.

3.2.7 Network proximity

As introduced in Sadegh et al. (40) we adopted the modified version of the originally proposed network-based proximity metric proposed by Guney et al. (42).

Network proximity is defined as the average length of shortest paths from target protein nodes of a drug to a set of disease nodes, called seeds. A background distribution is computed by creating a specified number (default 32) of random drug target sets and a

specified number of random seed sets (default 32). Drug target set sizes are chosen to represent the observed drug degrees in the drug-target network, differing from the original bin-based degree-preserving set randomization. The average shortest path between each observed drug-target set and the input seed set is compared to the background distribution to compute a statistical significance score (z-core) using mean and standard deviation of the background.

This algorithm was adopted in Drugst.One such that highest-scoring drugs are returned to the user as candidate drugs.

- Number Random Seed Sets: Default 32. Sets the number of random seed sets generated to establish the background distribution.
- Number Random Drug Target Sets: Default 32. Sets the number of random drug target sets generated to establish the background distribution.

3.2.8 TrustRank

TrustRank (43) is based on the same concepts as the Google PageRank algorithm and harmonic centrality (44, 45). A trust score is propagated through the network starting at the seeds, damping the score based on the distance traveled. The user can set the damping factor in a range from 0-1, with a higher damping factor causing the propagation to either stop at nodes in close proximity or in larger portions of the network. In Drugst.One, TrustRank is used to identify putative drug targets as well as drug candidates.

- Damping Factor: Correlates with the distance a trust score propagates through the network. The larger the factor, the larger the proportion of the network that is considered.

3.2.9 Algorithm applications

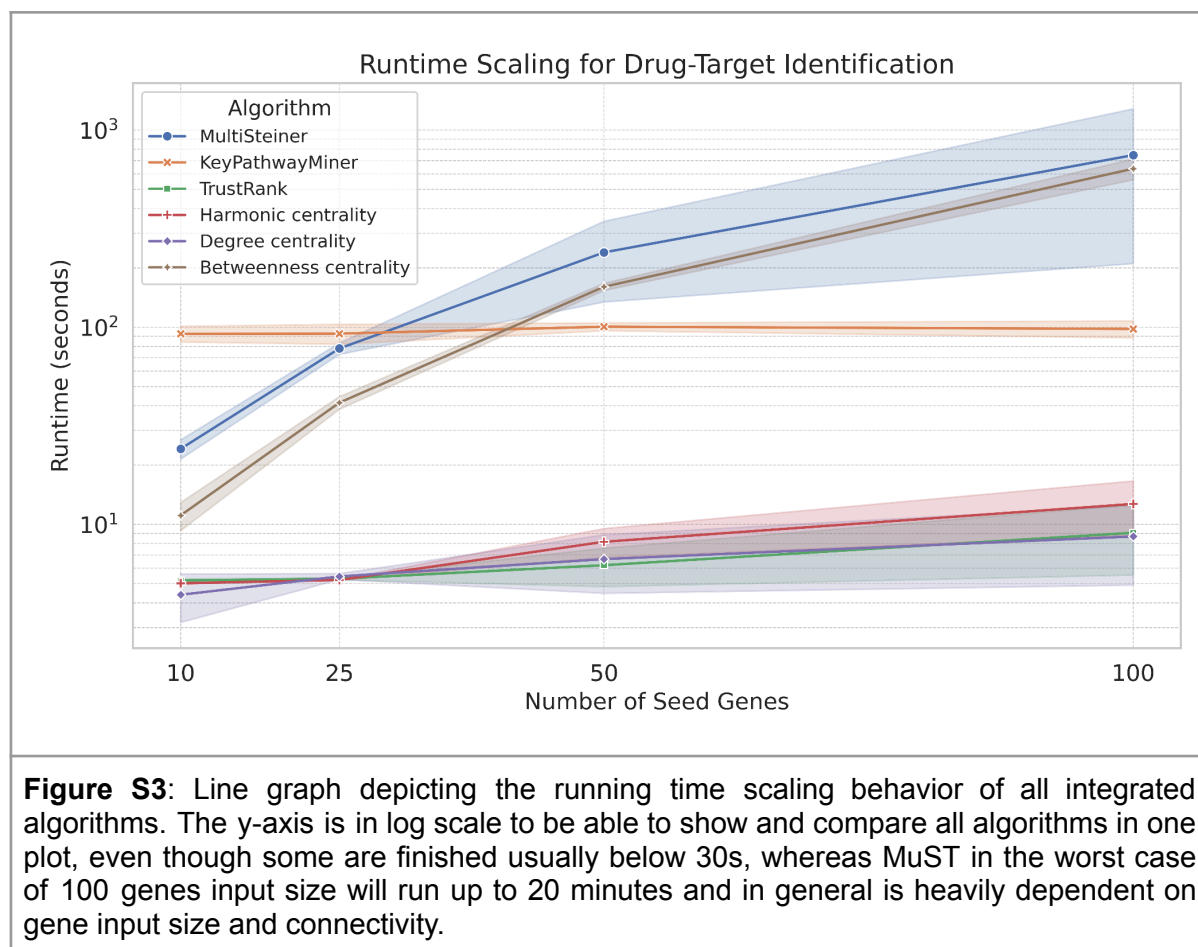
Table S3. An overview of all integrated algorithms regarding their availability in drug target and drug search functions.

Name	Drug target search	Drug search
Betweenness centrality	yes	no
Harmonic centrality	yes	yes
Degree centrality	yes	yes
KeyPathwayMiner	yes	no
Multi-Steiner tree	yes	no
Network proximity	no	yes
TrustRank	yes	yes

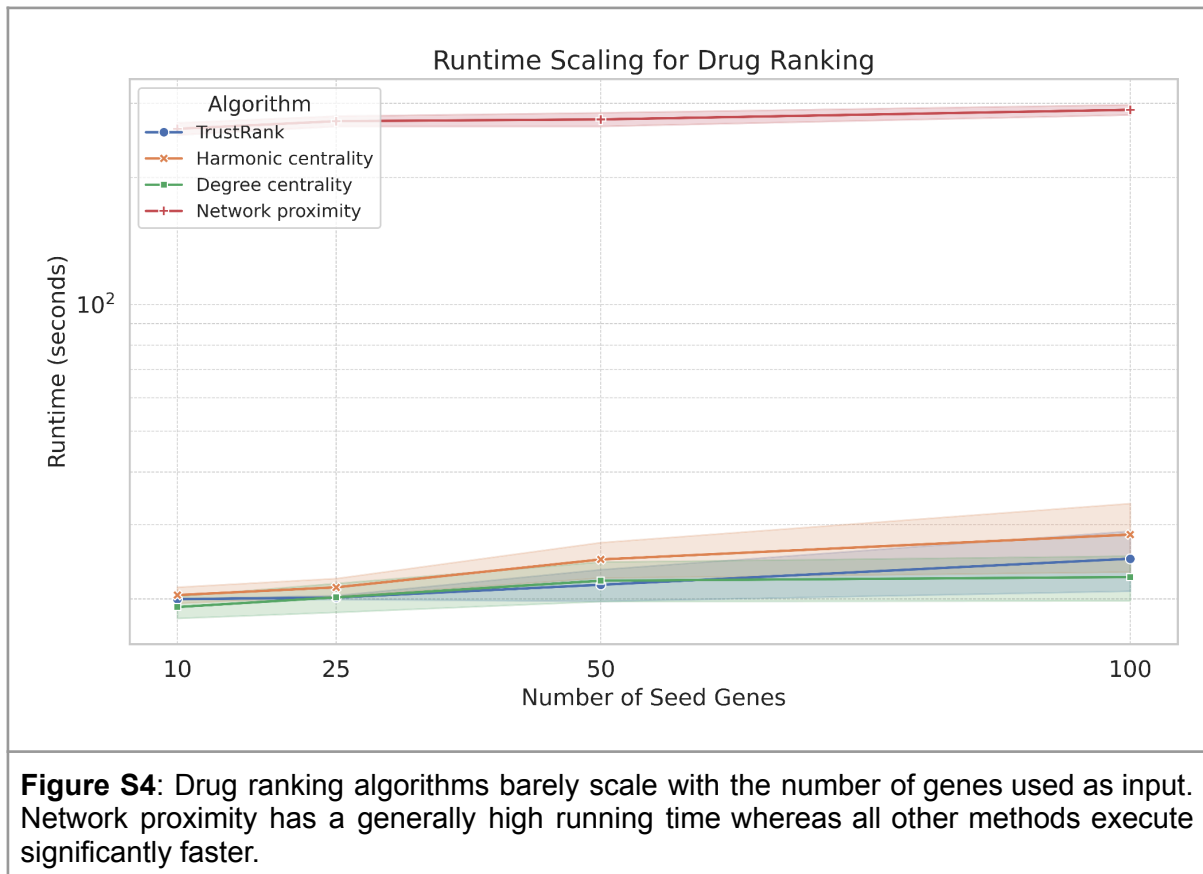
3.3 Algorithm Runtime Scaling

To be able to give users some additional guidance for selecting algorithms, we conducted some general runtime scaling behavior analysis for all integrated algorithms. For this, we selected different seed set sizes (10, 25, 50, 100 genes) as input complexity. To have realistic, yet diverse seed sets, we used NeDRexDB and identified 5 diseases for each step that have exactly or a little more genes annotated. We selected different types of diseases (cancerous, neurological, ...) to get a rough cross-section through different potential fields of applications including most likely different distribution of the seed genes within the interactome, influencing the running time of algorithms. To assume the worst case, we selected the largest available networks (NeDRex - licensed) and used the Drugst.One Python package to run and time the execution duration. The results are plotted as line graphs, with error areas indicating the variance between the different cases of the same seed set size.

For the drug-target identification step, quite different behaviors can be seen (Fig. S3). Harmonic centrality, Degree centrality, and TrustRank are slightly dependent on the input size, but changes are from ~4s for 10 seeds to ~15s for 100 seeds. KeyPathwayMiner runtime is steady at around 100s for any input size, whereas MuST and Betweenness centrality heavily scale with input size taking up to 10 or 20 minutes at 100 genes. Further, it can be seen that MuST experiences by far the greatest variation in execution speeds which tells us that it heavily depends on how seeds are distributed across the PPI.



The scaling behavior of drug ranking given different sizes of seed sets is mainly uneventful and generally independent of input size (Fig. S4). Network proximity always takes around 4-5 minutes to finish whereas the other algorithms, Closeness centrality, Degree centrality, and TrustRank, run only for ~20-30 seconds to return a ranked list.



All seed files, scripts, and results can be found in the following GitHub repository:
<https://github.com/drugst-one/algo-running-times>

4. Use Case: IBD

This use case was created on the weekly updated version of Drugst.One and thus results might differ slightly in the newest Drugst.One database version.

4.1 Repurposing of JAK inhibitors against IBD

A central aspect of Drugst.One is the focus on *in-silico* drug repurposing candidate prediction. To this end, we replicate an exemplary repurposing case study for inflammatory bowel disease (IBD).

Sadegh et al. (29) identified fostamatinib, ruxolitinib, and imatinib for application in IBD by starting from 30 seed genes associated with IBD according to DisGeNET (23) and OMIM (30) (Figure S1, Table S4). Multi-Steiner tree (MuST) (40) was applied to connect the seed genes in the network and the closeness centrality (CC) algorithm was employed to identify drug repurposing candidates (Figure S2).

To reproduce this example case in Drugst.One, we loaded the 30 seed genes associated with IBD into Drugst.One (Table S4). As protein-protein and drug-target interaction datasets the licensed NeDRex versions were used and we executed a 'Quick Drug Search', consisting of a 'Drug target search' using MuST (trees=5, tolerance=5, hub penalty=0.5) to connect all seeds (Table S5). Drugs were ranked using harmonic centrality, being the closest to CC used in the original paper, and the top 50 drugs were chosen as the most promising candidates. In the referenced paper, the authors identified fostamatinib, ruxolitinib, and imatinib on ranks 1, 5, and 12 respectively as drugs that have literature support for being relevant for IBD. With Drugst.One, the same drugs are found at ranks 4, 4, and 9, (Table S6). The difference might stem from small variations in the interactome and drug-target data or the use of harmonic centrality instead of closeness centrality. The drugs target *JAK2*, a gene added by MuST, with ruxolitinib and fostamatinib being known JAK inhibitors (JAKi), and their potential for IBD treatment is currently under investigation (46). Further, they inhibit *MKNK2*, another gene identified by MuST and investigated for its role in different types of colitis (47). JAKis interact with *MAPK*, a gene well-known for its role in IBD (48, 49), with which the observed effect of a dysregulated *MKNK2* might be explained. Tofacitinib, another JAKi and ranked second by HC, has the same targets as fostamatinib and ruxolitinib and is subject to studies investigating beneficial effects in IBD (50, 51).

Dasatinib, the tyrosine kinase inhibitor on rank one (Table S6), is known to induce ulcerative colitis, part of the IBD umbrella, in some patients (52, 53). Even though this drug does not have the desired effect, it is directly associated with the targeted disease and its pathways and may lead to the identification of impactful targets.

Rank two is shared by three immunomodulatory drugs thalidomide, pomalidomide and lenalidomide, as well as glucosamine. Thalidomide has known indications for ulcerative colitis, a subtype of IBD, while also lenalidomide and pomalidomide showed protective effects against IBD in mouse models (54). Lopez-Millan et al. identified lenalidomide as the more potent option, suggesting its usage as a therapeutic drug against inflammatory diseases. There is evidence of beneficial effects of these immunomodulators in human IBD, but their use in clinical practice is under discussion due to severe adverse side effects (55, 56).

Binimetinib is ranked third, which is mainly used as an anti-cancer drug. In this application, inflammatory colitis was observed as an adverse side effect (57), hinting towards a cause-effect relationship that can be studied further to research the mechanistic origins of IBD.

On rank four, together with ruxolitinib and fostamatinib, other inhibitors with the same targets (*MKNK2* and *JAK2*) are found, namely tofacitinib (FDA-approved for ulcerative colitis), sunitinib, midostaurin, erlotinib, and ceritinib.

On ranks two to five, tumor necrosis factor- α (*TNF*) inhibitors are found. *TNF* is a known target to treat IBD (58, 59) and thus effects of the drugs chloroquine on ulcerative colitis (60) and plecanatide on colitis symptoms in mice (61, 62). Further, epinephrine, pseudoephedrine, and clenbuterol share rank five, all have *TNF*-inhibiting effects (63–65) and appear to have general anti-inflammatory effects (66), most likely induced by downregulated or inhibited *IL-6* expression (65, 67).

Rank six terazosin, a drug inhibiting *TGFB1*, whose dysregulation is closely linked to IBD (68).

In summary, with Drugst.One we were not only able to re-identify three promising repurposing candidates from a previous study for IBD, but show that the first 5 ranks (that includes 20 drugs) contain valid candidates or already approved drugs for IBD treatment, including hypotheses about their molecular relationship with IBD.

The following 30 seeds are IBD-associated genes that were used by Sadegh et al. (29) to build their use case.

Table S4. List of the 30 IBD-associated genes used by Sadegh et al. (29) in their drug repurposing case study. This list serves as input for an example use case highlighting the potential of Drugst.One.

The parameters that can be used to replicate the use case, may be found in Table S5.

ATG16L1	ICAM1	TNF	SFRP2	APC2
IL10	CUL2	SFRP1	TNFSF15	ITGA4
DEFA5	MUC19	SLAMF8	APC	IL6
INAVA	CARD9	ITGB8	IL23R	NOD2
RASSF1	SLC11A1	IL18RAP	TGFB1	IRGM
PLCG2	PTPN22	PTGS2	VNN1	ITGAL

Table S5. The Drug-Protein, as well as the Protein-Protein datasets, used for the IBD drug repurposing use case, were set to the most complete one (NeDRex - licensed). For reproducibility, the exact parameters used in drug target and drug identification steps are listed.

Drugst.One	
Protein-Protein interaction dataset	NeDRex (licensed)
Drug-Protein (target) interaction dataset	NeDRex (licensed)
Drug target search	
Algorithm	MuST
Number of Steiner trees	5 (default)
Tolerance for trees	5
Hub penalty	0.5
Drug search	
Algorithm	Harmonic centrality
Result size	50

Table S6. The top 50 drugs resulting from the 'Quick Drug Search' in the IBD drug repurposing use case. The search was conducted on the NeDRex dataset (licensed). Listed are the drugs with their respective score (normalized) as returned by algorithm as well as the assigned rank.

Drug	Score	Rank
Dasatinib	1	1
Lenalidomide	0.9882636549013167	2
Thalidomide	0.9882636549013167	2
Glucosamine	0.9882636549013167	2
Pomalidomide	0.9882636549013167	2
Binimetinib	0.9824981860163579	3
Erlotinib	0.9600936780691096	4
Sunitinib	0.9600936780691096	4
Ruxolitinib	0.9600936780691096	4
Tofacitinib	0.9600936780691096	4
Midostaurin	0.9600936780691096	4
Fostamatinib	0.9600936780691096	4
Ceritinib	0.9600936780691096	4
Chloroquine	0.9600936780691095	5
Epinephrine	0.9600936780691095	5
Clenbuterol	0.9600936780691095	5
Pseudoephedrine	0.9600936780691095	5
Amrinone	0.9600936780691095	5
Plecanatide	0.9600936780691095	5
Terazosin	0.9492702822172273	6
Bortezomib	0.9334851363762997	7
Donepezil	0.9334851363762997	7
Lifitegrast	0.9283394387790588	8
Pralsetinib	0.9232501599628847	9
Nilotinib	0.9232501599628847	9
Niclosamide	0.9232501599628847	9
Fedratinib	0.9232501599628847	9
Imatinib	0.9232501599628847	9
Bosutinib	0.9232501599628847	9
Crizotinib	0.9232501599628847	9
Nintedanib	0.9232501599628847	9
Upadacitinib	0.9232501599628847	9

Entrectinib	0.9232501599628847	9
Pazopanib	0.9232501599628847	9
Zanubrutinib	0.9232501599628847	9
Axitinib	0.9232501599628847	9
Baricitinib	0.9232501599628847	9
Procainamide	0.9083117074801125	10
Decitabine	0.9083117074801125	10
Hydralazine	0.9083117074801125	10
Zinc acetate	0.9083117074801123	11
Vinflunine	0.8986184380435127	12
Ixabepilone	0.8986184380435127	12
Vinblastine	0.8986184380435127	12
Podofilox	0.8986184380435127	12
Paclitaxel	0.8986184380435127	12
Docetaxel	0.8986184380435127	12
Cabazitaxel	0.8986184380435127	12
Vindesine	0.8986184380435127	12
Vinorelbine	0.8986184380435127	12

4.2 Drug candidate identification and mechanism mining through microRNA targets

The database of human microRNA (miRNA) target predictions, mirDIP (version 5.3.0.1, database version 5.2.3.1) (6), is a resource for miRNA-based regulation information. mirDIP allows the identification of gene regulation through miRNAs while avoiding a prediction bias. Its unidirectional search function can be used to either find miRNAs targeting a given set of genes or to retrieve the set of targeted genes given a set of miRNAs. In both cases, mirDIP now offers the option to visualize and explore the used or found genes using the Drugst.One plugin.

We took 30 IBD-associated genes from Sadegh et al. (29) (Table S4) as mirDIP input to identify all known or predicted miRNA regulators ('miRNA-gene matrix' -> 'Search gene symbols'). Three microRNAs (*hsa-miR-142-3p*, *hsa-miR-3942-5p*, and *hsa-miR-574-3p*) were deemed to be main regulators of IBD genes because they each target three IBD-associated genes. Interestingly, *hsa-miR-142-3p* levels have been found to be elevated in the saliva of ulcerative colitis (UC) patients (69). For these microRNAs, we identified the targets using mirDIP unidirectional search and loaded the list of all targeted genes into the Drugst.One plugin to identify relevant drugs and disorders (Figure S5).

Using the first neighbor drug and first neighbor drug-disease association annotation in DrugSt, a number of drugs can be found, that have indications for UC but are not targeting any of the 30 IBD-associated genes but other genes that are regulated by at least one of the three miRNAs. The drugs are sulfadiazine, azathioprine, methylprednisolone, cortisone acetate, budesonide, prednisone/prednisolone, sulfasalazine, loperamide, and hydrocortisone. These drugs target seven genes, namely *NR3C1*, *CALM1*, *RAC1*, *SLC7A11*, *HTR2A*, *SCN3A*, and *GRIN2A*.

Comorbidity between UC, or any other IBD disorder, and diseases associated with those seven genes would give an indication for a shared underlying mechanism that can be dysregulated by the investigated microRNAs. Comorbidities could not be found, even though evidence for their participation in UC, Crohn's disease, or general IBD exists (*NR3C1* (70), *RAC1* (71), *SLC7A11* (72), *HTR2A* (73)). Some indications are given by carvedilol, a hypertensive drug. It has (pre-)clinical implications for IBD and targets both *HTR2A* and *SLC7A11* (74, 75).

This leaves room for further investigation of the mechanistic role of identified genes, especially *SLC7A11* and *HTR2A*, in IBD.

A

mirDIP : microRNA Data Integration Portal

[Search mirDIP](#)
[miRNA-gene matrix](#)
[Novel annotation](#)
[miRAnno](#)
[Search Tissues](#)
[Tissues matrix](#)
[Information](#)
[mirDIP API](#)
[About](#)

Human miRNA-gene interaction (adjacency) matrix
Search miRNAs targeting a group of genes.

Gene Symbols

ATG16L1
ICAM1
TNF
SFRP2
APC2
IL10
CUL2
SFRP1
TNFSF15

[Verify Genes](#)
[List all available Genes](#)

Input :

- You may enter a list of HUGO Gene symbols (minimum 2, case sensitive) delimited by: spaces, tabs, commas or semicolons. Search terms are not case sensitive.

Output :

- Search results will be presented as a table where "Y" indicates that a gene is targeted by a microRNA.
- Missed genes' column (rightmost column) shows the number of genes that are not targeted by the microRNA.

Tips :

- When you are searching 20+ genes, it is preferable to download results (CSV or TAB format) instead of displaying them in the browser.
- Retrieval times for sets of 200+ genes may exceed several minutes.

Example search :

- Gene symbol : [ADGRL2, AFF1, AHR, AKT3, BRD4, CFI3, OTX1, ZZZ](#)

Search Options

Score class
Minimum Score: Very High

Use this option to filter Score Class.

prioritizing

B

Search Results (74):

Link	miRNAs (Gene (Uniprot))	APC (P25054)	APC2 (Q95886)	ATG16L1 (Q67845)	CARD9 (Q9H257)	CUL2 (Q13617)	DEFB4 (Q61532)	...	IL23R (Q6VWK5)	IL8 (P06231)	ILNAU (Q3KPE6)	...	ITGAL (P29701)	ITGB6 (P26912)	MUC1B (Q7Z9P9)	...	PTN22 (Q9Y2R2)	RASSF1 (Q9HSZ3)	SFRP1 (Q84749)	SFRP2 (Q84HF2)
M	hsa-miR-142-3p	Y		Y										Y						
M	hsa-miR-3942-5p	Y				Y														Y
M	hsa-miR-574-3p					Y				Y							Y			

C

Gene Symbol

[Verify Genes](#)
[List all available Genes](#)

or

miRNAs

hsa-miR-142-3p
hsa-miR-3942-5p
hsa-miR-574-3p

[Verify miRNAs](#)
[List all available miRNAs](#)

Search :

- You may enter a list of HUGO Gene Symbols (case sensitive) or MicroRNAs (case sensitive) delimited by: spaces, tabs, commas or semicolons.

Integrated score :

Combines confidence scores from all available predictions.

Results download :

- We support Chrome, Edge or Firefox browsers.

Example search :

- Gene symbol : [APFMARK4, C17orf1, PRKY, ZLFP3, YG11](#)
- MicroRNA symbol : [hsa-miR-493, hsa-miR-7a-3p, hsa-miR-412-5p, hsa-miR-7582-3p, hsa-miR-409](#)

If you are looking for individual resources results, please use [Bibliographical Search](#) tab

Search Options

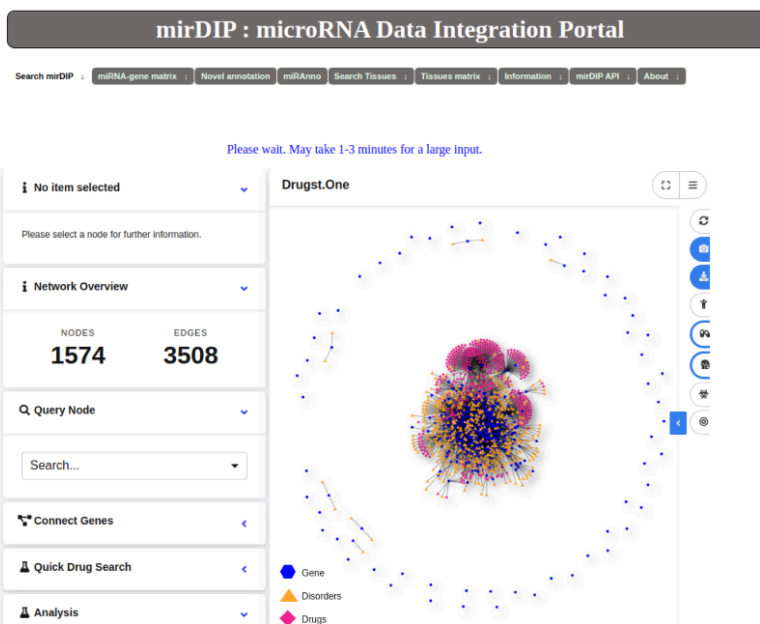
Score class
Minimum Score: Very High

Score class	
Top 1%	Very High
Top 5%	High
Top 1/3	Medium
Bottom 2/3	Low

Use this option to filter Score class.

filtering

D



context exploration

Figure S5: Workflow using mirDIP portal with Drugst.One plugin to identify microRNA targets for the 30 IBD genes (Table S4), and related drug targets. mirDIP was queried to identify regulating miRNAs, using all databases (**A**). 74 microRNAs were identified, of which only three regulate three IBD-associated genes each (**B**). Using the three microRNAs as a query to mirDIP (**C**) identifies 534 target genes. This set of genes is used to interrogate Drugst.One (**D**).

5. Use Case: Plugin integration with ROBUST-Web

ROBUST-Web (<https://robust-web.net>) presents a modified version of ROBUST (76) in an online web interface. It provides a network-based disease module identification algorithm based on prize-collecting Steiner trees that mitigates study bias using edge costs derived from study-attention or bait-usage information. Given a set of seed genes and a PPI network, ROBUST-Web constructs disease modules and passes nodes and edges to the Drugst.One plugin that takes care of result presentation and visualization in an interactive network view. Drugst.One also serves as a network explorer for the analysis of modules by offering an estimation of functional coherence with DIGEST (77), GO enrichment with g:Profiler, or a lookup in NDEX IQuery for identifying pathways with the same participants. Additionally, it adds disease annotations and drug repurposing functions to make the results of ROBUST-Web more actionable and derive hypotheses for follow-up research.

6. Other Collaborations

The Drugst.One initiative has collaborations with other projects to further extend its capabilities.

Table S7. Additional already initiated collaborations that are upcoming improvements to the Drugst.One platform. These are collaborations that concern the Drugst.One platform directly through the integration of standardized frameworks within the network medicine community.

Tool	URL	Tool Description	Collaboration
BioCypher (78)	https://biocyphe.r.org	Drugst.One uses BioCypher to facilitate the extensibility of the database and offer additional datasets, e.g. Omnipath	Integration in Drugst.One data warehouse
NDEx IQuery (8)	https://www.ndexbio.org/iquery/	Web tool for pathway and network-based gene set analysis. Allows Drugst.One users to search for curated pathways based on selected genes	Integration in Drugst.One plugin
NDEx (79, 80)	https://www.ndexbio.org/	Web platform for storing, sharing, and publishing user-created biological networks.	Integration of “Export to NDEx” function into Drugst.One plugin

References

1. Lazareva,O., Canzar,S., Yuan,K., Baumbach,J., Blumenthal,D.B., Tieri,P., Kacprowski,T. and List,M. (2020) BiCoN: Network-constrained biclustering of patients and omics data. *Bioinformatics*, 10.1093/bioinformatics/btaa1076.
2. Levi,H., Rahmanian,N., Elkon,R. and Shamir,R. (2022) The DOMINO web-server for active module identification analysis. *Bioinformatics*, **38**, 2364–2366.
3. Sadegh,S., Skelton,J., Anastasi,E., Maier,A., Adamowicz,K., Möller,A., Kriege,N.M., Kronberg,J., Haller,T., Kacprowski,T., *et al.* (2023) Lacking mechanistic disease definitions and corresponding association data hamper progress in network medicine and beyond. *Nat. Commun.*, **14**, 1662.
4. List,M., Schmidt,S., Christiansen,H., Rehmsmeier,M., Tan,Q., Mollenhauer,J. and Baumbach,J. (2016) Comprehensive analysis of high-throughput screens with HiTSeekR. *Nucleic Acids Res.*, **44**, 6639–6648.
5. Interactive-Enrichment-Analysis: A set of Shiny apps to provide interactive enrichment analysis and exploration of results Github.
6. Hauschild,A.-C., Pastrello,C., Ekaputeri,G.K.A., Bethune-Waddell,D., Abovsky,M., Ahmed,Z., Kotlyar,M., Lu,R. and Jurisica,I. (2023) MirDIP 5.2: tissue context annotation and novel microRNA curation. *Nucleic Acids Res.*, **51**, D217–D225.
7. Brown,K.R., Otasek,D., Ali,M., McGuffin,M.J., Xie,W., Devani,B., van Toch,I.L. and Jurisica,I. (2009) NAViGaTOR: Network Analysis, Visualization and Graphing Toronto. *Bioinformatics*, **25**, 3327–3329.
8. Pillich,R.T., Chen,J., Churas,C., Fong,D., Ideker,T., Liu,S.N., Gyori,B.M., Karis,K., Ono,K., Pico,A., *et al.* (2022) NDEx IQuery: a multi-method network gene set analysis leveraging the Network Data Exchange. *bioRxiv*, 10.1101/2022.10.24.513552.
9. Hoffmann,M., Poschenrieder,J.M., Incudini,M., Baier,S., Fitz,A., Maier,A., Hartung,M., Hoffmann,C., Trummer,N., Adamowicz,K., *et al.* (2023) Network medicine-based epistasis detection in complex diseases: ready for quantum computing. *bioRxiv*, 10.1101/2023.11.07.23298205.
10. Helmy,M., Mee,M., Ranjan,A., Hao,T., Vidal,M., Calderwood,M.A., Luck,K. and Bader,G.D. (2022) OpenPIP: An Open-source Platform for Hosting, Visualizing and Analyzing Protein Interaction Data. *J. Mol. Biol.*, **434**, 167603.
11. Pastrello,C., Kotlyar,M., Abovsky,M., Lu,R. and Jurisica,I. (2024) PathDIP 5: improving coverage and making enrichment analysis more biologically meaningful. *Nucleic Acids Res.*, **52**, D663–D671.
12. Hanspers,K., Riutta,A., Summer-Kutmon,M. and Pico,A.R. (2020) Pathway information extracted from 25 years of pathway figures. *Genome Biol.*, **21**, 273.
13. Adamowicz,K., Arend,L., Maier,A., Schmidt,J.R., Kuster,B., Tsoy,O., Zolotareva,O., Baumbach,J. and Laske,T. (2023) Proteomic meta-study harmonization, mechanotyping and drug repurposing candidate prediction with ProHarMeD. *npj Systems Biology and Applications*, **9**, 1–10.
14. Sarkar,S., Lucchetta,M., Maier,A., Abdrabbou,M.M., Baumbach,J., List,M., Schaefer,M.H. and Blumenthal,D.B. (2023) Online bias-aware disease module mining with

ROBUST-Web. *Bioinformatics*, 10.1093/bioinformatics/btad345.

15. Oubounyt,M., Adlung,L., Patroni,F., Wenke,N.K., Maier,A., Hartung,M., Baumbach,J. and Elkjaer,M.L. (2023) Inference of differential key regulatory networks and mechanistic drug repurposing candidates from scRNA-seq data with SCANet. *Bioinformatics*, **39**, btad644.
16. Wang,R.-S. and Loscalzo,J. (2018) Network-Based Disease Module Discovery by a Novel Seed Connector Algorithm with Pathobiological Implications. *J. Mol. Biol.*, **430**, 2939–2950.
17. Pico,A.R., Kelder,T., van Iersel,M.P., Hanspers,K., Conklin,B.R. and Evelo,C. (2008) WikiPathways: pathway editing for the people. *PLoS Biol.*, **6**, e184.
18. Alonso-López,D., Campos-Laborie,F.J., Gutiérrez,M.A., Lambourne,L., Calderwood,M.A., Vidal,M. and De Las Rivas,J. (2019) APID database: redefining protein–protein interaction experimental evidences and binary interactomes. *Database* , **2019**, baz005.
19. Stark,C., Breitkreutz,B.-J., Reguly,T., Boucher,L., Breitkreutz,A. and Tyers,M. (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.*, **34**, D535–9.
20. Mendez,D., Gaulton,A., Bento,A.P., Chambers,J., De Veij,M., Félix,E., Magariños,M.P., Mosquera,J.F., Mutowo,P., Nowotka,M., *et al.* (2019) ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.*, **47**, D930–D940.
21. Davis,A.P., Grondin,C.J., Johnson,R.J., Sciaky,D., Wiegiers,J., Wiegiers,T.C. and Mattingly,C.J. (2021) Comparative Toxicogenomics Database (CTD): update 2021. *Nucleic Acids Res.*, **49**, D1138–D1143.
22. Freshour,S.L., Kiwala,S., Cotto,K.C., Coffman,A.C., McMichael,J.F., Song,J.J., Griffith,M., Griffith,O.L. and Wagner,A.H. (2020) Integration of the Drug–Gene Interaction Database (DGIdb 4.0) with open crowdsourcing efforts. *Nucleic Acids Res.*, **49**, D1144–D1151.
23. Pinero,J., Queralt-Rosinach,N., Bravo,A., Deu-Pons,J., Bauer-Mehren,A., Baron,M., Sanz,F. and Furlong,L.I. (2015) DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database*, **2015**, bav028–bav028.
24. Wishart,D.S., Feunang,Y.D., Guo,A.C., Lo,E.J., Marcu,A., Grant,J.R., Sajed,T., Johnson,D., Li,C., Sayeeda,Z., *et al.* (2018) DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.*, **46**, D1074–D1082.
25. Avram,S., Bologa,C.G., Holmes,J., Bocci,G., Wilson,T.B., Nguyen,D.-T., Curpan,R., Halip,L., Bora,A., Yang,J.J., *et al.* (2021) DrugCentral 2021 supports drug discovery and repositioning. *Nucleic Acids Res.*, **49**, D1160–D1169.
26. Carithers,L.J., Ardlie,K., Barcus,M., Branton,P.A., Britton,A., Buia,S.A., Compton,C.C., DeLuca,D.S., Peter-Demchok,J., Gelfand,E.T., *et al.* (2015) A novel approach to high-quality postmortem tissue procurement: The GTEx project. *Biopreserv. Biobank.*, **13**, 311–319.
27. Kotlyar,M., Pastrello,C., Ahmed,Z., Chee,J., Varyova,Z. and Jurisica,I. (2022) IID 2021: towards context-specific protein interaction analyses by increased coverage, enhanced annotation and enrichment analysis. *Nucleic Acids Res.*, **50**, D640–D647.

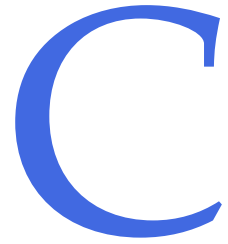
28. Orchard,S., Ammari,M., Aranda,B., Breuza,L., Briganti,L., Broackes-Carter,F., Campbell,N.H., Chavali,G., Chen,C., del-Toro,N., *et al.* (2014) The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.*, **42**, D358–63.
29. Sadegh,S., Skelton,J., Anastasi,E., Bennett,J., Blumenthal,D.B., Galindez,G., Salgado-Albarrán,M., Lazareva,O., Flanagan,K., Cockell,S., *et al.* (2021) Network medicine for disease module identification and drug repurposing with the NeDRex platform. *Nat. Commun.*, **12**, 6848.
30. Amberger,J.S., Bocchini,C.A., Scott,A.F. and Hamosh,A. (2019) OMIM.org: leveraging knowledge across phenotype-gene relationships. *Nucleic Acids Res.*, **47**, D1038–D1043.
31. Snel,B., Lehmann,G., Bork,P. and Huynen,M.A. (2000) STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Res.*, **28**, 3442–3444.
32. Junker,B.H., Koschützki,D. and Schreiber,F. (2006) Exploration of biological network centralities with CentiBiN. *BMC Bioinformatics*, **7**, 219.
33. Girvan,M. and Newman,M.E.J. (2002) Community structure in social and biological networks. *Proc. Natl. Acad. Sci. U. S. A.*, **99**, 7821–7826.
34. Rochat,Y. Closeness centrality extended to unconnected graphs: The harmonic centrality index.
35. Koschützki,D. and Schreiber,F. (2008) Centrality analysis methods for biological networks and their application to gene regulatory networks. *Gene Regul. Syst. Bio.*, **2**, 193–201.
36. Jalili,M., Salehzadeh-Yazdi,A., Gupta,S., Wolkenhauer,O., Yaghmaie,M., Resendis-Antonio,O. and Alimoghaddam,K. (2016) Evolution of Centrality Measurements for the Detection of Essential Proteins in Biological Networks. *Front. Physiol.*, **7**, 375.
37. Zhang,H., Fiszman,M., Shin,D., Miller,C.M., Rosemlat,G. and Rindflesch,T.C. (2011) Degree centrality for semantic abstraction summarization of therapeutic studies. *J. Biomed. Inform.*, **44**, 830–838.
38. Tang,X., Wang,J., Zhong,J. and Pan,Y. (2014) Predicting Essential Proteins Based on Weighted Degree Centrality. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **11**, 407–418.
39. List,M., Alcaraz,N., Dissing-Hansen,M., Ditzel,H.J., Mollenhauer,J. and Baumbach,J. (2016) KeyPathwayMinerWeb: online multi-omics network enrichment. *Nucleic Acids Res.*, **44**, W98–W104.
40. Sadegh,S., Matschinske,J., Blumenthal,D.B., Galindez,G., Kacprowski,T., List,M., Nasirigerdeh,R., Oubounyt,M., Pichlmair,A., Rose,T.D., *et al.* (2020) Exploring the SARS-CoV-2 virus-host-drug interactome for drug repurposing. *Nat. Commun.*, **11**, 3518.
41. Ahmed,R., Angelini,P., Sahneh,F.D., Efrat,A., Glickenstein,D., Gronemann,M., Heinsohn,N., Kobourov,S.G., Spence,R., Watkins,J., *et al.* (2019) Multi-level Steiner Trees. *ACM J. Exp. Algorithmics*, **24**, 1–22.

42. Guney,E., Menche,J., Vidal,M. and Barábasi,A.-L. (2016) Network-based in silico drug efficacy screening. *Nat. Commun.*, **7**, 10331.
43. Gyöngyi,Z., Garcia-Molina,H. and Pedersen,J. (2004) - Combating Web Spam with TrustRank. In Nascimento,M.A., Özsu,M.T., Kossmann,D., Miller,R.J., Blakeley,J.A., Schiefer,B. (eds), *Proceedings 2004 VLDB Conference*. Morgan Kaufmann, St Louis, pp. 576–587.
44. Page,L., Brin,S., Motwani,R. and Winograd,T. (1999) The PageRank Citation Ranking: Bringing Order to the Web.
45. Gyongyi,Z., Garcia-Molina,H. and Pedersen,J. Error - Stanford InfoLab Publication Server.
46. Agrawal,M., Kim,E.S. and Colombel,J.-F. (2020) JAK Inhibitors Safety in Ulcerative Colitis: Practical Implications. *J. Crohns. Colitis*, **14**, S755–S760.
47. Bjerrum,J.T., Nielsen,O.H., Riis,L.B., Pittet,V., Mueller,C., Rogler,G. and Olsen,J. (2014) Transcriptional analysis of left-sided colitis, pancolitis, and ulcerative colitis-associated dysplasia. *Inflamm. Bowel Dis.*, **20**, 2340–2352.
48. Broom,O.J., Widjaya,B., Troelsen,J., Olsen,J. and Nielsen,O.H. (2009) Mitogen activated protein kinases: a role in inflammatory bowel disease? *Clin. Exp. Immunol.*, **158**, 272–280.
49. Coskun,M., Olsen,J., Seidelin,J.B. and Nielsen,O.H. (2011) MAP kinases in inflammatory bowel disease. *Clin. Chim. Acta*, **412**, 513–520.
50. Sayoc-Becerra,A., Krishnan,M., Fan,S., Jimenez,J., Hernandez,R., Gibson,K., Preciado,R., Butt,G. and McCole,D.F. (2020) The JAK-Inhibitor Tofacitinib Rescues Human Intestinal Epithelial Cells and Colonoids from Cytokine-Induced Barrier Dysfunction. *Inflamm. Bowel Dis.*, **26**, 407–422.
51. Armuzzi,A., Hart,A., Cappelleri,J.C., Mammar,N., Hur,P., Hoskin,B., Hennessy,F., Milligan,G. and Dignass,A. (2023) Characteristics, clinical outcomes and patient-reported outcomes of patients with ulcerative colitis receiving tofacitinib: a real-world survey in the United States and five European countries. *BMC Gastroenterol.*, **23**, 17.
52. Del Sordo,R., Lupinacci,G., Tanzi,G., Bassotti,G. and Villanacci,V. (2022) Imatinib and Dasatinib-induced Ulcerative Colitis: Case Report. *Inflamm. Bowel Dis.*, **28**, e1–e2.
53. Yamauchi,K., Inaba,T., Colvin,H.S., Sakakihara,I., Yamamoto,K., Izumikawa,K., Takahashi,S., Tanaka,S., Ishikawa,S., Wato,M., *et al.* (2022) Dasatinib-induced colitis: clinical, endoscopic and histological findings. *Scand. J. Gastroenterol.*, **57**, 449–456.
54. Lopez-Millan,B., Diaz de la Guardia,R., Roca-Ho,H., García-Herrero,C.M., Lavoie,J.R., Rosu-Myles,M., Gonzalez-Rey,E., O’Valle,F., Criado,G., Delgado,M., *et al.* (2017) Therapeutic effect of the immunomodulatory drug lenalidomide, but not pomalidomide, in experimental models of rheumatoid arthritis and inflammatory bowel disease. *Exp. Mol. Med.*, **49**, e290.
55. Srinivasan,R. and Akobeng,A.K. (2009) Thalidomide and thalidomide analogues for induction of remission in Crohn’s disease. *Cochrane Database Syst. Rev.*
56. Millrine,D. and Kishimoto,T. (2017) A Brighter Side to Thalidomide: Its Potential Use in

Immunological Disorders. *Trends Mol. Med.*, **23**, 348–361.

57. Issac, A.G., Szafron, D., Wei, D., McQuade, J.L. and Wang, Y. (2022) Severe Inflammatory Colitis Related to Encorafenib and Binimetinib following Immune Checkpoint Inhibitor Therapy. *Case Rep. Gastroenterol.*, **16**, 388–393.
58. Gareb, B., Otten, A.T., Frijlink, H.W., Dijkstra, G. and Kosterink, J.G.W. (2020) Review: Local Tumor Necrosis Factor- α Inhibition in Inflammatory Bowel Disease. *Pharmaceutics*, **12**.
59. Cui, G., Fan, Q., Li, Z., Goll, R. and Florholmen, J. (2021) Evaluation of anti-TNF therapeutic response in patients with inflammatory bowel disease: Current and novel biomarkers. *EBioMedicine*, **66**, 103329.
60. Goenka, M.K., Kochhar, R., Tandia, B. and Mehta, S.K. (1996) Chloroquine for mild to moderately active ulcerative colitis: comparison with sulfasalazine. *Am. J. Gastroenterol.*, **91**, 917–921.
61. Bak, Y.-K., Lampe, J.W. and Sung, M.-K. (2014) Effects of dietary supplementation of glucosamine sulfate on intestinal inflammation in a mouse model of experimental colitis. *J. Gastroenterol. Hepatol.*, **29**, 957–963.
62. Shailubhai, K., Palejwala, V., Arjunan, K.P., Saykhedkar, S., Nefsky, B., Foss, J.A., Comiskey, S., Jacob, G.S. and Plevy, S.E. (2015) Plecanatide and dolcanatide, novel guanylate cyclase-C agonists, ameliorate gastrointestinal inflammation in experimental models of murine colitis. *World J. Gastrointest. Pharmacol. Ther.*, **6**, 213–222.
63. Izeboud, C.A., Monshouwer, M., van Miert, A.S. and Witkamp, R.F. (1999) The beta-adrenoceptor agonist clenbuterol is a potent inhibitor of the LPS-induced production of TNF- α and IL-6 in vitro and in vivo. *Inflamm. Res.*, **48**, 497–502.
64. Fiebich, B.L., Collado, J.A., Stratz, C., Valina, C., Hochholzer, W., Muñoz, E. and Bellido, L.M. (2012) Pseudoephedrine inhibits T-cell activation by targeting NF- κ B, NFAT and AP-1 signaling pathways. *Immunopharmacol. Immunotoxicol.*, **34**, 98–106.
65. Liao, J., Keiser, J.A., Scales, W.E., Kunkel, S.L. and Kluger, M.J. (1995) Role of epinephrine in TNF and IL-6 production from isolated perfused rat liver. *Am. J. Physiol.*, **268**, R896–901.
66. Griffin, É.W., Yssel, J.D., O'Neill, E., Ryan, K.J., Boyle, N., Harper, P., Harkin, A. and Connor, T. (2018) The β 2-adrenoceptor agonist clenbuterol reduces the neuroinflammatory response, neutrophil infiltration and apoptosis following intra-striatal IL-1 β administration to rats. *Immunopharmacol. Immunotoxicol.*, **40**, 99–106.
67. Wu, Z., Kong, X., Zhang, T., Ye, J., Fang, Z. and Yang, X. (2014) Pseudoephedrine/ephedrine shows potent anti-inflammatory activity against TNF- α -mediated acute liver failure induced by lipopolysaccharide/D-galactosamine. *Eur. J. Pharmacol.*, **724**, 112–121.
68. Ihara, S., Hirata, Y. and Koike, K. (2017) TGF- β in inflammatory bowel disease: a key regulator of immune cells, epithelium, and the intestinal microbiota. *J. Gastroenterol.*, **52**, 777–787.
69. Yarani, R., Shojaeian, A., Palasca, O., Doncheva, N.T., Jensen, L.J., Gorodkin, J. and Pociot, F. (2022) Differentially Expressed miRNAs in Ulcerative Colitis and Crohn's Disease. *Front. Immunol.*, **13**, 865777.

70. Skrzypczak-Zielinska,M., Gabryel,M., Marszalek,D., Dobrowolska,A. and Slomski,R. (2021) NGS study of glucocorticoid response genes in inflammatory bowel disease patients. *Arch. Med. Sci.*, **17**, 417–433.
71. Seinen,M.L., van Nieuw Amerongen,G.P., de Boer,N.K.H. and van Bodegraven,A.A. (2016) Rac Attack: Modulation of the Small GTPase Rac in Inflammatory Bowel Disease and Thiopurine Therapy. *Mol. Diagn. Ther.*, **20**, 551–557.
72. Xu,S., He,Y., Lin,L., Chen,P., Chen,M. and Zhang,S. (2021) The emerging role of ferroptosis in intestinal disease. *Cell Death Dis.*, **12**, 289.
73. Alvarado,D.M. and Ciorba,M.A. (2020) Serotonin Receptors Regulate Inflammatory Response in Experimental Colitis. *J. Nutr.*, **150**, 1678–1679.
74. Ramadan,N., Sami,S., Elmenshawy,S. and Salem,A.E. (2021) Possible protective effects of propranolol and carvedilol and their interaction with mesalazine on experimentally induced ulcerative colitis in male albino rats under supervision of. 10.13140/RG.2.2.22062.38724.
75. Murnane,K.S., Guner,O.F., Bowen,J.P., Rambacher,K.M., Moniri,N.H., Murphy,T.J., Daphney,C.M., Oppong-Damoah,A. and Rice,K.C. (2019) The adrenergic receptor antagonist carvedilol interacts with serotonin 2A receptors both in vitro and in vivo. *Pharmacol. Biochem. Behav.*, **181**, 37–45.
76. Bennett,J., Krupke,D., Sadegh,S., Baumbach,J., Fekete,S.P., Kacprowski,T., List,M. and Blumenthal,D.B. (2022) Robust disease module mining via enumeration of diverse prize-collecting Steiner trees. *Bioinformatics*, 10.1093/bioinformatics/btab876.
77. Adamowicz,K., Maier,A., Baumbach,J. and Blumenthal,D.B. (2022) Online in silico validation of disease and gene sets, clusterings or subnetworks with DIGEST. *Brief. Bioinform.*, **23**.
78. Lobentanzer,S., Aloy,P., Baumbach,J., Bohar,B., Carey,V.J., Charoentong,P., Danhauser,K., Doğan,T., Dreo,J., Dunham,I., *et al.* (2023) Democratizing knowledge representation with BioCypher. *Nat. Biotechnol.*, 10.1038/s41587-023-01848-y.
79. Pratt,D., Chen,J., Pillich,R., Rynkov,V., Gary,A., Demchak,B. and Ideker,T. (2017) NDEx 2.0: A Clearinghouse for Research on Cancer Pathways. *Cancer Res.*, **77**, e58–e61.
80. Pillich,R.T., Chen,J., Churas,C., Liu,S., Ono,K., Otasek,D. and Pratt,D. (2021) NDEx: Accessing Network Models and Streamlining Network Biology Workflows. *Curr Protoc*, **1**, e258.



NEEDL & EDA
SUPPLEMENTARY MATERIAL

This section contains the original supplementary material published alongside the article:

**“Network medicine-based epistasis detection in complex diseases:
ready for quantum computing”**

Network medicine-based epistasis detection in complex diseases: ready for quantum computing

Markus Hoffmann^{1,2,3,●,*}, Julian M. Poschenrieder^{1,4,●}, Massimiliano Incudini^{5,●}, Sylvie Baier^{1,●}, Amelie Fritz^{6,7,●},
Andreas Maier^{4,●}, Michael Hartung^{4,●}, Christian Hoffmann^{4,●}, Nico Trummer¹, Klaudia Adamowicz⁴, Mario
Picciani⁸, Evelyn Scheibling¹, Maximilian V. Harl¹, Ingmar Lesch¹, Hunor Frey¹, Simon Kayser¹, Paul Wissenberg¹,
Leon Schwartz¹, Leon Hafner^{1,2}, Aakriti Acharya^{9,10}, Lena Hackl⁴, Gordon Grabert^{9,10}, Sung-Gwon Lee^{3,11},
Gyuhyeok Cho¹², Matthew E. Cloward¹³, Jakub Jankowski³, Hye Kyung Lee³, Olga Tsoy⁴, Nina Wenke⁴, Anders
Gorm Pedersen⁶, Klaus Bønnelykke⁷, Antonio Mandarino¹⁴, Federico Melograna^{15,16}, Laura Schulz¹⁷, Héctor
Climente-González¹⁸, Mathias Wilhelm⁸, Luigi Iapichino¹⁷, Lars Wienbrandt¹⁹, David Ellinghaus¹⁹, Kristel Van
Steen^{15,16}, Michele Grossi²⁰, Priscilla A. Furth²¹, Lothar Hennighausen^{2,3}, Alessandra Di Pierro^{5,†}, Jan
Baumbach^{4,22,†}, Tim Kacprowski^{9,10,†}, Markus List^{1,23†,§}, and David B. Blumenthal^{24,†,*}

¹Data Science in Systems Biology, School of Life Sciences, Technical University of Munich, Freising, Germany

²Institute for Advanced Study (Lichtenbergstrasse 2 a, D-85748 Garching, Germany), Technical University of Munich, Germany

³National Institute of Diabetes, Digestive, and Kidney Diseases, Bethesda, MD 20892, United States of America

⁴Institute for Computational Systems Biology, University of Hamburg, Germany

⁵Dipartimento di Informatica, Università di Verona, Strada le Grazie 15 - 34137, Verona, Italy

⁶Department of Health Technology, Section for Bioinformatics, Technical University of Denmark, DTU, 2800 Kgs. Lyngby, Denmark

⁷Copenhagen Prospective Studies on Asthma in Childhood (COPSAC), Herlev and Gentofte Hospital, University of Copenhagen, Copenhagen, Denmark

⁸Computational Mass Spectrometry, Technical University of Munich, Freising, Germany

⁹Division Data Science in Biomedicine, Peter L. Reichertz Institute for Medical Informatics, Technische Universität Braunschweig and Hannover Medical School, Rebenring 56, 38106 Braunschweig, Germany

¹⁰Braunschweig Integrated Centre of Systems Biology (BRICS), Technische Universität Braunschweig, Rebenring 56, 38106 Braunschweig, Braunschweig, Germany

¹¹School of Biological Sciences and Technology, Chonnam National University, Gwangju, Korea

¹²Department of Chemistry, Gwangju Institute of Science and Technology, Gwangju, Korea

¹³Department of Biology, Brigham Young University, Provo, UT, USA

¹⁴International Centre for Theory of Quantum Technologies, University of Gdańsk, 80-309 Gdańsk, Poland

¹⁵BIO3 - Systems Genetics; GIGA-R Medical Genomics, University of Liège, Liège, Belgium

¹⁶BIO3 - Systems Medicine; Department of Human Genetics, KU Leuven, Leuven, Belgium

¹⁷Leibniz Supercomputing Centre of the Bavarian Academy of Sciences and Humanities (LRZ), Garching b. München, Germany

¹⁸RIKEN Center for Advanced Intelligence Project, Tokyo, Japan

¹⁹Institute of Clinical Molecular Biology, Christian Albrechts University of Kiel, Kiel, Germany.

²⁰European Organization for Nuclear Research (CERN), Geneva 1211, Switzerland

²¹Departments of Oncology & Medicine, Georgetown University, Washington, DC, USA

²²Computational BioMedicine Lab, University of Southern Denmark, Denmark

²³Munich Data Science Institute (MDSI), Technical University of Munich, Garching, Germany

²⁴Biomedical Network Science Lab, Department Artificial Intelligence in Biomedical Engineering, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

●The authors wish to be known that, in their opinion, the first eight authors should be considered as shared first authors

†The authors wish to be known that, in their opinion, the last five authors should be considered as shared last authors

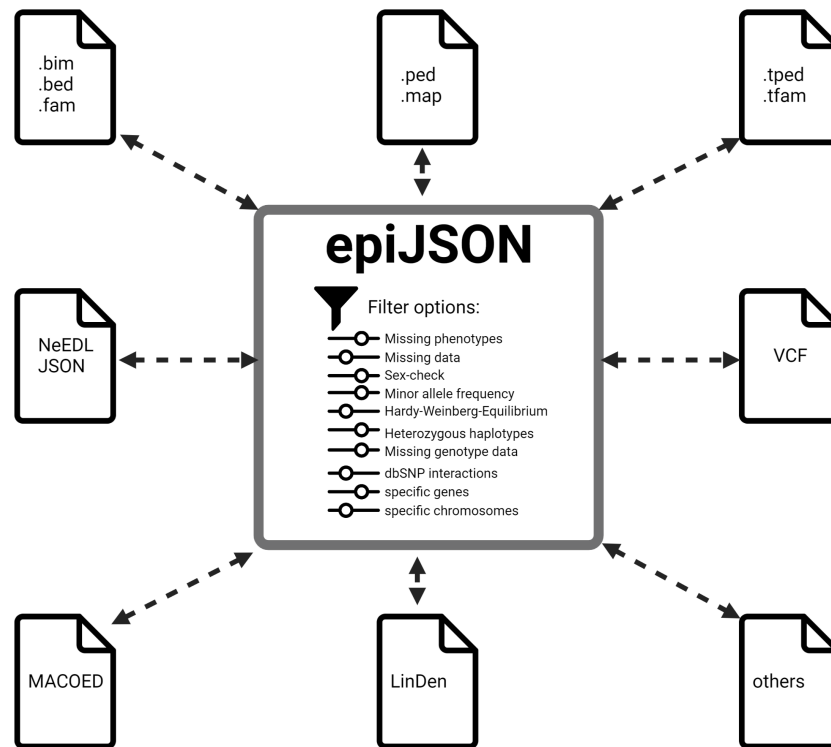
*Contact: markus.hoffmann@nih.gov, markus.list@tum.de, david.b.blumenthal@fau.de

Abstract

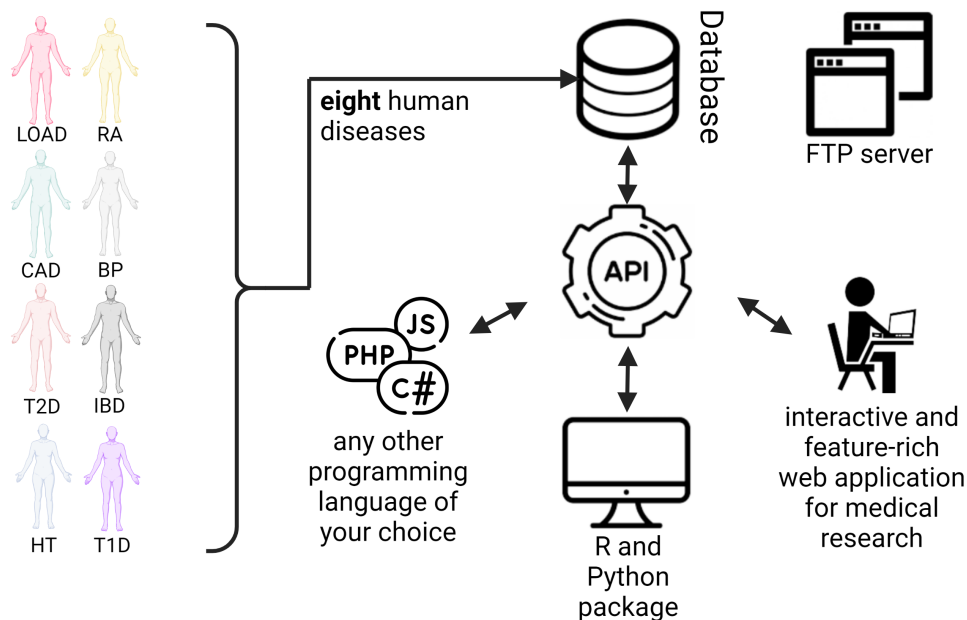
Most heritable diseases are polygenic. To comprehend the underlying genetic architecture, it is crucial to discover the clinically relevant epistatic interactions (EIs) between genomic single nucleotide polymorphisms (SNPs)^{1–3}. Existing statistical computational methods for EI detection are mostly limited to pairs of SNPs due to the combinatorial explosion of higher-order EIs. With NeEDL (**n**etwork-based **e**pistasis **d**etection via **l**ocal search), we leverage network medicine to inform the selection of EIs that are an order of magnitude more statistically significant compared to existing tools and consist, on average, of five SNPs. We further show that this computationally demanding task can be substantially accelerated once quantum computing hardware becomes available. We apply NeEDL to eight different diseases and discover

genes (affected by EIs of SNPs) that are partly known to affect the disease, additionally, these results are reproducible across independent cohorts. EIs for these eight diseases can be interactively explored in the Epistasis Disease Atlas (<https://epistasis-disease-atlas.com>). In summary, NeEDL demonstrates the potential of seamlessly integrated quantum computing techniques to accelerate biomedical research. Our network medicine approach detects higher-order EIs with unprecedented statistical and biological evidence, yielding unique insights into polygenic diseases and providing a basis for the development of improved risk scores and combination therapies.

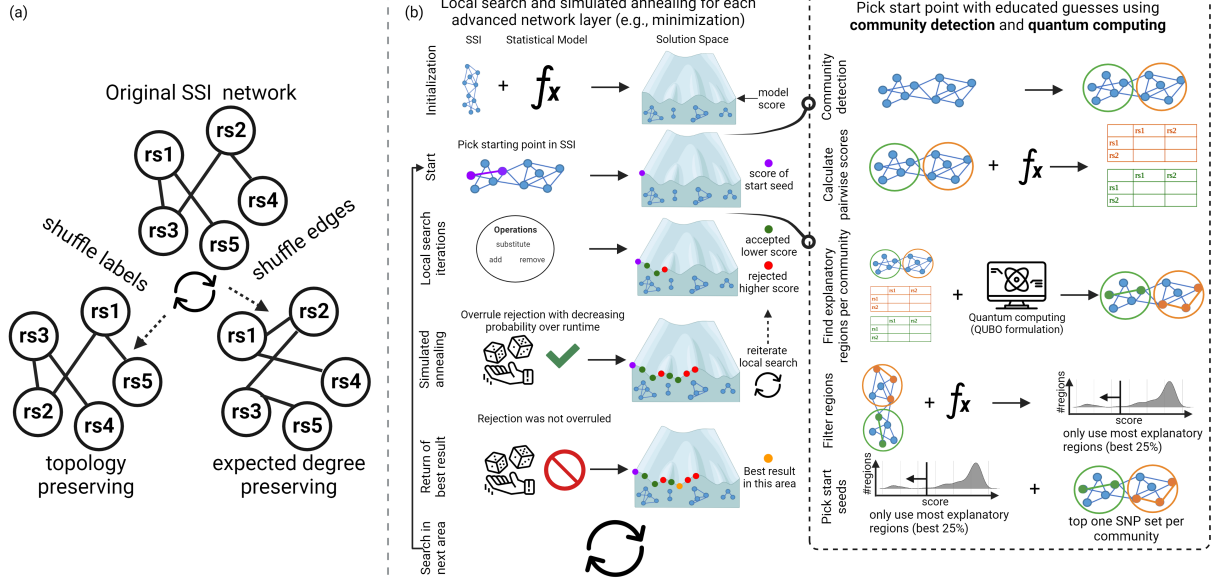
(a) epiJSON: a tool for format conversion and preprocessing of Genotype and Phenotype data



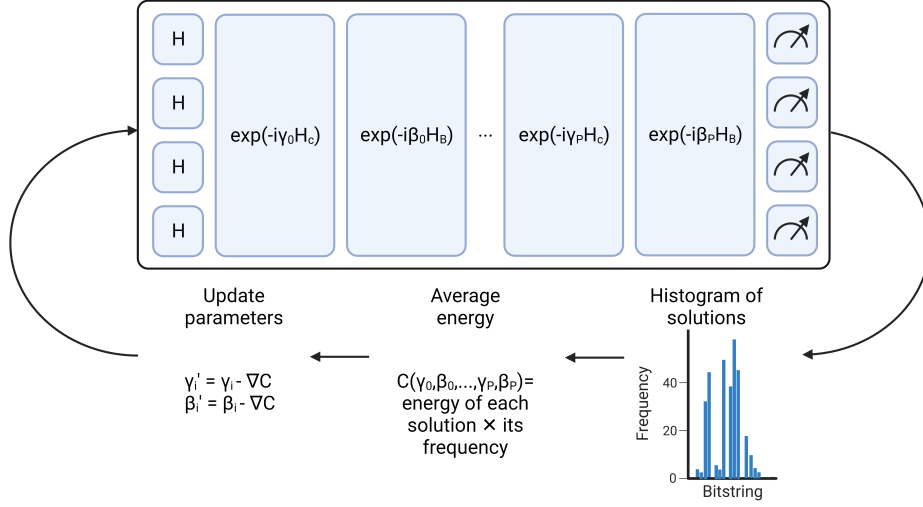
(b) Software architecture of the Epistasis Disease Atlas



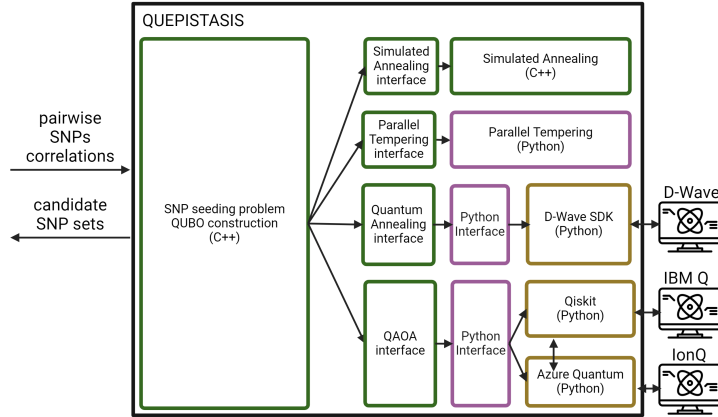
Supplementary Figure 1: (a) NeEDL supports with its built-in epiJSON tool various widely used GWAS data formats (e.g., the PLINK formats⁴) and provides several optional preprocessing routines⁵. (b) The software architecture of the Epistasis Disease Atlas. The web-based Epistasis Disease Atlas (<https://epistasis-disease-atlas.com>) contains the most promising SNP sets for eight diseases and can be accessed via various access routes.



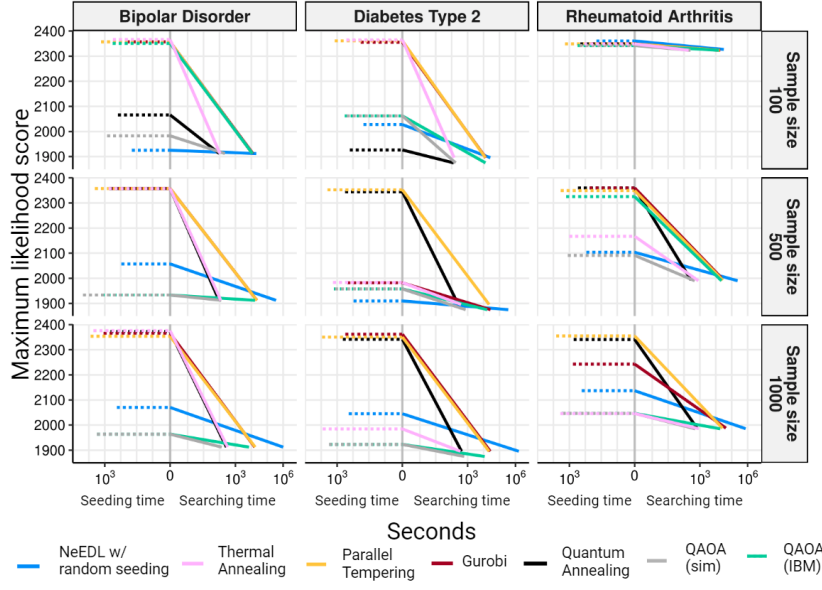
Supplementary Figure 2: Overview of (a) generation of random background models and (b) the heuristics of local search and simulated annealing with quantum computing seeding. (a) We generate two random background models to evaluate the information that can be gained by transforming a PPI network into an SSI network. We either shuffle the labels to obtain a topology-preserving network or shuffle the edges to obtain an expected degree-preserving network. (b) After initialization using the SSI network and a statistical model, we let the algorithm pick starting seeds (i.e., connected sets of SNPs either by random or by quantum computing). If we use quantum computing, we first identify densely connected communities in the network and calculate all pairwise scores of the community in a matrix. The matrix is then used in a QUBO formulation that can identify already promising regions in a community. We filter for the 25% most promising regions and use them as starting seeds. Next, we add, substitute, or remove SNPs from the set and the direct neighborhood in the SSI network until we cannot alleviate the score and rejects any neighboring solutions. To avoid being trapped in local minima, we employ simulated annealing that allows us to overrule a rejection with a decreasing probability over the time that is already being used in investigating this area. If we overrule the rejection, we reiterate the local search until the rejection was not overruled anymore. Lastly, we report the best result in this area and start to search for the most promising solution in the next area of the SSI.



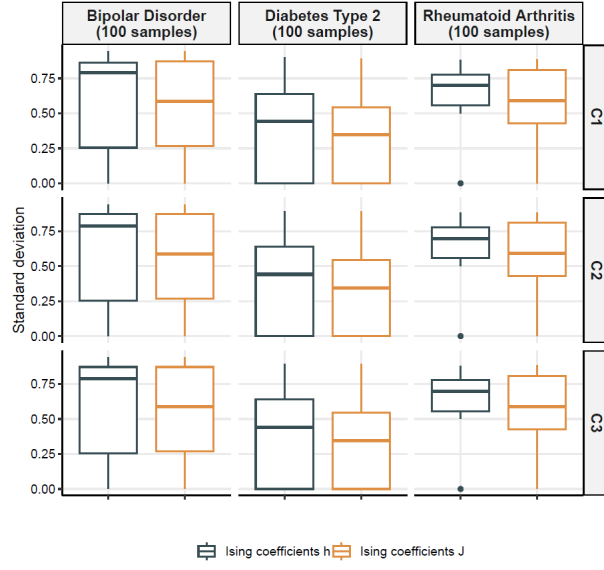
Supplementary Figure 3: Training of the QAOA circuit. The initial parameters are chosen randomly, and the execution of the circuit samples a histogram of possible solutions. The solutions are then used to define the average cost as the cost of each solution multiplied by their frequency. The cost function is then used to guide the optimization, often by gradient descent.



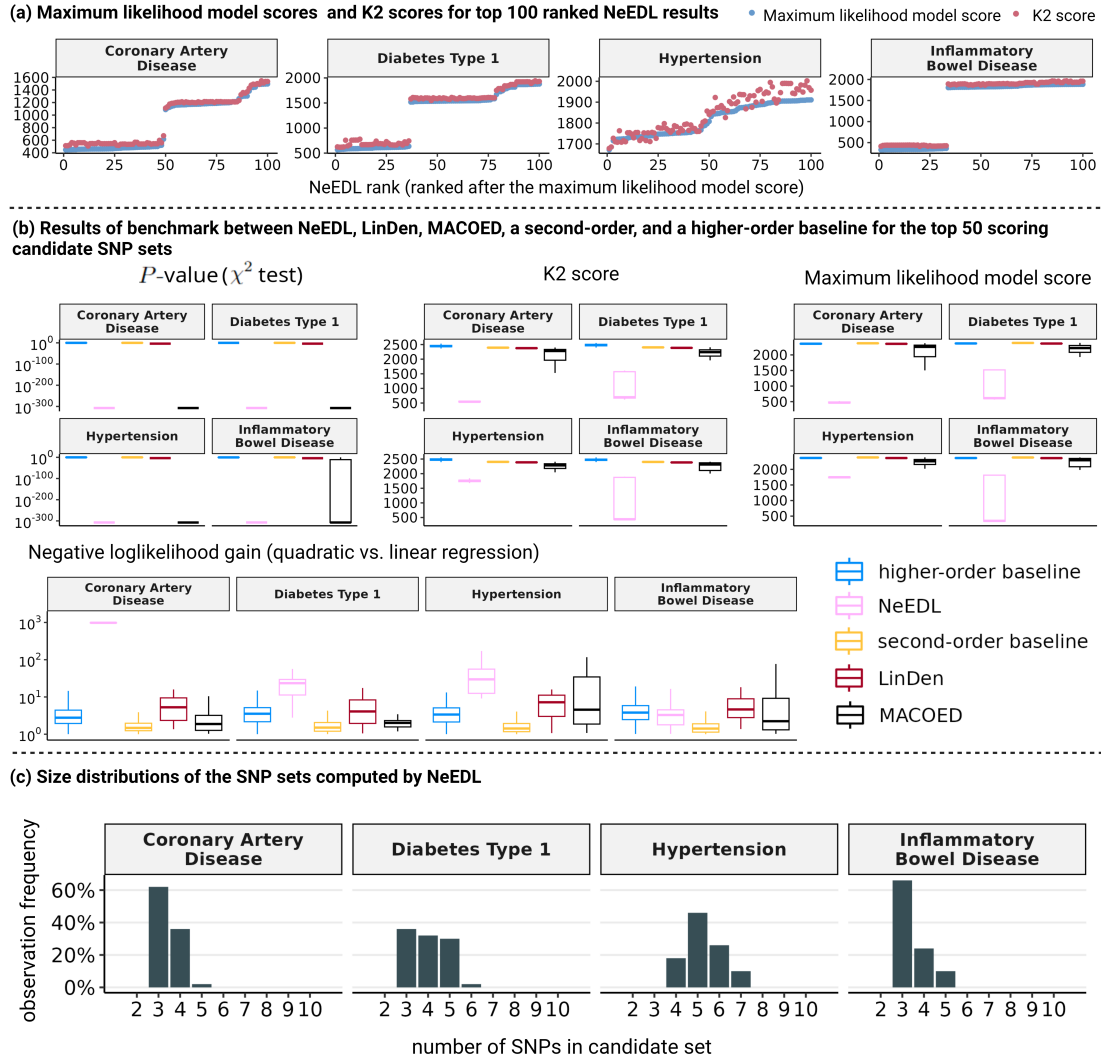
Supplementary Figure 4: High-level structure of *quepistasis*. The module takes as input a matrix of pairwise relations among SNPs, which is processed by a preliminary software layer to generate a QUBO formulation based on the input data. This QUBO instance can then be solved either by a thermal annealing algorithm on a classical device, or by using one of the available quantum devices such as D-Wave, IBM Quantum devices, or the devices accessible through the Azure Quantum cloud, which includes IonQ devices. The output is one or more SNP sets that will be used as seeds for the local search.



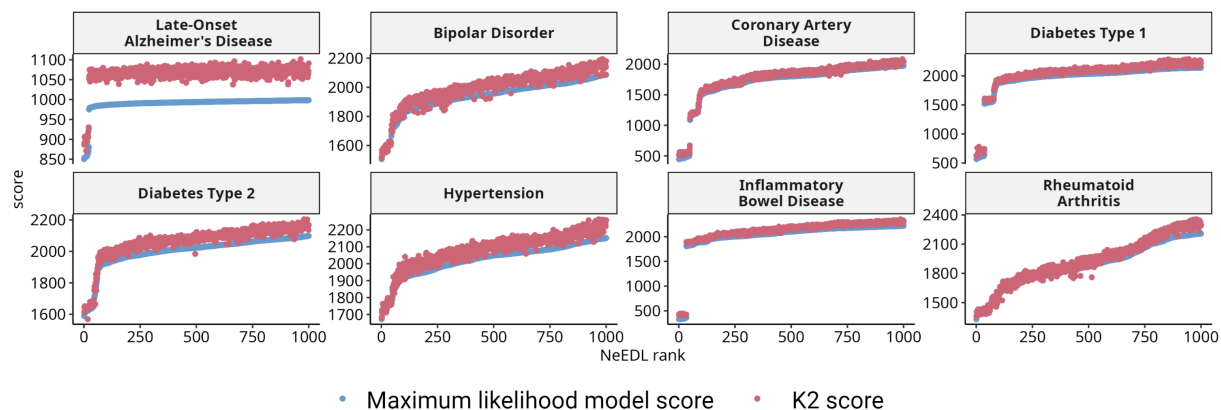
Supplementary Figure 5: Experiments on the quantum computers. The sample size refers to the number of SNPs. Comparison of the timing performance between Linear NeEDL and Quantum-computing enabled NeEDL. The negative time indicates the duration seeding process, and the positive time is the searching process. The seeding with quantum computing produces fewer but better seeds, speeding up the local search phase.



Supplementary Figure 6: Standard deviation of the Ising coefficients h and J in the various subproblems generated by the community detection tool for the three datasets shown and for some configuration. Legend: C1 is the configuration in Suppl. Table 12 having $\lambda_0 = 10.0, \lambda_1 = 0.0001$, C2 having $\lambda_0 = 10.0, \lambda_1 = 1.0$, and C3 having $\lambda_0 = 2.0, \lambda_1 = 1.0$.

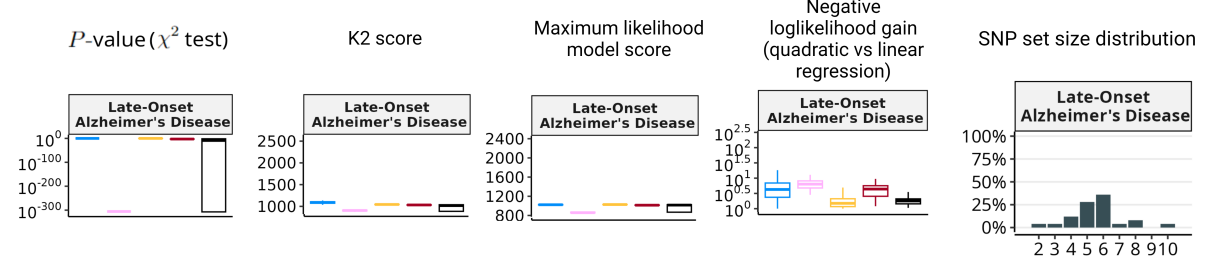


Supplementary Figure 7: Quantitative evaluation of the SNP sets computed by NeEDL. (a) Visualization of the maximum likelihood model score and the K2 score of the top 100 candidate SNP sets ranked by NeEDL. (b) A benchmark study between NeEDL, LinDen, MACOED, a second-order baseline, and a higher-order baseline shows that NeEDL outperforms in statistical significance existing epistasis detection tools with respect to four different evaluation metrics. (c) Analyzing the number of SNPs included in NeEDL's output, SNP sets reveal that the most promising SNP sets are typical of sizes between three to seven.

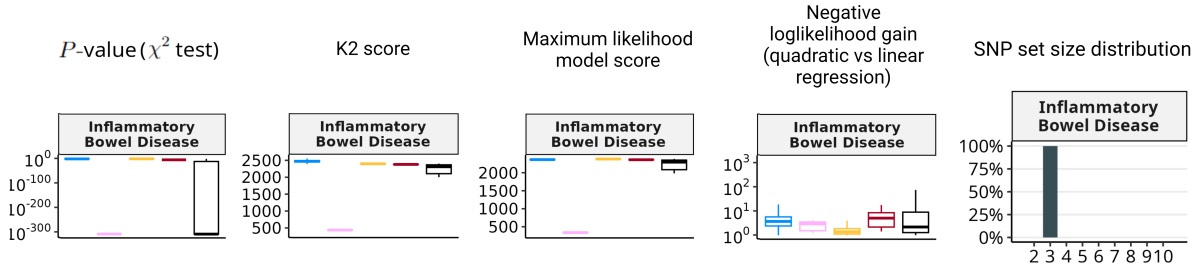


Supplementary Figure 8: Visualization of the Maximum likelihood model score and the K2 score of the top 1000 candidate SNP sets ranked by NeEDL.

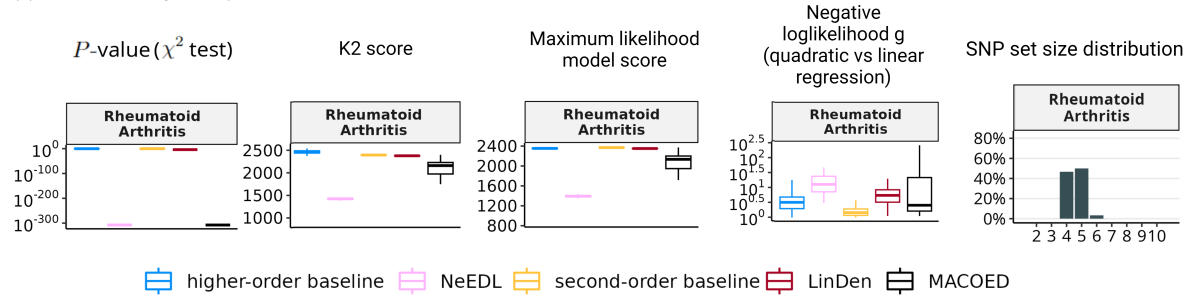
(a) Benchmarking for top 25 candidate SNP sets for Late-onset Alzheimer's Disease



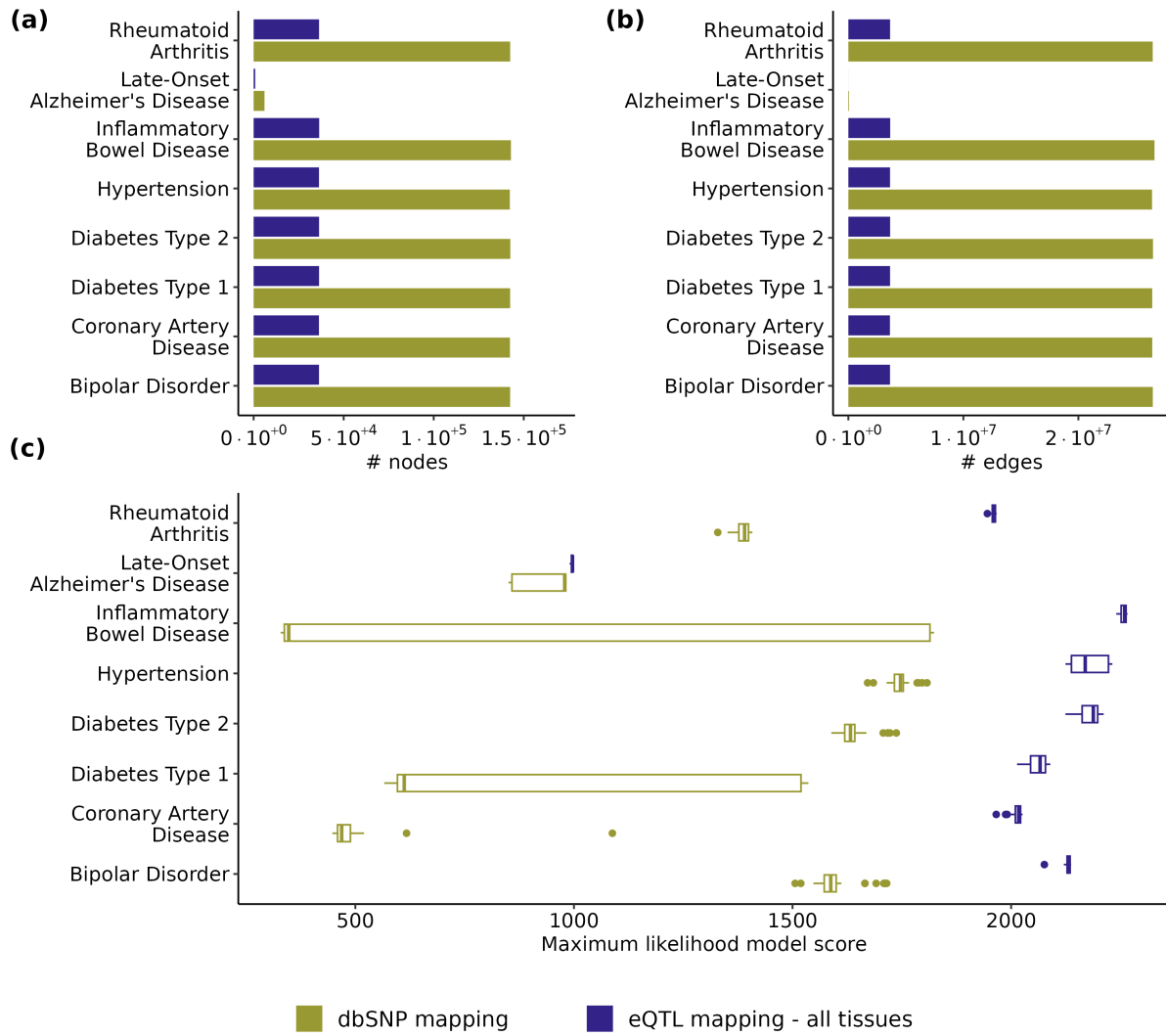
(b) Benchmarking for top 25 candidate SNP sets for Inflammatory Bowel Disease



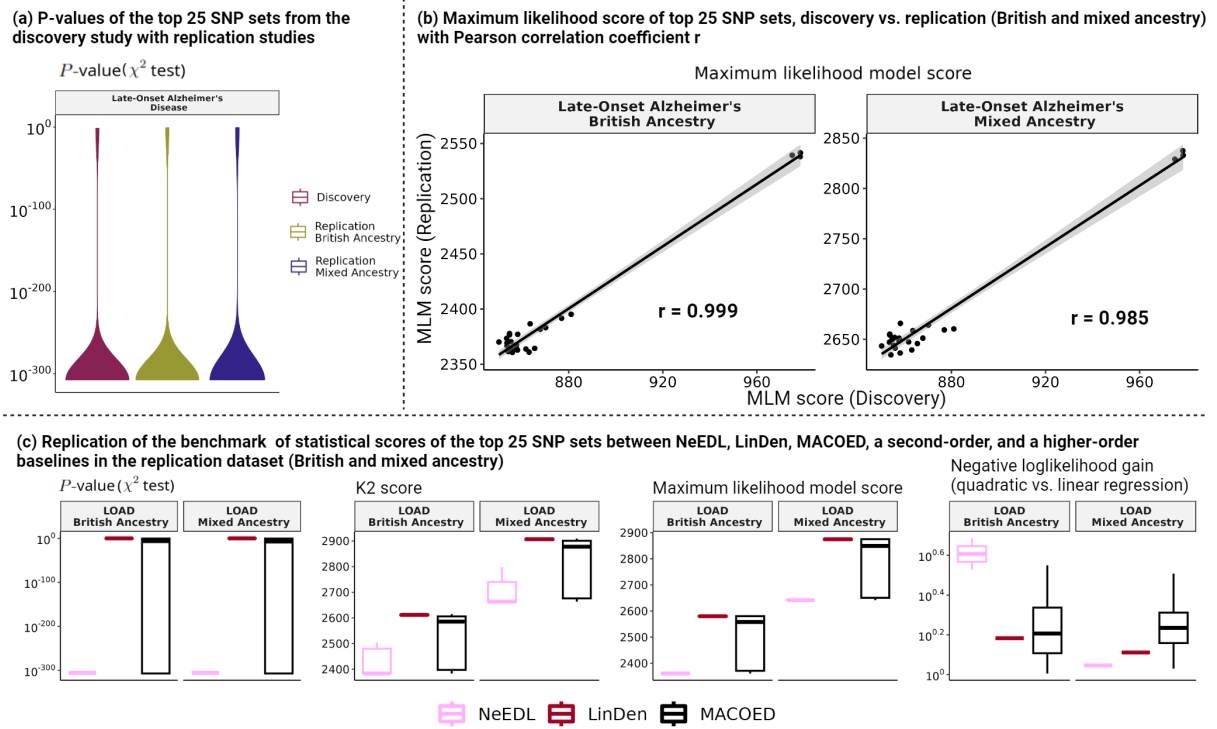
(c) Benchmarking for top 60 candidate SNP sets for Rheumatoid Arthritis



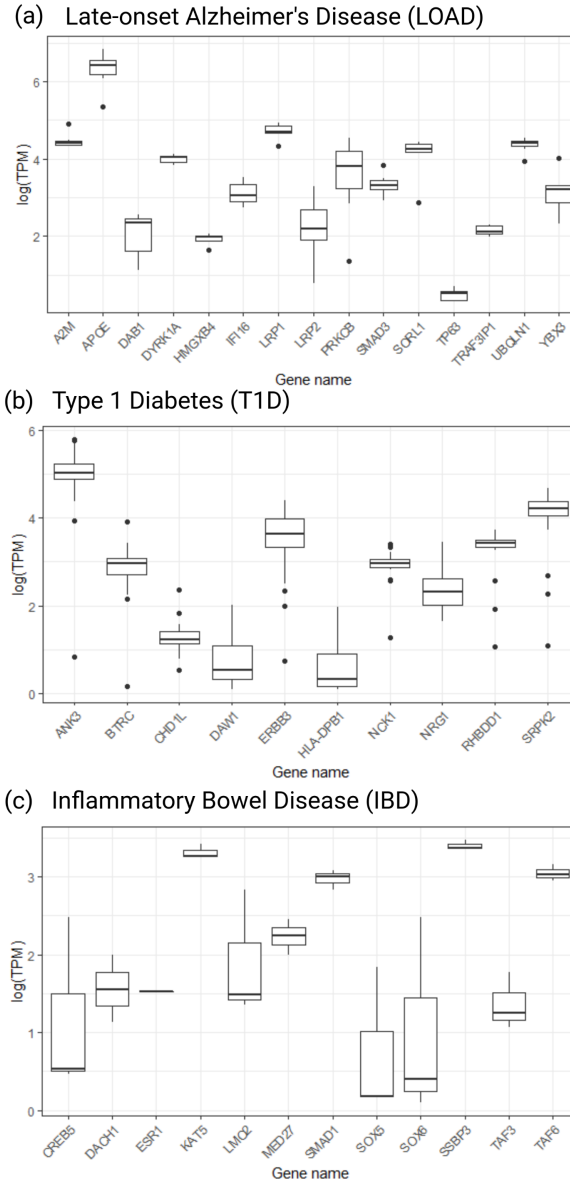
Supplementary Figure 9: Benchmarking study for other cutoffs for Late-onset Alzheimer's Disease, Inflammatory Bowel Disease, and Rheumatoid Arthritis due to differences of the phase transition gap. (a) A benchmark study for Late-onset Alzheimer's disease between the top 25 ranking candidate SNP sets of NeEDL, LinDen, MACOED, a second-order baseline, and a higher-order baseline shows that NeEDL outperforms LinDen overall metrics and MACOED with respect to the P -values and the NLL gain. NeEDL performs similarly or slightly better than MACOED in terms of the K2 score and the MLM score. The resulting SNP set size distribution is shown. (b) A benchmark study for Inflammatory Bowel Disease between the top 25 ranking candidate SNP sets of NeEDL, LinDen, MACOED, a second-order baseline, and a higher-order baseline shows that NeEDL outperforms LinDen in all but the NLL gain metric where it performs similarly. MACOED is outperformed with respect to P -values, K2 score, and MLM score but performs similarly in the NLL gain score. The resulting SNP set size distribution is shown. (c) A benchmark study for Rheumatoid Arthritis between the top 60 ranking candidate SNP sets of NeEDL, LinDen, MACOED, a second-order baseline, and a higher-order baseline shows that NeEDL outperforms LinDen in all metrics. MACOED is outperformed in the K2 score, the MLM score, and the NLL gain score but behaves similarly with respect to the P -values. The resulting SNP set size distribution is shown.



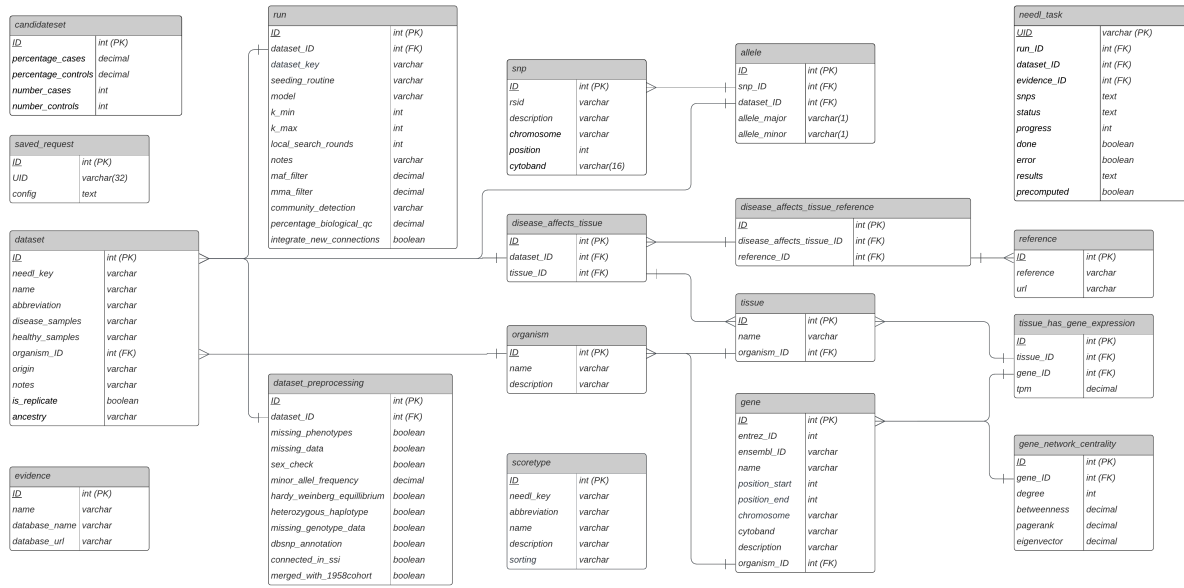
Supplementary Figure 10: Comparison of NeEDL runs using dbSNP data and eQTL data for mapping SNPs to genes. The dbSNP-based mapping leads to substantially larger PPI networks, containing more nodes (a) and more edges (b). The results of the dbSNP-based approach outperform the eQTL-based approach in the maximum likelihood model (c).



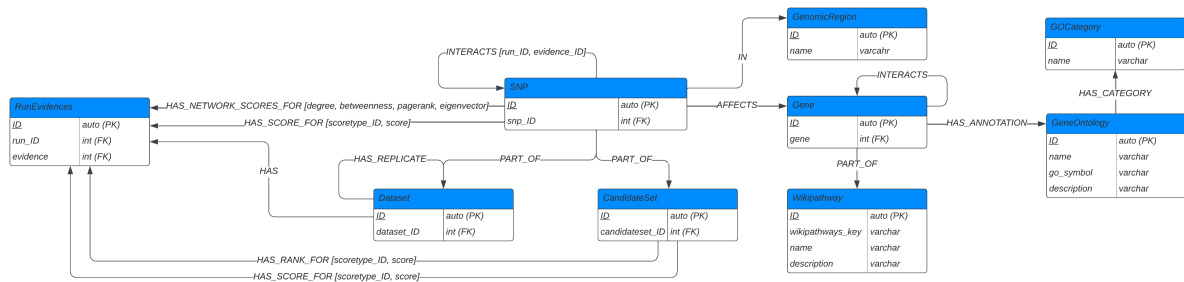
Supplementary Figure 11: Replication studies in two independent data sets from UK Biobank with British and mixed ancestry. (a) P -values across discovery and the two replication studies. (b) Correlation of the MLM score between the discovery study and the two replication studies. (c) Replication of the benchmarking between NeEDL, LinDen, MACOED, a second-order baseline, and a higher-order baseline in two replication data sets.



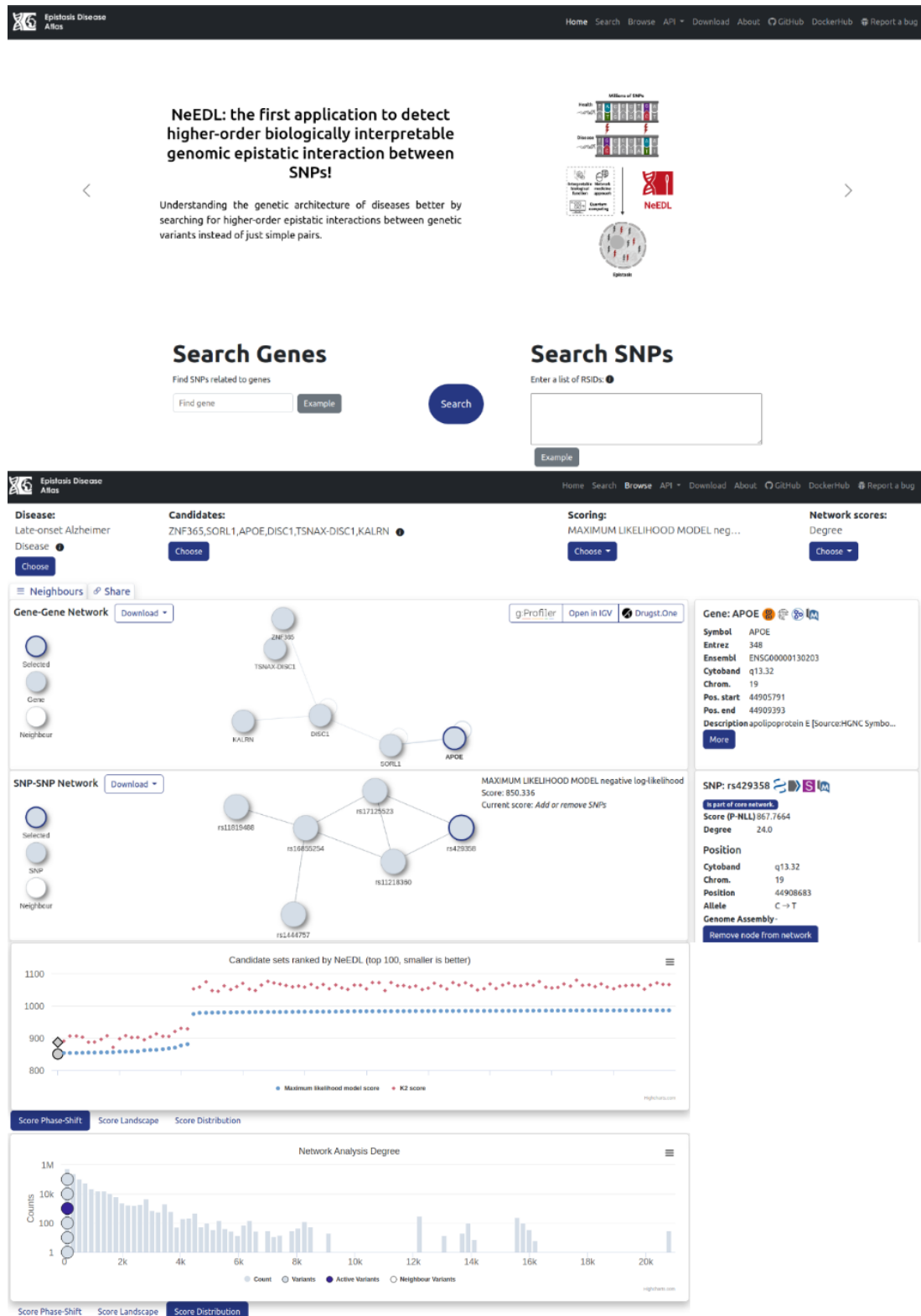
Supplementary Figure 12: Tissue-specific transcript per million (TPM) expression of genes enriched in critical pathways involved in disease development. (a) We included TPM values of the genes APOE, SORL1, SMAD3, LRP2, YBX3, TP63, PRKCB, DAB1, LRP1, IFI16, UBQLN1, A2M, DYRK1A, HMGXB4, and TRAF3IP1 of the tissues (Allocortex, Ammon's Horn, Brainstem, Cerebellum, Diencephalic nuclei, Entorhinal region, Hippocampus, Insular cortex, Neocortex, Temporal neocortex⁶⁻⁸) that were reported as affected in the literature in LOAD from the Human Protein Atlas⁹. (b) We included TPM values of the genes NCK1, ANK3, DAW1, BTRC, CHD1L, NRG1, ERBB3, HLA-DPB1, RHBDD1, SPRK1 of the tissues (Cerebellum, Endothelial cells, Heart, Kidney, Myelin, Pancreatic beta cells, Putamen, Retinal cells, Right parahippocampal gyrus, Thalamus¹⁰⁻¹⁶) that were reported as affected in the literature in T1D from the Human Protein Atlas. (c) We included TPM values of the genes TAF6, SMAD1, TAF3, KAT5, ESR1, MED27, SSBP3, LMO2, PASD1, DACH1, SOX5, SOX6, and CREB5 from tissues (Blood vessels, Colonic mucosa, Intestinal mucosa, Muscularis external, Serosa, Submucosa¹⁷⁻¹⁹) that were reported as affected in the literature in IBD from the Human Protein Atlas.



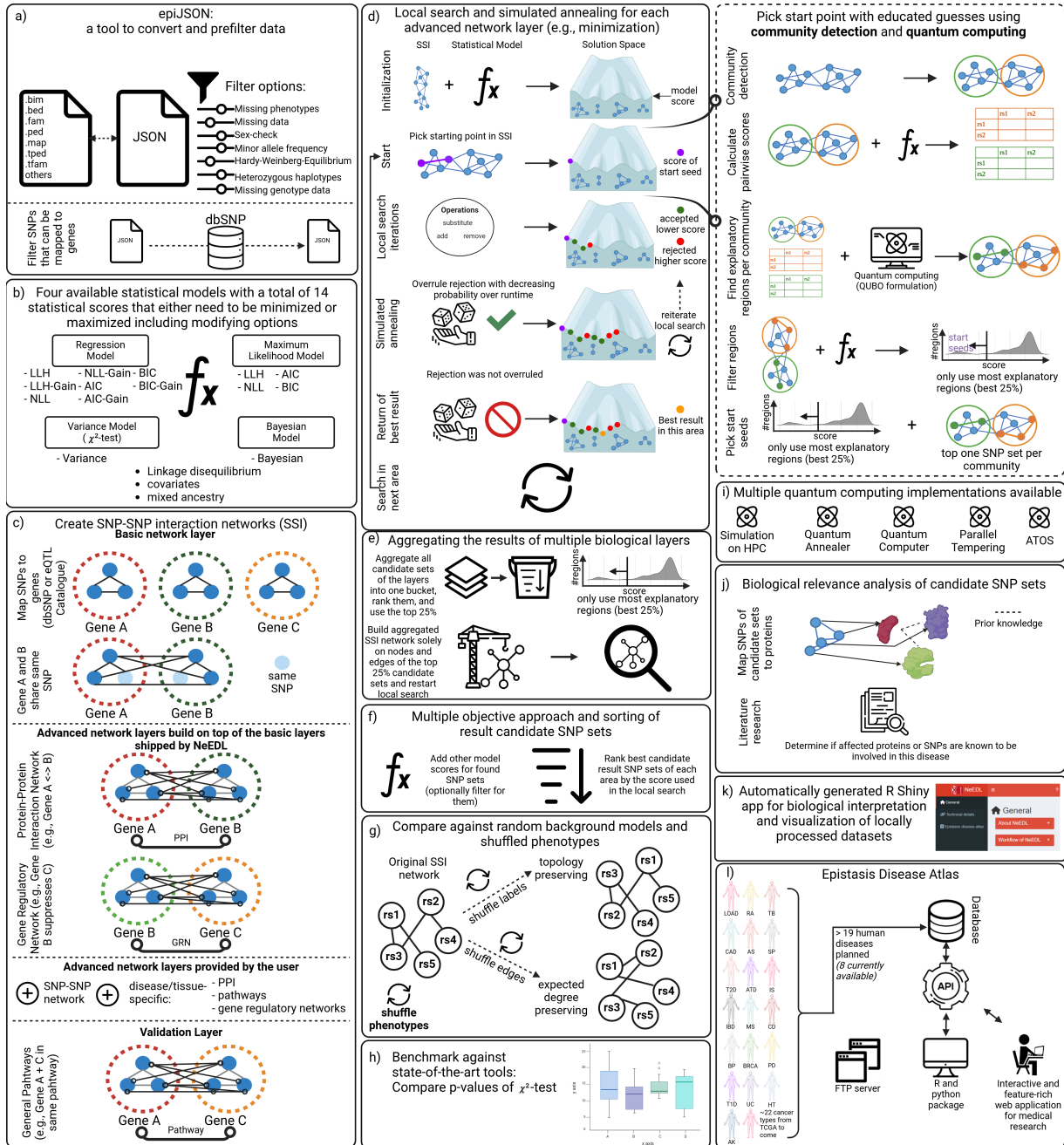
Supplementary Figure 13: An UML diagram of the Epistasis Disease Atlas SQL database. The layout roughly reflects the origin of information saved in each table. Left-oriented are tables with information specific to the project, whereas, to the right, generally available information to annotate given data is shown.



Supplementary Figure 14: An UML diagram of the Epistasis Disease Atlas Neo4j database. Nodes (entities) are represented as tables and edges (relationships) as connections between tables. The database design is fully focused on executing necessary queries time-efficiently.



Supplementary Figure 15: The Epistasis Disease Atlas browse page.



Supplementary Figure 16: Graphical overview of the whole project.

Supplementary Table 1: Feature comparison between NeEDL and existing epistasis detection tools which we initially aimed to include in our benchmark. We selected recently published methods and methods that were incorporated into recently published methods (LinDen into Potpourri). Due to the unresolvable problems summarized in Suppl. Table 16, the benchmark was ultimately run with MACOED and LinDen as the only competitors.

	NeEDL	PoCos	MACOED	LinDen	Potpourri	biologicalEpistasis
random subroutines	start seeds selection	none	ant colony optimization	none	none	none
statistical models	χ^2 -test, multiple quadratic regression scores, Bayesian K2-score, maximum likelihood model	quadratic regression	χ^2 -test, Akaike information criterion quadratic regression, Bayesian K2-score	χ^2 -test	χ^2 -test	Regr. for SNP interactions, truncated product method to derive pair P -value, simulated distribution with permutations under the null hypothesis
candidate SNP set size	variable (default: 2-10)	hundreds to thousands	2	2	2	2
implementation	C++ using boost and iGraph	MatLab	C++	C++	C++/MatLab	nextflow pipeline using R and bash

Supplementary Table 2: Number of cases and SNPs per disease that are discussed in this manuscript before and after the usage of epiJSON.

disease	Raw data set		Cleaned and filtered data set			
	SNPs	samples	SNPs	cases	controls	sum of samples
LOAD	759916	1599	6041	997	573	1570
BD	500568	3502	142441	1993	1498	3491
CAD	500568	3492	142347	1979	1498	3477
T1D	500568	3504	142432	1997	1498	3495
T2D	500568	3503	142546	1973	1498	3471
HT	500568	3505	142272	1996	1498	3494
IBD	500568	3509	142879	1993	1498	3491
RA	500568	3503	142466	1973	1498	3471

Supplementary Table 3: Effects of the filters of epiJSON (pre-filtering and pre-cleaning steps) on numbers of SNPs and samples in Late-Onset Alzheimer's Disease (LOAD).

filter step	#SNPs	#samples	#cases	#controls
raw data	759916	1599	1014	585
merge case and control	control were already included			
RS-ID and duplicate filter	756940	1599	1014	585
missing phenotype	756940	1599	1014	585
missingness >20%	756940	1599	1014	585
sex discrepancy	756940	1570	997	573
MAF <5%	756940	1570	997	573
HWE	756940	1570	997	573
heterozygosity	756940	1570	997	573
missingness >0%	17044	1570	997	573
dbSNP database	9550	1570	997	573
SNP-SNP connectivity	6041	1570	997	573

Supplementary Table 4: Effects of the filters of epiJSON (pre-filtering and pre-cleaning steps) on numbers of SNPs and samples in Bipolar Disorder (BD).

filter step	#SNPs	#samples	#cases	#controls
raw data	500568	3502	1998	1504
merge case and control	500566	3502	1998	1504
RS-ID and duplicate filter	500566	3502	1998	1504
missing phenotype	500566	3502	1998	1504
missingness >20%	500566	3502	1998	1504
sex discrepancy	500566	3491	1993	1498
MAF <5%	384955	3491	1993	1498
HWE	379007	3491	1993	1498
heterozygosity	379007	3491	1993	1498
missingness >0%	379007	3491	1993	1498
dbSNP database	207273	3491	1993	1498
SNP-SNP connectivity	142441	3491	1993	1498

Supplementary Table 5: Effects of the filters of epiJSON (pre-filtering and pre-cleaning steps) on numbers of SNPs and samples in Coronary Artery Disease (CAD).

filter step	#SNPs	#samples	#cases	#controls
raw data	500568	3492	1988	1504
merge case and control	500566	3492	1988	1504
RS-ID and duplicate filter	500566	3492	1988	1504
missing phenotype	500566	3492	1988	1504
missingness >20%	500566	3492	1988	1504
sex discrepancy	500566	3477	1979	1498
MAF <5%	384643	3477	1979	1498
HWE	378631	3477	1979	1498
heterozygosity	378631	3477	1979	1498
missingness >0%	378631	3477	1979	1498
dbSNP database	207105	3477	1979	1498
SNP-SNP connectivity	142347	3477	1979	1498

Supplementary Table 6: Effects of the filters of epiJSON (pre-filtering and pre-cleaning steps) on numbers of SNPs and samples in Diabetes Type 1 (T1D).

(d) Diabetes Type 1 (T1D)				
filter step	#SNPs	#samples	#cases	#controls
raw data	500568	3504	2000	1504
merge case and control	500566	3504	2000	1504
RS-ID and duplicate filter	500566	3504	2000	1504
missing phenotype	500566	3504	2000	1504
missingness >20%	500566	3504	2000	1504
sex discrepancy	500566	3495	1997	1498
MAF <5%	384838	3495	1997	1498
HWE	378856	3495	1997	1498
heterozygosity	378856	3495	1997	1498
missingness >0%	378856	3495	1997	1498
dbSNP database	207255	3495	1997	1498
SNP-SNP connectivity	142432	3495	1997	1498

Supplementary Table 7: Effects of the filters of epiJSON (pre-filtering and pre-cleaning steps) on numbers of SNPs and samples in Diabetes Type 2 (T2D).

filter step	#SNPs	#samples	#cases	#controls
raw data	500568	3503	1999	1504
merge case and control	500566	3503	1999	1504
RS-ID and duplicate filter	500566	3503	1999	1504
missing phenotype	500566	3503	1999	1504
missingness >20%	500566	3503	1999	1504
sex discrepancy	500566	3471	1973	1498
MAF <5%	384967	3471	1973	1498
HWE	379211	3471	1973	1498
heterozygosity	379211	3471	1973	1498
missingness >0%	379211	3471	1973	1498
dbSNP database	207379	3471	1973	1498
SNP-SNP connectivity	142546	3471	1973	1498

Supplementary Table 8: Effects of the filters of epiJSON (pre-filtering and pre-cleaning steps) on numbers of SNPs and samples in Hypertension (HT).

filter step	#SNPs	#samples	#cases	#controls
raw data	500568	3505	2001	1504
merge case and control	500566	3505	2001	1504
RS-ID and duplicate filter	500566	3505	2001	1504
missing phenotype	500566	3505	2001	1504
missingness >20%	500566	3505	2001	1504
sex discrepancy	500566	3494	1996	1498
MAF <5%	384400	3494	1996	1498
HWE	378553	3494	1996	1498
heterozygosity	378553	3494	1996	1498
missingness >0%	378553	3494	1996	1498
dbSNP database	207051	3494	1996	1498
SNP-SNP connectivity	142272	3494	1996	1498

Supplementary Table 9: Effects of the filters of epiJSON (pre-filtering and pre-cleaning steps) on numbers of SNPs and samples in Inflammatory Bowel Disease (IBD).

filter step	#SNPs	#samples	#cases	#controls
raw data	500568	3509	2005	1504
merge case and control	500566	3509	2005	1504
RS-ID and duplicate filter	500566	3509	2005	1504
missing phenotype	500566	3509	2005	1504
missingness >20%	500566	3509	2005	1504
sex discrepancy	500566	3491	1993	1498
MAF <5%	386025	3491	1993	1498
HWE	380215	3491	1993	1498
heterozygosity	380215	3491	1993	1498
missingness >0%	380215	3491	1993	1498
dbSNP database	207890	3491	1993	1498
SNP-SNP connectivity	142879	3491	1993	1498

Supplementary Table 10: Effects of the filters of epiJSON (pre-filtering and pre-cleaning steps) on numbers of SNPs and samples in Rheumatoid Arthritis (RA).

filter step	#SNPs	#samples	#cases	#controls
raw data	500568	3503	1999	1504
merge case and control	500566	3503	1999	1504
RS-ID and duplicate filter	500566	3503	1999	1504
missing phenotype	500566	3503	1999	1504
missingness >20%	500566	3503	1999	1504
sex discrepancy	500566	3471	1973	1498
MAF <5%	384923	3471	1973	1498
HWE	379039	3471	1973	1498
heterozygosity	379039	3471	1973	1498
missingness >0%	379039	3471	1973	1498
dbSNP database	207314	3471	1973	1498
SNP-SNP connectivity	142466	3471	1973	1498

Supplementary Table 11: Data from UK Biobank (www.ukbiobank.ac.uk, project ID 54273) used for running replication in Late-Onset Alzheimer’s Disease (LOAD). Individuals were selected based on ICD-10 code inclusion, controls were sampled excluding individuals with any of the ICD-10 code exclusion.

data set	#case/control	ICD-10 code inclusion	ICD-10 code exclusion
LOAD British ancestry	758/8055	G30	G, E, F
LOAD mixed ancestry	845/8954	G30	G, E, F

Supplementary Table 12: Table of parameters used for conducting experiments on NeEDL via the quantum computing seeding procedure. Legend: values $c_{\min}$ and $c_{\max}$ denote the minimum and maximum size of the subproblem, which the community detection tool utilizes to divide large instances into smaller ones. All other parameters are related to the QUBO formulation. Specifically, K represents the desired size of the SNPs set; N indicates the number of SNP sets returned; ν denotes the weight of the linear combination of statistical and biological information in the pairwise correlation metric between SNPs; $\lambda_0, \lambda_1, \lambda_2$ denote the weight of the different terms in the QUBO formulation.

Device	$c_{\min}$	$c_{\max}$	K	N	ν	λ_0	λ_1	λ_2
Simulated Annealing	2	10	5	1	0.5	5.0	1.0000	1.0000
Simulated Annealing	2	10	5	1	0.5	10.0	2.0000	1.0000
Simulated Annealing	2	10	5	1	0.5	10.0	1.0000	2.0000
Simulated Annealing	2	10	5	1	0.5	2.0	1.0000	1.0000
Simulated Annealing	2	100	5	1	0.5	5.0	1.0000	1.0000
Simulated Annealing	2	100	5	1	0.5	10.0	2.0000	1.0000
Simulated Annealing	2	100	5	1	0.5	10.0	1.0000	2.0000
Simulated Annealing	2	100	5	1	0.5	2.0	1.0000	1.0000
Simulated Annealing	2	100	5	1	0.5	10.0	0.1000	0.1000
Simulated Annealing	2	100	5	1	0.5	10.0	0.0500	0.0500
Simulated Annealing	2	100	5	1	0.5	10.0	0.0100	0.0100
Simulated Annealing	2	100	5	1	0.5	10.0	0.0010	0.0010
Simulated Annealing	2	100	5	1	0.5	10.0	0.0050	0.0050
Simulated Annealing	2	100	5	1	0.5	10.0	0.0010	0.0010
Simulated Annealing	2	100	5	1	0.5	10.0	0.0005	0.0005
Simulated Annealing	2	100	5	1	0.5	10.0	0.0001	0.0001
Parallel Tempering	2	100	5	1	0.5	10.0	0.0500	0.0500
Parallel Tempering	2	100	5	1	0.5	10.0	0.0100	0.0100
Parallel Tempering	2	100	5	1	0.5	10.0	0.0010	0.0010
Parallel Tempering	2	100	5	1	0.5	10.0	0.0050	0.0050
Gurobi	2	100	5	1	0.5	10.0	0.0500	0.0500
Gurobi	2	100	5	1	0.5	10.0	0.0100	0.0100
Gurobi	2	100	5	1	0.5	10.0	0.0010	0.0010
Gurobi	2	100	5	1	0.5	10.0	0.0050	0.0050
D-Wave Advantage 6.1	2	100	5	1	0.5	5.0	1.0000	1.0000
D-Wave Advantage 6.1	2	100	5	1	0.5	10.0	2.0000	1.0000
D-Wave Advantage 6.1	2	100	5	1	0.5	10.0	1.0000	2.0000
D-Wave Advantage 6.1	2	100	5	1	0.5	2.0	1.0000	1.0000
D-Wave Advantage 6.1	2	100	5	1	0.5	10.0	0.0000	0.0000
D-Wave Advantage 6.1	2	100	5	1	0.5	10.0	0.0100	0.0100
D-Wave Advantage 6.1	2	100	5	1	0.5	10.0	0.0050	0.0050
D-Wave Advantage 6.1	2	100	5	1	0.5	10.0	0.0010	0.0010
QAOA (simulated)	2	5	5	1	0.5	5.0	1.0000	1.0000
QAOA (simulated)	2	5	5	1	0.5	10.0	2.0000	1.0000
QAOA (simulated)	2	5	5	1	0.5	10.0	1.0000	2.0000
QAOA (simulated)	2	5	5	1	0.5	2.0	1.0000	1.0000
QAOA (simulated)	2	5	5	1	0.5	10.0	0.0100	0.0100
QAOA (simulated)	2	5	5	1	0.5	10.0	0.0050	0.0050
QAOA (simulated)	2	5	5	1	0.5	10.0	0.0010	0.0010
QAOA (simulated)	2	5	5	1	0.5	10.0	0.0001	0.0001
QAOA (IBM)	2	5	5	1	0.5	10.0	0.0005	0.0005

Supplementary Table 13: Detailed parameters and their settings in NeEDL software. (Page 1)

Parameter	Selected Value	Reason
--network-shuffle-method	Not selected for default runs, selected for network-shuffled runs: <code>TOPOLOGY_PRESERVING_WITHOUT_SNP_DEGREE</code> , <code>EXPECTED_DEGREE_KEEP_INDIVIDUAL_DEGREE</code>	Investigate information in the PPI-based SSI network using different algorithms.
--snp-annotate-dbSNP	Flag set	By default, dbSNP annotation is used to create the PPI network. eQTL mapping also available.
--network-BIOGRID	Flag set	Uses BIOGRID to connect SNPs with the same annotation to form the PPI network.
--ms-model	<code>PENETRANCE_NLL</code>	According to Blumenthal et al., it offers the best trade-off between accuracy and runtime among various statistical models.
--ms-max-rounds	300	Optimized to investigate local areas after 300 rounds of local search, suggesting a new area start is beneficial.
--ms-min-set	2	Minimum size to detect epistasis; must be more than 1.
--ms-max-set	10	A reasonable cut-off for the diseases studied, rarely exceeded.
--ms-annealing-type	<code>SIMULATED_ANNEALING</code>	Local search can easily get stuck at a local minimum; simulated annealing helps explore further.
--ms-cooling-factor	1.0	Standard value for simulated annealing cooling factor.
--ms-annealing-start-prob	0.8	Standard starting probability for simulated annealing.
--ms-annealing-end-prob	0.01	Standard ending probability for simulated annealing.

Supplementary Table 14: Detailed parameters and their settings in NeEDL software. (Continued, Page 2)

Parameter	Selected Value	Reason
--ms-seeding-routine	RANDOM.CONNECTED for normal runs, QUANTUM.COMPUTING for QC runs	Allows switching between random and quantum optimized seed selection.
--ms-rc-start-seeds	5000 only set for non-QC runs	Optimal number of start seeds for the current network size; adjust for larger networks.
--ms-qc-min-cluster-size	2	Minimum size for instance generation in community detection; cannot be less than 2 vertices.
--ms-qc-max-cluster-size	100 for D-Wave	Maximum instance size for community detection.
--ms-qc-n-clique	1	Number of solutions returned by the quantum procedure.
--ms-qc-k	5	Preferred solution size (cardinality of SNP set).
--ms-qc-nu	10.0	Constraint weight in QUBO formulation.
--ms-qc-lambda1	0.005	Constraint weight in QUBO formulation.
--ms-qc-lambda2	0.005	Constraint weight in QUBO formulation.
--ms-qc-mode	Optimization engine, one between 'SIMULATED_ANNEALING', 'PARALLEL_TEMPERING', 'GUROBI', 'QUANTUM_ANNEALING', 'QAOA'	Seeding mode (String)
--ms-qc-sa-num-samples	1000	Number of samples drawn. Only for Simulated Annealing (Positive integer)
--ms-qc-sa-num-sweeps	100000	Number of sweep (state updates) in a single execution. Only for Simulated Annealing (Positive integer)
--ms-qc-sa-seed	1356	Random seed. Only for Simulated Annealing (Positive integer)
--ms-qc-qa-token	reserved	Token of the D-Wave Leap account. Only for Quantum Annealing (String)
--ms-qc-qa-num-reads	1000	Number of samples of the solution. Only for Quantum Annealing (Positive integer) and Gurobi (number of iterations)
--ms-qc-qa-solver-idx	4	Integer from 0 to 6. Label of the solver namely 0 = 'hybrid_binary_quadratic_model_version2' 1 = 'hybrid_discrete_quadratic_model_version1' 2 = 'hybrid_constrained_quadratic_model_version1p' 3 = 'Advantage_system4.1' 4 = 'Advantage_system6.1' 5 = 'DW_2000Q_6' 6 = 'DW_2000Q_VFYC_6'. Only for Quantum Annealing

Supplementary Table 15: Detailed parameters and their settings in NeEDL software. (Continued, Page 3)

Parameter	Selected Value	Reason
--ms-qc-qa-fw-annealing-ramp-time	1.0	Ramp time in ms during forward annealing. Only for Quantum Annealing. (Positive real)
--ms-qc-qa-fw-annealing-pause-time	1.0	Pause time in ms during forward annealing. Only for Quantum Annealing (Positive real)
--ms-qc-qa-rev-annealing-ramp-time	1.0	Ramp time in ms during reverse annealing. Only for Quantum Annealing (Positive real)
--ms-qc-qa-rev-annealing-pause-time	1.0	Pause time in ms during reverse annealing. Only for Quantum Annealing (Positive real)
--ms-qc-qa-rev-annealing-s-target	0.5	Minimum reverse evolution. Only for Quantum Annealing (Positive real from 0.0 to 0.990)
--ms-qc-qaoa-vendor	'none' for normal QC run 'ibm' for QAOA IBM runs	Quantum vendor. Only for QAOA (String: 'none' or 'ibm' or 'azure')
--ms-qc-qaoa-azure-subscription-id	Azure platform not used for our experiments	Only for QAOA on Azure quantum the subscription id.(String)
--ms-qc-qaoa-azure-location	Azure platform not used for our experiments	Only for QAOA and Azure quantum the location of the Azure account (e.g. west-europe)
--ms-qc-qaoa-azure-backend	Azure platform not used for our experiments	Name of the quantum device for IBM or Azure quantum machines. Only for QAOA
--ms-qc-qaoa-optimizer	'ADAM'	Optimization technique for QAOA parameters. Only for QAOA (String: 'ADAM' 'COBYLA' 'NELDER_MEAD' or 'SLSQP')
--ms-qc-qaoa-reps	1	Number of repetition of the QAOA algorithm. Only for QAOA
--ms-qc-qaoa-n-shots	1024 for QAOA SIM 128 for QAOA IBM runs	Number of samples drawn from the quantum circuit suggested at least 1024. Only for QAOA
--ms-qc-qaoa-is-recursive-qaoa	0	If set to '1' enable to Recursive version of QAOA which is much more expensive and 0 otherwise. Only for QAOA
--disable-final-search	flag implicitly set as we are only using a single network	As we only have a single network we do not want to search inside the results again. This parameter was implemented for the integration of multiple networks which has not been discussed in this manuscript. However NeEDL is already fully functional for integrating multiple networks.
--disable-save-network	flag set only for shuffled-network runs	Network files are very large so it was intractable to save them for all the shuffled network runs.
--ms-search-time-limit	7d	The time limit was never reached.

Supplementary Table 16: Software problems encountered during benchmarking. Tools, where some problems could not be resolved after months of debugging, were excluded from the benchmark.

Tool	Problem	Resolved
PoCos	reported SNP sets are too big to be statistically evaluated (size > 2000 SNPs)	not resolvable
PoCos	the program does not report the found candidate SNP sets out of the box	resolved
PoCos	implemented in the proprietary language MatLab which complicates deploying it on a cluster	resolved
MACOED	configuration is not correctly parsed if entries are too long (buffer overflow potentially overwrites internal data structure)	resolved
MACOED	exhaustive memory consumption (1.3 TB)	resolved
MACOED	overwrites its own results	resolved
MACOED	no build dependencies	resolved
MACOED	no build pipeline available (e.g., CMake)	resolved
MACOED	computationally not parallelizable	not resolved, but not necessary for usage
Potpourri	crashes on input data without useful error messages (segmentation fault) despite correct formatting after their instructions and same format as example input data	not resolvable
Potpourri	user needs to write his/her own Matlab script alongside a demo script to run on custom data	resolved
Potpourri	is implemented in the proprietary language MatLab which complicates deploying it on a cluster	resolved
biologicalEpistasis	cannot be run on larger HPC clusters because the pipeline spawns slurm jobs exhaustively	not resolvable
biologicalEpistasis	did not terminate a single dataset after 3 months of runtime	not resolvable

Supplementary Table 17: Computational resources used for our benchmark study. Note that MACOED was run 100 times for each benchmark dataset to account for the fact that it contains randomized subroutines.

Tool	Main memory (estimates)	total runtime per benchmark dataset (estimates)
NeEDL	50 GB	2 hours
LinDen	500 MB	0.5 hours
MACOED	1.3 TB	48 hours

Supplementary Materials

Supplementary Materials 1: Functioning of quantum annealers

We describe the physical realization of a quantum annealer, in particular the D-Wave Advantage 6.1. This machine is composed of 5616 superconducting qubits. Each qubit is realized by means of superconducting quantum-interference (SQUID) implementation, that can be controlled via radio frequency pulses²⁰. Such pulses are also responsible for and used to manipulate the coupling between the qubits, in such a way that during the annealing time, the system evolves toward the Hamiltonian in which the optimization problem is encoded (see²¹ for more details). If the hypothesis of the adiabatic theorem holds, and if the annealing time is sufficiently long, the time-dependent Hamiltonian in Equation (5) will remain in its ground state during the annealing process.

A paramount aspect of quantum annealers is the topology of the QPUs, i.e. how many qubits the machine has and how they are connected to each other. Superconductor-based quantum hardware usually has restrictive topologies, implying that if two non-adjacent qubits must communicate, we will need to pass the information through a chain of qubits connecting them. D-Wave Advantage 6.1 has 40135 couplers to link its 5616 qubit, which largely outperforms its predecessor D-Wave 2000Q in terms of connectivity (with the latter having only 2048 qubits and 6016 couplers). When mapping an instance of QUBO problem onto the quantum hardware, in the best-case scenario, each QUBO variable will be mapped to a single qubit of the QPU. The mapping between QUBO variables to qubits is called *minor embedding* and consists concretely in finding an isomorphism between the graph underlying the QUBO instance and the QPU topology graph. As the best graph isomorphism algorithm's runtime is super-polynomial in the input size²², we have to rely on heuristics to find a suitable solution. Not only the minor embedding should be efficient in terms of asymptotic complexity but also extremely fast in order to add a negligible overhead to the optimization procedure. The D-Wave software framework, Ocean Software Development Kit, already implements several heuristics to perform minor embedding. A comparison of minor embedding techniques for D-Wave device's topologies is shown in²³. The degree of the nodes in the topology of D-Wave devices is bounded by some constant, we can expect Ising instances described by dense graphs won't be successfully mapped. However, we can artificially enlarge the connectivity of qubits by grouping several physical qubits into a *chain*, forming a single logical qubit. Then the graph of the Ising problem is mapped to the logical qubits. As the readout process will read all the physical qubits, there is the possibility that the readout finds a solution that has different values for the physical qubits associated with a single logical qubit. This problem, called *chain break*, is resolved using deterministic techniques, such as majority voting, which assigns the most frequent value of the chain to the logical qubit. The resource overhead of implementing chains of physical qubits results in the fact that, if theoretically a quantum annealer with m qubits can solve a QUBO problem on up to m binary variables, in practice, the number is much smaller.

The D-Wave QPU must be programmed using the coefficients h, J . Although Ising formulations allow coefficients to range from negative infinity to positive infinity, the quantum annealer requires them to be scaled to a finite range. For the D-Wave Advantage 6.1, the range for h coefficients is $[-4, 4]$ and for J coefficients, it is $[-1, 1]$. It is worth noting that these ranges can be slightly expanded for enforcing specific conditions, such as ensuring adherence to the physical-logical qubit mapping. The encoding precision can impact solution quality especially if there are large coefficients that fall at the extremes of the encoding range, causing smaller coefficients to be scaled to zero near the machine precision. In such a scenario, the D-Wave would effectively be solving a different problem than the one intended. By analyzing the distribution of coefficients h and J we can rule out that our specific instance is affected by the encoding precision problem. Additionally, the encoding precision is nonlinear, with a range of $[0.0005, 0.004]$ for h coefficients and $[0.0005, 0.002]$ for J coefficients, as specified in the technical documentation²⁴. This means that certain range coefficients are represented with greater accuracy, suggesting the QUBO formulation may be biased to have most coefficients in such favorable ranges.

The quantum annealer evolves its Hamiltonian according to the equation,

$$H(s) = A(s)H_0 + B(s)H_p \quad (1)$$

where $H_0 = \sum_{i=1}^n \sigma_x^{(i)}$ represents the free Hamiltonian, H_p denotes the Hamiltonian of the problem in question, and s is the time variable, normalized to fit within the range $[0, 1]$. The quantum annealer starts by

being in the ground state of H_0 , and as s moves from 0 to 1, $A(s)$ decreases while $B(s)$ increases. At the end of the process, the system is described by H_p and its state may be the ground state of H_p . The schedule of the annealer is determined by the equation $A(s)$ and $B(s)$, with $0 \leq A(s) \leq 1$ and $B(s) = 1 - A(s)$. The annealing schedule is defined by the total annealing time in microseconds and the shape of the curve, which is represented by a sequence of points $\{(s_i, p_i) \mid 0 \leq s_i < s_{i+1} \leq 1, 0 \leq p_i \leq p_{i+1} \leq 1 \forall i\}$ where s represents the normalized time and p the progress. The schedule can include mid-anneal quenches, which are changes in the slope, and mid-anneal pauses, which have been shown to potentially improve solution quality²⁵. Longer annealing times can have both positive and negative effects, as they facilitate the system to remain in the ground state during the whole evolution but also might cause larger decoherence errors.

The operation of a quantum annealer can be affected by multiple sources of noise due to the high number of interacting qubits, leading to a detrimental loss of quantum coherence, known as decoherence. The major sources of noise can be grouped into two categories. The first one, causing dephasing, is related to the interaction of the qubits with the external environment that can be modeled as a $1/f$ -like noise. Whereas, the second is due to imperfect tuning of the control fields and of the coupling between first-neighbor qubits that are responsible for effective next-neighbor hopping. In the case of the used quantum annealer, the second type of noise and the induced crosstalks due to imperfect tuning are referred to as *background susceptibility*. Apart from the two major ones, there are also other smaller issues that could affect the execution of a quantum annealer. These include the accuracy of the readout operation, which typically has a fidelity of over 99%, making the error close to negligible. Additionally, there is a low risk of programming errors, with an error rate below 10%.

Supplementary Materials 2: Functioning of quantum gate-based hardware

NeEDL is capable of interfacing with various quantum circuit-based hardware through its interface with both the Microsoft Azure Quantum Cloud and IBM Quantum platform. The Microsoft Azure Quantum Cloud provides access to ion-trap-based quantum processing units (QPUs) from IonQ, as well as other types of quantum devices added recently. The IBM Quantum platform offers access to a variety of superconducting-based QPUs. The hardware devices are briefly described to provide an overview of their respective features.

Ion-trap-based devices: IonQ Harmony

The IonQ Harmony quantum processor unit has a complete graph topology with 11 fully connected qubits, providing a significant advantage over superconducting-based QPUs that have limited sparse connectivity and require substantial overhead to operate on distant qubits. For detailed information on the technical specifications and performance benchmarks of IonQ devices, we refer to²⁶. Additionally, information about the specific native gate set and rotation precision of IonQ's Harmony device can be found in²⁷ and information about their system fidelity can be found on Microsoft Azure Documentation which provided access to these machines²⁸.

The IonQ Harmony Quantum Processing Unit operates using a native gate set composed of single-qubit gates $GPI(\phi)$, $GPI2(\phi)$, and $RZ(\phi)$, as well as the two-qubit gate $MS(\phi_1, \phi_2)$. The specific definitions of these gates are shown in²⁷. The IonQ Harmony QPU uses a set of native gates that is universal, i.e. can be used to create any unitary and therefore execute any quantum circuit. Careful design of the quantum circuit to follow this native set of gates can reduce the overhead of the compilation process. In particular, the operation RZ is physically more efficiently executed than GPI and $GPI2$ gates and should be favored when possible. For example, a Hamiltonian of $H = \sum_i X_i + \sum_{i,j} Z_i Z_j$ can be reformulated to $H = \sum_{i,j} X_i X_j + \sum_i Z_i$ for optimal execution on the IonQ Harmony QPU. The compiler can perform some of these transformations automatically, but for more complex cases, human intervention may be required.

The gates implemented may not perfectly match their theoretical counterparts. On average, the single-qubit gate fidelity is 99.35% and the two-qubit gate fidelity is 96.02%. Additionally, the rotational gates have a precision of $\pi 10^{-3}$.

The quantum state of the system remains coherent, i.e. it evolves correctly for a time frame that can be estimated using two constants: the T_1 relaxation time, which measures the time it takes for a state $|1\rangle$ to decay to state $|0\rangle$, with value $10^7 \mu s$, and the T_2 dephasing time, which measures the average time the system loses phase and is $2 \cdot 10^6 \mu s$. These values are much higher compared to superconducting-based

quantum hardware, indicating a significant performance advantage in the ion trap-based hardware. However, this advantage is partially worsened by the longer time it takes to perform a quantum operation, which is approximately $10\mu s$ for a single-qubit gate and $200\mu s$ for a two-qubit gate.

Quantum gate-based hardware: IBMQ Perth and Lagos

Most of the experiments performed with QAOA are carried out on IBM Quantum devices *Perth* and *Lagos*, which share several attributes such as their physical realization, architecture, and topology (refer to ²⁹ for additional information). These devices have a different technology than the ion-trap-based one, with the former being more conducive to scaling to a large number of qubits, but at the expense of lower fidelity.

The IBM QPUs used in some of our experiments is a superconducting-based quantum processing unit that has 7 qubits arranged in a specific topology. It's important to note that the results of a quantum circuit can vary greatly depending on how the qubits are mapped to the QPU's physical qubits, which is typically done by the software compiler. When qubits are mapped to non-physically adjacent qubits, additional SWAP gates must be added, which incurs a large overhead. A common, although sub-optimal, compiler strategy is to map the qubits with more two-qubit gates applied to them (e.g. CNOTs) to physical qubits with higher connectivity in a sequential, greedy manner. Additionally, it should be noted that some qubits within the architecture may have lower precision compared to others, both in terms of the accuracy of the gates applied to them and the coherence time. The T_1 and T_2 times have ranges of 24.50 to $132.70\mu s$ and 20.06 to $170.37\mu s$, respectively, whereas the gate time is approximately $0.5\mu s$. The single-qubit and two-qubit fidelity errors have ranges of 1.6 to 6.0×10^{-4} and 4.8 to 16.7×10^{-3} , respectively. The native gate set for both the IBM Perth and Lagos QPUs consists of CNOT, RZ, SX, and X gates, which together form a universal set.

Supplementary Materials 3: Quantum algorithm for SNP seeding

To solve the SNPs seeding problem, which involves finding the optimal set of candidate SNPs for local processing, we construct a SNPs graph, where vertices represent SNPs and edges represent the pairwise correlation (or other metric) between them. We assume that a group of SNPs is a strong candidate if the total pairwise correlation between the SNPs in the group is high. This enables us to model the problem as a *weighted-max-clique* problem, which consists in finding a *clique* with the highest sum of pairwise edge weight. As this is a NP-hard problem, it can be mapped, as explained before, into a QUBO formulation, which is the standard method for implementing combinatorial optimization on quantum hardware.

The Max-Clique problem

The Max-Clique problem is the computational problem of finding the maximum clique of an undirected graph and is used to model many real-world problems. A clique on a graph is a complete sub-graph of the given graph, that is, one where all the vertices are connected to all other vertices. To find a maximum clique of a graph, one could non-deterministically first try to determine a set of k distinct vertices and then test whether these vertices form a complete graph. In fact, this problem is NP-Complete, and it is conjectured that no polynomial time deterministic algorithm can be constructed to solve it.

Quadratic Unconstrained Binary Optimization (QUBO)

This is a mathematical formulation that can embrace an exceptionally large variety of important combinatorial optimization problems occurring both in industry and in science. The QUBO model has emerged as an underpinning of the quantum computing area known as quantum annealing. It can be shown to be equivalent to the Ising model, which plays a prominent role in physics, as highlighted in the paper by ³⁰, and underlies an important domain of problems arising in physics applications. QUBO models belong to a class of problems known to be NP-hard. Therefore, practically no exact solver designed to find optimal solutions (like the commercial CPLEX and Gurobi solvers) is likely to be successful except for very small problem instances. For our problem, we will use two meta-heuristic methods that are designed to find high-quality but not necessarily optimal solutions, namely *quantum annealing* and *QAOA*.

Mathematically, the QUBO problem requires minimizing a function $f : \{0, 1\}^n \rightarrow \mathbb{R}$,

$$f(x) = \sum_{i,j=1}^n x_i Q_{i,j} x_j \quad (2)$$

where Q is a $n \times n$ real matrix. The QUBO formulation allows Q to be equivalently either symmetrical or upper triangular, as every symmetric matrix is an upper triangular matrix in which the entries have been reflected on the diagonal. The former can be obtained from the latter using the transformation $Q' = Q + Q^T - \text{diag}(Q)$. Note the diagonal terms are linear since $x_i^2 = x_i$. If Q is sparse, we can equivalently describe the problem as a graph $G = (V, E, w : E \rightarrow \mathbb{R})$ where V is the set of n binary variables, E the set of edges such that $(u, v) \in E$ if and only if $Q_{u,v} \neq 0$, and $w((u, v)) = Q_{u,v}$. QUBO problem can deal with signed integer and fixed point arithmetic after appropriate encoding on a binary vector; for example, a positive integer can be encoded on d bits on the variable $x \in \{0, 1\}^d$ through the term $\sum_{i=0}^d 2^i x_i$. Any boolean function acting on n bits can be expressed using an n -th degree binary polynomial $f(x) = \sum_{i=1}^n a_i x_i + \sum_{i,j=1}^n b_{ij} x_i x_j + \sum_{i,j=1}^n c_{ij} x_i x_j + \dots$. Such a formula can be lead back to a quadratic polynomial with the use of auxiliary ('slack') binary variables $\alpha_k = x_i x_j$, although the number of necessary auxiliary variables grows super-polynomially with the degree of the polynomial. QUBO can be used to encode a constrained problem if we encode the constraint as a binary term $C(x)$, such that the formulation can be expressed as $f(x) = \lambda_0 \sum_{i,j=1}^n x_i Q_{i,j} x_j + \lambda_1 C(x)$ with $\lambda_0, \lambda_1 > 0$. For $\lambda_0 \ll \lambda_1$, the violation of the constraint results in a large cost $f(x)$, guiding the optimizer toward acceptable solutions. Authors in ³¹ suggest several strategies to effectively design constraints for QUBO formulations.

QUBO formulation for SNPs seeding

We can reduce Max-Clique to QUBO formulation, proving the NP-Harness of the QUBO: given the graph $G = (V = [1, n], E \subseteq V \times V)$ we consider the minimization of the QUBO function over the binary vector $x \in \{0, 1\}^n$,

$$f(x) = -\lambda_0 \sum_{i=1}^n x_i + \lambda_1 \sum_{(i,j) \notin E} x_i x_j \quad (3)$$

for $\lambda_0 \ll \lambda_1$: the first term reward x vectors with a large number of 1s, corresponding to larger cliques, while the second term penalizes solutions that violate the definition of the clique (vectors containing a vertex not fully connected with the others). The global minima of f is a solution for the max clique problem: the maximum clique is the set $\{v_i \in V \mid x_i = 1\}$.

To formulate epistasis detection as a QUBO problem, we consider a graph G of vertices V , edges E , and edge weight function $w : V \times V \rightarrow \mathbb{R}$ representing the relation between SNPs, i.e. the vertices are the SNPs and between each pair of vertices, there is a positive weight representing the relation between graphs according to some metrics (e.g. correlation).

We guide the optimization through the following assumption: an SNP set of an arbitrary number of elements is a good candidate solution if the sum of the pairwise metric values between SNPs is high.

Thus, the formulation solves and analyzes the graph to find several (N) candidate SNP sets, each having an arbitrary size K with the largest sum of edge weights. The formulation is named *the N max-weighted k -cliques*.

$$Q = \sum_{\ell=1}^N \left[\lambda_0 \left(\sum_i x_{i\ell} - K \right)^2 - \lambda_1 \sum_{(i,j) \notin E} w_{i,j} x_{i\ell} x_{j\ell} \right] + \lambda_2 \sum_{\ell=1}^N \sum_{m=\ell+1}^N \sum_{i=1}^{|V|} x_{i\ell} x_{im} \quad (4)$$

where:

- there are $N \times V$ binary variables $x_{i\ell}$, which is 1 if the i -th vertex (SNP) is associated with the ℓ -th solution set;
- the term $\lambda_0 (\sum_i x_{i\ell} - K)^2$ associate a penalty term for cliques that has a size different than K . The penalty is weighed by the hyper-parameter $\lambda_0 > 0$.

- the term $-\lambda_1 \sum_{1 \leq i, j \leq |V|} w_{i,j} x_{i\ell} x_{j\ell}$ reward the SNPs set identified with $\{x_{k\ell} : x_{k\ell} = 1, k = 1, \dots, |V|\}$ having the largest sum of weights. The penalty is weighed by the hyper-parameter $\lambda_0 > 1$.
- the term $\lambda_2 \sum_{\ell=1}^n \sum_{m=\ell+1}^N \sum_{i=1}^{|V|} x_{i\ell} x_{im}$ add a penalty for each pair of the N SPS set that are partially or totally overlapping, i.e. are composed of the same SNPs (vertices). The penalty is weighed by the hyper-parameter $\lambda_0 > 2$.

The formulation can be simplified to return $N = 1$ SNPs set and run the optimization technique multiple times with different random initial states. Such a simplification was necessary for any run on the quantum hardware due to the lack of resources.

Community detection

Due to quantum computing resource limitations, we could not run NeEDL on datasets containing approx. 140,000 SNPs. Hence, we need to subsample the datasets. We randomly subsample connected subnetworks from the PPI-based SSI network of the data of three diseases, namely, BD, RA, and T2D, to 100, 500, to 1,000 SNPs to showcase the speed-up that quantum computing could give in the future. We use the connected subnetworks within the PPI-derived SSI network and used it as input for the evaluation of the quantum-computing module.

When quantum computing hardware and substantial networks become available, it is not feasible to generate a matrix of all pairwise combinations of SNPs, which is required for quantum computing. Hence, we implemented the community-detection method Leiden³² to identify more densely connected parts of the network. We calculate the matrix for all pairwise SNP combinations for the subnetwork and use quantum computing on each of the subnetworks. We then choose the most promising regions identified by the quantum computer as starting seeds for the local search. See Suppl. Figure 2b on the right.

Supplementary Materials 4: Quantum Algorithms

We briefly review the quantum algorithms that we used to tackle epistasis detection on the currently available quantum computers: Quantum Annealing (QA) and Quantum Approximate Optimization Algorithms (QAOA). Quantum Annealing (QA) is the quantum counterpart to the simulated annealing algorithm³³, a metaheuristic employed to locate an approximate solution to a global optimization problem within a vast search space. Both QA and simulated annealing algorithms operate by identifying the state of minimum energy of spin glass systems studied in statistical mechanics, such as the random field Ising model³⁴. The primary distinction between the two methods lies in the adjustable parameter used to introduce the necessary fluctuations for exploring the solution space. In the case of quantum annealing, these fluctuations are generated by a field that permits quantum tunneling, as opposed to simulated annealing, which instead is based on thermal fluctuations³⁵. While annealing refers to an adiabatic model of quantum computation, QAOA is a hybrid quantum-classical algorithm that operates on circuit-based quantum computers. Both techniques are based on the total Hamiltonian of the system, H , which is a Hermitian operator acting on the Hilbert space $\mathcal{H}$ describing the system of N quantum spins ($\mathcal{H} = \mathbb{C}^{2^N}$). The eigenvalues of this operator correspond to the energy spectrum of the system. After an appropriate mapping of our problem, the strategy is to identify the lowest eigenvalue of the Hamiltonian, which is equivalent to identifying the lowest energy level or ground state of the system. To provide a more rigorous definition, we can describe the general structure of a Hamiltonian. This consists of a free part H_0 , which accounts for the free energy of all spins in the system, and an interaction term H_I , which represents the contribution to the total energy caused by the magnetic interaction among the spins. These two terms are typically non-commuting, i.e., $[H_0, H_I] = H_0 H_I - H_I H_0 \neq 0$, making it extremely difficult to find the exact eigenvalues of H either analytically or numerically, particularly when the number of spins in the system is high, i.e., $N \gg 1$. The total time-dependent Hamiltonian is written as

$$H(t) = (1 - \mu(t))H_0 + \mu(t)H_I, \quad (5)$$

where $\mu(t)$ is a control function that takes values in the interval $[0, 1]$ and allows switching between the two terms. Although each of the two terms could be, in principle, easy to diagonalize, the sum of free terms and interactions could lead to equations that are too complex to be solved exactly. In addition, if the computation time is limited to the interval $[0, \tau]$, a common procedure is to initialize the system in the ground state of H_0

with $\mu(0) = 0$, and then adjust the control parameter until reaching $\mu(\tau) = 1$. If the behavior of $\mu(t)$ satisfies the adiabatic theorem³⁶, both QA and QAOA can be seen as simplified versions of adiabatic quantum computation³⁷. As for the QAOA, it is implemented on gate-based quantum hardware by decomposing the unitary evolution $U(t_0, t) = \exp\left\{(-i/\hbar) \int_{t_0}^t H(t') dt'\right\}$ from the full Hamiltonian in Equation (5) into the Suzuki-Trotter decomposition³⁸.

Quantum annealing

Thermal and Quantum Annealing algorithms^{34,39} find their roots in the statistical mechanics' problem of finding the minimum-energy state of a generic Ising model. To explain these techniques, we introduce the random field Ising model, both classical and quantum. For the former we consider classical spin variables s_i that take values ± 1 , so that the energy of a system of N interacting spin arranged over a d -dimensional lattice is the classical Hamiltonian function:

$$H(s) = - \sum_{\langle i,j \rangle} J_{i,j} s_i s_j - \sum_i h_i s_i. \quad (6)$$

Here the notation $\langle i, j \rangle$ indicates that the sum is over all the pairs of neighboring sites, while the J_{ij} are the couplings strengths between two spins and h_i is a random external field. The quantum version of the Ising Hamiltonian in Equation (6)) is obtained by replacing the binary variable s_i with the corresponding Pauli matrix σ_i^α , $\alpha = \{x, y, z\}$:

$$H(\sigma) = - \sum_{\langle i,j \rangle} J_{i,j} \sigma_i^z \sigma_j^z - \sum_{i=1}^N h_i \sigma^x. \quad (7)$$

The above expression is not anymore a function but a Hermitian operator. It can be linked to the introductory discussion of this section, noting that the free term is $H_0 = \sum_{i=1}^N h_i \sigma^x$ and the interaction term coincides with $H_I = - \sum_{\langle i,j \rangle} J_{i,j} \sigma_i^z \sigma_j^z$. This type of Hamiltonian is used to describe the physics of spin glasses and their phases. Finding the minimal energy of such Hamiltonian is a NP-hard problem (specifically NP-complete) and the way it is linked to the solution of many NP problems is reported in⁴⁰. The Ising Hamiltonian formulation is tightly related to some mathematical optimization problems called Quadratic Binary Optimization (QUBO) problems, which belong to the NP-hard complexity class too. A QUBO problem can be mapped to an Ising Hamiltonian by transforming each variable x_i of the former as follows:

$$x_i \leftrightarrow \frac{1 + \sigma_i^z}{2}, \quad (8)$$

This transformation guarantees that the solution of the optimization task will coincide with the search for the ground state of the Ising model considered. We will investigate in the next section the relation between QUBO and the SNP seeding optimization problem. A hardware device dedicated to solving Ising problems is an *Ising machine*. Authors in⁴¹ have reviewed the many Ising machines proposed in the literature, based on different physical phenomena. Quantum annealers such as D-Wave are instances of Ising machines, as well as gate-based quantum hardware when used with the QAOA algorithm.

QAOA

The second algorithm, QAOA, was introduced in⁴² and can be implemented on the currently available noisy intermediate-scale quantum (NISQ) hardware supporting a gate-base model of computation. The algorithm, which is pictured in Suppl. Figure 3, corresponds to a quantum circuit that can be briefly described as follows. The N -qubit input state of the quantum circuit is initialized to $|+\rangle^{\otimes N} = (2^{-1/2} |0\rangle + 2^{-1/2} |1\rangle)^{\otimes N}$, where $|+\rangle$ is the eigenvector relative to the eigenvalue 1 of the Pauli matrix σ_x . Subsequently, a parametrized wavefunction $|\psi(\vec{\gamma}, \vec{\beta})\rangle$ is constructed by designing a suitable ansatz that takes into account the nature of the optimization problem one has to solve. The ansatz is constructed by acting on the initial state $|+\rangle^{\otimes N}$ with p repetitions of a sequence of two parametric unitary operation $U_C(\gamma_k), U_B(\beta_k)$. Therefore, the whole ansatz and the wavefunction will both depend on $2p$ parameters $\gamma_1, \beta_1, \dots, \gamma_{2p}, \beta_{2p} \in [-\pi, \pi]$. The link with the Hamiltonian operator defined in Equation (7) is given by setting $\mu(t) = \frac{1}{2}$, then the first unitary operator

reads $U_C(\gamma) = \exp\{-i\gamma H_I\}$ and H_I is the term in which the problem is encoded. Lastly, the second unitary operator is defined as $U_B(\beta) = \exp\{-i\beta H_0\}$. We observe that two Hamiltonian operators H_C, H_B (and consequently the associated unitary operators) do not have to commute to obtain meaningful results. Now, we can write the parameterized wavefunction, resulting from the circuit as

$$\left|\psi(\vec{\gamma}, \vec{\beta})\right\rangle = U_B(\beta_p)U_C(\gamma_p)\cdots U_B(\beta_1)U_C(\gamma_1)|+\rangle^{\otimes n}, \quad (9)$$

on which we compute the mean value of the energy of the system defined as

$$E(\vec{\gamma}, \vec{\beta}) = \left\langle \psi(\vec{\gamma}, \vec{\beta}) \left| H_C \right| \psi(\vec{\gamma}, \vec{\beta}) \right\rangle. \quad (10)$$

This expression is now a classical function that depends on $2p$ parameters and has to be minimized by means of classical optimization to find the optimal variational parameters $(\vec{\gamma}^*, \vec{\beta}^*)$ such that:

$$(\vec{\gamma}^*, \vec{\beta}^*) = \arg \min_{\vec{\gamma}, \vec{\beta}} E(\vec{\gamma}, \vec{\beta}). \quad (11)$$

Motivating the use of QA and QAOA

Quantum annealing and QAOA are heuristics and give no guarantees of speedup beforehand. However, the similarity they share with Grover’s algorithm, which gives us guarantees of polynomial speedup albeit being more demanding in terms of resources needed (can only be run on fault-tolerant quantum devices, in contrast with the heuristics being supported also on noisy devices), leads us to believe there is a benefit in using these heuristics anyway. In particular, note that the noiseless, limitation-free version of quantum annealing is the adiabatic quantum computing formalism which has proven to be able to offer the same speedup as Grover’s search in black box optimization tasks⁴³.

An important aspect to discuss is the implication in complexity theory of the use of quantum computers. It is widely believed in the scientific community that neither quantum nor classical computers can solve NP-hard combinatorial optimization tasks (including QUBO) in polynomial time. The best-known speedup in black-box optimization search, which is even less structured than most NP-hard problems, is quadratic concerning its classical counterpart. For more structured tasks, we can expect higher polynomial speedups or even super-polynomial ones.

Note that polynomial speedups are indeed important when studying epistatic interaction. Our experimental work suggests that epistasis interaction occurs with SNP sizes of k from 3 to 7. So, from a computational complexity point of view, we could think of the epistasis detection as a black-box search of a statistical score function in a search space of $\binom{n}{k} \in O(n^k)$ elements, which for $k = 3, \dots, 7$ makes a search space of $O(n^7)$ elements. This is already a polynomial, thus efficient in the computational complexity sense. However, from a practical point of view, analyzing billions of SNPs renders the problem intractable on any classical device, motivating the interest in quantum computation.

Supplementary Materials 5: *quepistasis* module

The purpose of the *quepistasis* module is to connect the NeEDL software with the quantum hardware. It requires the input of a pairwise correlation matrix between SNPs, where the correlation can be measured using statistical correlation or any other metric. The output is one or more sets of SNPs, which are then used in the local search procedure. The structure of *quepistasis* is schematized in Suppl. Figure 4.

The *quepistasis* module creates the QUBO formulation of the SNP seeding problem, which can be solved using the classical solver (thermal/simulated annealing, parallel tempering, Gurobi). Since Python 3 is the standard programming language for quantum computing and supports most quantum SDKs, we have written an interface between NeEDL and this software. Thus, the Python interface submodule takes the QUBO matrix and the target platform as inputs, along with some configuration parameters, and returns the binary vector solution found by the quantum device. This solution is then interpreted as an SNP set.

It is important to notice that the *quepistasis* is supported by the community detection tool: as the D-Wave quantum annealer and IBM/IonQ devices can deal with problems of approximately 100s and 10s of SNPs due to current technological constraints, a preprocessing phase that heuristically split any arbitrarily large instance of SNPs into smaller chunk is necessary.

Supplementary Materials 6: Detailed analysis of the quantum computing experiments

In this section, we provide a detailed account of the experiments conducted on quantum devices. Firstly, we discuss the effectiveness of an optimization-based seeding procedure as compared to the random sampling-based procedure of Linear NeEDL. We then compare various quantum and classical devices in terms of execution speed to complete the optimization task. The primary objective of our experiments was to demonstrate that the optimization-based technique expedites the costly local search phase, which is based on a multitude of randomly selected seeds, by carefully selecting fewer seeds, namely those resulting from solving our optimization problem. Secondly, we investigate how the optimization process would scale on different devices with respect to the size of the dataset. Thirdly, we analyze the distribution of coefficients in the QUBO matrices to ensure their effective encoding on the quantum hardware without any loss of precision.

Thanks to the community detection tool, we are able to break any dataset into SNPs set of arbitrary size, allowing *quepistasis* to be tested on any kind of dataset. In principle, this allows us to solve datasets of arbitrary size on even the smallest quantum computer. However, we could use subsampling only for analyzing the quantum computing procedure scaling, while its use for quantum resource reduction, by working on smaller sets of data, was not viable due to the long waiting time on the quantum platform available on the cloud. In fact, for large sets, community detection generates a large number of subproblems, which, although very small in dimensions, must be sent to a quantum computer one by one, each contributing an overhead generated by the sending the task, waiting for the scheduling and receiving the results.

Experimental setup

To determine the potential speedup achieved through the quantum technique, we assessed the performance of NeEDL with and without the quantum computing module in operation.

Firstly, we performed NeEDL in a linear fashion without the use of quantum computing. In this scenario, the only requirement is to disable parallel execution and no specific configuration is necessary. Secondly, we evaluated the software with the quantum computer module enabled, which formulates the SNPs seeding problem as an optimization problem that can be then tackled using various algorithms. To achieve this, the community detection software should be configured to generate subsets of the SNPs seeding problem that are compatible with the selected optimization algorithm. The subsets should contain at least two SNPs and the maximum size can be arbitrarily selected. The optimization problem can be solved using several available algorithms, such as classical simulated annealing, which serves as the baseline classical technique, quantum annealing through D-Wave devices, and QAOA algorithm via a noiseless simulator and IBM superconducting-based quantum processing unit. The parameters utilized for solving each dataset are summarized in (Suppl. Table 12). Additionally, there are specific configuration parameters for each algorithm.

Configuration of the thermal annealing

The thermal annealing requires a random seed and specifies the number of output samples (i.e. how many times the annealing process is stopped and repeated), which is always set to 1000, as well as the number of sweeps (i.e. how many times the state of a solution is perturbed during the annealing process), which is always set to 10^5 . The annealing cooling schedule is not configurable and follows a geometric schedule, which is determined by the formula $T(i) = T(0) \cdot \alpha^i$. The initial values for this schedule are automatically selected based on the standard procedure implemented in the Ocean SDK software.

Configuration of the parallel tempering

We have implemented the algorithm following [44, Section 2]. We initialize $M = 8$ chains in parallel, each having a different initial temperature β following a geometric schedule, i.e. evenly spaced in the log scale from a maximum value of 1 to a minimum initial value of 10^{-2} . In our case, $\{\beta_i\}_{i=1}^8 = \{1.000, 0.518, 0.268, 0.139, 0.072, 0.037, 0.019, 0.01\}$.

Configuration of Gurobi

Gurobi does not need any particular configuration to be set. The model used is `IsingModel` and the maximum number of iterations has been set to 10^6 .

Configuration of the quantum annealing

The quantum annealing process involves several specific settings. Firstly, the number of output samples is set to 1000. The solver utilized is Advantage 6.1, which is the latest hardware available from D-Wave at the time of writing. The annealing operation is performed twice - the first time as a forward annealing and the second time as a reverse annealing. During the forward annealing, the process starts at time 0 with the Hamiltonian H_I , then, after $1\mu s$, it implements the Hamiltonian $0.5H_I + 0.5H_P$, pauses for $1\mu s$, and completes the annealing process after an additional $1\mu s$. During the reverse annealing, the process begins at time 0 with the state corresponding to the solution found at the end of the forward annealing, using the Hamiltonian H_P . It then goes back to $0.5H_I + 0.5H_P$ after $1\mu s$, pauses for $1\mu s$, and converges to H_P again after another $1\mu s$. *quepistasis* allows to fully customize the quantum annealing schedule. It is worth noting that a longer schedule has the potential, in theory, to lead to better solutions by taking advantage of the adiabatic theorem. However, longer evolution periods also accumulate more noise on noisy devices, which can ultimately degrade the quality of the solutions. Therefore, it is generally more practical to increase the number of reads using shorter schedules rather than relying on a few reads with longer schedules. To utilize a D-Wave device, one must either purchase the hardware or rent it and access it through the cloud. We opted for the cloud-based access provided through our partnership with CINECA.

Configuration of the QAOA algorithm

The QAOA algorithm involves several configurable settings. Specifically, the quantum circuit of QAOA is dependent on the number of repetitions of the scheme $e^{-i\beta H_B} e^{-i\gamma H_C}$. Its training process requires a specific selection of optimizers, which in our case is the stochastic gradient descent algorithm ADAM, and a maximum number of epochs for the training. Furthermore, the statistical error in the output of the quantum circuit relies on the number of shots. The QAOA can be simulated on a classical device for instances of quantum circuits having a small number of qubits. The simulation employed is a linear algebra-based procedure that is noiseless (except for the statistical error in the sampling of solutions). For our simulated QAOA experiments, we executed the QAOA ansatz ten times, with random initialization of each parameter drawn from a uniform distribution ranging from $[-\pi, \pi)$. The circuit was trained for a maximum of ten epochs, although this number does not ensure convergence. Nonetheless, it is ten times larger than what can be accomplished in a realistic time frame using actual hardware, given that there are hundreds of independent subproblems to solve with QAOA for our datasets. We utilized 1024 shots. In order to use the IBM Perth and Lagos devices, one must subscribe to the IBM cloud-based service. Our partnership with CERN has provided us with access to this service. However, access to the machine is shared with a community of users, and requests are managed on a first-in, first-out (FIFO) basis, which can often lead to significant delays in obtaining results. In such case, we have repeated the QAOA ansatz once, and trained for 10, 5, and 1 number of epochs for datasets of 100, 500, and 1000 SNPs, respectively. The number of shots is 128. To utilize the IonQ Harmony device, it is required to have a subscription with the IonQ cloud-based provider. We have been granted access to this service through the Microsoft Azure Quantum platform. However, access to the IonQ device is also severely limited by a queue of users sharing the device. The QAOA algorithm software is based on the Qiskit library for Python, and it is made feasible to execute on both the simulator and quantum processing units through the use of various plug-ins.

Initialization strategies for QAOA

An important aspect of QAOA is the initialization strategy used to assign the starting values of the parameters. The most popular approach (which we also used for this study) is random initialization, i.e., drawing the parameters independently from the uniform distribution $[-\pi, \pi)$. Another popular technique is identity initialization, i.e., setting the parameter values to zero. This approach is used specifically to mitigate the barren plateau problem⁴⁵. Other popular choices include using a cheap and efficient procedure to find a

heuristically good initial solution, which can then be fed to QAOA as an initial point. The cheap procedure can be classical, for example, when using warm start⁴⁶, or quantum, as in the approach by Sack *et al.*⁴⁷. Specifically, warm start has been regarded as a promising approach due to the favourable guarantees when solving the maximum cut problem. Investigating how such alternative initialization strategies could be used in NeEDL may further improve NeEDL’s scalability.

Configuration of the QUBO formulation

There are some important considerations to be made about the parameters listed in (Suppl. Table 12). The value of parameter K is consistently set to 5, indicating that the optimizer is directed to return SNP sets of approximately the same size. The strength of this constraint varies based on the relationship between the values of λ_0 and λ_1, λ_2 . This implies that for subproblems of size 5 (used in QAOA experiments), the solution for the subproblem is expected to be the SNP set that corresponds to the entire subproblem. In such a case, the effectiveness of the approach depends almost entirely on the community detection tool. To minimize the size of each subproblem, the number of N of SNP sets returned is always set to 1. There is a trade-off between the use of the community detection tool and the value of N . A larger value of N increases the optimization problem size, necessitating a more aggressive community detection policy to compensate. Given more resources (qubits), this problem may be less significant. The interpretation of the ν parameters is related to the method of constructing weights for pairwise SNP correlation. Specifically, each weight between SNP i and j is determined using the formula $w_{i,j} = \nu w_{i,j}^{(1)} + (1 - \nu)w_{i,j}^{(2)}$, where $w_{i,j}^{(1)}$ denotes the statistical correlation between the SNPs, and $w_{i,j}^{(2)}$ represents biological information.

Normalization of QAOA seeding time on IBM devices

To perform our experiments using IBM devices, we had to choose between renting the machine for a fixed amount of time or submitting circuits to a job queue for processing. However, the time spent in the queue is not an accurate reflection of the actual time required to complete the job, since it depends on the usage of the machine at that particular moment. To account for this, we calculated the average execution time of a quantum program (which includes one quantum circuit estimated with one or several sets of parameters). We then multiplied this average time by the number of subproblems in the dataset, along with the initial startup time, to obtain a more accurate estimate of the total time required. To calculate the average execution time, we excluded the top 25% highest and lowest scores from any job execution.

Comment on the performance analysis

An analysis of our experiments is shown in Suppl. Figure 5. As anticipated, the seeding procedure of Linear NeEDL is generally faster than the optimization-based approach since the generation of random seeds is extremely rapid. However, the two techniques are comparable in terms of their order of magnitude. The little advantage achieved by Linear NeEDL is quickly lost during the searching phase, which requires checking a vast number of random seeds and takes up to three orders of magnitude longer than the optimization procedure of *quepistasis* to complete. This is because the latter produces fewer seeds, which are obviously quicker to verify and produce solutions with either the same (BD, RA datasets) or even superior (T2D dataset) statistical scores.

Although different techniques often result in seeds having varying statistical scores (except for RA with 100 SNPs, see Suppl. Figure 5, near values $x = 0$), the local search technique consistently (heuristically) recovers the best outcomes (see Suppl. Figure 5, near values $x = 10^6$). The classical baseline thermal annealing algorithm demonstrates comparable performances with datasets that have a smaller number of SNPs. However, as the number of SNPs increases, this advantage quickly dissipates, and thermal annealing becomes the slowest technique. Between the three classical baselines (thermal annealing, parallel tempering, and Gurobi), thermal annealing exhibits the best results in terms of producing high-quality seeds (lower NLL) and reducing the time needed for the local search. Parallel tempering consistently performs poorly. Gurobi’s performance fluctuates, sometimes almost matching parallel tempering (c.f. Suppl. Figure 5 datasets BD500, BD1000, T2D500) and at other times aligning more closely with thermal annealing. The simulated QAOA algorithm often produces the highest-quality initial seeds in terms of maximum likelihood, mainly due to

the community detection tool. QAOA subproblem instances consist of 5 qubits (5 SNPs), and according to our formulation and configuration, we expect QAOA to return the entire set of 5 SNPs for each subproblem instance. However, the classical simulation of QAOA is computationally intensive, resulting in seeding times comparable to the classical baseline (for smaller sets, simulated QAOA is slower, while for larger ones, it is faster). The performance of QAOA executed on IBM devices closely mirrors that of simulated QAOA, even though the noise of these devices leads to suboptimal solutions, which negatively affects the local search execution time. The seeding time of the QAOA on the IBM device had to be normalized to avoid misleading results due to the long wait in the queue for access to the shared device on the cloud platform. The details of this process are shown in the next paragraph. Finally, the D-Wave device appears to struggle to find high-quality seeds in terms of statistical score, even though the local search can still recover the best results. This technique is often the fastest in terms of both seeding time and local search time: starting with the D-Wave seeds, the local search converges quickly to the best solution.

Comment on the scaling analysis

In Figure 5, we propose a scaling model for our approach to studying the seeding time of the optimization-based seeding procedure with respect to the dataset size. We conducted experiments on classical and quantum devices using subsamples of 100, 500, and 1000 SNPs for each of the three diseases tested. This enabled us to show the linear growth of seeding time, as evidenced by lines with higher slopes indicating better scaling. Two of the classical baseline algorithms, thermal annealing and parallel tempering, performed the worst. In particular, parallel tempering exhibited much longer runtimes even for small datasets, and although its scaling is slightly more favorable than thermal annealing, this advantage is offset by the poorer quality of the seeds retrieved. The third classical baseline, Gurobi, also exhibited unfavorable scaling. Despite being the fastest approach for small sizes, this positive note is again mitigated by the poor quality of its seeds. The results are followed closely by the QAOA approach on both simulated and IBM devices. The QUBO formulation constrains the algorithm to return cliques of at least size 5, and since this is also the maximum number of SNPs for a subproblem in the configuration we tested using QAOA, the algorithm often returns the entire set of SNPs for the subproblem, resulting in similar outcomes (although QAOA (IBM) results are noisy). The poor scaling of QAOA is due to the community detection tool having to split the problem into many small sub-instances to make it feasible on current IBM quantum devices. D-Wave devices exhibited the best scaling: due to the larger number of qubits, the community detection tool was able to create instances as large as thermal annealing and much larger than QAOA. With fewer subproblems, the technique took much less time. Lastly, it is important to notice that while linear scaling is appropriate for datasets of modest size, there is no guarantee that this trend will continue for significantly larger datasets.

Range and sparsity of QUBO coefficients

The distribution of coefficients is a crucial factor in solving a QUBO or Ising formulation, particularly with respect to sparsity. Sparse QUBO matrices require less overhead on the D-Wave device. Additionally, coefficients should be uniformly distributed since they are rescaled to a specific interval (e.g., $[-1, 1]$). A few coefficients with a significantly larger magnitude than most others can result in a concentration of values around zero, which could cause problems with encoding precision. To illustrate this point, we have presented the coefficient distribution in Suppl. Figure 6. As shown in the figure, it can be observed that the average standard deviation of h and J coefficients is, once rescaled, always above the encoding precision of the devices (less than 10^{-3} for gate-based quantum devices, less than 4×10^{-3} for D-Wave Advantage device) and therefore we can assume the problem has been encoded with satisfactory precision on these devices.

Supplementary Materials 7: Biological plausibility studies for Diabetes Type 1 and Inflammatory Bowel Disease

For T1D, a gene associated with 19 high-prevalence SNP combinations was SRPK2. SRPK2 is involved in the regulation of mTORC1-driven metabolic disorders, which includes T1D^{48,49}. NeEDL identified 23 additional genes exhibiting potential epistatic SNP combinations with SRPK2 (Suppl. File 2). GSEA determined that 9 of these genes (NCK1, ANK3, DAW1, BTRC, CHD1L, NRG1, ERBB3, HLA-DPB1, RHBDD1) in addition to SRPK2 were significantly enriched in protein binding activity (GO:0030674; $k/K = 0.0144$; $P\text{-value} = 9.99e^{-07}$;

FDR q -value= $7.24e^{-03}$) or boosting gene expression (GO:0010628; k/K= 0.006; P -value= $1.83e^{-06}$; FDR q -value= $7.24e^{-03}$). One of these genes is NCK1, which is known to impact insulin biogenesis in pancreatic beta cells as well as their survival^{50,51}. Another one is NRG1, which is reported to regulate glucose tolerance⁵². We also identified ERBB3, which has been well recognized for its association with T1D risk^{53–56}.

For IBD, the most prevalent gene present in the top 42 candidate SNP sets is SOX5. A transcription factor analysis of the IBD-associated mucosal transcriptome showed SOX5 binding motifs were the most prevalent inflammation-associated sites for Crohn’s Disease (CD) and Ulcerative Colitis (UC) patients⁵⁷. NeEDL identified 42 additional genes exhibiting potential epistatic SNP combinations with SOX5 (Suppl. File 2). GSEA determined that 10 of these genes (TAF6, SMAD1, TAF3, KAT5, ESR1, MED27, SSBP3, LMO2, PASD1, DACH1) were significantly enriched in a gene set that is involved in building a protein complex that controls DNA transcription (GO:0005667; k/K=0.02; P -value= $3.13e^{-11}$; FDR q -value= $3.31e^{-07}$) with all but TAF3 significantly enriched along with SOX5, SOX6, and CREB5 in a gene set in the molecular function that regulates gene transcription for proper expression in cells and organisms (GO:0140110; k/K=0.0062; P -value= $1.34e^{-07}$; FDR q -value= $2.36e^{-04}$).

Supplementary Materials 8: Epistasis Disease Atlas

To make NeEDL available to the community and allow user-driven research of epistasis, we present the Epistasis Disease Atlas (<https://epistasis-disease-atlas.com/home>). It is a whole platform enabling users to browse precomputed findings within 8 human heritable diseases (see Section Datasets). We further offer real-time on-demand recomputation of epistasis scores for user-defined variant sets. The Epistasis Disease Atlas is an integrative resource providing filters, visualizations, and additional, relevant information about variants, genes, and pathways linking to widely used resources.

Database

For storage, a relational database (MySQL) and a graph database (Neo4j) are used in combination to get the best out of both systems. Any large n:m mapping between different entities, like “which genes are affected by a variant” or “which genes products interact with each other”, can be represented in the Neo4j instance and efficiently queried. The SQL database stores any other annotated information like common names for diseases, other protein IDs, or pathway descriptions. A complete UML representation depicting the database structures can be found in Suppl. Figure 13 and Suppl. Figure 14, respectively. The sources of the stored information can be separated into two categories. Generated or computed information is everything that is connected to executing NeEDL, this includes the evaluated datasets, runs, candidate sets, and scores. The second type is integrated data from 5 other sources with the goal of providing additional layers of information to the Epistasis Disease Atlas user. We further integrate the Human Protein Atlas (HPA)⁵⁸ to provide gene expression information of relevant tissues, WikiPathways⁵⁹ to show associated pathways, and Gene Ontology (GO)⁶⁰ to show GO classification for selected genes. The MV-Physical dataset of BioGrid⁶¹ provides a protein interaction network, translated to a gene interaction network for network visualization. Variant rsID and location information from dbSNP of UCSC⁶²(release 155) is imported using BigBed tools⁶³ and extended by allele and impact information using ClinVar data⁶⁴ (release 2023-03-05) and cytoband annotation from UCSC⁶² (cytoband.txt.gz). For the majority of the data, the import process is automatized; solely the list of tissues affected by the disease investigated in a dataset was manually curated by conducting a literature search.

Application programming interface (API)

The API server of the Epistasis Disease Atlas platform manages access to the databases, handles the asynchronous on-demand execution of the NeEDL algorithm, and makes results on the website reloadable. First-level applications like the website but also the R and Python packages utilize the API to provide functionalities, but documentation for the Epistasis Disease Atlas API is provided to be used by anyone in need of more control and functions (<https://epistasis-disease-atlas.com/api/schema/redoc/>, <https://epistasis-disease-atlas.com/api/schema/swagger/>).

All server applications are deployed using docker; the API is written in Django (v4.1.9) for Python 3.9, providing a RESTful API using djangorestframework (v3.13.1). An OpenAPI 3-based library called

drf-spectacular (v0.23.1) is used to render swagger and redoc documentation. Databases are MySQL (v5.7), and Neo4j (v4.4-enterprise), and queries are executed using mysqlclient (v2.1.0) and neomodel (v3.0.8) modules. Redis is used as the broker for the RabbitMQ queue handling tasks for Celery (v5.2.2) workers allowing execution of tasks in the background.

File download server, R and Python packages

We offer the full results calculated by NeEDL of 8 human heritable diseases (see Section Datasets) as a zip download via a static file server (<https://github.com/halverneus/static-file-server>) under <https://files.epistasis-disease-atlas.com/>. For convenience purposes for the user, we offer an R and Python package to directly load the results of NeEDL into the programming environment. The R and Python package can directly interact with the API of the Epistasis Disease Atlas and thus offers most data that is used for visualization in the web application.

Feature-rich web application

The main entry point for users, interested in data from the Epistasis Disease Atlas, is the web frontend. The website focuses on usability for researchers exploring epistasis research questions. Any precomputed results from available datasets are published there (see Section Datasets). There are three different ways to start the exploration of epistasis disease data: A query can be started by searching either for key genes or variants that should be investigated because of their hypothesized epistatic effects, leading to a list of candidate sets and thus runs these entities participated in, from which on entry can be selected. A third alternative is to start browsing the precomputed data directly. The browser page (<https://epistasis-disease-atlas.com/browse>) requires the selection of the dataset before listing all pre-computed sets of variants with the respective result scores. In all cases, upon selection, the result page for the given run is loaded, and results can be further explored, or on-demand scores can be requested.

In the result explorer page (Suppl. Figure 16), selected genes are integrated with the interactome (inferred from protein-protein interactions). Upon selecting a run, the respective variant (SNP-SNP) interaction network is constructed, as well as the energy landscape determined in the said run. Results are not just immutable, but functions are provided to add additional variants from the interaction neighborhood to investigate new epistasis hypotheses and expand on precalculated results. Candidate SNPs can be filtered by network properties in the variant interaction network and other scores to limit and only investigate relevant options. Recalculation of updated variant sets can be done live without great delay. Researched variant sets are internally associated with a unique identifier encoded in the explorer page URL. This allows results to be shared with fellow researchers or saved to a document to regenerate the exact same result page.

The platform cares about integrating into a researcher's common way of exploration by giving additional information about genes, variants, associated pathways, and go classification, as well as link-outs to popular follow-up analysis tools like g:Profiler or Drugst.One⁶⁵. For custom follow-up analysis, downloads (results in CSV format and networks as an SQLite file) for different parts of the analyzed variant set are provided.

The platform is designed to be easily updateable once more datasets become available and are analyzed by the NeEDL Algorithm.

The website was built with npm (v9.1.2), is written in AngularJS (v13.1.0) with typescript (v4.5.2) for node (v12.20.55), and styled with bootstrap (v5.1.3). For data visualization, vis.js (vis-data: v7.1.4, vis-network v9.1.2) is used for networks, datatables.net (v1.13.1) for tables, and highcharts (v10.1.0) for diagrams.

The Integrated Genome Browser plugin

To visualize SNPs in potentially important regions on the genome, we integrated an enhanced genome browser plug-in which includes the dependencies: npm 0.2.0, angular v13.1.0, typescript v4.5.2, Design: angular/material v13.3.9, Plots: @swimlane/ngx-charts v20.4.1, igv v2.15.8⁶⁶⁻⁶⁹, drugs.one v1.1.17⁶⁵.

This plugin includes the integration of the Human Protein Atlas⁷⁰, GeneOntology⁷¹, GeneCards⁷², WikiPathways⁵⁹, g:Profiler⁷³, Drugst.one⁶⁵, the Integrated Genome Browser⁶⁶⁻⁶⁹, ChIP-Atlas^{74, 75}, ReMap2022⁷⁶, ENA⁷⁷, ENCODE⁷⁸, GEO⁷⁹, JASPAR2022⁸⁰, EPDnew^{81, 82}, and Enhancer Atlas 2.0⁸³ for further seamless analysis.

Automatically generated R Shiny App that incorporates most functionalities of the Epistasis Disease Atlas for datasets that are processed on local machines or clusters

The Epistasis Disease Atlas can only visualize results and offer its functionalities for datasets that were preprocessed and stored in its database. However, GWAS data is mostly protected by strict licenses and cannot be shared, hence, we offer an R Shiny app that is automatically built after NeEDL successfully finished its analyses for data that is locally processed at the users' machine or cluster. The R Shiny app offers most functions of the Epistasis Disease Atlas for the convenience of the user. The R Shiny App is built on the following R Shiny packages (shiny v1.7.4, shinyWidgets v0.7.6, shinydashboard v0.7.2, shinyjs v2.1.0, and shinycssloaders v1.0.0).

References

- [1] Graham A Heap, Gosia Trynka, Ritsert C Jansen, Marcel Bruinenberg, Morris A Swertz, Lotte C Dinesen, Karen A Hunt, Cisca Wijmenga, David A Vanheer, and Lude Franke. Complex nature of SNP genotype effects on gene expression in primary human leucocytes. *BMC Med. Genomics*, 2:1, January 2009.
- [2] William S Bush and Jason H Moore. Chapter 11: Genome-wide association studies. *PLoS Comput. Biol.*, 8(12):e1002822, December 2012.
- [3] Jacqueline MacArthur, Emily Bowler, Maria Cerezo, Laurent Gil, Peggy Hall, Emma Hastings, Heather Junkins, Aoife McMahon, Annalisa Milano, Joannella Morales, Zoe May Pendlington, Danielle Welter, Tony Burdett, Lucia Hindorff, Paul Flicek, Fiona Cunningham, and Helen Parkinson. The new NHGRI-EBI catalog of published genome-wide association studies (GWAS catalog). *Nucleic Acids Res.*, 45(D1):D896–D901, January 2017.
- [4] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A R Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I W de Bakker, Mark J Daly, and Pak C Sham. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, 81(3):559–575, September 2007.
- [5] Andries T Marees, Hilde de Kluiver, Sven Stringer, Florence Vorspan, Emmanuel Curis, Cynthia Marie-Claire, and Eske M Derks. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *Int. J. Methods Psychiatr. Res.*, 27(2):e1608, June 2018.
- [6] H Braak and E Braak. Neuropathological staging of alzheimer-related changes. *Acta Neuropathol.*, 82(4):239–259, 1991.
- [7] Heiko Braak, Dietmar R Thal, Estifanos Ghebremedhin, and Kelly Del Tredici. Stages of the pathologic process in alzheimer disease: age categories from 1 to 100 years. *J. Neuropathol. Exp. Neurol.*, 70(11):960–969, November 2011.
- [8] Dietmar R Thal, Udo Rüb, Mario Orantes, and Heiko Braak. Phases of a beta-deposition in the human brain and its relevance for the development of AD. *Neurology*, 58(12):1791–1800, June 2002.
- [9] Peter J Thul and Cecilia Lindskog. The human protein atlas: A spatial map of the human proteome. *Protein Sci.*, 27(1):233–244, January 2018.
- [10] Pavel Filip, Antonietta Canna, Amir Moheet, Petr Bednarik, Heidi Grohn, Xiufeng Li, Anjali F Kumar, Evan Olawsky, Lynn E Eberly, Elizabeth R Seaquist, and Silvia Mangia. Structural alterations in deep brain structures in type 1 diabetes. *Diabetes*, 69(11):2458–2466, November 2020.
- [11] Maura Knapp, Xin Tu, and Rongxue Wu. Vascular endothelial dysfunction, a major mediator in diabetic cardiomyopathy. *Acta Pharmacol. Sin.*, 40(1):1–8, January 2019.
- [12] Dara P Schuster and Vani Duvuuri. Diabetes mellitus. *Clin. Podiatr. Med. Surg.*, 19(1):79–107, January 2002.

- [13] Décio L Eizirik, Lorenzo Pasquali, and Miriam Cnop. Pancreatic β -cells in type 1 and type 2 diabetes mellitus: different pathways to failure. *Nat. Rev. Endocrinol.*, 16(7):349–362, July 2020.
- [14] Kathleen M Gillespie. Type 1 diabetes: pathogenesis and prevention. *CMAJ*, 175(2):165–170, July 2006.
- [15] Louise Granlund, Anders Hedin, Olle Korsgren, Oskar Skog, and Marcus Lundberg. Altered microvasculature in pancreatic islets from subjects with type 1 diabetes. *PLoS One*, 17(10):e0276942, October 2022.
- [16] Adrian Stefański, Jacek Wolf, Joanna M Harazny, Eliza Miszkowska-Nagórna, Bogumił Wolnik, Joanna Murawska, Krzysztof Narkiewicz, and Roland E Schmieder. Impact of type 1 diabetes and its duration on wall-to-lumen ratio and blood flow in retinal arterioles. *Microvasc. Res.*, 147:104499, May 2023.
- [17] Elena Kiseleva, Maxim Ryabkov, Mikhail Baleev, Evgeniya Bederina, Pavel Shilyagin, Alexander Moiseev, Vladimir Beschastnov, Ivan Romanov, Grigory Gelikonov, and Natalia Gladkova. Prospects of intraoperative multimodal OCT application in patients with acute mesenteric ischemia. *Diagnostics (Basel)*, 11(4), April 2021.
- [18] Monjur Ahmed. Ischemic bowel disease in 2021. *World J. Gastroenterol.*, 27(29):4746–4762, August 2021.
- [19] Bryan T Green and David A Tendler. Ischemic colitis: a clinical review. *South. Med. J.*, 98(2):217–222, February 2005.
- [20] Morten Kjaergaard, Mollie E Schwartz, Jochen Braumüller, Philip Krantz, Joel I-J Wang, Simon Gustavsson, and William D Oliver. Superconducting qubits: Current state of play. *Annual Review of Condensed Matter Physics*, 11:369–395, 2020.
- [21] R. Harris, J. Johansson, A. J. Berkley, M. W. Johnson, T. Lanting, Siyuan Han, P. Bunyk, E. Ladizinsky, T. Oh, I. Perminov, E. Tolkacheva, S. Uchaikin, E. M. Chapple, C. Enderud, C. Rich, M. Thom, J. Wang, B. Wilson, and G. Rose. Experimental demonstration of a robust and scalable flux qubit. *Phys. Rev. B*, 81:134510, Apr 2010.
- [22] László Babai. Graph isomorphism in quasipolynomial time. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 684–697, 2016.
- [23] Stefanie Zbinden, Andreas Bärttschi, Hristo Djidjev, and Stephan Eidenbenz. Embedding algorithms for quantum annealers with chimera and pegasus connection topologies. In *High Performance Computing: 35th International Conference, ISC High Performance 2020, Frankfurt/Main, Germany, June 22–25, 2020, Proceedings*, pages 187–206. Springer, 2020.
- [24] D-Wave. Qpu-specific physical properties: Advantage system6.1, 2022. Available at https://docs.dwavesys.com/docs/latest/_downloads/1bfa16c9915114bdf8a37b14713c8953/09-1272A-A_QPU_Properties_Advantage_system6_1.pdf.
- [25] Neil G Dickson, MW Johnson, MH Amin, R Harris, F Altomare, AJ Berkley, P Bunyk, J Cai, EM Chapple, P Chavez, et al. Thermally assisted quantum annealing of a 16-qubit problem. *Nature communications*, 4(1):1903, 2013.
- [26] Kenneth Wright, Kristin M Beck, Sea Debnath, JM Amini, Y Nam, N Grzesiak, J-S Chen, NC Pimenti, M Chmielewski, C Collins, et al. Benchmarking an 11-qubit quantum computer. *Nature communications*, 10(1):5464, 2019.
- [27] IonQ. Best practices for using ionq hardware, 2022. Available at <https://ionq.com/best-practices> or <https://web.archive.org/web/20230209203834/https://ionq.com/best-practices>.
- [28] Microsoft. Ionq provider - microsoft azure quantum documentation, 2022. Available at <https://learn.microsoft.com/en-us/azure/quantum/provider-ionq> or <https://web.archive.org/web/20230210084454/https://learn.microsoft.com/en-us/azure/quantum/provider-ionq>.

- [29] IBM. Ibm quantum documentation, 2023. Available at <https://quantum-computing.ibm.com>.
- [30] Andrew Lucas. Ising formulations of many np problems. *Frontiers in physics*, 2:5, 2014.
- [31] Nike Dattani. Quadraticization in discrete optimization and quantum mechanics. *arXiv preprint arXiv:1901.04405*, 2019.
- [32] V A Traag, L Waltman, and N J van Eck. From louvain to leiden: guaranteeing well-connected communities. *Sci. Rep.*, 9(1):1–12, March 2019.
- [33] Scott Kirkpatrick, C Daniel Gelatt, and Mario P Vecchi. Optimization by simulated annealing. *science*, 220(4598):671–680, 1983.
- [34] Tadashi Kadowaki and Hidetoshi Nishimori. Quantum annealing in the transverse ising model. *Phys. Rev. E*, 58:5355–5363, Nov 1998.
- [35] Philipp Hauke, Helmut G Katzgraber, Wolfgang Lechner, Hidetoshi Nishimori, and William D Oliver. Perspectives of quantum annealing: methods and implementations. *Rep. Prog. Phys.*, 83(5):054401, may 2020.
- [36] Tosio Kato. On the adiabatic theorem of quantum mechanics. *Journal of the Physical Society of Japan*, 5(6):435–439, 1950.
- [37] Tameem Albash and Daniel A. Lidar. Adiabatic quantum computation. *Rev. Mod. Phys.*, 90:015002, Jan 2018.
- [38] Glen Bigan Mbeng, Rosario Fazio, and Giuseppe Santoro. Quantum annealing: a journey through digitalization, control, and hybrid quantum variational schemes, 2019.
- [39] Bruno Apolloni, C Carvalho, and Diego De Falco. Quantum stochastic optimization. *Stochastic Processes and their Applications*, 33(2):233–244, 1989.
- [40] Andrew Lucas. Ising formulations of many np problems. *Frontiers in Physics*, 2:5, 2014.
- [41] Naeimeh Mohseni, Peter L McMahon, and Tim Byrnes. Ising machines as hardware solvers of combinatorial optimization problems. *Nature Reviews Physics*, 4(6):363–379, 2022.
- [42] Edward Farhi, Jeffrey Goldstone, and Sam Gutmann. A quantum approximate optimization algorithm. *arXiv*, 2014.
- [43] Jérémie Roland and Nicolas J Cerf. Quantum search by local adiabatic evolution. *Physical Review A*, 65(4):042308, 2002.
- [44] David J Earl and Michael W Deem. Parallel tempering: Theory, applications, and new perspectives. *Physical Chemistry Chemical Physics*, 7(23):3910–3916, 2005.
- [45] Edward Grant, Leonard Wossnig, Mateusz Ostaszewski, and Marcello Benedetti. An initialization strategy for addressing barren plateaus in parametrized quantum circuits. *Quantum*, 3:214, 2019.
- [46] Reuben Tate, Jai Moondra, Bryan Gard, Greg Mohler, and Swati Gupta. Warm-started qaoa with custom mixers provably converges and computationally beats goemans-williamson’s max-cut at low circuit depths. *Quantum*, 7:1121, 2023.
- [47] Stefan H Sack and Maksym Serbyn. Quantum annealing initialization of the quantum approximate optimization algorithm. *quantum*, 5:491, 2021.
- [48] Gina Lee, Yuxiang Zheng, Sungyun Cho, Cholsoon Jang, Christina England, Jamie M Dempsey, Yonghao Yu, Xiaolei Liu, Long He, Paola M Cavaliere, Andre Chavez, Erik Zhang, Meltem Isik, Anthony Couvillon, Noah E Dephoure, T Keith Blackwell, Jane J Yu, Joshua D Rabinowitz, Lewis C Cantley, and John Blenis. Post-transcriptional regulation of de novo lipogenesis by mTORC1-S6K1-SRPK2 signaling, 2017.

- [49] Terunao Takahara and Tatsuya Maeda. Evolutionarily conserved regulation of TOR signalling. *J. Biochem.*, 154(1):1–10, July 2013.
- [50] George Kefalas, Nathalie Jouvet, Cindy Baldwin, Jennifer L Estall, and Louise Larose. Peptide-based sequestration of the adaptor protein nck1 in pancreatic β cells enhances insulin biogenesis and protects against diabetogenic stresses. *J. Biol. Chem.*, 293(32):12516–12524, August 2018.
- [51] Lama Yamani, Bing Li, and Louise Larose. Nck1 deficiency improves pancreatic β cell survival to diabetes-relevant stresses by modulating PERK activation and signaling. *Cell. Signal.*, 27(12):2555–2567, December 2015.
- [52] Gaël Ennequin, Nathalie Boisseau, Kevin Caillaud, Vivien Chavanelle, Monique Etienne, Xinyan Li, and Pascal Sirvent. Neuregulin 1 improves glucose tolerance in db/db mice. *PLoS One*, 10(7):e0130568, July 2015.
- [53] Jeffrey P Krischer, Xiang Liu, Åke Lernmark, William A Hagopian, Marian J Rewers, Jin-Xiong She, Jorma Toppari, Anette-G Ziegler, Beena Akolkar, and TEDDY Study Group. Predictors of the initiation of islet autoimmunity and progression to multiple autoantibodies and clinical diabetes: The TEDDY study. *Diabetes Care*, 45(10):2271–2281, October 2022.
- [54] Antti-Pekka Laine, Milla Valta, Jorma Toppari, Mikael Knip, Riitta Veijola, Jorma Ilonen, and Johanna Lempainen. Non-HLA gene polymorphisms in the pathogenesis of type 1 diabetes: Phase and endotype specific effects. *Front. Immunol.*, 13:909020, June 2022.
- [55] Jorma Ilonen, Antti-Pekka Laine, Minna Kiviniemi, Taina Härkönen, Johanna Lempainen, Mikael Knip, and Finnish Pediatric Diabetes Register. Associations between deduced first islet specific autoantibody with sex, age at diagnosis and genetic risk factors in young children with type 1 diabetes. *Pediatr. Diabetes*, 23(6):693–702, September 2022.
- [56] Julie Vandewalle, Bart J Van der Auwera, Henna Amin, Erik Quartier, Aster K Desouter, Sylvie Tenoutasse, Pieter Gillard, Christophe De Block, Bart Keymeulen, Frans K Gorus, Mark Van de Casteele, and Belgian Diabetes Registry. Genetic variation at ERBB3/IKZF4 and sexual dimorphism in epitope spreading in single autoantibody-positive relatives. *Diabetologia*, 64(11):2511–2516, November 2021.
- [57] Robert Häsler, Raheleh Sheibani-Tezerji, Anupam Sinha, Matthias Barann, Ateequr Rehman, Daniela Esser, Konrad Aden, Carolin Knecht, Berenice Brandt, Susanna Nikolaus, Sascha Schäuble, Christoph Kaleta, Andre Franke, Christoph Fretter, Werner Müller, Marc-Thorsten Hütt, Michael Krawczak, Stefan Schreiber, and Philip Rosenstiel. Uncoupling of mucosal gene regulation, mRNA splicing and adherent microbiota signatures in inflammatory bowel disease. *Gut*, 66(12):2087–2097, December 2017.
- [58] Mathias Uhlen, Per Oksvold, Linn Fagerberg, Emma Lundberg, Kalle Jonasson, Mattias Forsberg, Martin Zwahlen, Caroline Kampf, Kenneth Wester, Sophia Hober, Henrik Wernerus, Lisa Björling, and Fredrik Ponten. Towards a knowledge-based human protein atlas. *Nat. Biotechnol.*, 28(12):1248–1250, December 2010.
- [59] Marvin Martens, Ammar Ammar, Anders Riutta, Andra Waagmeester, Denise N Slenter, Kristina Hanspers, Ryan A. Miller, Daniela Digles, Elisson N Lopes, Friederike Ehrhart, Lauren J Dupuis, Laurent A Winckers, Susan L Coort, Egon L Willighagen, Chris T Evelo, Alexander R Pico, and Martina Kutmon. WikiPathways: connecting communities. *Nucleic Acids Res.*, 49(D1):D613–D621, November 2020.
- [60] M Ashburner, C A Ball, J A Blake, D Botstein, H Butler, J M Cherry, A P Davis, K Dolinski, S S Dwight, J T Eppig, M A Harris, D P Hill, L Issel-Tarver, A Kasarskis, S Lewis, J C Matese, J E Richardson, M Ringwald, G M Rubin, and G Sherlock. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat. Genet.*, 25(1):25–29, May 2000.
- [61] Chris Stark, Bobby-Joe Breitkreutz, Teresa Regul, Lorrie Boucher, Ashton Breitkreutz, and Mike Tyers. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.*, 34(Database issue):D535–9, January 2006.

- [62] Luis R Nassar, Galt P Barber, Anna Benet-Pagès, Jonathan Casper, Hiram Clawson, Mark Diekhans, Clay Fischer, Jairo Navarro Gonzalez, Angie S Hinrichs, Brian T Lee, Christopher M Lee, Pranav Muthuraman, Beagan Nguy, Tiana Pereira, Parisa Nejad, Gerardo Perez, Brian J Raney, Daniel Schmelter, Matthew L Speir, Brittney D Wick, Ann S Zweig, David Haussler, Robert M Kuhn, Maximilian Haeussler, and W James Kent. The UCSC genome browser database: 2023 update. *Nucleic Acids Res.*, 51(D1):D1188–D1195, November 2022.
- [63] W J Kent, A S Zweig, G Barber, A S Hinrichs, and D Karolchik. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics*, 26(17):2204–2207, September 2010.
- [64] Melissa J Landrum, Shanmuga Chitipiralla, Garth R Brown, Chao Chen, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Wonhee Jang, Kuljeet Kaur, Chunlei Liu, Vitaly Lyoshin, Zenith Maddipatla, Rama Maiti, Joseph Mitchell, Nuala O’Leary, George R Riley, Wenya Shi, George Zhou, Valerie Schneider, Donna Maglott, J Bradley Holmes, and Brandi L Kattman. ClinVar: improvements to accessing data. *Nucleic Acids Res.*, 48(D1):D835–D844, November 2019.
- [65] Andreas Maier, Michael Hartung, Mark Abovsky, Klaudia Adamowicz, Gary D Bader, Sylvie Baier, David B Blumenthal, Jing Chen, Maria L Elkjaer, Carlos Garcia-Hernandez, Markus Hoffmann, Igor Jurisica, Max Kotlyar, Olga Lazareva, Hagai Levi, Markus List, Sebastian Lobentanzer, Joseph Loscalzo, Noel Malod-Dognin, Quirin Manz, Julian Matschinske, Mhaned Oubounyt, Alexander R Pico, Rudolf T Pillich, Julian M Poschenrieder, Dexter Pratt, Nataša Pržulj, Sepideh Sadegh, Julio Saez-Rodriguez, Suryadipto Sakar, Gideon Shaked, Ron Shamir, Nico Trummer, Ugur Turhan, Ruisheng Wang, Olga Zolotareva, and Jan Baumbach. Drugst.One - a plug-and-play solution for online systems medicine and network-based drug repurposing. *Nucleic Acids Research*, Web Server Issue, in press.
- [66] James T Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S Lander, Gad Getz, and Jill P Mesirov. Integrative genomics viewer. *Nat. Biotechnol.*, 29(1):24–26, January 2011.
- [67] James T Robinson, Helga Thorvaldsdóttir, Aaron M Wenger, Ahmet Zehir, and Jill P Mesirov. Variant review with the integrative genomics viewer. *Cancer Res.*, 77(21):e31–e34, November 2017.
- [68] James T Robinson, Helga Thorvaldsdottir, Douglass Turner, and Jill P Mesirov. igv.js: an embeddable JavaScript implementation of the integrative genomics viewer (IGV). *Bioinformatics*, 39(1), January 2023.
- [69] Helga Thorvaldsdóttir, James T Robinson, and Jill P Mesirov. Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.*, 14(2):178–192, March 2013.
- [70] Mathias Uhlén, Linn Fagerberg, Björn M Hallström, Cecilia Lindskog, Per Oksvold, Adil Mardinoglu, Åsa Sivertsson, Caroline Kampf, Evelina Sjöstedt, Anna Asplund, Ingmarie Olsson, Karolina Edlund, Emma Lundberg, Sanjay Navani, Cristina Al-Khalili Szigarto, Jacob Odeberg, Dijana Djureinovic, Jenny Ottosson Takanen, Sophia Hober, Tove Alm, Per-Henrik Edqvist, Holger Berling, Hanna Tegel, Jan Mulder, Johan Rockberg, Peter Nilsson, Jochen M Schwenk, Marica Hamsten, Kalle von Feilitzen, Mattias Forsberg, Lukas Persson, Fredric Johansson, Martin Zwahlen, Gunnar von Heijne, Jens Nielsen, and Fredrik Pontén. Tissue-based map of the human proteome. *Science*, 347(6220):1260419, 2015.
- [71] David Binns, Emily Dimmer, Rachael Huntley, Daniel Barrell, Claire O’Donovan, and Rolf Apweiler. QuickGO: a web-based tool for gene ontology searching. *Bioinformatics*, 25(22):3045–3046, November 2009.
- [72] Gil Stelzer, Naomi Rosen, Inbar Plaschkes, Shahar Zimmerman, Michal Twik, Simon Fishilevich, Tsippi Iny Stein, Ron Nudel, Iris Lieder, Yaron Mazor, Sergey Kaplan, Dvir Dahary, David Warshawsky, Yaron Guan-Golan, Asher Kohn, Noa Rappaport, Marilyn Safran, and Doron Lancet. The GeneCards suite: From gene data mining to disease genome sequence analyses. *Curr. Protoc. Bioinformatics*, 54:1.30.1–1.30.33, June 2016.

- [73] Uku Raudvere, Liis Kolberg, Ivan Kuzmin, Tambet Arak, Priit Adler, Hedi Peterson, and Jaak Vilo. g:profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.*, 47(W1):W191–W198, July 2019.
- [74] Shinya Oki, Tazro Ohta, Go Shioi, Hideki Hatanaka, Osamu Ogasawara, Yoshihiro Okuda, Hideya Kawaji, Ryo Nakaki, Jun Sese, and Chikara Meno. ChIP-Atlas: a data-mining suite powered by full integration of public ChIP-seq data. *EMBO Rep.*, 19(12):e46255, December 2018.
- [75] Zhaonan Zou, Tazro Ohta, Fumihito Miura, and Shinya Oki. ChIP-Atlas 2021 update: a data-mining suite for exploring epigenomic landscapes by fully integrating ChIP-seq, ATAC-seq and bisulfite-seq data. *Nucleic Acids Res.*, March 2022.
- [76] Fayrouz Hammal, Pierre de Langen, Aurélie Bergon, Fabrice Lopez, and Benoit Ballester. ReMap 2022: a database of human, mouse, drosophila and arabidopsis regulatory regions from an integrative analysis of DNA-binding sequencing experiments. *Nucleic Acids Res.*, 50(D1):D316–D325, January 2022.
- [77] Rasko Leinonen, Ruth Akhtar, Ewan Birney, Lawrence Bower, Ana Cerdeno-Tárraga, Ying Cheng, Iain Cleland, Nadeem Faruque, Neil Goodgame, Richard Gibson, Gemma Hoad, Mikyung Jang, Nima Pakseresht, Sheila Plaister, Rajesh Radhakrishnan, Kethi Reddy, Siamak Sobhany, Petra Ten Hoopen, Robert Vaughan, Vadim Zalunin, and Guy Cochrane. The european nucleotide archive. *Nucleic Acids Res.*, 39(Database issue):D28–31, January 2011.
- [78] Anastasiya Kazachenka, Tessa M Bertozzi, Marcela K Sjöberg-Herrera, Nic Walker, Joseph Gardner, Richard Gunning, Elena Pahita, Sarah Adams, David Adams, and Anne C Ferguson-Smith. Identification, characterization, and heritability of murine metastable epialleles: Implications for non-genetic inheritance. *Cell*, 175(5):1259–1271.e13, November 2018.
- [79] Emily Clough and Tanya Barrett. The gene expression omnibus database. *Methods Mol. Biol.*, 1418:93–110, 2016.
- [80] Jaime A Castro-Mondragon, Rafael Riudavets-Puig, Ieva Rauluseviciute, Roza Berhanu Lemma, Laura Turchi, Romain Blanc-Mathieu, Jeremy Lucas, Paul Boddie, Aziz Khan, Nicolás Manosalva Pérez, Oriol Fornes, Tiffany Y Leung, Alejandro Aguirre, Fayrouz Hammal, Daniel Schmelter, Damir Baranasic, Benoit Ballester, Albin Sandelin, Boris Lenhard, Klaas Vandepoele, Wyeth W Wasserman, François Parcy, and Anthony Mathelier. JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, 50(D1):D165–D173, January 2022.
- [81] René Dreos, Giovanna Ambrosini, Rouayda Cavin Périer, and Philipp Bucher. The eukaryotic promoter database: expansion of EPDnew and new promoter analysis tools. *Nucleic Acids Res.*, 43(Database issue):D92–6, January 2015.
- [82] René Dreos, Giovanna Ambrosini, Romain Groux, Rouaïda Cavin Périer, and Philipp Bucher. The eukaryotic promoter database in its 30th year: focus on non-vertebrate organisms. *Nucleic Acids Res.*, 45(D1):D51–D55, January 2017.
- [83] Tianshun Gao and Jiang Qian. EnhancerAtlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species. *Nucleic Acids Res.*, 48(D1):D58–D64, January 2020.



This section contains the original supplementary material published alongside the article:

“Online *in silico* validation of disease and gene sets, clusterings or subnetworks with DIGEST”

Online in silico validation of disease and gene sets, clusterings, or subnetworks with DIGEST

Klaudia Adamowicz, Andreas Maier, Jan Baumbach, David B. Blumenthal

Supplement

1 Supplementary tables

Supplementary Table 1. Gene sets associated to Alzheimer's disease and hyperlipidemia used for exemplifying the reference-based gene set validation route.

Disease	Associated genes
Alzheimer's disease	A2M, ABCA7, ACE, ADAM10, APOE, APP, BAX, BCL2, BDNF, CD2AP, CLU, GSK3B, HFE, IGF1R, IGF2, IL1B, INS, INSR, LEP, MPO, NOS3, NPY, PICALM, PLA2, PSEN1, PSEN2, SORL1, TREM2, AD5, AD6, AD7, AD8, AD10, AD11, AD12, AD13, AD14, AD15, AD16, AD17
Hyperlipidemia	APOA2, APOA5, APOB, APOC2, APOE, CETP, EPHX2, GHR, GPIHBP1, LDLR, LDLRAP1, LIPC, LMF1, LPL, NOS3, PCSK9, PPP1R17, USF1, HYPLIP2

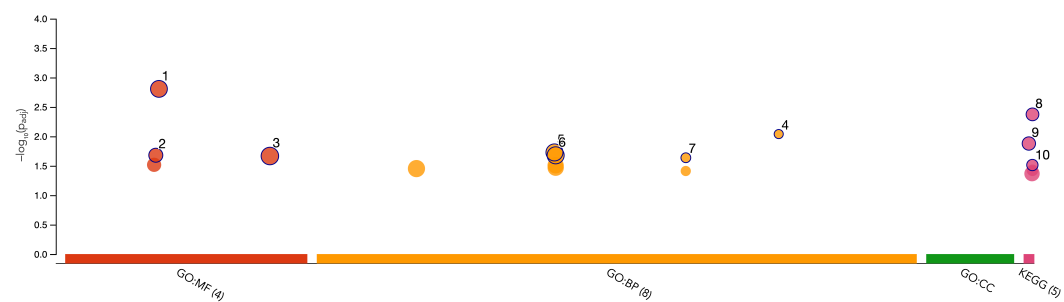
Supplementary Table 2. Gene and disease namespaces used by external databases queried by DIGEST.

Database	Gene namespace	Disease namespace
GO	UniProt	–
KEGG	NCBI-ProteinID, Entrez (NCBI-GeneID), UniProt	OMIM
DisGeNET	Entrez, Symbol	UMLS

2 Supplementary figures



Supplementary Figure 1. Results of g:Profiler gene set enrichment for the 10 genes with largest positive significance contributions on empirical P -values based on GO.MF annotations obtained for COVID-19 use case exemplifying the reference-free gene subnetwork validation route.

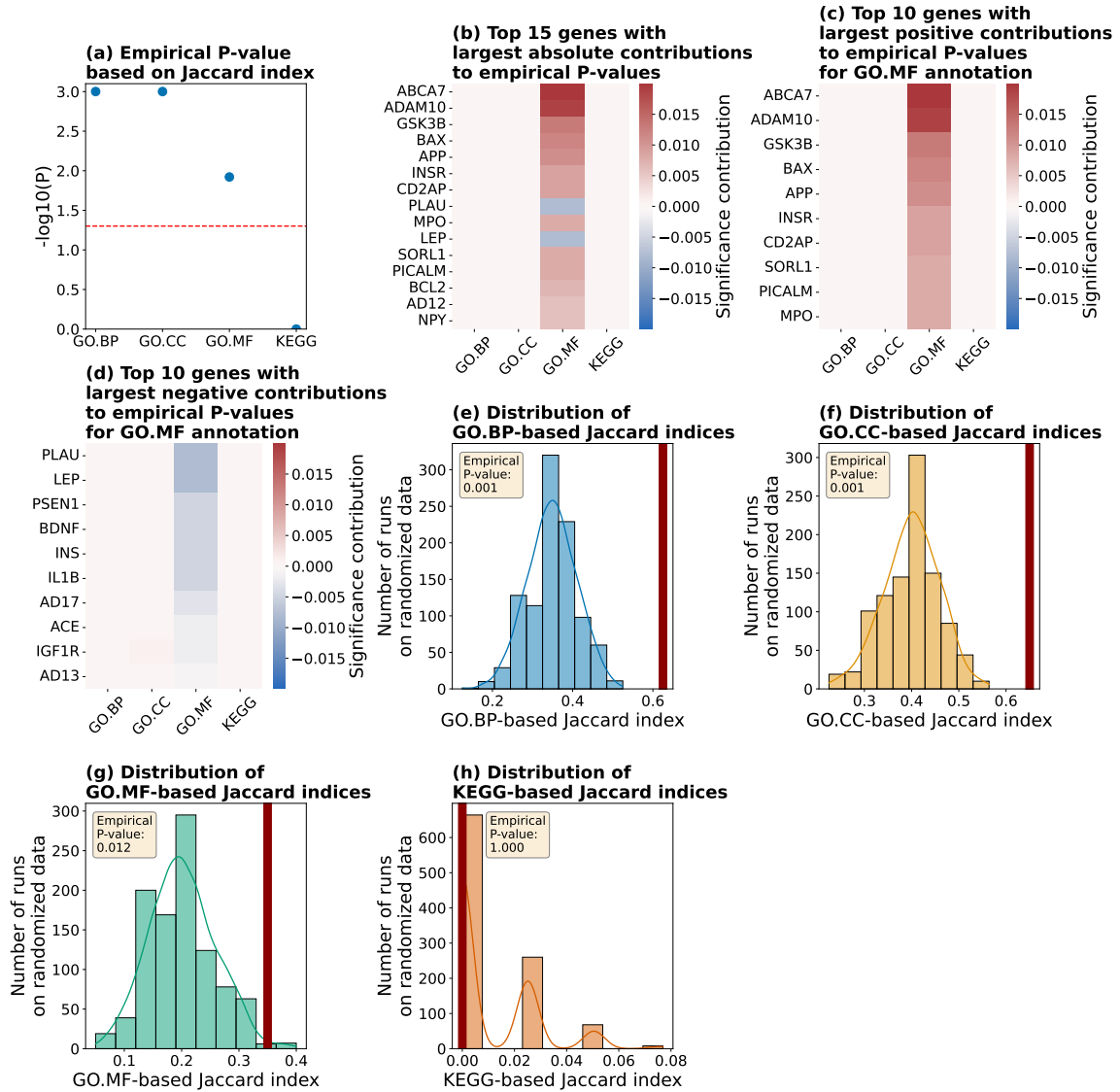


ID	Source	Term ID	Term Name	Padj (query_1)
1	GO:MF	GO:0030234	enzyme regulator activity	1.560×10 ⁻³
2	GO:MF	GO:0019887	protein kinase regulator activity	2.091×10 ⁻²
3	GO:MF	GO:0098772	molecular function regulator	2.157×10 ⁻²
4	GO:BP	GO:1900119	positive regulation of execution phase of apopt...	9.143×10 ⁻³
5	GO:BP	GO:0042981	regulation of apoptotic process	1.874×10 ⁻²
6	GO:BP	GO:0043067	regulation of programmed cell death	2.099×10 ⁻²
7	GO:BP	GO:0070234	positive regulation of T cell apoptotic process	2.308×10 ⁻²
8	KEGG	KEGG:05222	Small cell lung cancer	4.229×10 ⁻³
9	KEGG	KEGG:04210	Apoptosis	1.323×10 ⁻²
10	KEGG	KEGG:05216	Thyroid cancer	3.068×10 ⁻²

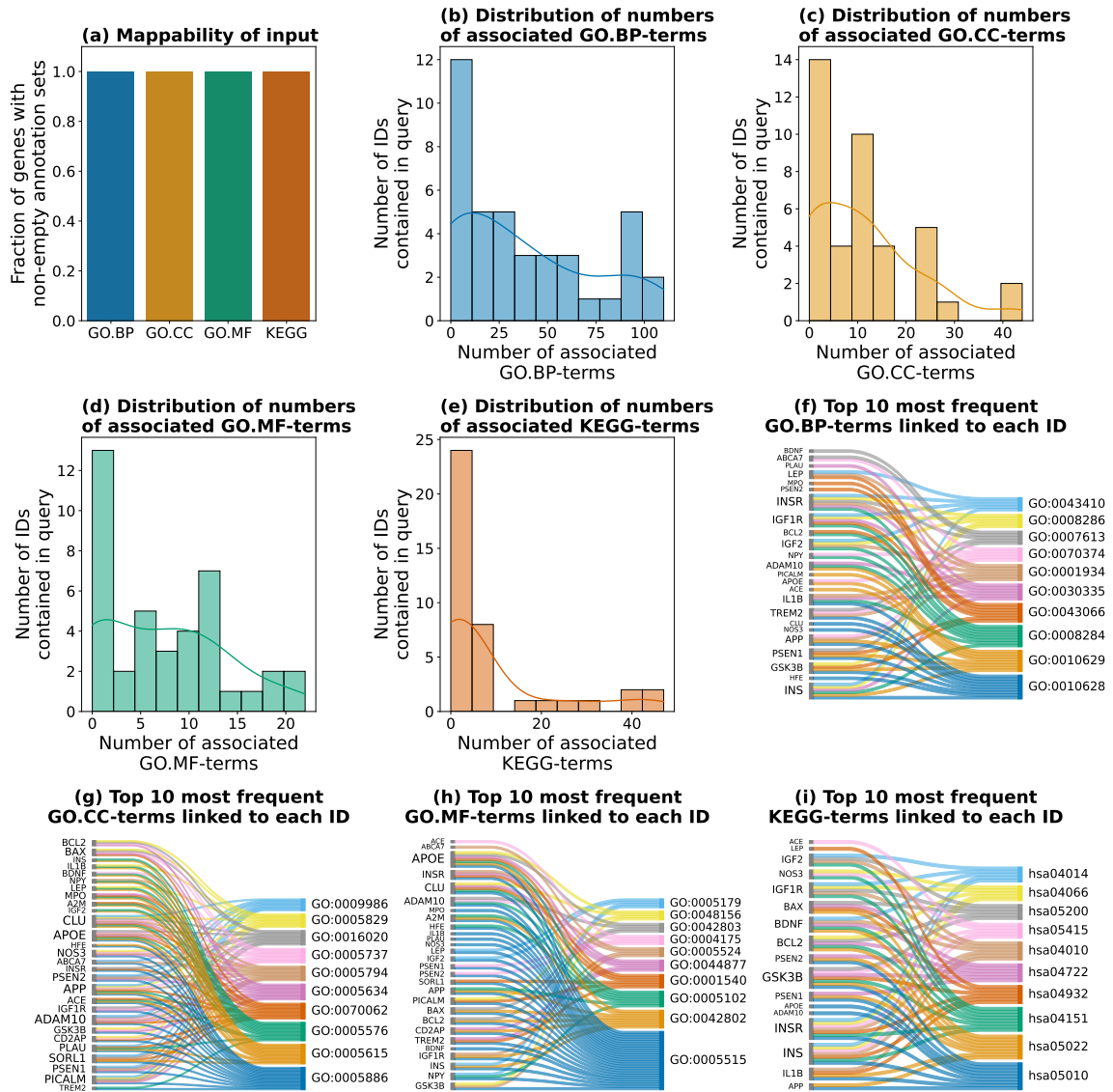
version e105_eg52_p16_e84549f
date 18/05/2022, 07:56:35
organism hsapiens

g:Profiler

Supplementary Figure 2. Results of g:Profiler gene set enrichment for the 10 genes with largest negative significance contributions on empirical *P*-values based on GO.MF annotations obtained for COVID-19 use case exemplifying the reference-free gene subnetwork validation route.



Supplementary Figure 3. Overview of all P -value plots generated for the results of the drug repurposing use case used to exemplify the reference-based gene set validation route. (a) Overview of empirical P -values. (b) Significance contributions of top 15 genes with largest absolute contributions to the empirical P -values. (c)–(d) Significance contributions of top 10 genes with, respectively, largest positive and largest negative significance contributions to the empirical P -values based on GO.MF annotations. (e)–(h) Background relevance score distribution plots with relevance scores for input displayed as red lines.



Supplementary Figure 4. Overview of all mappability plots generated for the results of the drug repurposing use case used to exemplify the reference-based gene set validation route. (a) Frequency of genes with non-empty annotation sets. (b)–(e) Distributions of gene annotation set sizes. (f)–(i) Top 10 most frequently occurring annotations linked to the associated genes in the input sets.

