Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Machine Learning Models for the Prediction of Frequent Hitters in Biochemical and Biological Assays

Cumulative Dissertation
with the aim of achieving a doctoral degree
Doctor rerum naturalium (*Dr. rer. nat.*)

at the Faculty of Mathematics, Informatics and Natural Sciences
Department of Chemistry
Universität Hamburg

submitted by
Conrad Stork
Master of Science, Universität Hamburg
born 02/28/1992
citizen of Germany

September 2021

Gutachter
Prof. Dr. Johannes Kirchmair
Prof. Dr. Andrew Torda
Tag der Disputation: 12.11.2021

I. List of publications originating from this work

[D1] Stork, C.; Kirchmair, J. PAIN(S) relievers for medicinal chemists: how computational methods can assist in hit evaluation, *Future Medicinal Chemistry* **2018**, *10*, 1533–1535.

[D2] Stork, C.; Wagner, J.; Friedrich, N.-O.; de Bruyn Kops, C.; Šícho, M.; Kirchmair, J. Hit Dexter: a machine-learning model for the prediction of frequent hitters, *ChemMedChem* **2018**, *13*, 564–571.

[D3] Stork, C.; Chen, Y.; Šícho, M.; Kirchmair, J. Hit Dexter 2.0: machine-learning models for the prediction of frequent hitters, *Journal of Chemical Information and Modeling* **2019**, *59*, 1030–1043.

[D4] Stork, C.; Embruch, G.; Šícho, M.; de Bruyn Kops, C.; Chen, Y.; Svozil, D.; Kirchmair, J. NERDD: A web portal providing access to in silico tools for drug discovery, *Bioinformatics* **2020**, *36*, 1291–1292.

[D5] Stork, C.; Mathai, N.; Kirchmair, J. Computational prediction of frequent hitters in target-based and cell-based assays, *Artificial Intelligence in the Life Sciences* **2021**, *1*, 100007.

II. Contents

I List of publications originating from this work **i**

II Contents **iii**

III List of abbreviations **v**

1 Zusammenfassung **1**

2 Abstract **5**

3 Introduction **9**
3.1 High-throughput assays for lead compound identification 9
3.2 Frequent hitters, nuisance compounds and dark chemical matter 10
3.2.1 Frequent hitters 12
3.2.2 Colloidal aggregators 14
3.2.3 Pan-assay interference compounds (PAINS) 15
3.2.4 Reactive compounds 16
3.2.5 Dark chemical matter 17
3.2.6 True promiscuous compounds (multi-target compounds) . 17
3.2.7 Assay technology-specific nuisance compounds 18
3.2.8 Applicability domain of existing computational models . . 18
3.2.9 Computational approaches for hit (de-)priorisation 19
3.3 Relevance of frequent hitters and nuisance compounds within the community and controversial discussion in the literature 24
3.4 Data sets for computational approaches 25
3.5 Machine learning approaches 26
3.5.1 Algorithms 27
3.5.2 Molecular descriptors 28
3.5.3 Performances metrics 29
3.6 Web servers as a tool for easy accessibility 30

4 Aims **31**

5 Results (cumulative part of this dissertation) **33**
5.1 Machine learning models for the prediction of frequent hitters based on target-based assay data sets 34

5.2 Refined machine learning models for the prediction of frequent hitters in primary screening assays and confirmatory dose-response assays 44
5.3 Machine learning models for the prediction of frequent hitters based on target-based and cell-based assay data sets 60
5.4 New e-resource for early drug discovery (NERDD) 75

6 Conclusions and future directions **79**

7 Bibliography **85**

8 Appendix **97**
A Gefahrstoffe nach GHS 97
B Supporting information for [D2] 98
C Supporting information for [D3] 107
D Supporting information for [D5] 110
E Scientific contribution 113

9 Danksagung **115**

10 Eidesstattliche Versicherung **117**

III. List of abbreviations

AD Applicability domain

ADME Absorption, distribution, metabolism and excretion

AI Artificial intelligence

ATR Active-to-tested ratio

AUC Area under the receiver operating characteristic curve

CDRA Confirmatory dose-response assay

CIAT Compound interfering with an assay technology

DCM Dark chemical matter

ET Extra tree

FN False negatives

FP False positives

FRET Fluorescence resonance energy transfer

HTS High-throughput screening

KNN K-nearest neighbors

MCC Matthews correlation coefficient

MLP Multilayer perceptron

MOE Molecular operating environment

NERDD New e-resource for drug discovery

NMR Nuclear magnetic resonance

PSA Primary screen assay

PAINS Pan-assay interference compounds

RF Random forest

SCAM Small, colloidally aggregating molecules

SMARTS Simplified molecular input line entry specification (SMILES) arbitrary target specification

SMILES Simplified molecular input line entry specification

TN True negatives

TP True positives

1. Zusammenfassung

"High-throughput screening" (HTS) ist eine Schlüsseltechnologie um große Mengen an Molekülen auf Aktivität gegenüber ausgewählten Biomakromolekülen und in zunehmendem Maße auch in Zellen zu testen. Weisen die Substanzen biologische Aktivität auf, können sie vielversprechende Kandidaten für die Entwicklung von Medikamenten, Agrochemikalien und Kosmetika sein. Eine große Herausforderung in HTS sind problematisch hohe Raten an falsch positiven Ergebnissen, die auf Assayinterferenz zurückgeführt werden können. Vor allem führen die falsch positiven Testergebnisse oft zu zwecklosen Folgeexperimenten, die signifikante Teile der Forschungskapazitäten blockieren.

Substanzen, die mit den Assays interferieren, werden "bad actors", "badly behaving compounds", "nuisance compounds" oder ungenau auch "pan-assay interference compounds" (PAINS) genannt. Dabei können die Gründe für die Assayinterferenz vielfältig sein: die Bildung von kolloidalen Aggregaten, chemische Reaktivität (z.B. Moleküle die kovalente Bindungen mit Proteine eingehen), Membranspaltungen, Substanzen die Metallkomplexe bilden, etc.

Viele, aber nicht alle, der "bad actors" fallen während biochemischer und biologischer Assays durch überdurchschnittliche Trefferquoten auf und werden daher auch "frequent hitters" genannt. Zum Beispiel weisen viele reaktive Substanzen in verschiedensten Assays falsch positive Testergebnisse auf, da sie unspezifisch viele verschiedene Proteine binden können. Allerdings sind nicht alle "frequent hitters" auch "bad actors", denn einige Substanzen können auch in spezifischer Art und Weise an verschiedene Proteine binden, da sie dafür geeignete Grundgerüste ("privileged Scaffolds") besitzen. Diese Substanzen sind im Bezug auf "polypharmacology" und "drug repurposing" von besonders großer Bedeutung.

Neueste Studien haben gezeigt, dass computergestützte Methoden das Potenzial haben "bad actors" zu identifizieren, sich jedoch noch in ihrem Anfangsstadium befinden. Ziel dieser Doktorarbeit ist es darum, auf maschinellem Lernen basierende Modelle zur Vorhersage von "frequent hitters" zu entwickeln.

Im ersten Teil dieser Doktorarbeit, wurde das Potenzial von Modellen für maschinelles Lernen zur Vorhersage von "frequent hitters", die auf einem öffentlich zugänglichen Bioaktivitätsdatensatz beruhen, geprüft. Dazu wurde ein

Datensatz mit ca. 311 000 Substanzen, für die Aktivitätsdaten von mindestens 50 verschiedener Proteine vorhanden sind, aus der PubChem Bioassay database extrahiert. Die Substanzen wurden abhängig von ihren Trefferquoten (Quotient aus Anzahl aktiver Messungen und Anzahl aller Messungen) mit dem "Durchschnitt plus Standardabweichung" Ansatz in drei Gruppen eingeteilt: nicht promiskuitive Substanzen (kleine Trefferquoten), promiskuitive Substanzen (überdurchschnittlich hohe Trefferquoten) und hoch promiskuitive Substanzen (unüblich hohe Trefferquoten). Auf diesem Datensatz wurden "extra tree" Klassifikatoren trainiert um (a) nicht promiskuitive von promiskuitiven Substanzen zu unterscheiden und (b) nicht promiskuitive von hoch promiskuitiven Substanzen zu unterscheiden. Die Modelle erreichten "Matthews correlation coefficients" (MCCs) von bis zu 0.67 und "area under the receiver operating characteristic curve" (AUC) Werte von bis zu 0.96. Die besten Modelle wurden unter dem Namen Hit Dexter im kostenlosen und frei zugänglichen Webserver "New E-Resource for Drug Discovery" (NERDD) veröffentlicht.

Im zweiten Teil dieser Doktorarbeit liegt der Fokus auf der Verbesserung der Modelle zur "frequent hitter" Vorhersage. Hierzu wurde der Trainingsdatensatz durch die Unterscheidung von "primary screen assays" (PSA) und "confirmatory dose-response assays" (CDRA) Daten verbessert. Zudem wurden Substanzen die auf strukturell ähnlichen Proteinen aktiv sind, und die dadurch fälschlicherweise als "frequent hitters" klassifiziert wurden, identifiziert und eliminiert. Diese neue Generation an "extra tree" Modellen, auch Hit Dexter 2.0 genannt, wurde auf verschiedenen qualitativ hochwertigen Datensätzen validiert. Zu diesen Testdatensätzen gehören ein Datensatz von zugelassenen Medikamenten, ein Datensatz mit ausschließlich inaktiv getesteten Substanzen (aus biochemisch und biologischen Assays), ein Datensatz mit Naturstoffen und ein Datensatz mit Substanzen, die für HTS Ansätze zusammengestellt wurden. Die besten Modelle erreichten auf dem zuvor unberücksichtigten Testdatensatz MCCs von bis zu 0.64 und AUC Werte von bis zu 0.96.

Im dritten Teil dieser Doktorarbeit wurde die Unterscheidung von Substanzen mit hohen Trefferquoten bezüglich biochemischen (Ziel-basierten) und biologischen (Zell-basierten) Assays vorgenommen und es wurden unterschiedliche Modelle für diese Assaytypen entwickelt. Dazu wurden Datensätze manuell aus der PubChem Bioassay database erstellt, die auf verschieden Assaytypen basieren: (i) Ziel-basierten Assaydaten, (ii) Zell-basierten Assaydaten, die ausgelegt sind spezifische Interaktionen in Zellen zu messen und (iii) Zell-basierten Assaydaten, die jegliche Art von Interaktion messen. Da für diese Datensätze mehr Datenpunkte zur Verfügung standen als bei der Entwicklung der ersten Hit Dexter Modelle, wurde die Anzahl an Proteinen, gegenüber denen die Substanzen mindestens getestet werden mussten, von 50 auf 100 erhöht. Dadurch basieren die neuen Modelle auf robusteren Daten. Die Vorhersagekraft verschiedener Algorithmen des maschinellen Lernens wurde auf diesen Datensätzen

untersucht. Dabei wurden “k-nearest neighbors” (KNN), “multilayer perceptron” (MLP) Klassifikatoren, “random forest” (RF) Klassifikatoren und “extra tree” (ET) Klassifikatoren verwendet. Die besten Modelle basieren auf MLP Klassifikatoren und wurden als Hit Dexter 3 veröffentlicht. Sie erreichen MCCs von bis zu 0.65.

Die Hit Dexter Modelle sind im öffentlichen und kostenlosen Webserver NERDD verfügbar gemacht worden. Zudem sind im Hit Dexter Webserver weitere regelbasierte Modelle und Modelle, die auf dem Ähnlichkeitsprinzip beruhen, zur Identifizierung von “bad actors” implementiert worden, um Substanzen aus HTS Experimenten mit verschiedenen Ansätzen priorisieren zu können.

Neben Hit Dexter sind sieben weitere Programme zur Beschleunigung des Wirkstoffdesigns in NERDD verfügbar. Unter anderem handelt es sich um Modelle zur Vorhersage von Ähnlichkeiten zu Naturstoffen (NP-Scout) beziehungsweise zur Vorhersage des Hautsensibilisierungspotentials (Skin Doctor CP) von Substanzen. Außerdem stehen fünf Programme zur Verfügung, die Vorhersagen über metabolisches Verhalten von Substanzen machen können: FAME3 ist ein Modell, basierend auf maschinellem Lernen, zur Vorhersage von wahrscheinlichen metabolisch-labilen Atompositionen im Phase 1 und Phase 2 Metabolismus. GLORY und GLORYx können wahrscheinliche metabolische Produkte von kleinen Molekülen vorhersagen. Zuletzt wurden in NERDD CYPstrate und CYPlebrity, zur Vorhersage von Substraten beziehungsweise Hemmstoffe von Cytochrom P450 Enzymen, verfügbar gemacht.

NERDD ist ein sicherer Webserver mit HTTPS Verschlüsselung. Außerdem können Rechnungen gestartet und zu einem späteren Zeitpunkt abgerufen werden, da die Ergebnisse temporär gespeichert werden. Trotzdem können jegliche Daten manuell vom Nutzer gelöscht werden, um auch die Arbeit mit vertraulichen Daten zu ermöglichen. NERDD kann auch mit größeren Rechnungen zum Beispiel mit Tausenden von Substanzen genutzt werden, da der Webserver an einen Hochleistungsrechencluster angeschlossen ist. Die Ergebnisse können im standard Dateiformat “csv” heruntergeladen werden, um sie weiter zu prozessieren oder zu evaluieren. Inzwischen ist NERDD zu einer etablierten Plattform für Wissenschaftler mit hoher Nachfrage geworden.

2. Abstract

High-throughput screening (HTS) approaches are key technologies for the identification of bioactive small molecules to be developed into drugs, agrochemicals and cosmetics. HTS allows the testing of large numbers of compounds for activity on biomacromolecules of interest and, to an increasing extent, also in cells. A major challenge in HTS are problematic rates of false assay readouts linked to assay interference. In particular false-positive assay outcomes regularly trigger futile follow-up experiments that can block significant resources in research.

Compounds that can cause assay interference are commonly referred to as "bad actors", "badly behaving compounds", "nuisance compounds", or, less accurately, "pan-assay interference compounds" (PAINS). The reasons underlying assay interference are manifold: formation of colloidal aggregates, chemical reactivity (e.g. covalent binders), membrane disruptors, metal complex-forming compounds, etc.

Many bad actors, but not all, show higher-than-expected hit rates in biochemical and biological assays and are therefore referred to as "frequent hitters". For example, many reactive compounds are frequent hitters as they may bind to multiple proteins and hence trigger (false) positive signals in different assays. Importantly, not all frequent hitters are bad actors; some of them are compounds that can bind, in a specific manner, to multiple proteins. These compounds are often based on "privileged scaffolds" that are compatible with multiple (protein) binding sites and may be particularly valuable in the context of "polypharmacology" and "drug repurposing".

Recent studies have shown that computational methods have the potential to identify bad actors but they are still in their infancy. The aim of this PhD study is to develop powerful machine learning approaches for the prediction of frequent hitters.

In the first part of this study, we explored the possibility to train machine learning models on large, publicly available bioactivity databases. More specifically, we compiled a set of approximately 311 000 compounds with measured activities on at least 50 proteins from the PubChem Bioassay database. The compounds were grouped into three classes depending on their hit rates with the averages plus standard deviation approach: non-promiscuous (low hit rates), promiscu-

ous (higher-than-expected hit rates) and highly promiscuous (uncommonly high hit rates) compounds. Based on these data sets we trained extra tree classifiers distinguishing (a) non-promiscuous from promiscuous compounds and (b) non-promiscuous from highly promiscuous compounds. These models obtained Matthews correlation coefficients (MCCs) and area under the receiver operating characteristic curve (AUC) values of up to 0.67 and 0.96, respectively. The set of best performing models, called Hit Dexter, have been made available to the public via a free web service New E-Resource for Drug Discovery (NERDD).

The second part of this study focused on the enhancement of frequent hitter prediction by extending the training data and taking differences between primary screen assays (PSA) and confirmatory dose-response assays (CDRA) explicitly into account. In addition, compounds that show activity on structurally similar protein targets were identified and eliminated in order to not be wrongly considered and classified as frequent hitters. Moreover, the new generation of extra tree models, distributed as Hit Dexter 2.0, were subjected to thorough validation with several high-quality data sets, including a data set of approved drugs, compounds that have consistently been measured as inactive across many different proteins, natural products and screening compounds. On holdout data these models reached MCCs and AUC values of up to 0.64 and 0.96, respectively.

The third part of this study focused on the differentiation of compound hit rates observed with biochemical (target-based) assays and biological (cell-based) assay and to build dedicated models for both assay types. A manually curated data set was compiled from the PubChem Bioassay database and dedicated models were developed for (i) target-based assays, (ii) cell-based assays measuring specific compound-target interactions, and (iii) an extended set of cell-based assays including also assays measuring nonspecific compound-target interactions. In order to obtain more robust models, the minimum number of available test data for the promiscuity label calculation of a compound was increased from 50 to 100 as more data became available. Multiple machine learning algorithms, including k-nearest neighbors (KNN) classifiers, multilayer perceptron (MLP) classifiers as well as random forest (RF) classifiers and extra tree (ET) classifiers were evaluated, and the best performing models (distributed as Hit Dexter 3) are based on MLP classifiers, reaching MCCs of up to 0.65.

The Hit Dexter models are available to the public via a free web service called NERDD. In addition to the machine learning models, the Hit Dexter web service offers a number of similarity-based and rule-based approaches for the identification of bad actors, making it a one-stop-shop for hit (de-)prioritization.

Besides Hit Dexter, NERDD features seven further modules which can be used to aid and accelerate drug discovery. These include models for the prediction of natural product likeness (NP-Scout) and skin sensitization potential of com-

pounds (Skin Doctor CP). Five programs are available for the prediction of the metabolic behavior of compounds: FAME3 is a machine learning model for the prediction of sites of metabolism in Phase 1 and Phase 2 metabolism. GLORY and GLORYx predict the likely metabolites of small molecules. The latest additions, CYPstrate and CYPlebrity, predict substrates and inhibitors of cytochrome P450 enzymes, respectively.

NERDD is a secure web server which comes along with HTTPS encryption. Calculations can be started within NERDD and can be retrieved at a later point in time as the results are temporarily saved on the server. However, manual deletion of the results is possible, which promotes the use of the web service also with confidential data. NERDD is easily scalable and thousands of compounds can be calculated due to its connection to high performance clusters. All results can be downloaded in standard csv file format for further processing and evaluation. NERDD has become an established platform for researchers and is in high demand.

3. Introduction

The development of a new drug can take longer than a decade and cost more than a billion dollar.[1] There is hence an urgent need to make drug discovery and development more efficient, and a main strategy to meet this need is the development and application of computational methods, e.g. for the identification of bioactive compounds and their optimization with regard to their bioactivity, toxicity as well as absorption, distribution, metabolism and excretion (ADME) profiles.[2–4]

3.1 High-throughput assays for lead compound identification

One of the most effective approaches to finding novel, bioactive compounds is high-throughput screening (HTS). HTS allows the screening of tens of thousands of compounds a day.[5–7] HTS technologies can be largely classified into cell- and target-based assays. While target-based assays are measuring the specific interaction of a compound with a purified protein, cell-based assays are determining the interactions of a compound in fully functional cells. For cell-based assays there is a higher degree of uncertainty about the target protein(s) of a compound of interest. However, this type of assays can provide particularly valuable insights and can be much more relevant to in vivo biology, pharmacology and toxicology than the target-based approaches.

Assays are commonly classified into primary screen assays (PSA) and confirmatory dose-response assays (CDRA). PSA are typically rapid approaches for the initial screening for active compounds. Any hits resulting from these screens are typically subjected to screening with CDRA. CDRA usually measure full concentration-dependent dose-response curves and are more precise regarding the activity of compounds than PSA. For a comprehensive overview of methods for assay screening setups (assay technologies) the reader is referred to Ref. [8].

A major challenge in HTS campaigns is to understand which of the initial hits are most promising to follow up on, and which ones are false positive outcomes as a result of nonspecific binding or interference of the compounds with the assay technology. The elimination of undesired compounds is done during the so-called risk assessment process of a screening campaign and is based on exper-

imental guidelines, strategies and recommendations.[9] The challenge resides in identifying false positive compounds that are observed only under specific conditions in some assays.[8] For example, autofluorescent compounds can trigger false positive outcomes in bioluminescence assays, whereas thiol-reactive compounds often trigger false positive readouts when the target involves thiol-containing amino acids. These false positive assay readouts entail an important waste of resources, since a lot of time and resources are invested into follow-up assays for compounds falsely identified as active. Although scientific journals are now making strong efforts to tackle this problem, e.g. by requiring a detailed examination of compounds containing problematic substructures, the problem still remains.[10]

For the purpose of supporting the process of hit (de-)prioritization (freely available) in silico methods have been developed that can provide guidance to researchers in drug discovery.[D1] Nevertheless, the expertise of medicinal chemists in experimental screening is needed for the design of assay campaigns as their experience and knowledge can detect potential pitfalls and avoid the report of false positive assay outcomes.[11]

3.2 Concepts and prediction of frequent hitters, nuisance compounds and dark chemical matter

Bad actors and multi-target compounds in HTS are often referred to as promiscuous compounds. However, this term is not well defined and is used in different ways. Therefore, in this work the single term "promiscuity" is not used, but the different categories of promiscuity are defined. A graphical overview of the different categories of compounds in biological and biochemical assay is shown in Figure 3.1. Note that the definitions of most of these terms are, to some extent, fuzzy and, in part, overlapping.

Frequent hitters are compounds that show higher-than-expected hit rates in assay panels due to interference with the screening technologies or, in a minority of cases, due to true promiscuity (i.e. the ability of compounds to bind to multiple targets) and were introduced by Roche et al.[12] Compounds causing assay interference are commonly referred to as "badly behaving compounds", "bad actors" or "nuisance compounds", and can be divided into three major categories: colloidal aggregators,[13] pan-assay interference compounds (PAINS)[14] and reactive compounds.[15] For further details of these categories see Section 3.2.2–3.2.4. The most prominent example of compounds showing the complexities and the connections of the promiscuity categories is curcumin and its derivatives.[16] Curcumin is a frequent hitter that has been repeatedly reported as active against many targets. However, it is known to be a membrane disruptor. This example

shows how important a closer look to the assay results is as curcumin might be a true active compound in some cases, whereas it might also be a false positive hit due to unwanted interactions with proteins. The different categories of compounds in HTS and how computational approaches can assist to detect them is described in the next Sections.

The converse of frequent hitters are so-called dark chemical matter (DCM). These are compounds that are extensively tested (at least 100 times) in target-based as well as cell-based assays and have never shown activity. These compounds were used for validating the predictions of the models of Ref. [D3] and Ref. [D5] and are a valuable source of potential, selective ligands, as discussed in Section 3.2.5.

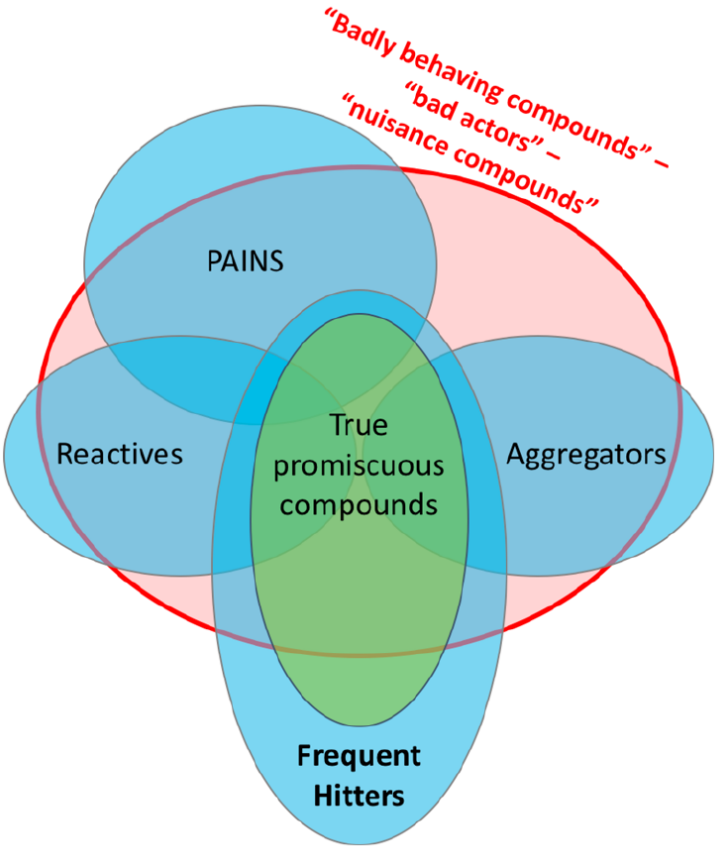


Figure 3.1: Promiscuous compounds can be divided into different groups which merge into each other and cannot be strictly separated from each other. Reprinted (adapted) with permission from Ref. [D3]. Copyright 2019 American Chemical Society.

The aim of the majority of the approaches discussed in the following sections is the identification of problematic compounds. Therefore, an overlap of the differ-

ent methods often exists. For example, for substructure-based approaches, which rely on SMARTS patterns and other collections of substructures. These algorithms match the known problematic substructures and the query compounds. The used substructure patterns are often similar to each other and computational algorithms exist to show the overlap of such substructure collections.[17, 18] This might be reasoned by the fact that the aim of these filters is often a similar one: detect problematic compounds.

3.2.1 Frequent hitters

For the identification of frequent hitters, first bioactivity data have to be analysed to find compounds with higher-than-expected hit rates.. Three main definitions for frequent hitters are published and used: i) an absolute measurement in which for each compound, the number of positive assay outcomes is counted. If a compound shows bioactivity (i.e. a positive outcome) in more assays than the defined cutoff, the compound is labeled as frequent hitter. For example, Pearce et al. set the cutoff to seven active assay outcomes and compounds above the cutoff were labeled as frequent hitters. Compounds were labeled as inactive if they were tested in at least 15 assays and never showed any activity.[19] ii) a relative measurement that is based on the fraction of active assay outcomes with respect to the total number of assay outcomes calculated for each compound in a certain data set. In this study such a measurement, called active-to-tested ratio (ATR), was used and is calculated according to Equation (3.1), where A reflects the number of active measurements and T the number of times a compound was tested overall.[D3, D5, D2]

$$ATR = \frac{A}{T} \tag{3.1}$$

Finally, iii) a binomial distribution that reflects the assay outcome of a compound as a statistical value that is normally distributed. Based on the normal distribution a probability can be predicted with statistical significance and biases are well catched out.[20] Typically, 11 – 13% of compounds show a higher-than-expected hit rate in standard HTS libraries and data sets.[D3] GlaxoSmithKline reported a data set containing around 13% “noisy” compounds (frequent hitters),[21] and AstraZeneca report a data set containing 6% frequent hitters.[20] Only a few of the compounds (around 2%) used in Hit Dexter 2.0 show a hit rate that is much higher (above hit rates of 5-10%) than the expected one. This is in line with a study of Pearce in which 0.1% of the active compounds show a high number of hits.[19] The study of AstraZeneca reported an overall hit rate average of 1.53% which is comparable with the values of Hit Dexter 2.0 of 0.8% - 1.5%.

The first models for the prediction of frequent hitters were developed by Roche et al. and reached Matthews correlation coefficients (MCCs) of up to 0.81. Problematic in this study is that the data and the models are not publically available.[12] The binomial function approach is an inhouse tool of AstraZeneca to evaluate newly reported bioactivity data in their inhouse library. Neither the tool nor the data are published, but the algorithm is fully described[20] and could be reimplemented for public usage.

A comprehensive study of the correlation of substructures and frequent hitters was performed by Charkravorty et al., who analyzed established rule-based approaches and their reported substructures. The substructures were classified into different frequent hitter classes depending on their hit rates. They analysed the GlaxoSmithKline HTS collection and showed which chemical substructures are suitable for frequent hitter prediction based on a relative frequent hitter measurement similar to the ATR (Equation (3.1)). The substructures that match most of the frequent hitting compounds were implemented in a new substructure filter set.[21]

Badapple is a tool for the identification of frequent hitters based on their molecular scaffolds. The model is derived from a hierarchical scaffold clustering (HierS[22]) analysis that scores each scaffold with a promiscuity score based on bioactivity data of a large data set (430k compounds measured in more than 800 assays).[23]

PrePeP is an algorithm based on decision trees using discriminative subgraph mining which can be used with any data set encoding a classification task. The algorithm balances the data, extracts subgraphs and trains decision trees to discriminate between two activity classes in any given data set. PrePep was validated for frequent hitter as well as PAINS prediction and is able to predict frequent hitters with an accuracy of 85% on a balanced test data set.[24, 25]

Hit Dexter is the result of these PhD studies, which are divided into three parts: Hit Dexter,[D2] Hit Dexter 2.0[D3] and Hit Dexter 3.[D5] The Hit Dexter models are based on a large data set extracted from PubChem and a machine learning algorithm trained on this data. PubChem is one of the largest publicly available data sources for bioactivity data containing around 300 million bioactivity data points (for further details on PubChem see Section 3.4). After a comprehensive data preprocessing procedure (for details see Method Sections of Section 5.1 - 5.3), three promiscuity classes of compounds were defined based on the aforementioned ATR (Eq.(3.1)), following the standard deviation plus mean approach: i) non-promiscuous compounds, ii) promiscuous compounds and iii) highly promiscuous compounds. Non-promiscuous molecules show an ATR of less than the average of the ATR, promiscuous compounds a higher ATR than the mean plus one standard deviation of the ATR and highly promiscuous com-

pounds a ATR higher than the mean plus three times the standard deviation of the ATR.

The Hit Dexter models were developed based on all PubChem data that were available after applying a number of prefiltering steps (for detail see Method section in Section 5.1). Hit Dexter 2.0 was refined for PSA and CDRA, and a recently developed tool, Hit Dexter 3, distinguishes cell-based and target-based assays. From the respective data sets (molecules encoded with Morgan2 fingerprints; for further details on the descriptors see Section 3.5.2) extra tree (ET) classifier and multilayer perceptron (MLP) classifier showing a performance of MCC up to 0.65 were derived. For the complete results see Sections 5.1 - 5.3.

3.2.2 Colloidal aggregators

Aggregators are compounds that generate colloidal aggregates under specific conditions.[26] These aggregates can be prone to false positive assay readouts (e.g. due to denaturation), which is an unwanted effect during assay screening.[27] Chemically, aggregators often contain planar ring systems with large aliphatic groups that undergo large van der Waal interactions with each other and therefore are easy to aggregate. Often at high concentrations and specific conditions (pH-value of the buffer; presence of macromolecules; temperature; etc.) micelles are formed. These micelles can interact (in a concentration-dependent and reversible fashion) with protein targets making aggregation difficult to detect. In Ref. [28] it was shown that compounds that show positive assay results due to aggregation on G-protein-coupled receptors can be found in the literature. Luckily, the addition of detergents often suppresses aggregation, which is the reason why most of the biochemical assays are run under detergent-containing conditions in their default setup.[8] Moreover, a study showed that around 93% of the false positive assay outcomes can be avoided by adding detergents.[29] The experimental detection of aggregation can be achieved by dynamic light scattering if the conditions are chosen carefully.[30]

Many in silico tools exist for the prediction of aggregators. One of the first attempts to predict aggregators with machine learning, more precisely by support vector machine, was reported in 2009 by Rao et al., based on a data set containing 1319 aggregators and 128 325 non-aggregators. Their model reached sensitivities of up to 78% during five-fold cross validation.[31] The most popular approach for potential colloidal aggregation detection is Aggregator Advisor. This is a similarity-based approach which also takes calculated logP values into account. The model is based on over 12600 known aggregators.[13]

Recently, two further approaches for predicting aggregators were published: ChemAgg[32] and small, colloidally aggregating molecules (SCAM) detective.[33]

The machine learning models of ChemAgg are freely available on a web server and based in part on the data set of the Aggregator Advisor (compounds labeled as aggregators). Non-aggregators were collected from drug/drug candidates comprising around 24 000 compounds. Accuracy values and area under the receiver operating characteristic curve (AUC) values of up to 95% and 99%, respectively, were obtained on the training data set. SCAM detective is also a freely available web server and based on random forest (RF) classifiers to predict the likelihood of compounds being aggregators in two different assay setups (one with β-lactamase as target and the other with the cysteine protease cruzain as target). Therefore six assays (three with detergent and three without) were used as training data (four with β-lactamase and two with cysteine protease cruzain as target). The models can predict aggregators for these two assay setups with a 53% and 46% better accuracy than former prediction tools (e.g. Aggregator Advisor).[33]

3.2.3 Pan-assay interference compounds (PAINS)

The most prominent example of badly behaving compounds in HTS are the PAINS.[14] PAINS are compounds containing at least one of 480 substructures that encode problematic compounds in AlphaScreen assays measuring protein protein interactions. Compounds containing such a substructure should be treated with extra caution as they are likely to trigger false positive assay readouts. Originally, the substructures were encoded using the SYBYL notation [34] and members of the scientific community translated them to the more commonly used SMARTS[35] patterns. Well-known examples of compounds encoded by the PAINS patterns are edox cyclers (e.g. Toxoflavins), covalent binders (e.g. isothiazolones or ene-rhodanine), membrane disruperts (e.g. curcumin), metal complexers (e.g. hydroxyphenyl hydrozones) and unstable compounds (e.g. phenol-sulphonamides).[36] Independent of the original PAINS study, it was shown that PAINS encode compounds that often have stability problems, are aggregators or cytotoxic.[37] However, it was also shown with X-ray structures that a lot of PAINS substructures show actual interactions with proteins (not false positive interactions).[38] A machine learning approach already exists to predict Compounds Interfering with an Assay Technology (CIAT) which conceptually are the same as PAINS. These machine learning models are based on a data set containing assays from the AlphaScreen, Fluorescence Resonance Energy Transfer (FRET) and time-resolved FRET assay reading technologies and can rank CIAT with AUC values of up to 0.81.[39]

One of the major problems of the PAINS approach is its narrow applicability domain (AD) as the substructures are built only on readouts from AlphaScreen assays. Nevertheless, compounds that were encoded by the PAINS patterns show also interactions with targets in other assays and have a higher degree of

promiscuity than other compounds.[40] Further, it was confirmed that the PAINS patterns work well for AlphaScreen but are not suitable for FRET assays.[41] Often the PAINS patterns are not only used for vetting outcomes of AlphaScreen assays but also for those of other assay technologies (for example, cell-based assays). However, their application to other assay setups is not recommended as these assay types are outside the AD.[42] For a data set containing nuisance compounds of cell-based assay screenings see Ref. [43]. Another problem of the PAINS patterns is that they are derived from a proprietary data set that is not accessible to the scientific community.

The concept of PAINS is controversial in the scientific community.[44–46] One problem, for example, are the so-called phantom PAINS which are compounds that match at least one PAINS pattern but show no or low activity hit rates in biochemical assays. In the mentioned study a data set with over 70 000 random PAINS compounds from PubChem was analysed.[47] The findings corroborate the narrow AD of the PAINS concept. Further, these findings may lead to the conclusion that the awareness of the scientific community regarding PAINS has already increased and that these compounds may be validated in a more sophisticated way resulting in inactive reports. The blind use of the PAINS (and any other) filters is not an adequate way of hit validation.[48] Nevertheless, the use of filters as decision support for (de-)prioritizing hits might be a good idea. Jasial et al. have shown that the combination of the PAINS patterns with machine learning algorithms can enhance the performance of the PAINS substructure filters.[49]

3.2.4 Reactive compounds

Reactive compounds are likely to produce false outcomes in biological assays due to interactions and/or reactions with the protein target and/or with components of the assay screening technology. Among the mostost prominent examples of reactive groups are Michael acceptors or α-halocarbonyl compounds.[15] In general, most electrophiles[50] are problematic, but there are also other groups that should be treated with caution (e.g. epoxides).[51] It was shown by density-functional theory calculations that reactive frequent hitter compounds often have an electrophilic character.[52]

For the experimental identification of reactive compounds several approaches exist. One of the most prominent approaches is ALARM nuclear magnetic resonance (NMR) in which the unwanted reactivity of compounds can be observed by NMR spectrometry. Other NMR methods exist for identifying specific thiol reactivity[53] and an overview of the existing methods is given in Ref. [54].

The computational prediction of reactive groups is mainly based on substructures encoding reactive groups that are unwanted in HTS campaigns. For example, Hann et al. published 55 substructures that are undesired for lead optimization due to their reactivity and are used as hard filters (i.e. compounds containing one of the substructures are not further considered).[55] First attempts are made to build machine learning models to enhance the prediction of reactive groups in biochemical assays.[56]

3.2.5 Dark chemical matter

The opposite of frequent hitters is so-called DCM. DCM represents compounds which have been tested in at least 100 biochemical and biological assays and have not shown activity in any of the assays.[57] The idea behind the concept of DCM was to build a data set for HTS in which compounds that result in a positive outcome are likely to be specific, selective and true positive results.[58] The DCM concept is explored in several studies such as the one from Ballante et al. who built a DCM data set for docking in a virtual screening approach against several targets. Positive results from the docking approach can then be screened experimentally.[59] In the present study, the DCM data set was used to validate the models.[D3]

3.2.6 True promiscuous compounds (multi-target compounds)

True promiscuous compounds are multi-target compounds interacting with a number of distinct biomacromolecules in a specific manner. These compounds are also called master key compounds[60] and privileged structures, which correspond to scaffolds that are likely to interact with multiple targets.[61, 62] True promiscuous compounds are normally frequent hitters, which implies that frequent hitters can trigger specific positive readouts in assay screenings (and not necessarily only false positive readouts).[63] An overview of data sets for detecting true promiscuous compounds is given in Ref. [64].

Polypharmacology uses exactly this multi-target approach for drug discovery as compounds can be much more effective when they interact with more than a single target. However, compounds that interact with multiple targets often show side-effects (off-target effects). Before the start of clinical trials a lot of these toxicity and ADME properties are checked, for example in toxicity assay panels supported by in silico methods.[2] High drug safety standards are ensured during this step in drug development.[65, 66] These studies can be supported by target prediction to identify potentially true promiscuous compounds that have the potential to act with off-targets.[67]

3.2.7 Assay technology-specific nuisance compounds

Most nuisance compounds are frequent hitters and show up in several different assay screening setups. Compounds that trigger false positive assay results only under specific experimental conditions in some assay detection technologies and assay setups are problematic, as they cannot be detected as frequent hitters. For example, a compound that shows autofluorescence or absorbance will only give false positive results in light-dependent assay reading technologies (e.g. in bioluminescence assays). On a standard 70 000 sample large screening library of NIH was shown that 5% of compounds show false activity in fluorescence assays.[68]

Bioluminescence assays (often using luciferase) are one of the best examples of the efforts that are being made for the prediction of badly behaving compounds specific to an assay detection technology. In luciferase assays compounds that show autofluorescence or inhibit the luciferase enzyme are likely to give false positive results in such assays.[69] The use of orthogonal assays is one way to detect false positive assay outcomes. However, performing complementary assay setups showing activity through another mode of action is time consuming and expensive. Computational approaches like Luciferase Advisor,[70] ChemFluc[71] and InterPerd[72] are cheaper and can identify compounds eliciting false positive assay readouts in luciferase assays with balanced accuracies of up to 89,7%, 86% and accuracies of around 80%, respectively.

In principle fluorescence compounds, compounds with light absorbance properties and compounds with quenching activities are only problematic in specific assay detection technologies.[73] For example, in luciferase assays compounds with specific properties (i.e. autofluorescence) often show false positive assay readouts and for glutathione S-transferase–glutathione interaction assays other compounds are problematic.[74] But also specific targets containing particular amino acids (especially cysteine) can be problematic during assay screening. The oxidation of cysteine protease can occur and lead to false positive assay readouts. This effect can be measured using a liquid chromatography – mass spectrometry/tandem mass spectrometry approach.[75] Another cause of false positive assay readouts can be triggered by the use of dimethyl sulfoxide as an assay component.[76] Some of these false positive outcomes can be detected with already available computational approaches for hit prioritization in specific cell-based assays, such as the reporter gene assay.[77]

3.2.8 Applicability domain of existing computational models

One challenge of the discussed approaches is the definition of their applicability domain as not all approaches might be suitable for all purposes.[78] The blind use of filters such as the PAINS filter, which have a clearly defined applicabil-

ity domain,[14] results in controversial discussions about the usefulness of such filters.[48] Beside the applicability domain the advantage of substructure filters (like the PAINS filters) is their good interpretability of the predefined substructures, which are easy to understand for a medicinal chemist. Without an applicability domain quantitative structure–activity relationship models, which are often used as "black boxes", are more difficult to interpret and will not be applied as a user will have problems understanding when to use a certain model and for what.[79]

3.2.9 Computational approaches for hit (de-)priorisation

In an editorial for Future Medicinal Chemistry (Ref. [D1]) we discussed the most important computational methods for biochemical assay hit (de-)priorisation. We structured the work into four parts: rule-based approaches, similarity-based approaches, statistical approaches and machine learning approaches.

[D1] PAIN(S) relievers for medicinal chemists: how computational methods can assist in hit evaluation
Conrad Stork and Johannes Kirchmair
Future Medicinal Chemistry, 2018

Available at `https://doi.org/10.4155/fmc-2018-0116`.

Contribution:
C. Stork wrote the manuscript, with contributions from J. Kirchmair. J. Kirchmair supervised the work.

The following article was reprinted with permission from:
Stork, C. and Kirchmair, J. PAIN(S) relievers for medicinal chemists: how computational methods can assist in hit evaluation, *Futur Med.Chem.* **2018**, *10*, 1533–1535.
Copyright 2018 Johannes Kirchmair & Conrad Stork

Editorial

For reprint orders, please contact: reprints@future-science.com

Future Medicinal Chemistry

PAIN(S) relievers for medicinal chemists: how computational methods can assist in hit evaluation

Conrad Stork[1] & Johannes Kirchmair*,[1]

[1]Department of Computer Science, Center for Bioinformatics, Faculty of Mathematics, Informatics & Natural Sciences, Universität Hamburg, Hamburg, 20146, Germany

*Author for correspondence: Tel.: +49 0 40 42838 7303; kirchmair@zbh.uni-hamburg.de

"Today, several *in silico* methods are at our disposal and can provide guidance to medicinal chemists on potential nuisance compounds."

First draft submitted: 10 April 2018; Accepted for publication: 10 April 2018; Published online: 29 June 2018

Keywords: aggregators • frequent hitters • *in silico* prediction • *in vitro* screening • machine learning • molecular similarity • PAINS • promiscuous compounds • rule-based approaches • statistical methods

Modern high-throughput screening technologies allow for the testing of tens of thousands of compounds per day. However, a substantial proportion of the initial hits can be artifacts related to aggregate formation [1], chemical reactivity, photoreactivity, redox activity, metal chelation, interference with assay spectroscopy, membrane disruption, decomposition in buffers and other mechanisms [2–4].

The seminal works by the Shoichet group on aggregators [1] and by Baell and Holloway on pan-assay interference compounds (PAINS) [2] have greatly increased the scientific community's awareness of the pollution of medicinal chemistry and chemical biology literature with 'bad actors' and 'frequent hitters'. Less present in discussions but not of lower significance are impurities and decomposition products as sources of assay interference [5,6].

Recently, the editors-in-chief of nine ACS journals have teamed up to define best practice guidelines for how to identify assay artifacts and reject such hits [4]. Recommendations include the measurement and publication of full concentration response curves as well as the use of reporter-free methods such as surface plasmon resonance.

At this point, it is important to note that frequent hitters are not necessarily bad actors and vice versa. Frequent hitters are compounds which have a higher-than-expected activity rate recorded in historical screening data. Bad actors, on the other hand, are compounds that trigger false assay readouts under specific conditions and therefore often, but by far not always, show a high frequency of false readouts. In addition to some bad actors, frequent hitters also include true promiscuous compounds (sometimes related to privileged scaffolds) that may in fact be of interest in the context of polypharmacology and drug repurposing.

Computational methods can make a significant contribution to the identification of potential bad actors and/or frequent hitters. These computational techniques include rule-based and similarity-based methods, statistical approaches and machine learning. Here, we will briefly discuss the most relevant approaches that are publicly accessible.

Rule-based approaches

Rule-based approaches aim to encode existing empirical knowledge from *in vitro* data in rule sets. Most widely applied is a publicly available set of rules defined by 480 patterns that encode substructures present in classes of compounds (PAINS) that have been linked to assay interference under specific conditions. Importantly, the term PAINS is not a synonym of nuisance compounds, bad actors or frequent hitters, although it is often used as such. For the appropriate application of PAINS (and any other) patterns, it is of utmost importance to consider their definitions, scope and limitations. The PAINS patterns have been derived from 100 k compounds screened generally at high concentrations against six protein–protein interactions with a single screening technology (i.e., AlphaScreen assay). The use of high concentrations may have emphasized assay interference. All compounds

newlands press

had previously passed a garbage filter and were screened under detergent-containing conditions, meaning that the PAINS patterns do not account for many of the generally undesirable functional groups (e.g., electron-deficient and reactive epoxides) or for aggregate formation (with exceptions). Importantly, the definition of PAINS is class-based, meaning the presence of PAINS patterns in individual compounds does not necessarily imply pan-assay interference. Moreover, the individual PAINS patterns are derived from differing numbers of experimental observations [2]. On a more technical note, users should be aware that different implementations of PAINS patterns and matching algorithms exist, which may lead to different results. Baell, Holloway and Nissink have pointed out these facts and limitations [2,3]. However, this set of rules is all too often used, without the necessary diligence, as a hard filter to reject compounds. Such use will almost certainly result in a loss of true hits (~5% of US FDA-approved drugs raise a PAINS alert [3]). Equally bad, false hits may be selected for follow-up studies because they do not trigger a PAINS alert.

Several further rule sets are in existence and have been incorporated into databases such as ChEMBL [7]. Most of them originate from major pharmaceutical companies and were developed with the aim to deprioritize compounds with undesired chemical features from *in vitro* screening. As in the case of the PAINS patterns, these rule sets may be useful as guidance but not necessarily as hard filters. One valid criticism about most of these sets of rules is that they have been derived from proprietary data, meaning that they cannot be directly verified or reproduced [8]. Scientists from GSK have recently reported a critical analysis of the usefulness of published filters based on two million unique compounds that have been tested in several hundred in-house screening assays [5]. They also introduced rules for some new classes of nuisance compounds. One of their main conclusions is that a variety of filter strategies need to be employed in order to properly account for the different types of nuisance.

Similarity-based approaches

Aggregator Advisor [9] compares the molecular structure of compounds of interest to over 12,600 known aggregators. Compounds exceeding a defined similarity or log P threshold are recommended for testing of aggregate formation. Due to the intrinsic nature of this approach, the absence of a structurally related molecule in the reference set does not imply a compound's benignity in the context of screening. Aggregator Advisor is available as a free web service, and the dataset of known aggregators is available for free download [10].

Statistical approaches

All methods discussed so far focus on the prediction of one or several different types of assay artifacts. A different approach is followed by Badapple [11], which uses a statistical model to identify compounds that are likely frequent hitters based on their scaffolds. The regression model was derived from a set of over 430 k compounds measured in 822 different assays. For compounds of interest, Badapple performs a hierarchical scaffold analysis to compute a promiscuity score which corresponds to the likelihood that a compound containing a certain scaffold is promiscuous. A further example of a statistical model for the prediction of frequent hitters is an in-house tool from AstraZeneca derived from their corporate database [12]. This tool is not publicly available but has recently been compared by Baell and Nissink [3] with Badapple and the PAINS patterns. Based on the results obtained for 16 of the most highly populated PAINS substructures (i.e., filter family A as defined in [2]), the authors concluded that the individual approaches consistently recognize a substantial number of problematic substructures. Badapple is available as a free web service [13] and as a plug-in of the Bioassay Research Database [14].

Machine learning approaches

The most recent approach for the prediction of frequent hitters is Hit Dexter [15], which was developed in our lab in Hamburg. The idea behind Hit Dexter is the development of a robust machine learning model able to identify frequent hitters independent of the underlying interference mechanism (or chemical pattern conferring true ligand promiscuity). Although Hit Dexter and Badapple are trained on similar datasets, the methodologies are clearly different. Firstly, Hit Dexter is based on two extremely randomized trees classifiers rather than a statistical model, and secondly, Hit Dexter was designed to account for subtle differences in compound structure rather than focusing only on scaffolds. Using Morgan2 fingerprints to encode molecular structures, the classifiers reached Matthews correlation coefficients and area under the receiver operating characteristic curve values of up to 0.67 and 0.96, respectively, on a large, independent test set. Hit Dexter is also available as a free web service [16].

Conclusion

The research of assay-interfering compounds is still in an early stage. There is much left to be done in order to establish a community-wide standard for vetting measured bioactivity data. The recent developments and ongoing discussions clearly point in the right direction [3–5,8]. In particular, researchers in academic drug discovery are increasing efforts to implement hit validation procedures established in the industry. These procedures are based on the principle that screening hits are only considered valuable, if a strong structure–activity relationship can be established.

Today, several *in silico* methods are at our disposal and can provide guidance to medicinal chemists on potential nuisance compounds. Clearly, there is room for improvement of these methods. Our biggest concern, however, is that the limitations of these methods, with respect to applicability and accuracy are all too often not considered appropriately. Predictions of all methods and models discussed herein may be used carefully for flagging or (de-)prioritizing compounds but should not be applied blindly as hard filters to reject compounds. Used wisely, these computational models can help in the design of screening libraries and in making better-informed decisions during hit triage and follow-up.

Financial & competing interests disclosure

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - project number KI 2085/1–1. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

Open access

This work is licensed under the Attribution-NonCommercial-NoDerivatives 4.0 Unported License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-nd/4.0/

References

1. McGovern SL, Caselli E, Grigorieff N, Shoichet BK. A common mechanism underlying promiscuous inhibitors from virtual and high-throughput screening. *J. Med. Chem.* 45(8), 1712–1722 (2002).
2. Baell JB, Holloway GA. New substructure filters for removal of pan-assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *J. Med. Chem.* 53(7), 2719–2740 (2010).
3. Baell JB, Nissink JWM. Seven year itch: pan-assay interference compounds (PAINS) in 2017-utility and limitations. *ACS Chem. Biol.* 13(1), 36–44 (2017).
4. Aldrich C, Bertozzi C, Georg GI *et al.* The ecstasy and agony of assay interference compounds. *J. Med. Chem.* 60(6), 2165–2168 (2017).
5. Chakravorty SJ, Chan J, Greenwood MN *et al.* Nuisance compounds, PAINS filters, and dark chemical matter in the GSK HTS collection. *SLAS Discov.* 23(6), 532–545 (2018).
6. Hermann JC, Chen Y, Wartchow C *et al.* Metal impurities cause false positives in high-throughput screening campaigns. *ACS Med. Chem. Lett.* 4(2), 197–200 (2013).
7. The ChEMBL-og. http://chembl.blogspot.de/2015/02/chembl-20-released.html
8. Kenny PW. Comment on the ecstasy and agony of assay interference compounds. *J. Chem. Inf. Model.* 57(11), 2640–2645 (2017).
9. Irwin JJ, Duan D, Torosyan H *et al.* An aggregation advisor for ligand discovery. *J. Med. Chem.* 58(17), 7076–7087 (2015).
10. Aggregator Advisor. http://advisor.bkslab.org
11. Yang JJ, Ursu O, Lipinski CA, Sklar LA, Oprea TI, Bologa CG. Badapple: promiscuity patterns from noisy evidence. *J. Cheminform.* 8, 29 (2016).
12. M Nissink JW, Blackburn S. Quantification of frequent-hitter behavior based on historical high-throughput screening data. *Future Med. Chem.* 6(10), 1113–1126 (2014).
13. Badapple. http://pasilla.health.unm.edu/tomcat/badapple/badapple
14. Howe EA, de Souza A, Lahr DL *et al.* BioAssay Research Database (BARD): chemical biology and probe development enabled by structured metadata and result types. *Nucleic Acids Res.* 43(Database issue), D1163–D1170 (2015).
15. Stork C, Wagner J, Friedrich N-O, de Bruyn Kops C, Šícho M, Kirchmair J. Hit Dexter: a machine-learning model for the prediction of frequent hitters. *ChemMedChem.* 13(6), 564–571 (2017).
16. Hit Dexter. http://hitdexter.zbh.uni-hamburg.de

3.3 Relevance of frequent hitters and nuisance compounds within the community and controversial discussion in the literature

The relevance of frequent hitters, PAINS, reactive compounds, aggregators and true promiscuous compounds is not only shown by the huge amount of citations of the original study for the PAINS substructure patterns, but also by the controversial discussion in the scientific literature. A suggestion of the editors of the ACS journals to use PAINS filters as a quality criterion during the review process[10] was published with controversial reactions.[80] The problem was stated within the reaction, that there are also other problematic compounds that are not detected by the PAINS filters and that PAINS can show real interactions with targets (as evidenced by X-ray structures). Another observation is that other scientific fields like biology are also becoming aware of the problem of false positive assay outcomes.[81] At some point the discussion reached the point where "PAINS shaming" (discussions in blog entries that show PAINS strucuters that were actually not acting as PAINS) was practised and the concept of PAINS had to be defended.[82] Nevertheless, this topic should not be treated in a binary way, as PAINS can be useful in some cases but only when applied in a conscious manner.

Industry and larger academia institutions as well as large screening facilities use many different substructure filters and machine learning models for different tasks. For example, the filtering of screening libraries is performed by Glaxo Wellcome using a "hard filter" for filtering undesired groups,[55] Pfizer developed the "LINT" rules for finding undesired groups[83] and the NIH Molecular Libraries Small Molecule Repository removes compounds containing one of 116 patterns.[84] The University of Dundee uses 105 patterns for removing assay interference compounds,[85] similar to the ones used by Bristol-Myers Squibb (180 patterns),[19] whereas the 480 PAINS patterns,[14] which are often used in academic research, and the frequent hitter pattern detection by Chakravorty et al.[21] developed and used by GlaxoSmithKline have been discussed in Section 3.2.3. A different issue is addressed by the ChEMBL ToxAlter rules, which are more specialized for toxic compounds.[86]

Further tools used for hit evaluation in industry (discussed above; Section 3.2.1) are the tool developed and used by AstraZeneca,[20] which uses the binomial function to evaluate the likelihood of a compound to be an outlier, and the machine learning algorithm of Roche[12] (which was at least used in their company for some time).

It can be concluded that this is a hot topic in academia and in industry which has caused some controversial discussion over the last decade. In the end all developed tools can only be as good as the way they are used.

3.4 Data sets for the development of computational approaches for frequent hitter prediction

In this work several data sets are used for model development as well as for model validation. The most relevant ones are discussed in this Section.

PubChem is the largest freely available data source of biochemical activity data.[87–89] It consists of several connected databases sharing bioactivity data of substances. The most important PubChem database for compiling data sets for the work presented in this PhD thesis is the PubChem Bioassay database.

The ChEMBL[90] database is a comprehensive resource on bioactivity data for small molecules. Contrary to PubChem it contains mainly active compound-target bioactivity records.[91] In this work, the ChEMBL database was mainly used in this study for the comparison of chemical spaces (see Section refsec:HD2 for details). The chemical space that is displayed by the ChEMBL database is large and drug-like. This database contains 2% of compounds with multi-target activity against structurally distinct targets, which makes it suitable for searching for promising polypharmacological drugs.[92]

Drugbank contains approximately 2000 food and drug association approved and 206 withdrawn drugs, as well as information about drugs and drug-like molecules.[93] For example, Drugbank contains information about the metabolism and toxicity and ADME properties of drugs. In this work Drugbank was mainly used as an external data set for model evaluation.

Natural products are interesting for drug discovery as they were evolutionarily designed to be active against some targets. Since many groups of natural products are known to be frequent hitters they were used in this study for model evaluation.[94] For that purpose, a well-curated, large dataset of natural products, published by Chen et al.,[95] was used.

Different data sets are suitable for different tasks and applications. Most data sets are biased depending on the purpose for which the database was developed for. The overall chemical space is inconceivably large and our chemical libraries are normally biased into the direction of drug-like space.[96] Another challenge for data analysis on the existing and biased databases is that, depending on the data set and data selection criteria, different results can be drawn.[97]. A good overview of multi-target compound data sets is given in Ref. [98] It can be observed that, in general, bioactive compounds (i.e. compounds that were tested active) show less multi-target activity than drugs. This effect might

be caused by the fact that drugs are more often tested than other compound classes.[99] Nevertheless, a higher number of performed experiments does not directly translate in a higher true promiscuity rate, as this rate remained stable over the last 40 years.[100]

3.5 Machine learning approaches

Computational algorithms that can detect patterns in unseen data after learning from past data are called artificial intelligence (AI). AI is a large field that includes machine learning. Machine learning is a statistical approach of generating information from data in a linear and/or nonlinear manner. One big area of machine learning is deep learning which includes neural networks with multiple hidden layers aiming to simulate the human brain. In this work several machine learning algorithms, nonlinear statistical models for generating knowledge from data were used.

A lot of machine learning algorithms exist in several different architectures. In principle machine learning algorithms can be divided into two main approaches: unsupervised learning and supervised learning. Unsupervised learning uses unlabeled data, meaning that the outcome of the data is not predefined (i.e. there is no right or wrong answer to the problem). These algorithms structure the data and try to find patterns in it with the aim of, for example, dividing the data points into two (or multiple) classes. In supervised learning, however, the outcomes of the training data are already defined. For example, it is known if a compound results in a positive or negative outcome for a specific task or not. In this case the machine learning algorithm learns from training data (with known class labels) and can make predictions for unseen (and unlabeled) data.

The labeled data used for supervised learning can have mainly two different structures. On the one hand, the data can have different labels (classifications). Binary classifications, which were used during this Ph.D. study, are the easiest classification problems and have simple “yes” and “no” (1 and 0, true and false) labels. On the other hand, the data can take any float number as a label (for example any float between 0 and 1), in which case the models have to solve a regression task. Regression tasks are more difficult than binary classification models as a continuous value has to be predicted. As these high requirements could not be reached by the data sets binary classification models were mainly used in this work.

Both classification and regression models can be further divided into linear and nonlinear machine learning algorithms. Linear models are suitable for linear data.

In this work unsupervised learning algorithms were applied for data analysis, and all models for frequent hitter prediction were based on supervised machine learning algorithms.

3.5.1 Algorithms

The most relevant machine learning algorithms in the context of this work are RF classifier, ET classifier and MLP classifier. The implementation of the machine learning algorithms applied here belong to scikit-learn, a python library with a large collection of tools and models.[101, 102]

RF classifier and ET classifier algorithms are both based on the decision tree algorithm. The aim of a decision tree is the classification of instances by their descriptors, separating the data based on the value of one descriptor at each node (i.e. decision) of the tree. This decision tree is similar to a flow diagram which always has two possibilities for a decision and no loops. After several decisions based on the different values of the descriptors a final classification is done. The below discussed algorithms use a collection of trees which are called forests, to make the final prediction.

The RF classifier is a classification model that takes several randomized decision trees into account to make a prediction. Thereby each tree has a randomized setup regarding the descriptor values, regarding the ordering of the decisions and regarding the samples used for training in a tree. In the end a majority vote of all trees in the forest results in a prediction value between zero and one, whereas the cutoff for binary classification is normally set to 0.5. Hence, all samples with a prediction of 0.5 or larger are classified as one (or active), all samples with a prediction below 0.5 are predicted as zero (or inactive).[103] In ET classifier the randomization is increased by also using a random threshold for the descriptor values at each decision node of the tree.[104]

MLP classifiers are based on multiple layers of perceptrons that interact with each other. The design of the network and its components are inspired by the human brain. A perceptron takes any number of input values, adds a bias to them by applying a nonlinear function on the linear combination of all the values, and returns the resulting value of these operations. In a simple MLP classifier, multiple perceptrons in a layer are optimized to minimize the error of the predictions. Using multiple layers of perceptrons with different numbers of perceptrons per layer, deep neural networks that can perform extremely complex classification tasks, can be built.[105–108]

3.5.2 Molecular descriptors

Molecular descriptors are an essential part of training of machine learning models, since they should describe all important characteristics (in most cases they do cover all characteristics) of a molecule in a computer readable format from which the models can learn. The selected descriptors used to train the models will therefore have a strong impact on the performance of the models themselves.

Starting from these molecular representations such as SMILES and SMARTS, molecular descriptors can be calculated. One of the most prominent cheminformatic toolkits for deriving the above-mentioned descriptors is RDKit.[109, 110] With RDKit molecules can be processed with a well written application programming interface (API) within a python program or script. Biopython[111] is a python package that can be used for database searches, for example, within the NCBI protein database. Another tool for molecule processing which can be used with a graphical user interface and also via the command line is molecular operating environment (MOE).[112] This tool can also be used for descriptor calculation or for performing more challenging tasks like simulations involving macromolecules (e.g. proteins).

The most important descriptors in the context of this work are derived from MOE and RDKit. Physicochemical descriptors like the molecular weight, logP values or the number of carbon atoms within a molecule were calculated with MOE and bit vector descriptors were calculated with RDKit. Physicochemical descriptors can have a wide range of values (often with orders of magnitude differing between each other) and therefore have to be standardized (to have a normal distribution) before machine learning algorithms are used on these descriptors. In this work a better performance was obtained with the bit vector descriptors that contain a vector of a given length where each value can be one or zero. For example, the 166 bits of a MACCS key encode different structural properties which can be the question if there are more than three oxygens within the molecule or if a sulfur bond is present. If one of these conditions is met, the bit encoding this feature is set to one, otherwise it remains at zero.[113] The descriptors that worked best in this work were Morgan or extended connectivity fingerprint-like fingerprints.[114, 115] These circular fingerprints encode different sized substructures or fragments (depending on the selected radius) that reflect relevant regions of molecules. For example, a Morgen fingerprint with a radius of two is called a Morgan2 fingerprint and encodes substructures with a diameter of four atoms. Normally these bit vectors are large and need a lot of memory when used in a program. The solution to this problem is hashing, during which the bits are “folded” into a fixed length (e.g. into 1024 bits). However, sometimes different bits are hashed to the same position, causing bit collision, which means that two features are hashed in the same bit and are not distinguishable anymore.

3.5.3 Performances metrics

Classification and regression models can be evaluated with different types of performance metrics. Depending on the classification task (i.e. binary classification or multi-class classification) and the ratio of the different classes some metrics are better suited than others. Regression models are usually evaluated with regard to the distance of the predicted values to the true value. In binary classification models (for multi-class classifications different metrics are necessary and are outside the scope of this work) the prediction can only take two values (i.e. true or false; 1 or 0; active or inactive) and thus distance measurements are not well suited for their evaluation. In the following the most important binary classification performance metrics are discussed.

As in binary classification models the answer can only be true or false, only four types of outcomes can result from a binary classification model: false negatives (FN), false positives (FP), true negatives (TN) and true positives (TP). FN are instances (in our case normally compounds) that were predicted as active but are inactive compounds, whereas FP were predicted as inactive but are actually active compounds. TN and TP are the correctly predicted instances as inactive and active, respectively. A matrix which displays all the four categories is called a confusion matrix. The measurements, described next, are normally calculated from these (or a subset of these) four categories.

For the selection of the most suitable performance metric for a problem, one of the important parameters to consider is whether the underlying data are balanced or imbalanced (i.e. containing different numbers of instances in each class). Some metrics are not suited for imbalanced data, since prediction errors in the minority class are disregarded. Further, it is important which classes are the most important ones in the prediction. For example, in some tasks (like toxicity prediction) false positives could be regarded as a minor problem, while false negatives need to be avoided. Depending on these factors often different measurements are used.

In this work mainly the MCC (Eq.(3.2)) was used to evaluate model performance.[116] MCC is a balanced performance metric (i.e. suitable for imbalanced data sets) that takes all the four aforementioned classes (FN, FP, TN and TP) into account. It ranges from -1 to 1 whereas 1 means perfect prediction, 0 random prediction and -1 consistent classification of the opposite class.

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}} \tag{3.2}$$

Accuracy (Eq.(3.3)) is the fraction of correctly predicted classes among all predicted instances. This metric is only suitable for balanced data sets, since a model predicting all instances with the class label of the majority class would still get a high accuracy (e.g. accuracy of 0.90 if 90% of the data belongs to the majority class). However, such a model would classify all instances in the minority class wrongly.

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \tag{3.3}$$

Sensitivity or Recall (Eq.(3.4)) is the fraction of active compounds correctly identified (among all active predictions). Whereas specificity (Eq.(3.5)) is the fraction of inactive compounds correctly identified (among all inactive predictions).

$$Sensitivity = \frac{TP}{TP + FN} \tag{3.4}$$

$$Specificity = \frac{TN}{TN + FP} \tag{3.5}$$

The receiver operating characteristic is a performance measurement that shows the ability of a classification model to rank the compounds according to their classes. The AUC values range from 0 to 1 where 1 is a perfect classification and 0 a consistent classification of the opposite class. A random classification is denoted with a AUC value of 0.5. The ROC curve has more information: for example a steep early slope shows that under the instances ranked with high probabilities there are only few false positive samples.

3.6 Web servers as a tool for easy accessibility

The models developed within the scope of this PhD thesis are available via the free public web service called New E-Resource for Drug Discovery (NERDD). NERDD is based on django[117] and also uses the javascript molecule editor toolbox[118] for graphical input of molecule structures. Results are visualised by similarity maps.[119] One challenge of setting up the web server was to enable complex calculations on an high performance cluster since otherwise parts of the calculations (e.g. AD calculation) are too time consuming.

4. Aims

Some compounds trigger positive signals in biological assays more often than others. These compounds are therefore called "frequent hitters". A lot of frequent hitter compounds, but by far not all, may trigger false positive activity outcomes by interacting, in an undesired manner, with the assay screening technology or protein target. Publication of such compounds as bioactive substances often triggers follow-up studies without the possibility of success, blocking valuable resources in research. Experimental detection of bad actors is time-consuming and computer-based models for the prediction of bad actors are still in their infancy. Hence, the development of in silico methods for quick and easy prediction of frequent hitters based on the latest developed machine learning algorithms are needed.

The main objective of this PhD study is the development of machine learning models for the prediction of frequent hitters in biochemical and biological assays. The models are derived from a large data set extracted from the PubChem Bioassay database. Therefore three generations of machine learning models were developed called Hit Dexter, Hit Dexter 2.0 and Hit Dexter 3. Whereas the first part of this Ph.D. study, including the first generation of Hit Dexter models, was mainly focused on the data preparation, which includes taking only statistically relevant data into account and the preparation of a well-curated data set as well as a proof of concept for the machine learning models on such data sets, Hit Dexter 2.0 includes and distinguishes two different assay domains (i.e. primary screen assays (PSA) and confirmatory dose-response assays (CDRA)) and is much more focused on the evaluation of the developed machine learning models. The validation was performed on several data sets including drugs and compounds consistently measured as inactive across a broad protein space, as well as several others. Finally, Hit Dexter 3 progresses further and distinguishes target- and cell-based assay screenings. As the different assay screening types differ in the way compounds interact with the protein target, for example in target-based assays with a purified protein and in cell-based with protein targets within a cell, it is important to not mix them up.

Another important aim of this PhD thesis is the development of New E-Resource for Drug Discovery (NERDD). NERDD is a web server, which makes in silico tools easily accessible to the scientific community and also includes the Hit Dexter models. NERDD was developed as an easily extensible and maintainable

web server. Besides the Hit Dexter models, six other tools, namely CYPstrate, CYPlebrity, FAME3, GLORY, GLORYx, NP-Scout and Skin Doctor CP, are accessible via NERDD.

5. Results (cumulative part of this dissertation)

In this section the results of this cumulative dissertation are presented in the form of the publications that are derived from this work, along with a short summary of each publication.

5.1 Machine learning models for the prediction of frequent hitters based on target-based assay data sets

Frequent hitters (compounds with higher-than-expected hit rates in biological assays) need to be treated with extra caution during high-throughput screening (HTS) campaigns as they are likely bad actors. In an attempt to develop machine learning models for the prediction of frequent hitters in target-based assays a key requirement is a large data set for model building. Therefore a computational approach for the automated extraction and compilation of assay data from the PubChem Bioassay database was developed. Based on the compounds hit rates (i.e. fraction of times a compound was tested active and times a compound was tested), compounds are assigned to three groups: non-promiscuous compounds (hit rates below the average), promiscuous compounds (hit rates above the average plus one standard deviation) and highly promiscuous compounds (hit rates above the average plus three standard deviations). For the sake of robustness, only compounds tested against at least 50 different protein targets were considered in this work. Based on these large data sets, machine learning models were derived for discriminating non-promiscuous from promiscuous compounds, as well as non-promiscuous compounds from highly promiscuous compounds reaching Matthews correlation coefficients (MCCs) and area under the receiver operating characteristic curve (AUC) values of up to 0.67 and 0.96, respectively. The best performing classification models (based on extra tree classifiers and Morgan2 fingerprints) are distributed as “Hit Dexter” and available via a free web server, called New E-Resource for Drug Discovery (NERDD; for detail see below).

[D2] Hit Dexter: A Machine-Learning Model for the Prediction of Frequent Hitters
Conrad Stork, Johannes Wagner, Nils-Ole Friedrich, Christina de Bruyn Kops, Martin Šícho and Johannes Kirchmair
ChemMedChem, 2018

Available at `https://doi.org/10.1002/cmdc.201700673`.

Contribution:
C. Stork, J. Wagner and J. Kirchmair conceptualized the research. C. Stork compiled the data sets and developed the machine learning approach based on the findings from exploratory studies conducted by J. Wagner, N.-O. Friedrich, C. de Bruyn Kops and M. Šícho. C. Stork validated the models on holdout data and developed a web server to make the machine learning models available to the public. C. Stork wrote the manuscript, with contributions of J. Wagner, N.-O. Friedrich, C. de Bruyn Kops, M. Šícho and J. Kirchmair. J. Kirchmair supervised the work.

The following article was reprinted with permission from:
Stork, C.; Wagner, J.; Friedrich N.-O.; de Bruyn Kops, C.; Šícho, M. and Kirchmair, J. Hit Dexter: A Machine-Learning Model for the Prediction of Frequent Hitters, *ChemMedChem* **2018**, *13*, 564–571.
Copyright 2018 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

The supplementary information for this work can be found in Section B.

ChemPubSoc Europe

DOI: 10.1002/cmdc.201700673

CHEMMEDCHEM Full Papers

VIP *Very Important Paper*

SPECIAL ISSUE

Hit Dexter: A Machine-Learning Model for the Prediction of Frequent Hitters

Conrad Stork,[a] Johannes Wagner,[a] Nils-Ole Friedrich,[a] Christina de Bruyn Kops,[a] Martin Šícho,[a, b] and Johannes Kirchmair*[a]

False-positive assay readouts caused by badly behaving compounds—frequent hitters, pan-assay interference compounds (PAINS), aggregators, and others—continue to pose a major challenge to experimental screening. There are only a few in silico methods that allow the prediction of such problematic compounds. We report the development of Hit Dexter, two extremely randomized trees classifiers for the prediction of compounds likely to trigger positive assay readouts either by true promiscuity or by assay interference. The models were trained on a well-prepared dataset extracted from the PubChem Bioassay database, consisting of approximately 311 000 compounds tested for activity on at least 50 proteins. Hit Dexter reached MCC and AUC values of up to 0.67 and 0.96 on an independent test set, respectively. The models are expected to be of high value, in particular to medicinal chemists and biochemists who can use Hit Dexter to identify compounds for which extra caution should be exercised with positive assay readouts. Hit Dexter is available as a free web service at http://hitdexter.zbh.uni-hamburg.de.

Introduction

Biochemical assays are of considerable importance for early drug discovery, and modern high-throughput screening technologies allow the testing of over one hundred thousand compounds within one day.[1,2] However, high rates of false-positive readouts caused by various types of assay interference remain a major issue. A substantial number of false hits continue to appear as valid active compounds in the peer-reviewed literature.[3] As a consequence, efforts to characterize badly behaving compounds[4,5] (frequent hitters, pan-assay interference compounds, aggregators and others) and develop good practice guidelines on how to identify assay artifacts and reject such hits[6,7] have recently been gaining traction.

False-positive results can be related to the chemical reactivity of a compound.[8] In particular, electrophiles can bind covalently (and non-discriminately) to various proteins, thereby changing the function of the bio-macromolecule that is measured by the assay.[9] A wide range of in silico approaches for the identification of reactive compounds are available. Reactive compounds can be identified using models based on sets of rules, quantum chemical methods, and other linear and nonlinear modeling techniques.[10]

Besides chemical reactivity, false-positive readouts in biochemical assays may be related to a variety of other effects and processes, such as redox cycling, interference with assay spectroscopy, membrane disruption, decomposition in buffers and metal complexation.[3] Baell et al.[11] have devised a set of 480 substructures from high-throughput screening data that encode the molecular substructures of pan-assay interference compounds (PAINS). These substructures can be encoded as SMARTS patterns to use as a filter for flagging compounds that are likely PAINS. However, their applicability domain is narrow and they also match (potentially) benign moieties.[11,12] A recent study showed that the patterns match a substantial number of compounds that do not show any assay activity (i.e., "Dark Chemical Matter").[13,14] They should therefore be used as indicators rather than hard filters.

Colloidal aggregators are a further and possibly the most abundant type of compounds that may cause false-positive signals in biochemical assays.[15] These compounds are related to the formation of micelles at specific concentrations and generally not covered by the SMARTS patterns discussed above. An in silico approach for flagging likely colloidal aggregators based on molecular similarity with over 12 600 known aggregators (taking calculated log*P* values into account) is available.[16]

Most of the available computational approaches are limited to the identification of a specific type of badly behaving compounds.[6] An exception is Badapple,[17] which assigns a promiscuity score to compounds based on their molecular scaffolds. Badapple is derived from more than 430 000 compounds mea-

[a] C. Stork, J. Wagner, N.-O. Friedrich, C. de Bruyn Kops, M. Šícho, Prof. Dr. J. Kirchmair
Center for Bioinformatics, Universität Hamburg, Bundesstraße 43, 20146 Hamburg (Germany)
E-mail: kirchmair@zbh.uni-hamburg.de

[b] M. Šícho
National Infrastructure for Chemical Biology, Laboratory of Informatics and Chemistry, Faculty of Chemical Technology, University of Chemistry and Technology Prague, 166 28 Prague 6 (Czech Republic)

Supporting information and the ORCID identification number(s) for the author(s) of this article can be found under: https://doi.org/10.1002/cmdc.201700673.

SPECIAL ISSUE This article is part of a Special Issue on Cheminformatics in Drug Discovery. To view the complete issue, visit: http://onlinelibrary.wiley.com/doi/10.1002/cmdc.v13.6/issuetoc.

© 2018 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

sured in 822 different assays. While this model can in principle provide valuable indications of compound promiscuity, the reduction of molecular structures to scaffolds limits its capacity to account for subtle differences in compound structure.

In this work we explored machine learning approaches to develop classifiers for the prediction of frequent hitters (also referred to as promiscuous molecules) based on a large, curated dataset extracted from the PubChem Bioassay database.[18,19] The best performing models resulting from this work are available as a web service at http://hitdexter.zbh.uni-hamburg.de.

Results and Discussion

Compilation of datasets for model development

Bioactivity data on 468 260 small molecules measured in 2266 confirmatory dose–response assays were retrieved from the PubChem Bioassay database. The dataset was subjected to a multi-step data preparation process (Figure 1) resulting in 427 657 unique compounds with activity data on a total of 653 unique proteins.

From this dataset, two subsets were extracted for model development, consisting of 391 552 and 311 491 compounds that have been tested for activity on at least 20 and 50 different proteins, respectively (Figure 2). These two cutoff values were found to produce datasets covering a broad range of biological activities and a large chemical space. The latter was analyzed by principal component analysis (PCA) as reported in Figure 3. We refer to these datasets as the PC20 and the PC50 datasets, where PC is used as an abbreviation for "protein count".

The diversity of the PC20 and PC50 datasets was tested with a clustering approach. For each of the two datasets, 20 subsets were compiled, each consisting of 50 000 randomly selected compounds. These subsets were clustered with the Butina (unsupervised non-hierarchical) clustering algorithm[20] based on

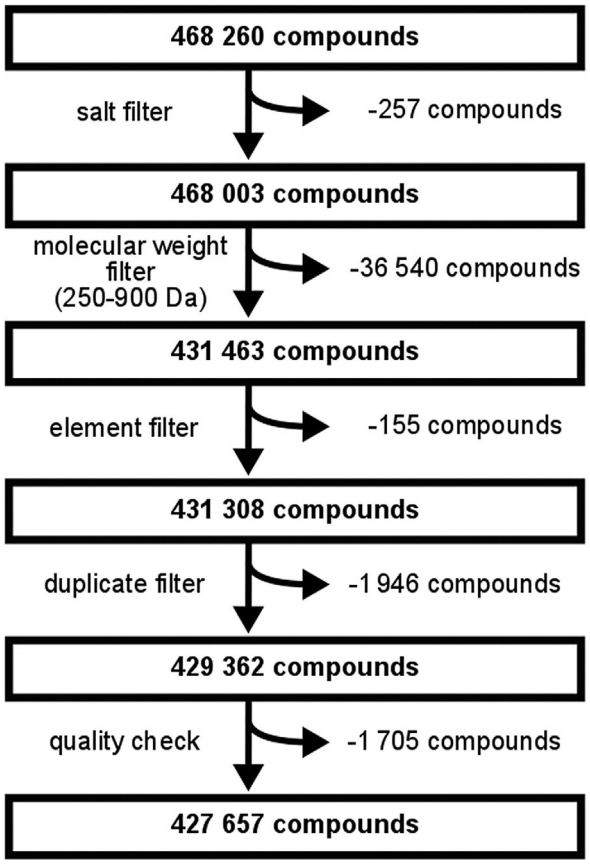


Figure 1. Overview of the data preparation pipeline. The numbers of ligands that survived each filtering step are reported in boxes, and the number of removed compounds are shown on the right. After the treatment of multi-component compounds (salts), molecules with a molecular weight below 250 or above 900 Da and molecules consisting of elements other than those commonly observed in drug-like molecules were removed. A duplicate filter was applied to the remaining compounds, followed by quality checks to discard, e.g., contradicting bioassay data. See Experimental Section for detail.

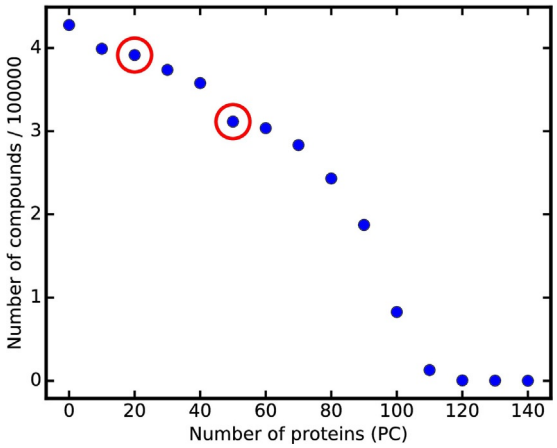


Figure 2. Number of compounds in the curated dataset that have bioactivity data reported for at least the given number of proteins (PC). The numbers of compounds relevant for the PC20 and PC50 dataset are indicated by red circles.

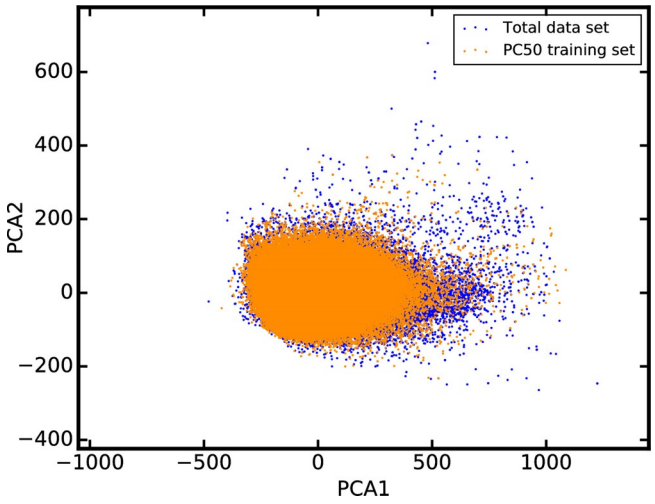


Figure 3. The scatter plot of the second against the first component based on 45 molecular descriptors (Table S1) shows that the PC50 training set covers the total (i.e., processed PubChem Bioassay) dataset well.

© 2018 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

Morgan2 fingerprints[21,22] and a Tanimoto similarity threshold of 0.75. Among all subsets of both datasets, the largest clusters contained only 32 (PC20) and 15 (PC50) molecules, respectively. The lowest number of clusters for all subsets was 44649 (PC20) and 44514 (PC50), respectively. Based on these results we deemed the datasets sufficiently diverse for modeling.

The hit rates of the individual compounds in biochemical assays were quantified with the active-to-tested ratio (ATR), which is calculated as in Equation (1):

$$ATR = \frac{A}{T} \quad (1)$$

where A is the number of proteins for which a compound was measured as active and T is the total number of proteins on which a compound has been tested. The ATR is low for most compounds in the dataset, but a significant number of frequent hitters are also present (Figure 4).

Compounds were assigned one of three different promiscuity labels according to the ATR thresholds reported in Table 1: a "non-promiscuous" (NP) label for any compounds with an $ATR < ATR_{mean}$, a "promiscuous" (P) label for any compounds with $ATR > ATR_{mean} + 1\sigma$, and a "highly promiscuous" (HP) label for any compounds with $ATR > ATR_{mean} + 3\sigma$. Note that, according to this definition, highly promiscuous compounds are a subset of promiscuous compounds.

Prior to any modeling experiments, the datasets were each split into a training set and an independent test set using a 9:1 ratio (Table 1). This resulted in a training set with up to 246331 instances for each of the promiscuity classes. The test sets consisted of up to 27450 instances for each promiscuity class.

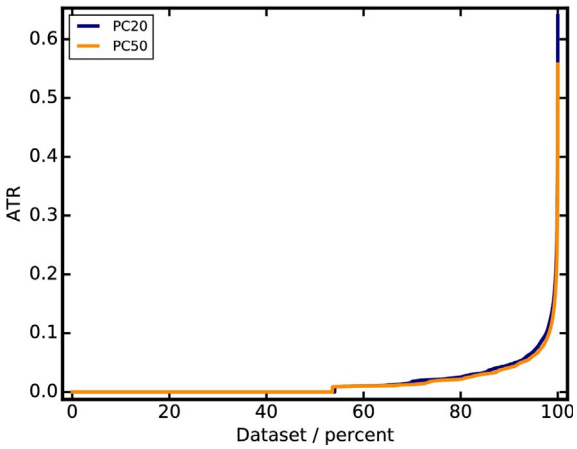


Figure 4. ATR distributions among compounds of the PC20 and PC50 datasets.

Analysis of the physicochemical properties of non-promiscuous and promiscuous molecules

The ability of a compound to trigger a positive signal in biochemical assays can be related to its physicochemical properties or to the presence of specific chemical patterns. We computed characteristic physicochemical properties to probe whether a link to compound promiscuity can be established.

As shown in Figure 5, the molecular weight distribution of NP, P and HP compounds is similar. However, P (and also HP) compounds tend to be more lipophilic than NP compounds (Table 2). Their calculated log$P$ is on average one log unit higher than that of NP compounds. This is consistent with the general observation that nonspecific compound binding is correlated with hydrophobicity. In addition, a higher proportion of

Table 1. Composition of the datasets used for model training and validation.

Assigned promiscuity class	Dataset	Number of compounds in PC20	Number of compounds in PC50	Threshold definition[a]	Threshold value PC20[b]	Threshold value PC50[b]
Non-promiscuous (NP)	Total:	273781	226710	$ATR < ATR_{mean}$	0.017	0.015
	Training set:	246331	203992			
	Test set 1:[c]	27450	22718			
	Test set 2:[d]	16872	14611			
	Test set 3:[e]	6569	5863			
Promiscuous (P)	Total:	35438	29112	$ATR > ATR_{mean} + 1\sigma$	0.049	0.043
	Training set:	31915	26201			
	Test set 1:[c]	3523	2911			
	Test set 2:[d]	2303	2060			
	Test set 3:[e]	1090	965			
Highly promiscuous (HP): a subset of compounds labeled P	Total:	7371	5527	$ATR > ATR_{mean} + 3\sigma$	0.112	0.100
	Training set:	6653	4970			
	Test set 1:[c]	718	557			
	Test set 2:[d]	496	409			
	Test set 3:[e]	283	203			

[a] Compounds with ATRs between ATR_{mean} and the given standard deviation were not assigned a promiscuity label and were effectively removed from the datasets. [b] ATR threshold values calculated for the individual datasets according to the ATR threshold definition. [c] Independent test set obtained by random split of the curated dataset prior to model development. [d] Subset of the independent test set consisting only of molecules showing a Morgan2 fingerprint-based maximum Tanimoto coefficient of 0.8 to any compounds in the training data. [e] Subset of the independent test set consisting only of molecules showing a Morgan2 fingerprint-based maximum Tanimoto coefficient of 0.7 to any compounds in the training data.

© 2018 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

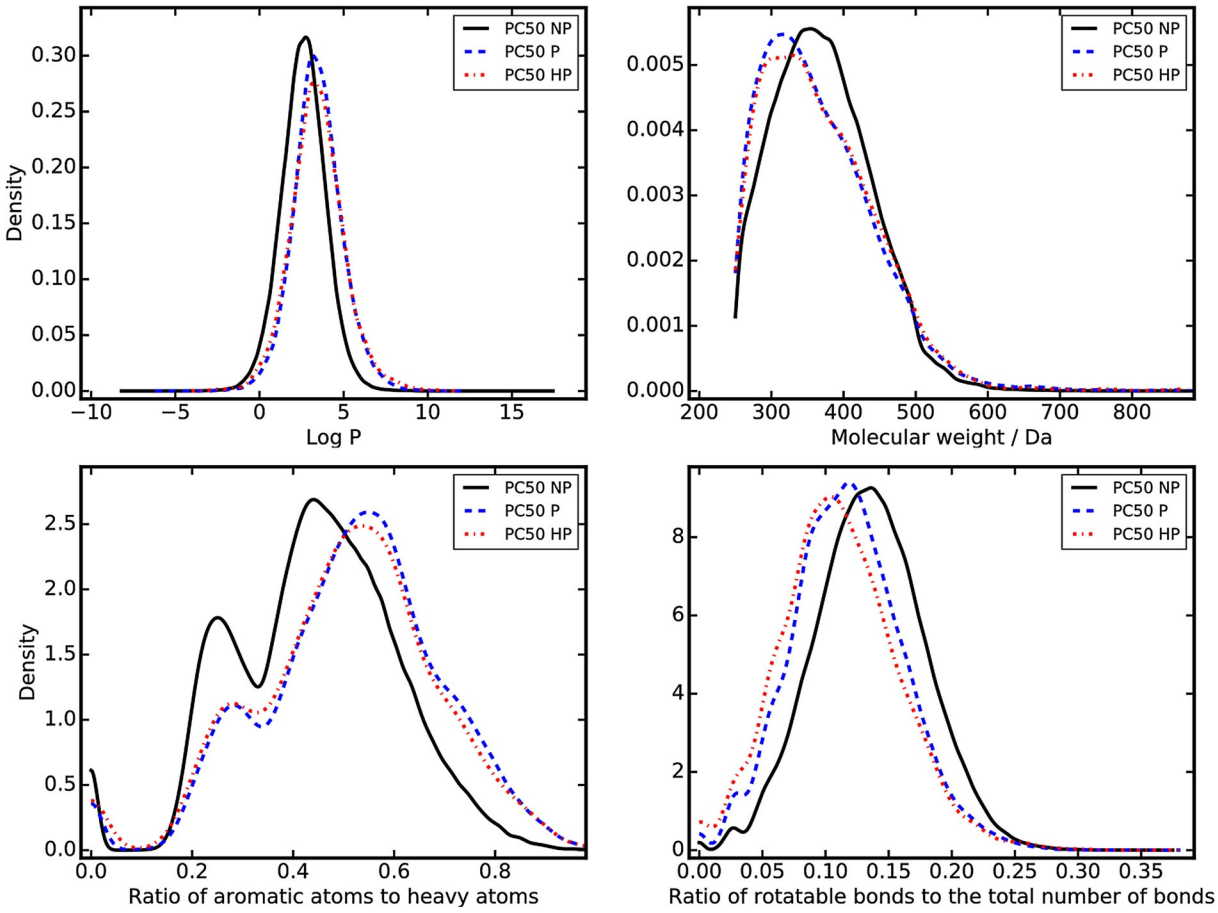


Figure 5. Density plots of the log*P*, molecular weight, ratio of aromatic atoms to heavy atoms and ratio of rotatable bonds among all bonds for the PC50 dataset. HP compounds in red; P compounds in blue; NP compounds in black.

Table 2. Physicochemical properties and their correlations with the ATR for the PC50 dataset.

Property	NP[a,d]	P[b,d]	HP[c,d]	Correlation with ATR
Ratio of aromatic atoms to heavy atom	0.44 ± 0.16	0.52 ± 0.18	0.50 ± 0.18	0.13
Ratio of rotatable bonds among all bonds	0.14 ± 0.04	0.12 ± 0.04	0.11 ± 0.05	−0.13
log*P*	2.62 ± 1.33	3.46 ± 1.47	3.46 ± 1.60	0.20
Molecular weight	371.19 ± 7.92	364.45 ± 77.49	366.74 ± 78.19	−0.01

[a] Non-promiscuous compounds. [b] Promiscuous compounds. [c] Highly promiscuous compounds. [d] Data are the average ± standard deviation.

aromatic atoms as well as a lower proportion of rotatable bonds were found among P (and also HP) compounds. Both of these properties are related to planarity and flexibility, which themselves are known to be associated with a compound's ability to form colloidal aggregates. Whether these differences in physicochemical properties are sufficient to yield accurate classifiers will be explored in the subsequent sections.

Model development

Two different types of binary classification models were developed: one to discriminate promiscuous from non-promiscuous compounds (P-NP) and one to discriminate highly promiscuous from non-promiscuous compounds (HP-NP). In combination with the two datasets, PC20 and PC50, this gave rise to a total of four different models.

Identification of the most suitable machine learning algorithm and descriptor sets

In initial experiments we explored the performance of random forest classifiers[23] (RFCs) and extremely randomized tree classifiers[24] (ETCs) trained on 1) all 206 2D physicochemical property descriptors implemented in MOE,[25] 2) MACCS key fingerprints (166 bits), and 3) Morgan2 fingerprints (1024 bits), both implemented in RDKit.[26]

All models were trained with scikit-learn[27] and evaluated by 10-fold cross-validation. Default values were used for the hy-

© 2018 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

perparameters, except for the number of estimators, which was increased to 50, and the class weights, which were set to "balanced".

The Matthews correlation coefficient (MCC) was used as the primary measure of model performance. The MCC is a balanced measure of prediction quality which not only takes true positives (TP) and false positives (FP) into account, but also true negatives (TN) and false negatives (FN). It is calculated according to Equation (2).

$$MCC = \frac{TP \bullet TN - FP \bullet FN}{\sqrt{(TP + FP) \bullet (TP + FN) \bullet (TN + FP) \bullet (TN + FN)}} \quad (2)$$

The area under the receiver operating characteristic curve (AUC) served as an additional measure of how well the model was able to rank the compounds for promiscuity according to the probabilities given by the machine learning algorithms.

Models derived from the combination of the extremely randomized tree algorithm with Morgan2 fingerprints consistently obtained the best performance for all combinations of promiscuity thresholds and datasets. The models' MCC and AUC values ranged up to 0.61 and 0.94, respectively (Table 3). The random forest classifier in combination with Morgan2 fingerprints obtained comparable results (MCC of up to 0.56 and AUC of up to 0.94). Models based on molecular fingerprints clearly outperformed those based on physicochemical property descriptors. This result was expected because assay interference is often linked to specific molecular substructures, and Morgan2 fingerprints are the most suitable (among those tested) to capture these substructures. Differences in performance with respect to promiscuity thresholds and datasets were small. As a result of these experiments, the combination of the ETC with Morgan2 fingerprints was identified as the most suitable starting point for further optimization of the models.

Optimization of model hyperparameters

The number of estimators and the maximum fraction of features considered per split were optimized using a grid search with 10-fold cross-validation (Table 4). Model performance was evaluated based on the average MCC obtained over all folds.

Table 4. Hyperparameters optimized by grid search.

Parameter	Tested values[a]
Number of estimators (*n_estimators*)[b]	10,[c] 50, **100**, 150, 200, 250, 300, 400, 500, 600
Maximum fraction of features considered per split (*max_features*)[b]	"sqrt",[c] **0.2**, 0.4, 0.6, 0.8, None[d]

[a] Bold values were used for final model development. [b] Parameter name in the scikit-learn implementation. [c] Default value. [d] All features are used.

For all combinations of datasets and promiscuity thresholds, minor performance improvements corresponding to increasing numbers of estimators were observed (Tables S2–5). For example, the MCC values of models with 100 estimators were up to 0.04 higher than those of models with only 10 estimators. Marginal, if any, improvements in performance beyond 100 estimators did not justify the additional computational cost. The effect of the maximum fraction of features considered per split (*max_features*) on model performance was small (up to around 0.01 in MCC and AUC). The best models were achieved with *max_features* set to 0.2 for all combinations of datasets and promiscuity thresholds.

Overall, the best-performing classifier that emerged from the grid search was able to distinguish HP from NP compounds with an MCC and AUC of 0.62 and 0.95 for the PC20 dataset, and 0.61 and 0.95 for the PC50 dataset, respectively (*n_estimators* = 100, *max_features* = 0.2, for both datasets; Figure 6). The

Table 3. Performance of models derived from different combinations of machine learning algorithms and descriptor sets during 10-fold cross-validation.[a]

	MOE physicochemical property descriptors				MACCS fingerprints				Morgan2 fingerprints			
Algorithm:	ETC		RFC		ETC		RFC		ETC		RFC	
Metric:	MCC	AUC	MCC	AUC	MCC	AUC	MCC	AUC	MCC	AUC	MCC	AUC
P-NP with PC20	$0.47 \pm 0.3 \times 10^{-4}$	0.89	$0.47 \pm 0.4 \times 10^{-4}$	0.89	$0.53 \pm 0.3 \times 10^{-4}$	0.87	$0.52 \pm 0.3 \times 10^{-4}$	0.89	$0.58 \pm 0.9 \times 10^{-4}$	0.91	$0.55 \pm 0.5 \times 10^{-4}$	0.91
HP-NP with PC20	$0.44 \pm 4.1 \times 10^{-4}$	0.93	$0.43 \pm 3.5 \times 10^{-4}$	0.92	$0.56 \pm 1.6 \times 10^{-4}$	0.92	$0.53 \pm 1.1 \times 10^{-4}$	0.93	$0.61 \pm 1.5 \times 10^{-4}$	0.94	$0.56 \pm 1.9 \times 10^{-4}$	0.94
P-NP with PC50	$0.46 \pm 1.2 \times 10^{-4}$	0.89	$0.46 \pm 1.1 \times 10^{-4}$	0.89	$0.52 \pm 1.0 \times 10^{-4}$	0.87	$0.51 \pm 1.1 \times 10^{-4}$	0.89	$0.57 \pm 0.8 \times 10^{-4}$	0.91	$0.54 \pm 1.2 \times 10^{-4}$	0.91
HP-NP with PC50	$0.41 \pm 4.0 \times 10^{-4}$	0.92	$0.40 \pm 1.7 \times 10^{-4}$	0.92	$0.56 \pm 1.9 \times 10^{-4}$	0.92	$0.52 \pm 1.5 \times 10^{-4}$	0.92	$0.61 \pm 2.4 \times 10^{-4}$	0.94	$0.55 \pm 2.3 \times 10^{-4}$	0.94

[a] ETC, extra tree classifier; RFC, random forest classifier; P-NP, discrimination of promiscuous from non-promiscuous compounds; HP-NP, discrimination of highly promiscuous from non-promiscuous compounds. MCC (with standard deviations) and AUC values averaged over all folds of the cross-validation.

© 2018 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

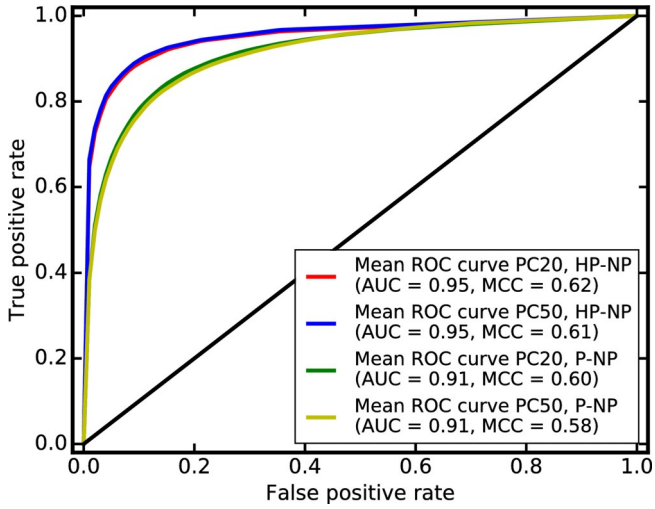


Figure 6. Mean ROC curves obtained during 10-fold CV for the best-performing, optimized models (i.e., the ETC derived with Morgan2 fingerprints, *n_estimators* = 100, *max_features* = 0.2).

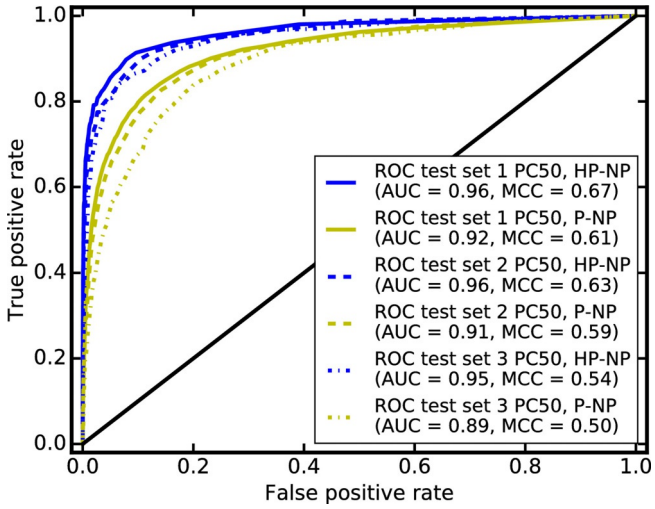


Figure 7. ROC curves obtained with the final models for the three test sets extracted from the PC50 dataset prior to model development.

P-NP classifiers performed slightly worse than the HP-NP classifiers, obtaining MCC and AUC values of 0.60 and 0.91 for the PC20 dataset and 0.58 and 0.91 for the PC50 dataset, respectively (Figure 6). The observed differences in model performance were expected, as the ATR margin between the HP and NP classes (3σ) is broader than the margin between the P and NP classes (1σ). Because the performance of models derived from the PC20 and PC50 dataset was comparable, further discussion will focus on models derived from the latter.

Model evaluation on independent test sets

The final models were trained with the above-mentioned, optimized hyperparameters (*n_estimators* 100; *max_features* 0.2) on the complete PC50 training set balanced with the synthetic minority over-sampling technique (SMOTE) algorithm.[28] Performance data on the models derived from the PC20 dataset are provided in Figure S1.

The HP-NP model was able to predict compound promiscuity for the independent test set 1 with MCC and AUC values of 0.67 and 0.96, respectively (Figure 7). Consistent with the trends observed in the cross-validation, slightly lower values were obtained with the P-NP model (MCC 0.61; AUC 0.92). The MCC and AUC values for the independent test sets were slightly better (up to 0.06 MCC and 0.01 AUC) than those for cross-validation on the training set. The increase in performance is likely a result of the over-sampling approach and the fact that more data were available and used for training than during the cross-validation approach.

To explore the robustness of the models, two subsets of the independent test set were generated consisting only of molecules showing a Morgan2 fingerprint-based maximum Tanimoto coefficient of 0.8 (test set 2) and 0.7 (test set 3) to any compounds in the training data (Table 1). As expected, the MCC and AUC values obtained for the test sets 2 and 3 were lower than for test set 1 (Figure 7). The HP-NP classifier obtained an MCC and AUC of 0.63 and 0.96, respectively, for test set 2. Both values were approximately 0.05 lower for the P-NP classifier on the same dataset. For test set 3, MCC and AUC values of 0.54 and 0.95, respectively, were obtained for the HP-NP classifier. The respective values for the P-NP classifier were again around 0.05 lower than those of the HP-NP classifier.

In addition, the HP-NP and P-NP classifiers were also tested on the Dark Chemical Matter (DCM) dataset,[13] which consists exclusively of compounds that have been tested in a minimum of 100 different assays and have not shown any activity. Prior to testing, any compounds present in the PC50 training set (341 compounds in total) and any compounds outside the applicability domain of the models (13672 compounds that did not pass the filters applied for molecular weight and element types; see the Experimental Section for details) were removed from the DCM dataset. This resulted in a test set of 125339 compounds, of which 99.9% and 98.4% were correctly classified as not promiscuous by the HP-NP and the P-NP models, respectively (Figure 8).

Hit Dexter web service

A web service called "Hit Dexter" is accessible free of charge via http://hitdexter.zbh.uni-hamburg.de. The web service offers an easy and quick way to make predictions for individual molecules and sets of molecules with the best-performing classifier (i.e., the ETC derived from the SMOTE-balanced PC50 dataset, Morgan2 fingerprints, *n_estimators* = 100, *max_features* = 0.2). Users upload molecular structures as SMILES or a list of SMILES and initiate the calculations. After a few seconds the user is presented a tabular overview of results, including the molecule name and the calculated probabilities of a compound to be a frequent hitter (Figure 9). The results and a log file can be downloaded for further use. There is an option to also retrieve the five nearest neighbors of query molecules present in the training set, which will give users a better estimate of how reli-

© 2018 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

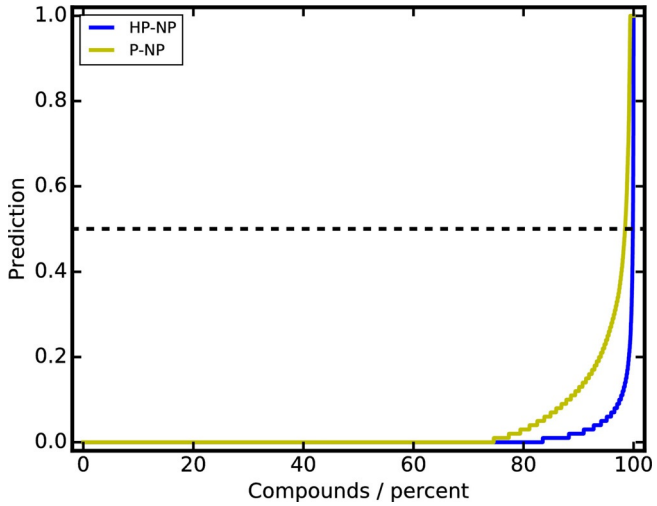


Figure 8. Likelihood of promiscuity predicted for over 125 000 compounds of a preprocessed subset of the DCM dataset. The figure shows that the HP-NP and P-NP models correctly classify the vast majority of compounds as not promiscuous. They obtained an overall accuracy of 99.9% and 98.4%, respectively.

able the predictions are for their particular compounds of interest.

Conclusions

Assay interference continues to present a significant challenge in early drug discovery. Current computational approaches attempting to identify frequent hitters, including reactive compounds, PAINS and aggregators, are clearly limited in their applicability. In this work we report on the development of Hit Dexter, a machine learning approach for the identification of compounds likely to trigger positive assay signals. The classification models included in Hit Dexter reached MCC and AUC values of up to 0.67 and 0.96 on an independent test set, respectively. A free, public web service includes functionality to retrieve the five nearest neighbors present in the training data for each query molecule, in order to support users in estimating the reliability of the predictions for their particular compounds of interest. Importantly, besides reactive compounds, PAINS and aggregators, Hit Dexter also identifies compounds with particular pharmacophores that allow them to bind to multiple proteins.

We believe that Hit Dexter will help scientists to flag compounds that have an increased likelihood of triggering positive signals in biochemical assays. Compounds flagged by Hit Dexter should not be regarded as being of lower value for drug discovery but rather as having more uncertainty regarding their activity. In fact, frequent hitters may even be desirable, for example, in the context of polypharmacology and drug repurposing,[29] provided they are true promiscuous binders.

The purpose of Hit Dexter is to raise awareness and motivate further investigations of the flagged compounds in orthogonal assays. In particular, we also hope that these models will contribute to the effort to decrease the amount of false hits in the scientific literature.

Experimental Section

Activity data for chemical substances (*substance type* = "chemical") measured with 2266 confirmatory dose–response assays (*screening stage* = "confirmatory, dose–response") for single protein targets (*target* = "single" and *target type* = "Protein Targets") were downloaded from the PubChem Bioassay database.[18,19,30] The SMILES notations for all 468 260 compounds in this dataset were retrieved via the PubChem Identifier Exchange Service.[31] Compounds consisting of multiple components (salts) were split and the components sorted by decreasing number of heavy atoms ("size"). If the second-largest component was significantly smaller than the largest one (i.e., number of heavy atoms less than 70% of the largest component), the largest component was defined as the active component and all others were discarded. If this was not the case, the compounds were removed from the dataset (as no clear as-

Figure 9. Screenshot of the Hit Dexter result page.

© 2018 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

signment of the main components could be made) unless the two largest components were identical, in which case one of these was preserved and all others discarded (salt filter in Figure 1). Compounds with the same unique SMILES were treated as one compound (duplicate filter in Figure 1).

Compounds with a molecular weight below 250 or above 900 Da (molecular weight filter in Figure 1), as well as compounds consisting of any element other than H, B, C, N, O, F, Si, P, S, Cl, Se, Br and I (element filter in Figure 1) were removed from the dataset. The InChIs were retrieved for all remaining compounds via the PubChem Identifier Exchange Service.[31] The 240 compounds for which the InChI could neither be retrieved via the PubChem Identifier Exchange Service nor the PubChem PUG REST interface[32] were also discarded.

All downloaded bioactivity records in the PubChem Bioassay database have one of the following four activity values (*activity outcomes*): "Includes Probe", "Active", "Inactive" or "Unspecified/Inconclusive". Any assays not having at least one "Active" and one "Inactive" record were removed from the dataset. Any compounds (i.e., all instances having the same InChI after application of the salt filter) with contradicting activity values for one and the same assay were discarded (quality check in Figure 1). Following this step, any compounds reported by at least one assay as active on a particular protein were labeled active on that protein. This procedure resulted in a total of 405 399 compounds with assigned bioactivities.

All PubChem Bioassays are linked to a "gene identifier" (GI), a unique identifier for genes in the NCBI Protein database.[33] This identifier was retrieved for the individual assays via the PubChem PUG REST[32] interface to link assays to proteins. A total of 712 unique GIs were retrieved. Using these GIs, the protein sequences were retrieved in FASTA file format from the NCBI Protein database. The protein sequences were checked for sequence identity with *cd-hit*[34] (structure equality = 100%), resulting in 653 unique proteins.

All calculations are performed on Linux workstations running openSUSE 42.2 and equipped with Intel i5 processors (3.2 GHz) and 16GB of main memory.

Acknowledgements

Rainer Fährrolfes, Florian Flachsenberg, and Gerd Embruch are thanked for technical support and discussions. This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation; grant KI 2085/1-1) and by the Ministry of Education of the Czech Republic (grants NPU I-LO1220 and LM2015063). M.S. was supported by the Erasmus+ Programme of the European Commission.

Conflict of interest

The authors declare no conflict of interest.

Keywords: cheminformatics · compound promiscuity · frequent hitters · PAINS · high-throughput screening

[1] P. Szymański, M. Markowicz, E. Mikiciuk-Olasik, *Int. J. Mol. Sci.* **2012**, *13*, 427 – 452.
[2] R. Macarron, M. N. Banks, D. Bojanic, D. J. Burns, D. A. Cirovic, T. Garyantes, D. V. S. Green, R. P. Hertzberg, W. P. Janzen, J. W. Paslay, U. Schopfer, G. S. Sittampalam, *Nat. Rev. Drug Discovery* **2011**, *10*, 188 – 195.
[3] J. Baell, M. A. Walters, *Nature* **2014**, *513*, 481 – 483.
[4] E. Gilberg, D. Stumpfe, J. Bajorath, *F1000Res.* **2017**, *6*, 1505.
[5] J. W. M. Nissink, S. Blackburn, *Future Med. Chem.* **2014**, *6*, 1113 – 1126.
[6] P. W. Kenny, *J. Chem. Inf. Model.* **2017**, *57*, 2640 – 2645.
[7] C. Aldrich, C. Bertozzi, G. I. Georg, L. Kiessling, C. Lindsley, D. Liotta, K. M. Merz, Jr., A. Schepartz, S. Wang, *ACS Med. Chem. Lett.* **2017**, *8*, 379 – 382.
[8] G. M. Rishton, *Drug Discovery Today* **1997**, *2*, 382 – 384.
[9] C. Jöst, C. Nitsche, T. Scholz, L. Roux, C. D. Klein, *J. Med. Chem.* **2014**, *57*, 7590 – 7599.
[10] M. P. Gleeson, S. Modi, A. Bender, R. L. M. Robinson, J. Kirchmair, M. Promkatkaew, S. Hannongbua, R. C. Glen, *Curr. Pharm. Des.* **2012**, *18*, 1266 – 1291.
[11] J. B. Baell, G. A. Holloway, *J. Med. Chem.* **2010**, *53*, 2719 – 2740.
[12] J. B. Baell, *Future Med. Chem.* **2010**, *2*, 1529 – 1546.
[13] A. M. Wassermann, E. Lounkine, D. Hoepfner, G. Le Goff, F. J. King, C. Studer, J. M. Peltier, M. L. Grippo, V. Prindle, J. Tao, A. Schuffenhauer, I. M. Wallace, S. Chen, P. Krastel, A. Cobos-Correa, C. N. Parker, J. W. Davies, M. Glick, *Nat. Chem. Biol.* **2015**, *11*, 958 – 966.
[14] S. J. Capuzzi, E. N. Muratov, A. Tropsha, *J. Chem. Inf. Model.* **2017**, *57*, 417 – 427.
[15] S. L. McGovern, E. Caselli, N. Grigorieff, B. K. Shoichet, *J. Med. Chem.* **2002**, *45*, 1712 – 1722.
[16] J. J. Irwin, D. Duan, H. Torosyan, A. K. Doak, K. T. Ziebart, T. Sterling, G. Tumanian, B. K. Shoichet, *J. Med. Chem.* **2015**, *58*, 7076 – 7087.
[17] J. J. Yang, O. Ursu, C. A. Lipinski, L. A. Sklar, T. I. Oprea, C. G. Bologa, *J. Cheminf.* **2016**, *8*, 29.
[18] S. Kim, P. A. Thiessen, E. E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B. A. Shoemaker, J. Wang, B. Yu, J. Zhang, S. H. Bryant, *Nucleic Acids Res.* **2016**, *44*, D1202 – D1213.
[19] Y. Wang, S. H. Bryant, T. Cheng, J. Wang, A. Gindulyte, B. A. Shoemaker, P. A. Thiessen, S. He, J. Zhang, *Nucleic Acids Res.* **2017**, *45*, D955 – D963.
[20] D. Butina, *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 747 – 750.
[21] H. L. Morgan, *J. Chem. Doc.* **1965**, *5*, 107 – 113.
[22] D. Rogers, M. Hahn, *J. Chem. Inf. Model.* **2010**, *50*, 742 – 754.
[23] L. Breiman, *Mach. Learn.* **2001**, *45*, 5 – 32.
[24] P. Geurts, D. Ernst, L. Wehenkel, *Mach. Learn.* **2006**, *63*, 3 – 42.
[25] *Molecular Operating Environment (MOE)*, version 2016.08; Chemical Computing Group ULC, 1010 Sherbooke St. West, Suite #910, Montreal, QC, H3A 2R7 (Canada), **2018**B.
[26] RDKit version 2016.09.4: Open-Source Cheminformatics Software: http://www.rdkit.org.
[27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, *J. Mach. Learn. Res.* **2011**, *12*, 2825 – 2830.
[28] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, *J. Artif. Intell. Res.* **2002**, *16*, 321 – 357.
[29] P. Schneider, M. Röthlisberger, D. Reker, G. Schneider, *Chem. Commun.* **2016**, *52*, 1135 – 1138.
[30] *PubChem Bioassay*, https://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi (accessed: August 29, **2017**).
[31] *PubChem Identifier Exchange Service*, https://pubchem.ncbi.nlm.nih.gov/idexchange/idexchange.cgi (accessed August 30, **2017**).
[32] S. Kim, P. A. Thiessen, E. E. Bolton, S. H. Bryant, *Nucleic Acids Res.* **2015**, *43*, W605 – W611.
[33] NCBI Resource Coordinators, *Nucleic Acids Res.* **2016**, *44*, D7 – D19.
[34] L. Fu, B. Niu, Z. Zhu, S. Wu, W. Li, *Bioinformatics* **2012**, *28*, 3150 – 3152.

Manuscript received: October 30, 2017
Revised manuscript received: December 21, 2017
Accepted manuscript online: December 29, 2017
Version of record online: February 1, 2018

© 2018 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

5.2 Refined machine learning models for the prediction of frequent hitters in primary screening assays and confirmatory dose-response assays

The frequent hitters observed with primary screen assays (PSAs) may differ from those observed with confirmatory dose-response assays (CDRAs). The differentiation of these two assay domains is therefore important towards the understanding of frequent hitters in target-based assays. Two large data sets were compiled from the PubChem Bioassay database, based on the automated extraction pipeline that was established as part of our previous work. A further improvement during data preparations in this work was the consideration of protein evolutionary relationships by clustering proteins according to their protein sequence. All compounds considered in this work have been tested on at least 50 distinct protein clusters. Dedicated models for the PSA data set and the CDRA data set were built to distinguish non-promiscuous and promiscuous, as well as non-promiscuous and highly promiscuous compounds. On holdout data, these models reached MCCs and AUC values of up to 0.64 and 0.96, respectively. The best performing models (extra tree classifiers using Morgan2 fingerprints) based on the PSA data set and CDRA data sets are distributed as the Hit Dexter 2.0 models. In addition, several other established rule-based approaches (e.g. the well known pan-assay interference compounds (PAINS) approach) and similarity-based approaches (e.g. AggregatorAdvisor) were implemented within the web server which makes it to a one-stop-shop for hit (de-)priorisation. During model evaluation several use cases for Hit Dexter 2.0, based on data sets derived from approved drugs, natural products, potential PAINS (compounds that were detected by the PAINS rules), consistently inactive tested compounds (also known as DCM), as well as several screening libraries, were explored.

[D3] Hit Dexter 2.0: Machine-Learning Models for the Prediction of Frequent Hitters

Conrad Stork, Ya Chen, Martin Šícho and Johannes Kirchmair
J. Chem. Inf. Model., 2019

Available at `https://doi.org/10.1021/acs.jcim.8b00677`.

Contribution:
C. Stork and J. Kirchmair conceptualized the research. C. Stork developed the machine learning models and compiled the data sets. C. Stork validated the models, with contributions from M. Šícho and Y. Chen. C. Stork wrote the manuscript, with contributions from Y. Chen, M. Šícho, and J. Kirchmair. J. Kirchmair supervised the work.

The following article was reprinted with permission from:
Stork, C.; Chen, Y.; Šícho, M. and Kirchmair, J. Hit Dexter 2.0: Machine-Learning Models for the Prediction of Frequent Hitters, *J. Chem. Inf. Model.* **2018**, *59*, 1030–1043.
Copyright 2019 American Chemical Society

The supplementary information for this work can be found in Section C.

JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING
Cite This: J. Chem. Inf. Model. 2019, 59, 1030−1043
Article
pubs.acs.org/jcim

Hit Dexter 2.0: Machine-Learning Models for the Prediction of Frequent Hitters

Conrad Stork,[†] Ya Chen,[†] Martin Šícho,[†,‡] and Johannes Kirchmair*[,†,§,∥]

[†]Center for Bioinformatics (ZBH), Department of Computer Science, Faculty of Mathematics, Informatics and Natural Sciences, Universität Hamburg, Hamburg, 20146, Germany

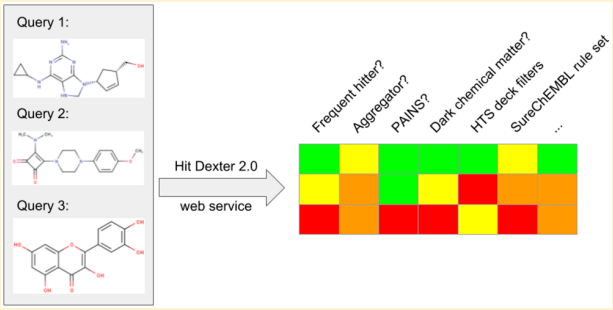
[‡]CZ-OPENSCREEN: National Infrastructure for Chemical Biology, Laboratory of Informatics and Chemistry, Faculty of Chemical Technology, University of Chemistry and Technology Prague, 166 28 Prague 6, Czech Republic

[§]Department of Chemistry, University of Bergen, N-5020 Bergen, Norway

[∥]Computational Biology Unit (CBU), University of Bergen, N-5020 Bergen, Norway

Supporting Information

ABSTRACT: Assay interference caused by small molecules continues to pose a significant challenge for early drug discovery. A number of rule-based and similarity-based approaches have been derived that allow the flagging of potentially "badly behaving compounds", "bad actors", or "nuisance compounds". These compounds are typically aggregators, reactive compounds, and/or pan-assay interference compounds (PAINS), and many of them are frequent hitters. Hit Dexter is a recently introduced machine learning approach that predicts frequent hitters independent of the underlying physicochemical mechanisms (including also the binding of compounds based on "privileged scaffolds" to multiple binding sites). Here we report on the development of a second generation of machine learning models which now covers both primary screening assays and confirmatory dose−response assays. Protein sequence clustering was newly introduced to minimize the overrepresentation of structurally and functionally related proteins. The models correctly classified compounds of large independent test sets as (highly) promiscuous or nonpromiscuous with Matthews correlation coefficient (MCC) values of up to 0.64 and area under the receiver operating characteristic curve (AUC) values of up to 0.96. The models were also utilized to characterize sets of compounds with specific biological and physicochemical properties, such as dark chemical matter, aggregators, compounds from a high-throughput screening library, drug-like compounds, approved drugs, potential PAINS, and natural products. Among the most interesting outcomes is that the new Hit Dexter models predict the presence of large fractions of (highly) promiscuous compounds among approved drugs. Importantly, predictions of the individual Hit Dexter models are generally in good agreement and consistent with those of Badapple, an established statistical model for the prediction of frequent hitters. The new Hit Dexter 2.0 web service, available at http://hitdexter2.zbh.uni-hamburg.de, not only provides user-friendly access to all machine learning models presented in this work but also to similarity-based methods for the prediction of aggregators and dark chemical matter as well as a comprehensive collection of available rule sets for flagging frequent hitters and compounds including undesired substructures.



INTRODUCTION

Biochemical assays are a core component of early drug discovery.[1−3] Some small molecules however can pose significant challenges to biochemical assays as they may trigger false outcomes. Whereas false negative results may lead to a loss of valuable bioactive compounds, false positive outcomes can, if they remain undetected, tie up and consume significant resources and time without prospect of success. In the worst case, these "badly behaving compounds", "bad actors", or "nuisance compounds" get reported as bioactive compounds and pollute the medicinal chemistry and chemical biology literature. Once published, invalid assay outcomes may propagate and trigger follow-up studies based on false grounds, which hampers the global drug discovery effort.

Nuisance compounds (Figure 1) include compounds that can form colloidal aggregates,[4,5] compounds with reactive groups,[6] and pan assay interference compounds (PAINS).[7] Note that the PAINS substructures by design do not cover aggregators because they were derived from the outcomes of high-throughput screening campaigns run in the presence of a detergent and casein in order to minimize phenomena related

Special Issue: Machine Learning in Drug Discovery

Received: October 1, 2018
Published: January 9, 2019

Downloaded via UNIV HAMBURG on December 2, 2020 at 07:10:51 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

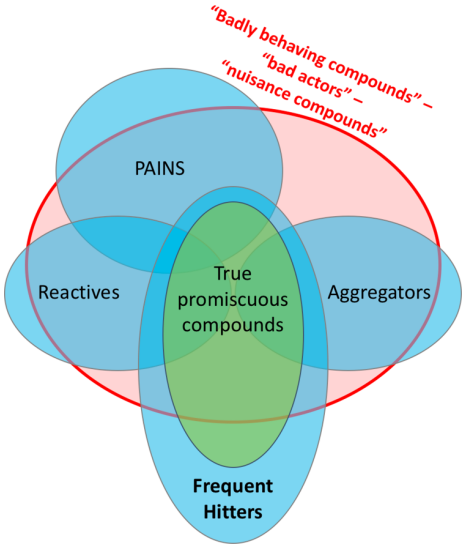


Figure 1. Schematic overview of key concepts and terms used in the context of assay interference and small-molecule drug discovery. Note that by design PAINS exclude aggregators and many types of reactive compounds. However, overlaps between the different types of compounds other than those depicted in this scheme certainly exist.

to aggregate formation. They also do not cover many types of reactive compounds because compounds with reactive functional groups had been removed from the screening library prior to assaying.[7]

The different types of badly behaving compounds discussed so far involve no definite assertions about the frequency by which they cause assay interference. Rather than interfering with all different kinds of assays, they trigger false outcomes only under specific (and by far not all) conditions. However, a tendency of badly behaving compounds to have higher hit rates is apparent.

Compounds for which a higher than expected hit rate is recorded in historical assay data are referred to as "frequent hitters".[8] Many of these compounds are aggregators, reactive compounds, or PAINS, but importantly, a significant proportion of frequent hitters are true promiscuous compounds. True promiscuity is often related to "privileged scaffolds"[9] or "master key compounds",[10] which have the ability to bind to multiple binding sites. True promiscuous compounds are not necessarily nuisance compounds. In fact, they can be valuable in the context of drug repurposing and polypharmacology.[11,12]

Computational methods for predicting nuisance compounds and frequent hitters are still in an early stage of development.[13] The most established approaches for identifying problematic compounds are rule-based methods, which flag compounds containing substructures that have been linked to assay interference. In recent years, the 480 patterns encoding substructures derived from PAINS have become not only one of the best known but also one of the most misused rule sets in medicinal chemistry. All too often, the limitations of the PAINS concept, most of which have been pointed out clearly by its inventors, are not paid the necessary attention.[14,15] Further approaches for the prediction of nuisance compounds and frequent hitters include similarity-based, statistical and machine learning approaches, an overview of which is provided in ref 13.

An important statistical approach for the prediction of frequent hitters is Badapple,[16] which performs a hierarchical scaffold analysis to derive a promiscuity score ("pScore"). The pScore corresponds to the likelihood of a compound based on a specific scaffold being promiscuous. Badapple was derived from a large public data set of more than 430 000 compounds measured in a total of more than 800 different assays.

We recently reported two machine learning models for the prediction of frequent hitters which are accessible via a free web service called "Hit Dexter".[17] Hit Dexter was developed with the idea of creating a reliable model for the prediction of frequent hitters independent of the underlying physicochemical mechanisms (including also the binding of compounds based on "privileged scaffolds" to multiple binding sites). Such a model could advice scientists for which compounds to exercise extra caution with positive assay readouts. The initial Hit Dexter models were trained on more than 235 000 compounds measured in at least 50 different confirmatory dose–response assays (CDRAs). They reached a high level of accuracy on independent test data, with Matthews correlation coefficients[18] (MCCs) of up to 0.67 and area under the receiver operating characteristic curve (AUC) values of up to 0.96.

Here we report on the development of a second generation of machine learning models for the prediction of frequent hitters, which are accessible via the new Hit Dexter 2.0 web service.[19] The models are a result of several major refinements and extensions of the data collection, data processing and modeling procedures. For example, a clustering approach was introduced in order to avoid an overrepresentation of structurally and functionally related proteins such as protein kinases. Hit Dexter 2.0 also includes models trained on data measured with primary screening assays (PSAs). In contrast to CDRAs, PSAs are primarily high-throughput screening assays measuring single-dose inhibition. The inclusion of these models in Hit Dexter 2.0 will allow a better representation of assays employed for primary screening.

In addition to method and model refinement, we also report on comprehensive tests of Hit Dexter 2.0 with various types of compounds, including dark chemical matter (DCM),[20] approved drugs, and natural products. Last but not least, we present a direct comparison of Hit Dexter 2.0 with Badapple and introduce the new Hit Dexter 2.0 web service.

RESULTS AND DISCUSSION

Data Set Compilation and Analysis. Two large data sets were compiled from PubChem Bioassay,[21,22] one consisting of 803 898 compounds measured in 931 PSAs and the other one consisting of 468 258 compounds measured in 2273 CDRAs. During data preprocessing, filtering in particular, 20 921 and 18 003 compounds were removed from the PSA and CDRA data sets, respectively (Table 1).

Following this procedure, the proteins covered by the PSA and CDRA data sets were clustered based on sequence similarity: The 429 proteins covered by the PSA data set were assigned to 388 unique protein clusters, and the 712 proteins covered by the CDRA data set were assigned to 537 unique protein clusters (see Methods for details).

The definition of whether a compound is (highly) promiscuous or not is based on the active-to-tested ratio (ATR), which is calculated according to eq 1:

DOI: 10.1021/acs.jcim.8b00677
J. Chem. Inf. Model. 2019, 59, 1030–1043

Table 1. Number of Compounds Removed During Filtering and Quality Checks

reason for removal	PSA data set [cpds]	CDRA data set [cpds]
invalid SMILES notation	1	3
salt filter with ambiguous outcome[a]	770	231
molecular weight outside the range of 200 to 900 Da	11231	10815
elements other than H, B, C, N, O, F, Si, P, S, Cl, Se, Br, and I	116	175
duplicate molecules[b]	8460	5171
rejected during quality checks[c]	343	1608
sum	**20921**	**18003**

[a]Multicomponent compounds for which the main component could not be unequivocally defined (see Methods and ref 17 for detail). [b]Identified based on canonicalized SMILES. Associated data were merged as outlined in ref 17. This led to the effective reduction of the number of compounds as reported in the table. [c]Compounds with conflicting data (e.g., activity data). See Methods and ref 17 for detail.

$$\mathrm{ATR} = \frac{A}{T} \tag{1}$$

where A is the number of protein clusters for which a compound was measured as active on at least one protein of that cluster, and T is the total number of protein clusters a compound was measured on.

The ideal data set to derive ATRs from would consist of a large number of compounds measured on a large number of protein clusters. Obviously, with the available data a compromise needs to be found between the number of instances available for training and testing the models (i.e., size of the data set in terms of the number of compounds) and the minimum number of protein clusters for which activity data are recorded for the individual compounds.

Figure 2 shows the relationship between data set size and the minimum number of protein clusters for which assay results

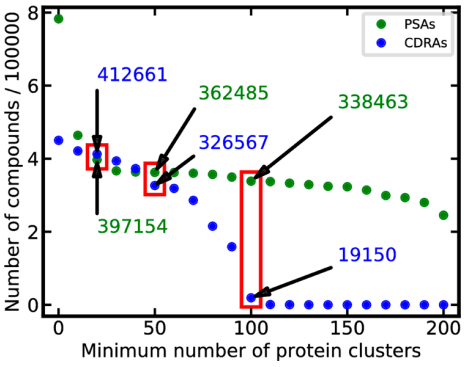


Figure 2. Data set size (number of compounds) as a function of the minimum number of protein clusters for which measured data are available. The rectangles mark the PSA20/CDRA20, PSA50/CDRA50, and PSA100/CDRA100 data sets.

have been recorded. For example, the processed data set includes more than 362 000 compounds which have been measured with PSAs representing at least 50 different protein clusters. Likewise, over 326 000 compounds are represented by the respective data collected from CDRAs. One of the most obvious differences between the PSA and CDRA data sets is that for the vast majority of compounds of the PSA data set measured data are available for 150 and more protein clusters, whereas for the CDRA data set a steep decline in the number of compounds for which measured data are available for 70 and more protein clusters is observed.

We explored the use of data sets containing all compounds for which bioactivity data has been recorded for at least 20, 50, and (only for PSAs) 100 protein clusters (data not shown). In agreement with previous results,[17] we found the data sets containing all compounds for which bioactivity data has been recorded for at least 50 protein clusters to be highly diverse and most suitable for modeling. We refer to these data sets as the PSA50 and CDRA50 data sets.

Next, all compounds were assigned a promiscuity label based on their ATR: "NP" for nonpromiscuous compounds, "P" for promiscuous compounds and "HP" for highly promiscuous compounds. Note that, according to the definitions of promiscuity summarized in Table 2, highly promiscuous compounds are a subset of promiscuous compounds.

Suitable cutoffs for the assignment of promiscuity labels were calculated for the PSA50 and CDRA50 data sets according to the definitions derived as part of our previous work (recited in Table 2, column "threshold definition").[17] According to these definitions, any compounds with an ATR greater than 0.024 for PSAs and 0.043 for CDRAs were labeled promiscuous, accounting for 11% and 13% of all compounds, respectively. These proportions of promiscuous compounds are in good agreement e.g. with the findings of a recent study from GlaxoSmithKline, which reported a fraction of 13% of all compounds as "noisy",[23] and higher than the averaged incidence of frequent hitters reported for the AstraZeneca screening library (which is 6%).[24] The mean ATRs for the PSA and CDRA data sets were 0.008 and 0.015, respectively (see Table 2 for more detail). These mean ATRs correspond well to the findings of other studies, such as that on the AstraZeneca compound collection, which reported an overall hit rate of 1.53%.[24]

CDRAs tend to have higher hit rates than PSAs. This is observed in the ATR distributions reported in Figure 3 and also reflected in the higher mean ATR for CDRAs (Table 2). The differences in hit rates can be explained by the fact that CDRAs are often used to measure compounds which have previously been reported as active by a PSA and are hence more likely to also show activity in CDRAs than a random set of screening compounds.

Comparison of the Chemical Space of the PSA and CDRA Data Sets. The chemical space of the individual data sets used for modeling was determined and compared using (i) principal component analysis (PCA) on 44 physically meaningful 2D descriptors computed with MOE[25] (listed in Table S1 of ref 17) and (ii) pairwise similarity analysis based on the Tanimoto coefficient calculated from Morgan2 fingerprints.[26,27]

In a first experiment, we analyzed whether the reduction of the PSA and CDRA data sets to subsets of compounds annotated with measured data for at least 50 protein clusters leads to a substantial loss of coverage. As shown by the PCA scatter plots and histograms reported in Figure 4, the chemical space of compounds covered by the PSA50 and CDRA50 data sets is—to a large extent—comparable with that of the PSA0 and CDRA0 data sets, respectively. Only about 11% of all compounds of the PSA0 data set and about 9% of all compounds of the CDRA0 data set have a maximum Tanimoto coefficient of less than 0.5 measured against any

DOI: 10.1021/acs.jcim.8b00677
J. Chem. Inf. Model. 2019, 59, 1030−1043

Table 2. Composition of the Data Sets Used for Model Training and Validation

assigned promiscuity class	number of unique compounds in data set	PSA50	CDRA50	threshold definition[a]	threshold value PSA50[b]	threshold value CDRA50[b]
nonpromiscuous (NP)	total:	247110	234811	ATR < ATR_{mean}	0.008	0.015
	training set:	222272	211264			
	test set[c]:	24881	23574			
promiscuous (P)	total:	29042	33982	ATR > ATR_{mean} + 1σ[d]	0.024	0.043
	training set:	26117	30478			
	test set[c]:	2930	3507			
highly promiscuous (HP)—a subset of compounds labeled P	total:	6625	6246	ATR > ATR_{mean} + 3σ[d]	0.054	0.100
	training set:	5956	5609			
	test set[c]:	670	637			

[a]Derived as part of our previous work.[17] Compounds with ATRs between ATR_{mean} and ATR_{mean} + 1σ were not assigned a promiscuity label and removed from all data sets. [b]ATR threshold values calculated for the individual data sets according to the ATR threshold definition. [c]Independent test set obtained by random split of the curated data set prior to model development. [d]Standard deviation.

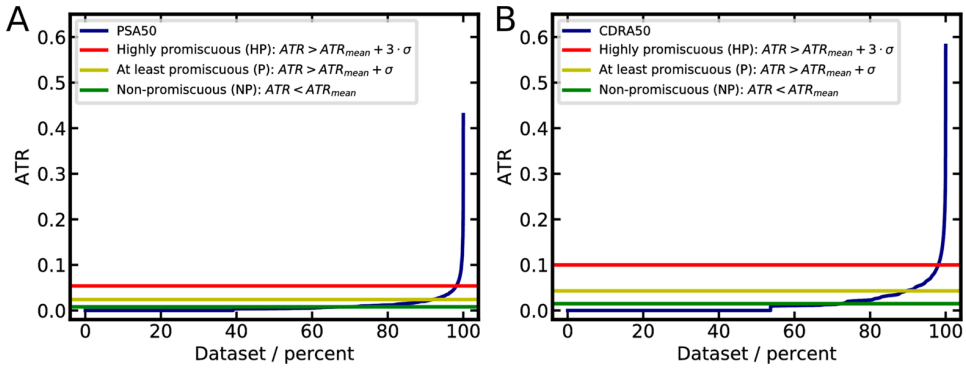


Figure 3. ATR distributions calculated for the (A) PSA50 and (B) CDRA50 data sets. The vertical lines mark the cutoffs applied for the assignment of promiscuity labels. Any compounds below the green line were labeled "NP", any above the yellow line "P", and any above the red line "HP".

compounds present in the PSA50 and CDRA50 data sets, respectively. In other words, this means that by constraining the data used for model development to compounds for which measured data has been recorded for at least 50 protein clusters does not lead to a substantial reduction of chemical space coverage as compared to the complete (processed) PubChem Bioassay data sets.

Following the same protocol, we also compared the chemical space covered by the PSA50 and CDRA50 data sets. As shown in Figure 5A and D, no substantial differences in coverage between the two data sets are apparent from the PCA and pairwise similarity analysis. Last but not least we compared the PSA50 and CDRA50 data sets to the complete ChEMBL database.[28,29] The ChEMBL database was prepared following the identical data preprocessing protocol (without considering any biological data) and consists of about 1.5 million compounds. From this comparison it can be seen that chemical space of the ChEMBL database is wider than that of the PSA50 and CDRA50 data sets (Figure 5B and C). This is an expected result, since the ChEMBL database contains substantially more compounds from a large number of diverse sources. Nevertheless, the plots in Figure 5E and F show that approximately 50% of all ChEMBL compounds are well represented by the PSA50 and CDRA50 data sets.

Model Development. Prior to model development, the PSA50 and CDRA50 data sets were randomly split into a training and a test set with a ratio of 9:1. In contrast to our previous work,[17] an additional data preprocessing step was implemented which checks for the presence of any compounds with distinct canonicalized SMILES but identical Morgan2 fingerprints (as in the case of stereoisomers, for example) because this can lead to inconsistent predictions. In order to address these issues, any instances with identical Morgan2 fingerprints were merged if their promiscuity labels were identical. If their labels differed, all instances with identical Morgan2 fingerprints were removed from the training data. Table 3 lists the number of compounds removed from the training and test sets as part of this process.

For assays of both screening stages (i.e., PSAs and CDRAs), two binary classifiers were developed: one to distinguish promiscuous from nonpromiscuous compounds (P-NP classifier) and another one to distinguish highly promiscuous from nonpromiscuous compounds (HP-NP classifier). An overview of the size of the training and test set is reported in Table 2.

As a first step in the model building process, the most suitable machine learning algorithm and descriptor set were identified. In addition to the extra tree classifiers (ETC) and random forest classifiers (RFC) explored in our previous work,[17] we also tested several meta classifiers such as the AdaBoost Classifier[30,31] and Bagging Classifier,[32,33] both in combination with the ETC and RFC. With respect to descriptors, we explored all 206 2D descriptors available in MOE, Morgan2 fingerprints (1024 bit), and MACCS keys (166 bits).

The various combinations of machine learning algorithms and descriptors were tested with 10-fold cross-validation (see Methods for more detail). The performance of the individual classifiers was compared based on the MCC, which quantifies

DOI: 10.1021/acs.jcim.8b00677
J. Chem. Inf. Model. 2019, 59, 1030−1043

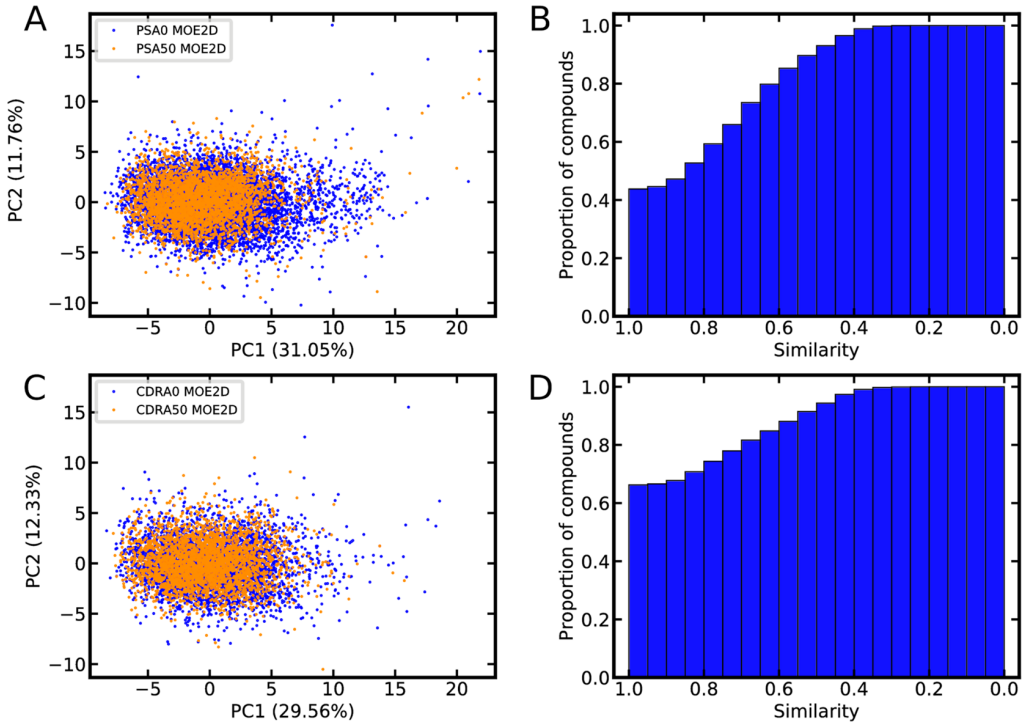


Figure 4. Comparison of the chemical space of the (A, B) PSA0 and PSA50 data sets and (C, D) CDRA0 and CDRA50 data sets. The PCA scatter plots in parts A and C are based on the first against the second component and derived from 44 physicochemically meaningful 2D descriptors calculated with MOE (see Table S1 in ref 17). PCA was performed on the full data set. For the sake of clarity, only a randomly selected 1% of all data points are shown. The axis labels report the percentage of the total variance explained by the respective principal component (PC). The histogram in part B shows the proportion of compounds of the PSA0 data set represented by the PSA50 data set at a given minimum similarity (Tanimoto coefficient calculated from Morgan2 fingerprints). Likewise, the histogram in part D shows the proportion of compounds of the CDRA0 data set represented by the CDRA50 data set.

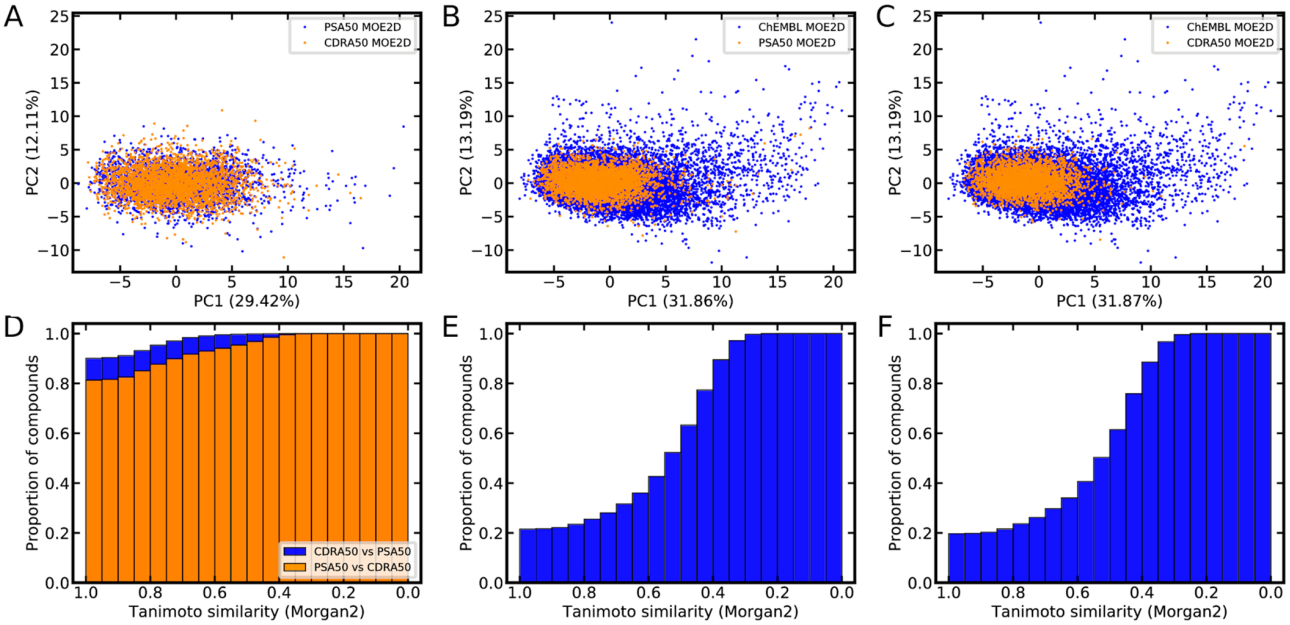


Figure 5. Comparison of the chemical space of the PSA50, CDRA50, and ChEMBL data sets in the (A, B, C) MOE 2D descriptor space and (D, E, F) Morgan2 fingerprint space. The PCAs were performed on the full data sets. For the sake of clarity, only a randomly selected 1% of all data points are shown. The axis labels report the percentage of the total variance explained by the respective principal component. The histograms in part D show the proportion of PSA50 compounds represented by the CDRA50 data set at a given minimum similarity (Tanimoto coefficient calculated from Morgan2 fingerprints) in orange and vice versa in blue. Likewise, the histograms in parts E and F show the proportion of compounds of the ChEMBL database represented by the (E) PSA50 and (F) CDRA50 data sets.

DOI: 10.1021/acs.jcim.8b00677
J. Chem. Inf. Model. 2019, 59, 1030−1043

Table 3. Number of Compounds Filtered Due to Duplicate Morgan2 Fingerprints

data set	no. of compounds merged due to identical fingerprints and promiscuity labels	no. of compounds with identical fingerprints removed due to contradicting promiscuity labels
PSA50 training set	2522	303
PSA50 test set	41	8
CDRA50 training set	1664	281
CDRA50 test set	26	4

the correlation between the predictions and their true value by taking into account the true positive, false positive, true negative, and true positive predictions. MCC values range from −1 to +1, where a value of +1 indicates perfect prediction, a value of 0 a performance equal to random prediction, and a value of −1 total disagreement of the prediction. In addition, we generated receiver operating characteristic (ROC) curves and calculated the area under the ROC curves (AUCs). By considering these three components, a solid understanding of the goodness of the models can be obtained: Whereas the MCC quantifies the capability of a model to correctly classify a compound of interest, ROC curves and (to some extent also) AUC values quantify a model's ability to identify (highly) promiscuous compounds by assigning them high probabilities as compared to nonpromiscuous compounds (i.e., ranking (highly) promiscuous compounds early in a list).

For all data sets the best performance during 10-fold cross-validation was obtained by ETCs in combination with Morgan2 fingerprints (MCC values between 0.56 and 0.58). These observations are consistent with the observations made during our previous work.[17]

Following these experiments, the hyperparameters (Table 4) for the ETCs (in combination with Morgan2 fingerprints)

Table 4. Hyperparameters Optimized by Grid Search[a]

parameter	option
number of estimators (estimators)[b]	10[c], 50, **100**, 150, 200, 250, 300, 400, 500, 600
maximum fraction of features considered per split (max_features)[b]	sqrt[c], **0.2**, 0.4, 0.6, 0.8, none[d]

[a]Bold numbers indicate settings used for the production of the final models. [b]Parameter name in the scikit-learn[35] implementation. [c]Default value. [d]All features are used.

were optimized, again with 10-fold cross-validation and with MCC as performance measure. Optimization of the hyperparameters did not yield substantial improvements of the models. The most suitable settings for the number of estimators and the maximum fraction of features considered per split were 100 and 0.2, respectively. Classifiers using these hyperparameters obtained MCC values between 0.57 and 0.60 and AUC values between 0.91 and 0.96 during 10-fold cross-validation (Figure 6). The final models (with optimized parameters) were built on the complete training sets balanced with the synthetic minority oversampling technique (SMOTE).[34]

Model Evaluation on Independent Test Data. The final models (ETC; Morgan2 fingerprints; SMOTE for balancing the training data; n_estimators = 100; max_features = 0.2), referred to as Hit Dexter 2.0 models, were tested on an independent test set (Table 2) derived by random split of the preprocessed PSA50/CDRA50 data sets prior to model building. The MCC values obtained by the four classifiers (i.e., HP-NP and P-NP classifiers, trained on PSA or CDRA data) for the independent test set was between 0.60 and 0.64 (Figure 7A), whereas their AUC values ranged from 0.91 to 0.96. Both HP-NP classifiers performed on average slightly better than the P-NP classifiers. This is expected because the margin between the cutoffs utilized to assign compounds to either of the two promiscuity classes is larger for the HP-NP classifier.

The robustness of the Hit Dexter 2.0 models was further probed by iteratively removing compounds from the test set that are similar to any of the compounds in the training data. More specifically, the maximum allowed similarity between the compounds of the test set and any compounds in the training set (measured as Tanimoto coefficient calculated from Morgan2 fingerprints) was reduced by 0.02 during each iteration (Figure 8). For example, for the subset of test compounds with a maximum Tanimoto coefficient of 0.8, MCC values of 0.55 to 0.58 were obtained, whereas AUC values were between 0.90 and 0.95 (Figure 7B). For the subset of test compounds with a maximum Tanimoto coefficient of 0.7, MCC values were between 0.44 and 0.50, and AUC values between 0.87 and 0.92 (Figure 7C). Decent performance was observed for subsets of test compounds with a maximum Tanimoto coefficient as low as 0.6 (MCC values between 0.34 and 0.42; AUC values between 0.82 and 0.88). Overall, the MCC values obtained for the initial Hit Dexter models[17] are slightly higher than those obtained for Hit Dexter 2.0. This may be a result of the protein clustering procedure, as compounds active on several related proteins (which therefore may contain characteristic structural patterns that can be more easily recognized by machine learning algorithm) may no longer be part of the P (and HP) data set.

Application of Hit Dexter 2.0 to Different Data Sets. In order to obtain a better understanding of the scope and limitations of Hit Dexter 2.0, we analyzed its predictions for a number of data sets with distinct characteristics:

- The dark chemical matter (DCM) data set:[20] a library of compounds which have been tested in at least one hundred different biochemical assays and have never shown activity. This data set originates from Novartis and PubChem assay data collected for more than 139 000 compounds (all size indications in this list referring to the unprocessed data sets).
- The aggregators data set:[36] a set of more than 12 600 compounds known to form colloidal aggregates. This library serves as data resource for Aggregator Advisor.[37]
- The Enamine HTS Collection:[38] Enamine is a leading provider of screening compounds and screening blocks. The Enamine HTS collection was selected as a representative library widely used in high-throughput screening. It contains 1.9 million compounds.
- The ChEMBL database:[29] a curated chemical database of 1.7 million (mainly) drug-like compounds, richly annotated with measured bioactivity data.
- The approved drugs subset of DrugBank:[39] a set of 2158 drugs approved in at least one jurisdiction, at some point in time.

DOI: 10.1021/acs.jcim.8b00677
J. Chem. Inf. Model. 2019, 59, 1030−1043

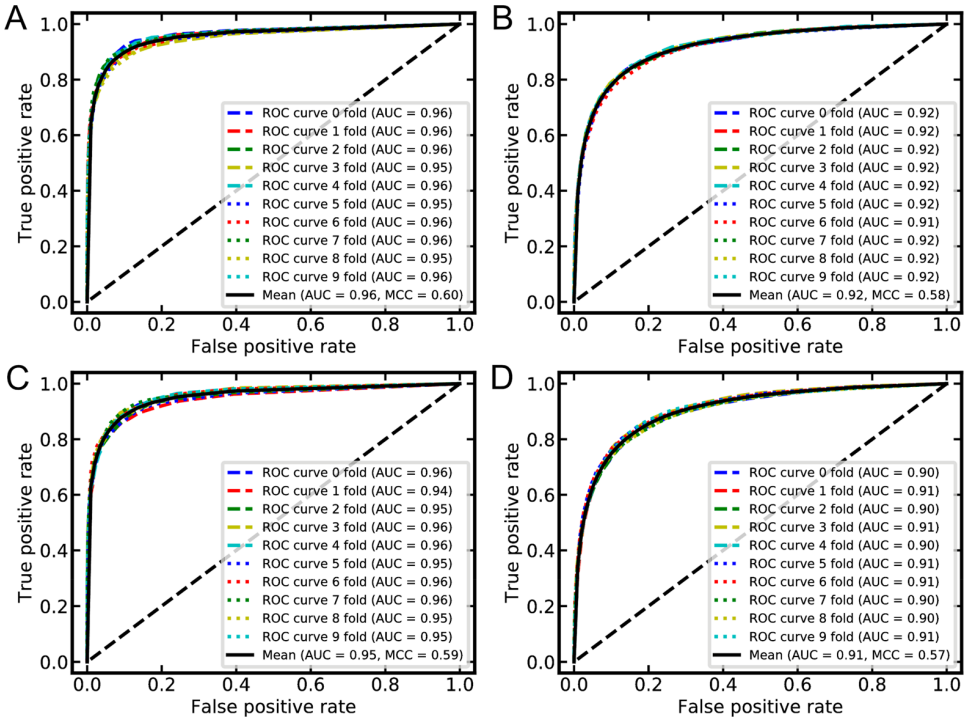


Figure 6. ROC curves obtained during 10-fold cross-validation for the selected models (i.e., ETC in combination with Morgan2 fingerprints; n_estimators = 100; max_features = 0.2). (A) HP-NP classifier for PSAs, (B) P-NP classifier for PSAs, (C) HP-NP classifier for CDRAs, and (D) P-NP classifier for CDRAs.

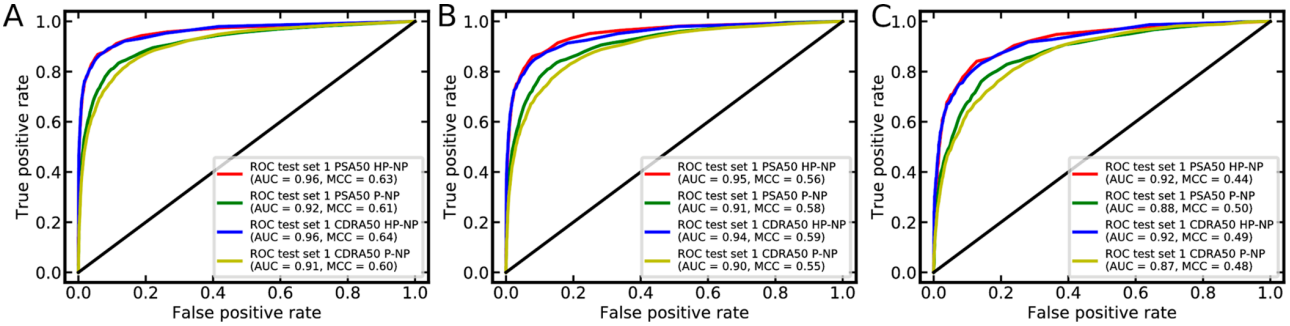


Figure 7. ROC curves obtained for the four classifiers on the independent test set (A) and subsets thereof, consisting of compounds with a maximum Tanimoto coefficient of (B) 0.8 or (C) 0.7 measured against any compound of the training set. MCC and AUC values are also reported.

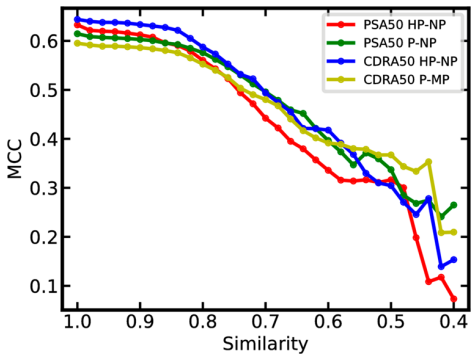


Figure 8. Classification performance (measured as MCC values) as a function of the maximum molecular similarity (Tanimoto coefficient calculated from Morgan2 fingerprints) between any pair of training and test compounds.

- A "potential PAINS" data set: a subset of over 51 600 compounds of the Enamine HTS Collection that match PAINS patterns (see Methods for details).
- A natural products data set: a comprehensive set of more than 208 000 known natural products compiled from 18 different sources as part of our previous work.[40]

The chemical structures included in these data sets were prepared following the identical procedure employed for preprocessing the Hit Dexter 2.0 training data. Any compounds present in the training data were removed from the individual data sets. The size of the filtered data sets is listed in Table 5.

In the previous section we showed that Hit Dexter 2.0 performs well on compounds represented by at least one instance in the training data with a minimum fingerprint-based Tanimoto coefficient of 0.6 (Figure 8). Considering this threshold, a large proportion of synthetic compounds is

DOI: 10.1021/acs.jcim.8b00677
J. Chem. Inf. Model. 2019, 59, 1030−1043

Table 5. Agreement of Predictions of the PSA and CDRA Classifiers

data set	no. of cpds in the HP-NP data set	no. of compounds (%) predicted as HP by the PSA HP-NP classifier	no. of compounds (%) predicted as HP by the CDRA HP-NP classifier	agreement of predictions [%]	no. of compounds in the P-NP data set	no. of compounds (%) predicted P by the PSA P-NP classifier	no. of compounds (%) predicted P by the CDRA P-NP classifier	agreement of predictions [%]
DCM	11116	79 (0.7)	69 (0.6)	26.5	10 944	306 (2.8)	361 (3.3)	19.1
aggregators	5786	225 (3.9)	272 (4.7)	34.0	4183	514 (12.3)	596 (14.3)	37.6
Enamine HTS collection	1856964	5883 (0.3)	7961 (0.4)	33.2	1853518	31068 (1.7)	46190 (2.5)	38.1
ChEMBL	1194343	27643 (2.3)	26447 (2.2)	37.0	1166478	88527 (7.6)	95031 (8.2)	46.0
approved drugs	972	48 (4.9)	58 (6.0)	39.5	813	93 (11.4)	102 (12.6)	36.4
potential PAINS	49498	1670 (3.4)	2246 (4.5)	45.9	49044	4867 (9.9)	6925 (14.1)	50.2
natural products	167873	8010 (4.8)	7919 (4.7)	36.4	167557	24046 (14.4)	22641 (13.5)	57.0
BADAPPLE_NP	110624	1575 (1.4)	1873 (1.7)	32.0	108620	7239 (6.7)	9853 (9.1)	42.4
BADAPPLE_P	346	82 (23.7)	87 (25.1)	55.1	330	170 (51.5)	203 (61.5)	69.6
BADAPPLE_HP	104	53 (51.0)	52 (50.0)	75.0	98	66 (67.4)	70 (71.4)	88.9

covered well by the training data (Figure S1). Taking the PSA50 training set of the P-NP classifier as an example, almost all DCM compounds, more than 80% of all aggregators, and approximately 60% of all approved drugs are represented by at least one instance in the training data with a Tanimoto coefficient of 0.6 or higher. The percentage of compounds from ChEMBL and the Enamine HTS collection covered at this level of (minimum) similarity is about 40%. In contrast to synthetic compounds, only for approximately 15% of all natural products the maximum pairwise similarity (measured as Tanimoto coefficient based on Morgan2 fingerprints) with all instances of the training data is 0.6 or higher (Figure 9).

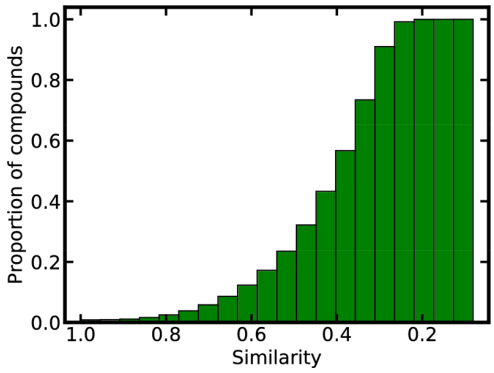


Figure 9. Example of the distribution of maximum pairwise similarities (Tanimoto coefficient calculated from Morgan2 fingerprints) between the training set (in this case, the PSA50 training set of the P-NP classifier) and the natural products data set.

Considering the limited prediction accuracy of the classifiers below a Tanimoto coefficient of 0.6 (Figure 8), the distance to the nearest neighbor(s) in the training data should be closely monitored. Therefore, in the following sections, in addition to the percentages of compounds referring to the complete, processed data sets, we report in brackets the percentages of compounds referring to the respective subsets consisting exclusively of compounds with a minimum Tanimoto coefficient of 0.6. In other words, the percentages reported in brackets are likely more accurate but they are based on less-representative subsets. Predictions on the data sets listed above are reported in Figures 10 and 11. From the graphs obtained for the DCM data set (Figure 10A), it can be seen that any of the four models (i.e., HP-NP and P-NP classifiers, each for PSAs and CDRAs) classify at least 96% [96%] of all compounds of the DCM data set as nonpromiscuous (both HP-NP classifiers obtained 99% [99%] correct classifications). This is an encouraging result since any of these compounds have been tested in a large number of assays and have never shown activity, for which reason they are unlikely frequent hitters (note that all reported numbers are based on a decision threshold of 0.5, which is the default value). In contrast to the observations made for the DCM data set, a substantial number of compounds of the aggregators data set (approximately 15% [18%]) are predicted as promiscuous and approximately 4% [5%] as highly promiscuous (with classifiers derived from assays of either screening stage (Figure 10B). This again is a plausible result because aggregators are known to cause false positive assay readouts under, importantly, specific assay conditions. It is hence expected that not all known aggregators will be flagged by Hit Dexter 2.0. The distribution of class probabilities among compounds from the Enamine HTS Collection is similar to that of the DCM data set (Figure 10C). This suggests that the Enamine HTS Collection is a well-curated screening library. For the ChEMBL database, the distributions of class probabilities are located somewhere in between those of the DCM data set and the aggregators data set (Figure 10D), meaning that there is a relevant fraction of compounds predicted as promiscuous (approximately 8% [17%]) or highly promiscuous (approximately 2% [6%]). The results obtained for the approved drugs data set may seem surprising: Hit Dexter 2.0 predicts approximately 13% [26%] of approved drugs as promiscuous and 6% [12%] as highly promiscuous (Figure 10E), which is generally even higher than the rates predicted for aggregators. Several approved drugs are known to form colloidal aggregates under specific assay conditions. However, a substantial part of the predicted frequent hitter behavior is likely linked to true promiscuity. Given the challenges involved in designing selective small molecules, this is not only plausible but also forms the basis for drug repurposing and polypharmacology. Note that the percentages of compounds predicted as (highly) promiscuous differ substantially between the (processed) approved drugs data set and the respective subset of compounds well-represented by the training data. These differences are plausible because of substantial differences in the composition of the two data sets: the processed approved drugs data set consists of around 1000 compounds, and the subset consists of only approximately three hundred compounds.

Interestingly, the distributions of class probabilities for approved drugs are even slightly steeper than those calculated for potential PAINS (Figure 10F). This supports the case that

DOI: 10.1021/acs.jcim.8b00677
J. Chem. Inf. Model. 2019, 59, 1030−1043

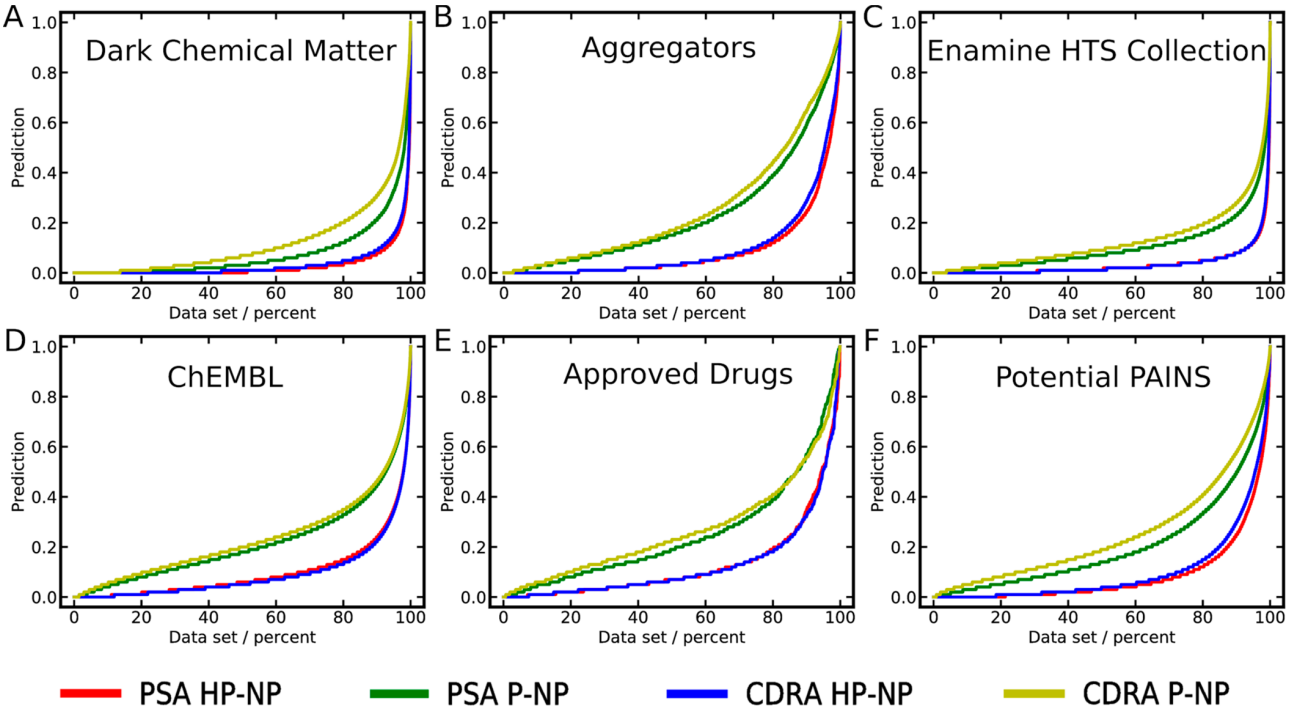


Figure 10. Distribution of class probabilities predicted with classifiers derived from PSA and CDRA data for (A) dark chemical matter, (B) known aggregators, (C) screening compounds from the Enamine HTS Collection, (D) the ChEMBL database, (E) approved drugs from DrugBank, and (F) subset of compounds of the Enamine HTS collection that match at least one PAINS pattern.

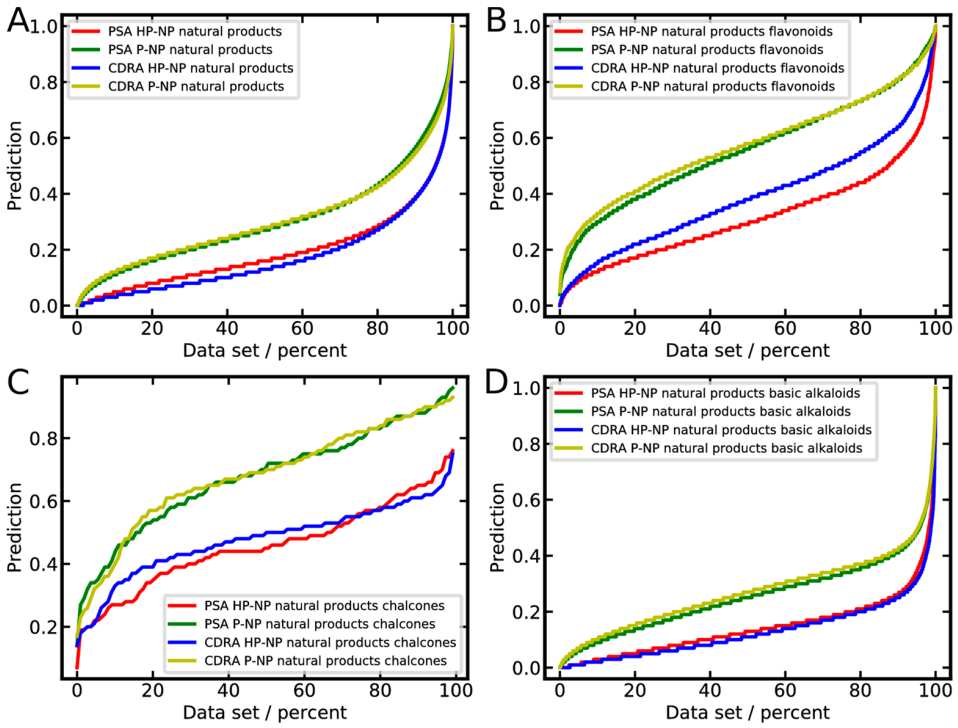


Figure 11. Distribution of class probabilities obtained with classifiers trained on PSA and CDRA data for (A) all natural products, (B) flavonoids, (C) chalcones, and (D) basic alkaloids.

compounds matching PAINS patterns are not necessarily frequent hitters.

Last but not least we utilized Hit Dexter 2.0 to predict the promiscuity of 208 000 natural products. Natural products can be challenging to screen in vitro, and it is known that several classes of natural products are prone to interfere with biochemical assays for different reasons.[41] This is reflected by the predictions of Hit Dexter 2.0. The class probability

DOI: 10.1021/acs.jcim.8b00677
J. Chem. Inf. Model. 2019, 59, 1030−1043

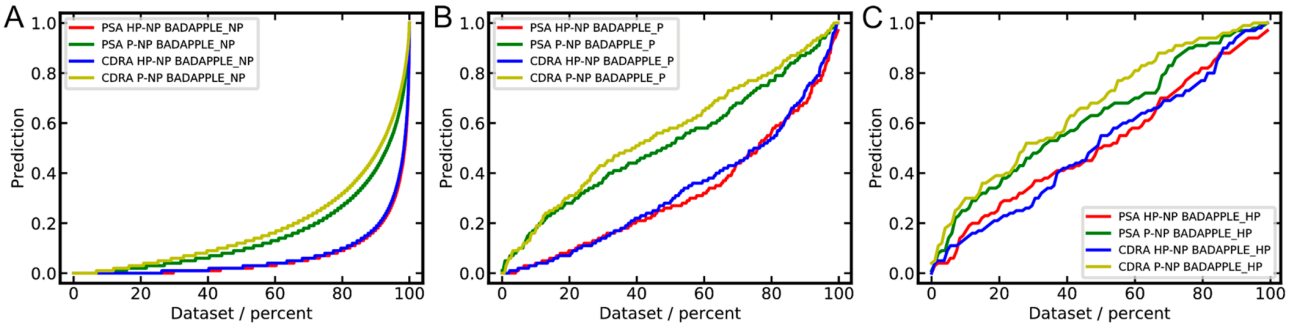


Figure 12. Hit Dexter 2.0 predictions of scaffold promiscuity for the (A) BADAPPLE_NP, (B) BADAPPLE_P, and (C) BADAPPLE_HP data sets.

distribution curves (in particular of those of the P-NP classifiers) show a steeper increase than for any other investigated data set (Figure 11A). Particularly noticeable is the high percentage of (highly) promiscuous compounds among flavonoids (the natural product classes were assigned with an automated approach presented previously),[40] with approximately 65% [73%] of all flavonoids predicted as promiscuous and 20% [31%] as highly promiscuous (Figure 10B). Among the investigated flavonoid subclasses (anthocyanidins, chalcones, flavandioles, flavanoles, flavanones, flavanonoles, flavones, flavonoles, and isoflavones), chalcones showed the highest rates of highly promiscuous (~42% [50%]) and promiscuous (~85% [86%]) compounds (Figure 11C; note that anthocyanidins are not represented in sufficient numbers in the data set that would allow to draw definite conclusions on their hit rates in assays). In contrast to the observations with flavonoids, Hit Dexter 2.0 reports less than 2% [9%] of all basic alkaloids (see ref 40 for the exact definition) as highly promiscuous and less than 6% [21%] as promiscuous (Figure 10D; note that the subset of compounds covered by the training data according to the above-mentioned criterion is just 10%).

Flavonoids have been reported in the literature to exhibit bioactivity on a large number of different proteins.[41] For example, according to a recent analysis,[41] by the year 2016 more than 680 distinct activities had been reported for quercetin, which is one of the most widely distributed flavonoids in nature but also a known aggregator and PAINS. For this particular flavonoid, PubChem Bioassay currently lists conclusive testing results of more than 1000 distinct assays, with quercetin reported as active in close to one out of two of these assay outcomes.

Whereas the health-promoting benefits of quercetin and other flavonoids are undisputed, it is reasonable to assume that many of the recorded activities are likely a result of assay interference. It is important to reemphasize at this occasion that the potential of a compound to interfere with assays does not per se lower its value as a bioactive compound, but it may make the rational optimization of its activity a difficult or, in some cases, even impossible task.

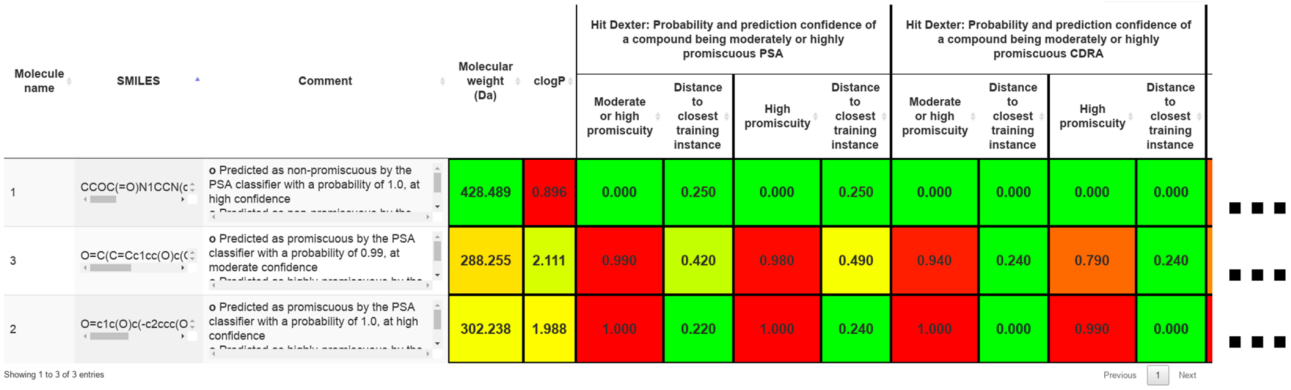
Exploration of the Badapple Data Sets with Hit Dexter 2.0. In contrast to Hit Dexter 2.0, which is trained on complete molecular structures, the Badapple model is derived from molecular scaffolds, each of which was assigned a promiscuity score. In order to explore to what extent predictions of Hit Dexter 2.0 based on molecular scaffolds are in agreement with Badapple (and the underlying data sets), we compiled sets of 142 468 nonpromiscuous scaffolds ("BADAPPLE_NP"), 610 promiscuous scaffolds ("BADAPPLE_P″), and 231 highly promiscuous scaffolds ("BADAPPLE_HP") from the Badapple data sets (published in ref 16).

As shown in Figure 12, Hit Dexter 2.0 is able to recognize compound promiscuity based on molecular scaffolds even though it was trained on complete molecular structures. Hit Dexter 2.0 correctly predicted the vast majority (91−99% [90−99%]) of nonpromiscuous scaffolds (Figure 12A). Both P-NP classifiers detected about 57% [72%] of all promiscuous scaffolds as such (Figure 12B), and both HP-NP classifiers predicted around 50% [79%] of the highly promiscuous scaffolds as such (Figure 12C). This can be considered a good agreement for two reasons: First, Hit Dexter 2.0 and Badapple use distinct thresholds for labeling compounds, and second, the BADAPPLE_HP and BADAPPLE_P data sets are small in size and contain only around 300 and 100 scaffolds (after preprocessing and removal of duplicates), respectively.

Comparison of the PSA and CDRA Models. As shown in the previous section, the overall behavior and performance of the PSA and CDRA models are comparable. In particular, the numbers of compounds assigned by the different models to one of the three promiscuity classes are similar. One interesting aspect to investigate is the agreement between predictions of the PSA and CDRA classifiers. Table 5 provides an overview of this for each of the above-mentioned data sets. Taking the Approved Drugs subset of DrugBank as an example, the PSA and CDRA models predicted 48 compounds (~5%) and 58 compounds (~6%) of this data set as highly promiscuous. The agreement between both predictions (defined as the fraction of compounds predicted as highly promiscuous by both classifiers as compared to those predicted as highly promiscuous by either classifier) was approximately 40%. Given the fact that only a small number of compounds was predicted as highly promiscuous, this can be considered a good agreement. For the BADAPPLE_HP data set, which consists entirely of highly promiscuous scaffolds, the agreement between the predictions made by the PSA HP-NP and CDRA HP-NP classifiers was 75%. Nevertheless, both classifiers have different sensitivity and for this reason the use of both predictors in parallel is recommended.

Hit Dexter 2.0 Web Service. The previously introduced Hit Dexter web service[17] was extended substantially. In addition to all models described in this work, we also implemented capabilities to predict aggregators and DCM based on molecular similarity, and to flag nuisance and undesired compounds based on several established collections of SMARTS patterns:

DOI: 10.1021/acs.jcim.8b00677
J. Chem. Inf. Model. 2019, 59, 1030−1043



Molecule name	SMILES	Comment	Molecular weight (Da)	clogP	Hit Dexter: Probability and prediction confidence of a compound being moderately or highly promiscuous PSA				Hit Dexter: Probability and prediction confidence of a compound being moderately or highly promiscuous CDRA			
					Moderate or high promiscuity	Distance to closest training instance	High promiscuity	Distance to closest training instance	Moderate or high promiscuity	Distance to closest training instance	High promiscuity	Distance to closest training instance
1	CCOC(=O)N1CCN(c	o Predicted as non-promiscuous by the PSA classifier with a probability of 1.0, at high confidence	428.489	0.896	0.000	0.250	0.000	0.250	0.000	0.000	0.000	0.000
3	O=C(C=Cc1cc(O)c(C	o Predicted as promiscuous by the PSA classifier with a probability of 0.99, at moderate confidence	288.255	2.111	0.990	0.420	0.980	0.490	0.940	0.240	0.790	0.240
2	O=c1c(O)c(-c2ccc(O	o Predicted as promiscuous by the PSA classifier with a probability of 1.0, at high confidence	302.238	1.988	1.000	0.220	1.000	0.240	1.000	0.000	0.990	0.000

Showing 1 to 3 of 3 entries

Previous 1 Next

Figure 13. Example of a heat map generated with the Hit Dexter 2.0 web service for three query compounds.

- The "hard filters" rule set developed at Glaxo Wellcome,[42] consisting of 55 patterns of undesired functional groups.
- A rule set developed at the University of Dundee,[43] consisting of 105 patterns of unwanted functional groups and substructures that likely cause interference with HTS assays.
- The "HTS deck filters" rule set developed at Bristol-Meyers Squibb,[44] consisting of 180 patterns of unwanted functional groups derived from intuition and experience.
- The SureChEMBL rule set of ToxAlert,[45] consisting of 166 patterns of toxicophores.
- The "excluded functionality filters" rule set of the NIH Molecular Libraries Small Molecule Repository,[46] consisting of 116 patterns for removing unwanted functional groups.
- The "Lint" rule set developed at Pfizer,[47] consisting 57 patterns of problematic functional groups during drug optimization.
- The PAINS set of substructures linked to assay interference,[7] consisting of 480 patterns. Note that the original PAINS patterns were encoded by Sybyl line notation[7] whereas the Hit Dexter 2.0 web service utilizes SMARTS patterns in combination with the substructure search implemented in RDKit.[48] This may lead to differing results in some cases.
- A set of 28 substructures derived from undesirable compounds. This is a subset of rules recently introduced by investigators from GlaxoSmithKline. The 28 substructures are listed in Table S2 of ref 23 (value "remove" in column "GSK Recommendation").

Search queries can either be sketched with the JSME Molecule Editor,[49] pasted as individual SMILES, or uploaded as a list of SMILES. Predictions are presented as a heat map (Figure 13) and include the results from all the machine learning models and similarity-based and rule-based approaches. Importantly, also the distance to the nearest neighbor in the training data is reported, which gives an indication of the reliability of predictions. A column with comments summarizes the conclusions that may be drawn from the predictions. We believe that these comments will be helpful in particular to occasional users of Hit Dexter 2.0.

The processing of a single compound takes few seconds. Predictions for 1000 compounds take approximately 4 h. The authors plan to increase the capacity of the web service should the need arise.

CONCLUSIONS

In this work we report on the second generation of machine learning models for the prediction of frequent hitters independent of the underlying physicochemical mechanisms, including the binding of compounds based on "privileged scaffolds" or of "master key compounds" to multiple binding sites. These models are, among others, accessible via the Hit Dexter 2.0 web service.[19]

In addition to a number of refinements of the data preparation and modeling strategy, substantial improvements presented in this work include the implementation of a protein clustering method in order to avoid an overrepresentation of structurally and functionally related proteins in the training data, and the utilization of PSA data, in addition to CDRA data, for machine learning. During comprehensive tests on independent data, models based on either PSA or CDRA data were shown to predict frequent hitters with high accuracy and robustness. While predictions from both model types were generally in good agreement, the parallel use of both types of classifiers can support the interpretation of results and is recommended.

Hit Dexter 2.0 was used for profiling compounds with specific biological and physicochemical properties, such as dark chemical matter, aggregators, compounds from a high-throughput screening library, drug-like compounds, approved drugs, potential PAINS, and natural products. The predictions obtained with Hit Dexter 2.0 confirm common observations and knowledge but also led to some less anticipated observations, such as the high fraction of frequent hitters predicted among approved drugs. A further encouraging observation made was the good agreement between predictions of Hit Dexter 2.0 and the Badapple data sets of molecular scaffolds and their observed promiscuity.

Since its initial launch in late 2017, the Hit Dexter web service has evolved from a small web presence with rudimentary features into a one-stop shop for the interrogation of compounds regarding their likelihood to exhibit frequent hitter behavior and/or interfere with biochemical assays and their general desirability in the context of drug discovery. More specifically, the new Hit Dexter 2.0 web service provides user-friendly access to machine learning approaches for the prediction of compound promiscuity, similarity-based methods for the prediction of aggregators and dark chemical matter, and a comprehensive collection of established and new rule sets for flagging frequent hitters and compounds with undesired substructures.

DOI: 10.1021/acs.jcim.8b00677
J. Chem. Inf. Model. 2019, 59, 1030−1043

We believe that Hit Dexter 2.0 will enable investigators to make better-informed decisions during hit triage and follow-up. However, the models should not be used as the sole basis for the acceptance or rejection of hits.

METHODS

Data Sets. Activity data measured for chemical substances (*substance type* = "chemical") on single protein targets (*target* = "single" and *target type* = "Protein Targets") in 932 primary screening assays (*screening stage* = "primary screening") and 2266 confirmatory dose−response assays (*screening stage* = "confirmatory, dose−response") were separately downloaded from the PubChem Bioassay database[21,22,50] via the PUG REST interface.[51] The download of BioAssay record AID 1224865 failed permanently and was therefore not considered in this work. The SMILES notations for all 803 898 compounds of the primary screening assays (PSAs) and all 468 260 compounds of the confirmatory dose−response assays (CDRAs) were retrieved via the PubChem Identifier Exchange Service.[52] Salt components, compounds with unsupported elements, and conflicting bioactivity data were identified and removed following the procedure described in ref 17. Also compounds with a molecular weight below 200 Da and above 900 Da were removed. In addition, all molecular structures were neutralized and tautomers merged using the "canonalize" method implemented in the "tautomer" class of MolVS.[53] Subsequently, duplicate compounds were removed based on identical SMILES. In order to ensure the consistency of predictions, compounds with identical Morgan2 fingerprints and differing promiscuity labels (e.g., stereoisomers) were removed from the training sets. This concerned a total of 1945 compounds for the PSA and 2825 compounds for the CDRA training sets. For any compounds with identical Morgan2 fingerprints only one instance was kept in the training sets.

For each PubChem Bioassay record the unique identifier for genes of the NCBI Protein database[54] ("gene identifier", GI) was obtained via the PubChem PUG REST interface. In total, 429 and 712 unique GIs were retrieved for the PSA and CDRA records, respectively. Subsequently, using these GIs, the protein sequences of all proteins of interest were downloaded in FASTA file format from the NCBI Protein database. Clustering of all protein sequences with *cd-hit*[55] (*sequence identity* = 60%; *tolerance* = 3) resulted in 388 and 537 protein clusters for the PSA and CDRA records, respectively.

For model development, each data set was split randomly into an external test set (10%) and a training set (90%) with the "train_test_split" method of the "model_selection" module of scikit-learn (version: 0.19.1).[35] Only the training set was used for model selection. Initial experiments for selecting the most suitable machine learning algorithm and descriptor sets were performed with default parameters for ETCs and RFCs, except for the number of estimators, which was set to 50, the class weight, which was set to "balanced", and bootstrapping, which was disabled. Default parameters were used for all meta classifiers (with ETCs and RFCs parametrized as described above). Stratified splitting was performed as part of cross-validation.

All data sets used to explore and determine the performance of Hit Dexter 2.0 were prepared and filtered according to identical protocol as outlined for the training data.

The Badapple data sets were compiled from the original source[16] by merging scaffolds with identical SMILES and removing any instances with contradicting promiscuity labels. Scaffolds assigned a pScore above 300 were included in the BADAPPLE_HP data set, scaffolds assigned a pScore between 100 and 299 were included in the BADAPPLE_P data set, and scaffolds assigned a pScore between 0 and 99 were included in the BADAPPLE_NP data set.

Hardware and Software. All calculations are performed on Linux workstations running openSUSE 42.2 and equipped with Intel i5 processors (3.2 GHz) and 16 GB of main memory.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.8b00677.

Additional figures and tables: Maximum pairwise similarity distributions between the PSA50 training set and six chemical data sets (PDF)

AUTHOR INFORMATION

Corresponding Author

*E-mail: johannes.kirchmair@uib.no. Tel.: +47 55 58 34 64.

ORCID

Conrad Stork: 0000-0002-5499-742X
Ya Chen: 0000-0001-5273-1815
Martin Šícho: 0000-0002-8771-1731
Johannes Kirchmair: 0000-0003-2667-5877

Funding

C.S. and J.K. are supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—project number KI 2085/1-1. J.K. is also supported by the Bergen Research Foundation (BFS)—grant no. BFS2017TMT01. Y.C. is supported by the China Scholarship Council (201606010345). M.S. is supported by the Ministry of Education of the Czech Republic—project numbers NPU I-LO1220 and LM2015063.

Notes

The authors declare no competing financial interest.

The Hit Dexter 2.0 web service is available at the following address: http://hitdexter2.zbh.uni-hamburg.de.

ACKNOWLEDGMENTS

Rainer Fährrolfes, Florian Flachsenberg, Robert Schmidt, and Gerd Embruch from the Center of Bioinformatics (ZBH) of the University of Hamburg are thanked for technical support and discussions.

ABBREVIATIONS

ATR, active to tested ratio; AUC, area under the ROC curve; CDRA, confirmatory dose−response assay; DCM, dark chemical matter; ETC, extra tree classifier; HP, highly promiscuous; HTS, high-throughput screening; MCC, Matthews correlation coefficient; MOE, Molecular Operating Environment; NP, nonpromiscuous; P, promiscuous; PAINS, pan-assay interference compounds; PCA, principal component analysis; PSA, primary screen assay; RFC, random forest classifier; ROC, receiver operating characteristic; SMARTS, SMILES arbitrary target specification; SMILES, simplified molecular input line entry specification; SMOTE, synthetic minority oversampling technique

DOI: 10.1021/acs.jcim.8b00677
J. Chem. Inf. Model. 2019, 59, 1030−1043

REFERENCES

(1) Szymański, P.; Markowicz, M.; Mikiciuk-Olasik, E. Adaptation of High-Throughput Screening in Drug Discovery-Toxicological Screening Tests. *Int. J. Mol. Sci.* **2012**, *13*, 427−452.

(2) Macarron, R.; Banks, M. N.; Bojanic, D.; Burns, D. J.; Cirovic, D. A.; Garyantes, T.; Green, D. V. S.; Hertzberg, R. P.; Janzen, W. P.; Paslay, J. W.; Schopfer, U.; Sittampalam, G. S. Impact of High-Throughput Screening in Biomedical Research. *Nat. Rev. Drug Discovery* **2011**, *10*, 188−195.

(3) Janzen, W. P. Screening Technologies for Small Molecule Discovery: The State of the Art. *Chem. Biol.* **2014**, *21*, 1162−1170.

(4) Ganesh, A. N.; Donders, E. N.; Shoichet, B. K.; Shoichet, M. S. Colloidal Aggregation: From Screening Nuisance to Formulation Nuance. *Nano Today* **2018**, *19*, 188−200.

(5) McGovern, S. L.; Caselli, E.; Grigorieff, N.; Shoichet, B. K. A Common Mechanism Underlying Promiscuous Inhibitors from Virtual and High-Throughput Screening. *J. Med. Chem.* **2002**, *45*, 1712−1722.

(6) Rishton, G. M. Reactive Compounds and in Vitro False Positives in HTS. *Drug Discovery Today* **1997**, *2*, 382−384.

(7) Baell, J. B.; Holloway, G. A. New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. *J. Med. Chem.* **2010**, *53*, 2719−2740.

(8) Roche, O.; Schneider, P.; Zuegge, J.; Guba, W.; Kansy, M.; Alanine, A.; Bleicher, K.; Danel, F.; Gutknecht, E.-M.; Rogers-Evans, M.; Neidhart, W.; Stalder, H.; Dillon, M.; Sjögren, E.; Fotouhi, N.; Gillespie, P.; Goodnow, R.; Harris, W.; Jones, P.; Taniguchi, M.; Tsujii, S.; von der Saal, W.; Zimmermann, G.; Schneider, G. Development of a Virtual Screening Method for Identification of "Frequent Hitters" in Compound Libraries. *J. Med. Chem.* **2002**, *45*, 137−142.

(9) Evans, B. E.; Rittle, K. E.; Bock, M. G.; DiPardo, R. M.; Freidinger, R. M.; Whitter, W. L.; Lundell, G. F.; Veber, D. F.; Anderson, P. S.; Chang, R. S.; et al. Methods for Drug Discovery: Development of Potent, Selective, Orally Effective Cholecystokinin Antagonists. *J. Med. Chem.* **1988**, *31*, 2235−2246.

(10) Medina-Franco, J. L.; Giulianotti, M. A.; Welmaker, G. S.; Houghten, R. A. Shifting from the Single to the Multitarget Paradigm in Drug Discovery. *Drug Discovery Today* **2013**, *18*, 495−501.

(11) Anighoro, A.; Bajorath, J.; Rastelli, G. Polypharmacology: Challenges and Opportunities in Drug Discovery. *J. Med. Chem.* **2014**, *57*, 7874−7887.

(12) Peters, J.-U. Polypharmacology − Foe or Friend? *J. Med. Chem.* **2013**, *56*, 8955−8971.

(13) Stork, C.; Kirchmair, J. PAIN(S) Relievers for Medicinal Chemists: How Computational Methods Can Assist in Hit Evaluation. *Future Med. Chem.* **2018**, *10*, 1533−1535.

(14) Baell, J. B.; Nissink, J. W. M. Seven Year Itch: Pan-Assay Interference Compounds (PAINS) in 2017-Utility and Limitations. *ACS Chem. Biol.* **2018**, *13*, 36−44.

(15) Kenny, P. W. Comment on The Ecstasy and Agony of Assay Interference Compounds. *J. Chem. Inf. Model.* **2017**, *57*, 2640−2645.

(16) Yang, J. J.; Ursu, O.; Lipinski, C. A.; Sklar, L. A.; Oprea, T. I.; Bologa, C. G. Badapple: Promiscuity Patterns from Noisy Evidence. *J. Cheminf.* **2016**, *8*, 29.

(17) Stork, C.; Wagner, J.; Friedrich, N.-O.; de Bruyn Kops, C.; Šícho, M.; Kirchmair, J. Hit Dexter: A Machine-Learning Model for the Prediction of Frequent Hitters. *ChemMedChem* **2018**, *13*, 564−571.

(18) Matthews, B. W. Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme. *Biochim. Biophys. Acta, Protein Struct.* **1975**, *405*, 442−451.

(19) Hit Dexter 2.0 web service. http://hitdexter2.zbh.uni-hamburg.de (accessed Nov 23, 2018).

(20) Wassermann, A. M.; Lounkine, E.; Hoepfner, D.; Le Goff, G.; King, F. J.; Studer, C.; Peltier, J. M.; Grippo, M. L.; Prindle, V.; Tao, J.; Schuffenhauer, A.; Wallace, I. M.; Chen, S.; Krastel, P.; Cobos-Correa, A.; Parker, C. N.; Davies, J. W.; Glick, M. Dark Chemical Matter as a Promising Starting Point for Drug Lead Discovery. *Nat. Chem. Biol.* **2015**, *11*, 958−966.

(21) Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H. PubChem Substance and Compound Databases. *Nucleic Acids Res.* **2016**, *44*, D1202−D1213.

(22) Wang, Y.; Bryant, S. H.; Cheng, T.; Wang, J.; Gindulyte, A.; Shoemaker, B. A.; Thiessen, P. A.; He, S.; Zhang, J. PubChem BioAssay: 2017 Update. *Nucleic Acids Res.* **2017**, *45*, D955−D963.

(23) Chakravorty, S. J.; Chan, J.; Greenwood, M. N.; Popa-Burke, I.; Remlinger, K. S.; Pickett, S. D.; Green, D. V. S.; Fillmore, M. C.; Dean, T. W.; Luengo, J. I.; Macarrón, R. Nuisance Compounds, PAINS Filters, and Dark Chemical Matter in the GSK HTS Collection. *SLAS Discov* **2018**, *23*, 532−544.

(24) M Nissink, J. W.; Blackburn, S. Quantification of Frequent-Hitter Behavior Based on Historical High-Throughput Screening Data. *Future Med. Chem.* **2014**, *6*, 1113−1126.

(25) *Molecular Operating Environment (MOE)*, Version 2016.08; Chemical Computing Group, Montreal, QC.

(26) Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5*, 107−113.

(27) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.

(28) Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magariños, M. P.; Overington, J. P.; Papadatos, G.; Smit, I.; Leach, A. R. The ChEMBL Database in 2017. *Nucleic Acids Res.* **2017**, *45*, D945−D954.

(29) ChEMBL 23. http://www.ebi.ac.uk/chembl (accessed Dec 8, 2017).

(30) Freund, Y.; Schapire, R. E. A Decision-Theoretic Generalization of on-Line Learning and an Application to Boosting. *Lecture Notes Comput. Sci.* **1995**, *904*, 23−37.

(31) Hastie, T.; Rosset, S.; Zhu, J.; Zou, H. Multi-Class AdaBoost. *Stat. Interface* **2009**, *2*, 349−360.

(32) Breiman, L. Pasting Small Votes for Classification in Large Databases and On-Line. *Mach. Learn.* **1999**, *36*, 85−103.

(33) Breiman, L. Bagging Predictors. *Mach. Learn.* **1996**, *24*, 123−140.

(34) Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; Kegelmeyer, W. P. SMOTE: Synthetic Minority Over-Sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321−357.

(35) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.

(36) Irwin, J. J.; Duan, D.; Torosyan, H.; Doak, A. K.; Ziebart, K. T.; Sterling, T.; Tumanian, G.; Shoichet, B. K. An Aggregation Advisor for Ligand Discovery. *J. Med. Chem.* **2015**, *58*, 7076−7087.

(37) Aggregator Advisor web service. http://advisor.bkslab.org (accessed Oct 1, 2018).

(38) Enamine HTS collection. https://enamine.net (accessed May 23, 2018).

(39) 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.; Maciejewski, A.; Gale, N.; Wilson, A.; Chin, L.; Cummings, R.; Le, D.; Pon, A.; Knox, C.; Wilson, M. DrugBank 5.0: A Major Update to the DrugBank Database for 2018. *Nucleic Acids Res.* **2018**, *46*, D1074−D1082.

(40) Chen, Y.; Garcia de Lomana, M.; Friedrich, N.-O.; Kirchmair, J. Characterization of the Chemical Space of Known and Readily Obtainable Natural Products. *J. Chem. Inf. Model.* **2018**, *58*, 1518−1532.

(41) Bisson, J.; McAlpine, J. B.; Friesen, J. B.; Chen, S.-N.; Graham, J.; Pauli, G. F. Can Invalid Bioactives Undermine Natural Product-Based Drug Discovery? *J. Med. Chem.* **2016**, *59*, 1671−1690.

DOI: 10.1021/acs.jcim.8b00677
J. Chem. Inf. Model. 2019, 59, 1030−1043

(42) Hann, M.; Hudson, B.; Lewell, X.; Lifely, R.; Miller, L.; Ramsden, N. Strategic Pooling of Compounds for High-Throughput Screening. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 897−902.
(43) Brenk, R.; Schipani, A.; James, D.; Krasowski, A.; Gilbert, I. H.; Frearson, J.; Wyatt, P. G. Lessons Learnt from Assembling Screening Libraries for Drug Discovery for Neglected Diseases. *ChemMedChem* **2008**, *3*, 435−444.
(44) Pearce, B. C.; Sofia, M. J.; Good, A. C.; Drexler, D. M.; Stock, D. A. An Empirical Process for the Design of High-Throughput Screening Deck Filters. *J. Chem. Inf. Model.* **2006**, *46*, 1060−1068.
(45) Sushko, I.; Salmina, E.; Potemkin, V. A.; Poda, G.; Tetko, I. V. ToxAlerts: A Web Server of Structural Alerts for Toxic Chemicals and Compounds with Potential Adverse Reactions. *J. Chem. Inf. Model.* **2012**, *52*, 2310−2316.
(46) NIH Molecular Libraries Small Molecule Repository. https://grants.nih.gov/grants/guide/notice-files/not-rm-07-005.html (accessed Oct 1, 2018).
(47) Blake, J. Identification and Evaluation of Molecular Properties Related to Preclinical Optimization and Clinical Fate. *Med. Chem.* **2005**, *1*, 649−655.
(48) RDKit version 2016.09.4: Open-Source Cheminformatics Software. http://www.rdkit.org (accessed Nov 23, 2018).
(49) Bienfait, B.; Ertl, P. JSME: A Free Molecule Editor in JavaScript. *J. Cheminf.* **2013**, *5*, 24.
(50) PubChem Bioassay database. https://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi (accessed Sep 10, 2018).
(51) Kim, S.; Thiessen, P. A.; Cheng, T.; Yu, B.; Bolton, E. E. An Update on PUG-REST: RESTful Interface for Programmatic Access to PubChem. *Nucleic Acids Res.* **2018**, *46*, W563−W570.
(52) PubChem Identifier Exchange Service. https://pubchem.ncbi.nlm.nih.gov/idexchange/idexchange.cgi (accessed Sep 11, 2018).
(53) MolVs version 0.1.1. https://github.com/mcs07/MolVS (accessed Jul 12, 2018).
(54) NCBI Resource Coordinators. NCBI Resource Coordinators. *Nucleic Acids Res.* **2016**, *44*, D7−D19.
(55) Fu, L.; Niu, B.; Zhu, Z.; Wu, S.; Li, W. CD-HIT: Accelerated for Clustering the next-Generation Sequencing Data. *Bioinformatics* **2012**, *28*, 3150−3152.

DOI: 10.1021/acs.jcim.8b00677
J. Chem. Inf. Model. 2019, 59, 1030−1043

5.3 Machine learning models for the prediction of frequent hitters based on target-based and cell-based assay data sets

Further important refinements of the previously developed machine learning models were performed in the third part of this Ph.D. study. These include the differentiation of target-based and cell-based assay screenings. In target-based assays a compound interacts with a purified protein which involves different action modes than in cell-based assays where a compound is interacting with a complete cell. Regarding the aim of dedicated models for target-based and cell-based assays, we manually extracted three large data sets from the PubChem Bioassay database, including a target-based assay data set, a cell-based assay data set (excluding assays measuring nonspecific interactions like cytotoxicity) and an extended cell-based assay data set (which also includes assay measuring nonspecific interactions). As more data are available for these three data sets it was possible to consider only compounds that were tested against at least 100 distinct proteins, which makes the calculated hit rates (i.e. fraction of times a compound was tested active and times a compound was tested) more robust and meaningful. Several machine learning models were generated, including k-nearest neighbors (KNN), random forest (RF), extra tree (ET) and multilayer perceptron (MLP) classifiers. Dedicated models, called Hit Dexter 3 models and which are available in NERDD, were built for the three data sets for the differentiation of non-promiscuous and promiscuous, as well as non-promiscuous and highly promiscuous compounds. The best performing classification models (i.e. neural network classifiers trained on Morgan2 fingerprints) reached MCCs of up to 0.65. In addition, it was shown that the separation of target-based and cell-based assay screenings is necessary as models based on the target-based assay data set cannot predict frequent hitters derived from the cell-based assay data set and vice versa.

[D5] Computational prediction of frequent hitters in target-based and cell-based assays

Conrad Stork, Neann Mathai and Johannes Kirchmair

Artificial Intelligence in the Life Sciences, 2021

Available at `https://doi.org/10.1016/j.ailsci.2021.100007`.

Contribution:
C. Stork, N. Mathai and J. Kirchmair conceptualized the research. C. Stork developed the machine learning models. C. Stork validated the models, with contributions from N. Mathai. C. Stork wrote the manuscript, with contributions from N. Mathai and J. Kirchmair. J. Kirchmair supervised the work.

The following article was reprinted with permission from:
Stork, C.; Mathai, N. and Kirchmair, J. Computational prediction of frequent hitters in target-based and cell-based assays, *Artificial Intelligence in the Life Sciences* **2021**, *1*, 100007.
Copyright 2021 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

The supplementary information for this work can be found in Section D.

Artificial Intelligence in the Life Sciences 1 (2021) 100007

Contents lists available at ScienceDirect

Artificial Intelligence in the Life Sciences

journal homepage: www.elsevier.com/locate/ailsci

Computational prediction of frequent hitters in target-based and cell-based assays

Conrad Stork [a], Neann Mathai [b], Johannes Kirchmair [a,b,c,*]

[a] Universität Hamburg, Faculty of Mathematics, Informatics and Natural Sciences, Department of Informatics, Center for Bioinformatics, 20146 Hamburg, Germany
[b] Department of Chemistry and Computational Biology Unit (CBU), University of Bergen, N-5020 Bergen, Norway
[c] Department of Pharmaceutical Sciences, Division of Pharmaceutical Chemistry, Faculty of Life Sciences, University of Vienna, 1090 Vienna, Austria

ARTICLE INFO

Keywords:
Machine learning
Frequent hitters
Nuisance compounds
PAINS
Biological assays
High-throughput screening

ABSTRACT

Compounds interfering with high-throughput screening (HTS) assay technologies (also known as "badly behaving compounds", "bad actors", "nuisance compounds" or "PAINS") pose a major challenge to early-stage drug discovery. Many of these problematic compounds are "frequent hitters", and we have recently published a set of machine learning models ("Hit Dexter 2.0") for flagging such compounds.

Here we present a new generation of machine learning models which are derived from a large, manually curated and annotated data set. For the first time, these models cover, in addition to target-based assays, also cell-based assays. Our experiments show that cell-based assays behave indeed differently from target-based assays, with respect to hit rates and frequent hitters, and that dedicated models are required to produce meaningful predictions. In addition to these extensions and refinements, we explored a variety of additional setups for modeling, including the combination of four machine learning classifiers (i.e. k-nearest neighbors (KNN), extra trees, random forest and multilayer perceptron) with four sets of descriptors (Morgan2 fingerprints, Morgan3 fingerprints, MACCS keys and 2D physicochemical property descriptors).

Testing on holdout data as well as data sets of "dark chemical matter" (i.e. compounds that have been extensively tested in biological assays but have never shown activity) and known bad actors show that the multilayer perceptron classifiers in combination with Morgan2 fingerprints outperform other setups in most cases. The best multilayer perceptron classifiers obtained Matthews correlation coefficients of up to 0.648 on holdout data. These models are available via a free web service.

Introduction

High-throughput screening (HTS) assay technologies are a cornerstone of modern drug discovery. They allow the biological testing of large numbers of compounds on targets of interest within a short period of time [1]. A major challenge faced in high-throughput screening is false-positive hits resulting from different types of assay interference [2].

Compounds causing assay interference are referred to as "badly behaving compounds", "bad actors" or "nuisance compounds". Many of them, but by far not all, are "frequent hitters" (i.e. compounds which show higher-than-expected hit rates in biological assays). This is because not all types of assay interference are frequent events. In fact, many types of assay interference are triggered only by specific conditions.

Importantly, not all frequent hitters are nuisance compounds. Quite on the contrary: frequent hitter behavior can be a result of true promiscuity mediated by "privileged scaffolds" [3]. Privileged scaffolds enable compounds to bind, in a specific manner, to a number of distinct proteins. Such compounds can be particularly useful in the context of polypharmacology and drug repurposing.

An established experimental strategy to discriminate genuine hits from false-positive results is the use of orthogonal and counterscreen assays [4], but even with such an advanced experimental setup some cases of assay interference may not be captured because the underlying mechanisms are manifold.

Given the complexities involved in the conduction and analysis of experimental screens, computational tools to aid the discrimination of genuine hits from false ones are in high demand. Today, a variety of in silico approaches for cherry-picking the most promising hits for follow-up studies are at our disposal [5–10]. We will discuss these briefly in the context of the individual types of assay interference.

The most prominent cause of interference in biological assays (biochemical assays in particular) is related to the formation of aggregates by small molecules that engage in nonspecific interactions with

* Corresponding author.
E-mail address: johannes.kirchmair@univie.ac.at (J. Kirchmair).

https://doi.org/10.1016/j.ailsci.2021.100007
Received 14 June 2021; Received in revised form 5 August 2021; Accepted 6 August 2021
Available online 8 August 2021
2667-3185/© 2021 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

biomacromolecules [5]. Several computational approaches have been reported for the assessment of small molecules with regard to their risk of forming colloidal aggregates. These tools include Aggregator Advisor [11], ChemAgg [12] and SCAM detective [13]. Aggregator Advisor flags potential aggregators based on their molecular similarity to a set of 12,000 known aggregators, taking logP into account. ChemAgg and SCAM Detective are machine learning models for the classification of small molecules into aggregators and non-aggregators. Whereas ChemAgg is based on a XGBoost model, SCAM detective utilizes a set of random forest models.

A second important cause of assay interference is the chemical reactivity of compounds, in particular that related to electrophilicity [14]. Chemically reactive compounds may bind covalently to biomacromolecules or interact with the assay screening technology in an undesired way. Computational approaches for identifying reactive compounds are mostly based on sets of rules which describe substructures that have been linked to chemical reactivity [15].

Further types of assay interference are covered under the umbrella of the well-known pan-assay interference compounds (PAINS) concept [16]. PAINS are compounds based on molecular scaffolds that have been associated with various types of assay interference. PAINS include redox cycling compounds (e.g. toxoflavins), covalent binders (e.g. isothiazolones or ene-rhodanines), membrane disruptors (e.g. curcumin), metal complex-forming compounds (e.g. hydroxyphenyl hydrazones) and unstable compounds (e.g. phenol-sulfonamides) [17]. The molecular fragments linked to PAINS have been compiled in a collection of several hundred structural patterns, and this collection has been implemented in various in silico platforms and software libraries to offer means for flagging potentially problematic compounds [18]. An alternative approach to flagging potential PAINS was recently presented by Koptelov et al. [19]. They use discriminative subgraph mining to identify characteristic patterns in PAINS and non-PAINS, and utilize these patterns, in combination with numerical descriptors, to derive decision tree models for PAINS prediction.

A number of focused machine learning models have been devised for the identification of compounds that likely cause specific types of assay interference. For example, Luciferase Advisor [20] and ChemFluc [21] are models for the prediction of compounds (luciferase inhibitors) that may interfere with luciferase-based assays. InterPred [22] includes a set of QSAR models for the prediction of luciferase inhibitors and autofluorescence compounds in cell-based and target-based assays.

Several computational tools are in existence that predict frequent hitters independent of the underlying mechanisms (genuine promiscuity; various types of assay interference). For example, researchers at AstraZeneca have derived a statistical model for the prediction of frequent hitters based on their in-house historical bioactivity data [23]. Another statistical model for the prediction of frequent hitters is BADAPPLE [24]. In contrast to the AstraZeneca model, the BADAPPLE model is derived from molecular scaffolds rather than complete molecular structures.

More recently, machine learning has been moved into the focus also in the field of frequent hitter and assay interference prediction. For example, Hit Dexter 2.0 [25], developed by some of us, predicts frequent hitters utilizing a set of extra tree models that are trained on large sets of data extracted from the PubChem Bioassay database [26]. More recently, Feldmann et al. [27] reported a machine learning approach for the prediction of true promiscuous compounds (multi-target compounds) in which they removed likely aggregators and other types of assay interference compounds from the training sets in an effort to work with cleaner sets of promiscuous and non-promiscuous compounds.

Whereas a sizable number of in silico models for the prediction of frequent hitters and badly behaving compounds are at our disposal today, most of them have clear limitations with respect to the coverage of mechanisms of interference and assay technologies. In particular, the existing approaches are focused on, or limited to, biochemical (i.e. target-based) assays and do not adequately represent cell-based assays, which can behave very differently with respect to assay interference.

Table 1
Definitions of values for the manually assigned label "target type".

Label value	Description
target-based	Assays generating readouts from purified proteins or peptides
cell-based	Assays generating readouts from cells
other	Any other assays such as tissue-based and organism-based assays

In continuation of the further development of Hit Dexter, we present here a refined set of machine learning models for frequent hitter prediction that cover biochemical assays and, for the first time, also cell-based assays. More specifically, we have developed three types of models: (i) models for target-based assays, (ii) models for cell-based assays designed to measure a specific protein-compound interaction, and (iii) models for an extended selection of cell-based assays, covering also cell-based assays designed to measure nonspecific interactions such as toxicity.

Each of the models is derived from a new, large, high-quality data set that we extracted from the PubChem Bioassay database and annotated manually. In addition to the extra tree (ET) classifiers employed previously, we are now exploring also k-nearest neighbors (KNN) classifiers as baseline models, as well as random forest (RF) and multilayer perceptron (MLP) classifiers. The best models presented in this work are available via a free web service at https://nerdd.univie.ac.at/hitdexter3/ and information on the assay data sets is provided as Supporting Information.

Materials and methods

Data set compilation

PubChem Bioassay data selection and annotation

The PubChem Bioassay Database [28–30] was queried for all assays with measured bioactivity data reported for at least 10,000 compounds (i.e., compounds with unique PubChem Compound IDs, CIDs). The data for the selected assays were downloaded and the labels "target type" and "bioactivity type" were assigned manually to each of these assays according to the definitions provided in Tables 1 and 2.

Following manual assay labeling, three different data sets were compiled:

- target-based assay data set: includes all data from assays with "target type" = "target-based" (which implies "bioactivity type" = "specific bioactivity")
- cell-based assay data set: includes all data from assays with "target type" ="cell-based" AND "bioactivity type" = "specific bioactivity"
- extended assay data set: includes, in addition to the data included in the cell-based assay data set, all data from assays with "target type" = "cell-based"

The individual assays of the target-based assay data set were checked for the availability of Protein Gene Identifier (GI) information, which is utilized to retrieve protein sequence information from the NCBI Protein database [31] (the protein sequence information will be required, in a later step, for protein clustering and to ensure a diverse protein set). Sixty-six assays of the target-based assay data set had no GI or multiple GI annotations and were hence removed. In addition, seven assays of the target-based assay data set, four assays of the cell-based assay data set, and seven assays of the extended cell-based assay data set were removed because of disproportionally high hit rates (i.e. hit rates in excess of the average hit rate plus three standard deviations (σ), calculated over all assays of the respective data set). For the target-based assay data set, the six assays with the highest hit rates are all measuring CYP P450 enzyme activity. In the case of the cell-based assay data set, this concerns four assays, with hit rates of 59%, 55%, 17% and 15% (note that for approximately three quarters of the assays included in the cell-based assay data set their hit rates are below 1%). For the extended cell-based assay data set the seven assays with hit rates above 16% were removed.

Table 2
Definitions of values for the manually assigned label “bioactivity type".

Label value	Description
specific bioactivity	Assays designed to measure a specific biological property such as the activity of an enzyme. Cytotoxicity assays are not included in this category. Counterscreen assays are included if they measure a specific biological effect. An example of a counterscreen assigned this label value is a luciferase counterscreen that is commonly employed to identify compounds which can cause interference in luciferase-based (bioluminescence) assays
nonspecific bioactivity	Assays that measure cell growth, cell viability, cytotoxicity, cell growth inhibition, or other nonspecific assay readouts
other	Assays that measure physicochemical processes (not bioactivities), DNA or RNA binding, etc.

Table 3
Data set sizes and compounds removed during chemical structure processing.

	Target-based assay data set	Cell-based assay data set	Extended cell-based assay data set
No. compounds in the data set prior to chemical structure processing	1,545,406	1,421,472	1,858,887
No. compounds removed due to invalid SMILES	1	3	9
No. compounds removed due to lack of a single, valid activity outcome[1]	45,184	23,259	53,984
No. compounds removed due to presence of elements uncommon to drug-like compounds	331	381	3151
No. compounds removed by the molecular weight filter	10,847	11,120	22,106
No. compounds in the final data set	1,489,043	1,386,709	1,779,637

[1] Compounds that were removed because of the lack of a valid activity outcome that can be derived from the raw data (i.e. compounds without a single annotated “Active” or “Inactive” assay outcome)

As the last filtering criterion, any assays without at least one compound measured as active and one compound measured as inactive were removed from the data set. For a complete overview of all assays removed during data preparation see Table SI_1.

Chemical structure processing

The SMILES notations of the 1,545,406 compounds covered by the target-based assay data set, the 1,421,472 compounds covered by the cell-based assay data set, and the 1,858,887 compounds covered by the extended cell-based assay data set were retrieved from the PubChem Bioassay database via the PubChem PUG REST interface [32]. The ChEMBL Structure Pipeline [33] (also known as “ChEMBL Compound Curation Pipeline”), was utilized to (i) neutralize charged molecules, (ii) remove salt and solvent components, and (iii) neutralize charged molecules once more (to cover cases where a charged component was removed during step ii). The technical description of this chemical structure preparation procedure is reported in Ref. [33].

Any compounds with molecular weight below 180 or above 900 Da were removed from the data set, as well as any compounds composed of any elements other than H, B, C, N, O, F, Si, P, S, Cl, Se, Br and I. Molecules represented by more than one tautomer were merged to a single representation using the “canonalize” method implemented in the “TautomerEnumerator” class of RDKit [34] (version 2020.09.1). During this procedure the compounds were represented as RDKit molecules and were in a last step converted to canonical SMILES. Further duplicate compounds were removed based on identical SMILES. For an overview of the removed compounds see Table 3. For all additional data sets used within this study, including the ChEMBL 23 database [35], the dark chemical matter (DCM) data set compiled by Wassermann et al. [36], the data set of Dahlin et al. [37] (containing compounds that are known to cause interference in biological assays), and the data set of Borrel et al. [22] (containing compounds that were experimentally confirmed to cause false positive readouts in bioluminescence assays due to luciferase inhibition and/or autofluorescence), the same chemical structure standardization process was performed. Since the data set of Borrel et al. contains only CAS numbers as compound identifiers, the SMILES notations were fetched via the Chemical Identifier Resolver [38].

Extraction of activity data from the selected assays

For each of the selected assays, any compounds consistently (i.e. one or several times) labeled as “Active” were defined as active, and any compounds consistently labeled as “Inactive” were defined as inactive. Any compounds with contradicting assay outcomes (e.g. “Active” and “Inactive”, or “Active” and “Inconclusive”) were removed. A compound is treated as active on a cluster of proteins (see “Protein clustering”) if it is active on at least one protein of that cluster.

In order to ensure the consistency of predictions, compounds with identical Morgan2 fingerprints [39,40] (1024 bits) but differing promiscuity labels (e.g., symmetric molecules) were removed from the respective training set. For any compounds with identical Morgan2 fingerprints only one instance was kept in the respective training set.

Definition of the active-to-tested ratio (ATR)

The hit rate of a compound in biological assays is described as the active-to-tested ratio (ATR; Eq. (1)):

$$ATR = \frac{A}{T}, \tag{1}$$

where A is the number of assays a compound was tested active and T is the total number of assays a compound was tested in. For compounds, the terms hit rate and ATR are used interchangeably in this work.

Protein clustering

Based on the GIs assigned to the individual proteins, the FASTA sequences of the respective proteins were retrieved from the NCBI using the “Entrez” package of Biopython [41] (version 1.78). Protein clustering was performed using cd-hit [42] with the same parameters described in Ref. [25] (sequence identity= 60%; tolerance= 3). This resulted in 273 protein clusters using 296 unique proteins for the target-based assay data set.

Model development and hyperparameter optimization

Prior to model development, a random, stratified split of the data into a training set (90%) and a testing set (10%) was performed with the “train_test_split” method of the “model_selection” module of scikit-learn [43] (version 0.23.2). All models were trained and optimized on the training set. The final models were tested on the test set.

Morgan fingerprints and MACCS keys were calculated with RDKit, whereas 206 2D physicochemical property descriptors (meaning the complete set of available 2D descriptors) were calculated with the Molecular Operating Environment [44] (MOE; version 2020.09).

Default parameters were employed for generating machine learning models for the selection of a suitable set of descriptors, with the following exceptions: For the KNN classifier, the number of nearest neighbors

to be taken into account for prediction (n_neighbors) was set to 1; for the RF and the ET classifiers, the class weight (class_weight) was set to "balanced"; for the MLP classifier (implemented in scikit-learn), the number of iterations was set to 1000 as some of the calculations did not converge within the default, 200 iterations.

The generation of the individual models was repeated for ten times, with different random states (i.e. 42 to 51), in order to compute the median and the variance of the performance metrics (details provided in the Results section). The final models were generated with random state=42 and the application of the synthetic minority oversampling technique (SMOTE version 0.7.0) [45].

Performance measurements and variance estimation

The MCC (Eq. (2)) was used as the primary measure of model performance. The MCC is a balanced metric that takes the true positive (TN), false positive (FP), true negative (TN) and false negative (FN) instances into account:

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}} \quad (2)$$

The MCC returns values between -1 (total disagreement between prediction and observation) and +1 (perfect agreement).

The area under the (receiver operating characteristic) curve (AUC) was used as an indicator of the ranking performance of the models.

The tests for statistical significance were performed with the "ttest_rel" function of the "scipy.stats" module. The variance in the performance of the models (on the test data) was estimated by testing the models on ten randomly compiled subsets (80%) of the original test set.

Results

Analysis, annotation and refinement of PubChem Bioassay data

In order to develop a better understanding of the relevance of the data available from the PubChem Bioassay database for modeling the frequent hitter behavior of small molecules, we conducted a comprehensive analysis of the chemical and biological data.

With more than 297 Million measured bioactivities, the PubChem Bioassay database is the world's largest, public collection of bioassay data [30]. It is also one of only a few data resources offering access to a large amount of high-throughput screening data. The number of measured bioactivities recorded per assay varies greatly across the individual assay data sets, from a single compound to 646,275 compounds (Table 4).

We decided to base our work on the 1180 (i.e. 474 + 706) assay data sets containing measurements for at least 10,000 compounds because these data sets offer a good trade-off between data quality and coverage. The vast majority of these data sets have been generated by the most reputable HTS facilities (including the Scripps Research Institute, the Sanford-Burnham Medical Research Institute, The Broad Institute of MIT and Harvard, and the NIH/National Center for Advancing Translational Sciences (NCATS)), for which reason a high standard in HTS can be expected.

Table 4
Size of the PubChem Bioassay data sets.[1]

Number of assays in the PubChem Bioassay database	Number of measured compounds
587,477	1
633,294	2 to 99
5082	100 to 999
1403	1000 to 9999
474	10,000 to 99,999
706	100,000 to 646,275 (maximum)

[1] Numbers referring to the raw, unprocessed PubChem Bioassay database.

In preparation of model development, we manually annotated the 1180 assay data sets according to the "assay type" (i.e. target-based, cell-based, other; see Table 1 for exact definitions) and "bioactivity type" (i.e. specific bioactivity, nonspecific bioactivity, other; see Table 2 for exact definitions). Models for the prediction of frequent hitters in biochemical (i.e. target-based) assays will be built on all (359) assay data sets labeled as "target-based" (which implies the "bioactivity type" value "specific bioactivity") and annotated with exactly one Protein Gene Identifier (GI; the GI will be utilized later to obtain protein sequence information to quantify the relatedness of proteins; the requirement for assays to be assigned exactly one GI ensures that the assay is designed to measure one particular protein of interest). Similarly, models for cell-based assays designed to measure a specific activity will be built on all (369) assay data sets labeled as "cell-based" AND "specific bioactivity". Models will also be derived from an extended set of cell-based assays that includes data from an additional 250 cell-based assays labeled "nonspecific bioactivity". These additional, cell-based assays measure non-specific properties such as cell viability or cytotoxicity. A list of the Assay Identifiers (AIDs) for the three assay data sets is provided in Table SI_2.

We also set steps to address two important biases in the assay data set collection. The first bias results from assays with unusually high hit rates. In target-based assays, high hit rates are often related to the measurement of highly promiscuous proteins such as CYP enzymes. In cell-based assays, high hit rates can be related, for example, to cytotoxicity or high assay sensitivity. Compounds which have been measured, for whatever reason, in several of these assays may, in consequence, be identified as frequent hitters, regardless of whether their activities are focused on a number of closely related proteins or observed across a range of distinct proteins.

The average hit rate of the 359 target-based assays is 0.009. However, a small number of assays has much higher hit rates, up to 0.252 (Fig. 1). Similarly, the average hit rate for the 369 cell-based assays is 0.014, with a small number of assays having much higher hit rates, up to 0.588. For the extended set of 619 cell-based assays, the average hit rate is 0.023, with the maximum at 0.588. For the reasons discussed above we decided to remove any assays with hit rates exceeding the average hit rate plus three σ. This concerned seven, four and seven assays of the target-based, cell-based and extended cell-based assay data set, respectively.

The second bias is introduced by groups of assays measuring related proteins. Related proteins have a high likelihood of binding the same small molecules, meaning that, for example, assay data sets with a strong representation of protein kinase targets will likely show high hit rates for protein kinase inhibitors. Models for frequent hitter prediction that are trained on such data would likely flag any kinase inhibitor as a frequent hitter, which is not the intended behavior of these models.

In order to address the bias introduced by the overrepresentation of groups of related proteins, we clustered the target-based assay data set according to the amino acid sequences of the target proteins (note that the clustering was not performed for the cell-based assays data sets because cell-based assays may report activities for a number of different proteins). More specifically, all data sets related to proteins with an amino acid sequence identity exceeding 60% were merged into a cluster (see Materials and Methods for details). This clustering procedure resulted in 296 protein clusters (starting from 352 proteins covered by the target-based assay data set).

After addressing the two important biases, in the final processing step the molecular structures contained in the data sets were processed and checked for correctness. Any problematic instances were removed, as outlined in Table 3 and described in the Materials and Methods section in full detail. This resulted in a target-based, a cell-based and an extended cell-based assay data set consisting of 1,489,043, 1,386,709 and 1,779,637 unique compounds with at least one confirmed target protein, respectively.

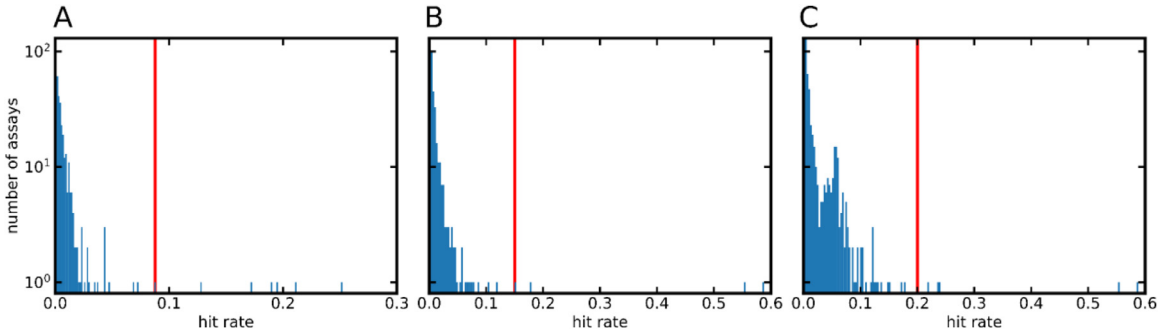


Fig. 1. Histograms (200 bins each) showing the hit rates of the assays included in the (A) target-based, (B) cell-based, and (C) extended cell-based assay data sets. The red line marks the mean hit rate + 3σ. Note that the scales of the x-axes differ for the three diagrams.

Table 5
Composition of the training and test sets.

Data set	Promiscuity class	Class definitions	No. compounds in the training set	No. compounds in the test set
target-based assay data set	HPROM[1]	ATR > 0.053	4614	550
	PROM	ATR > 0.022	20274	2303
	NPROM	ATR < 0.007	219061	24483
cell-based assay data set	HPROM[1]	ATR > 0.058	5578	616
	PROM	ATR > 0.025	24913	2825
	NPROM	ATR < 0.008	226382	25427
extended cell-based assay data set	HPROM[1]	ATR > 0.070	5135	538
	PROM	ATR > 0.030	24673	2776
	NPROM	ATR < 0.010	235241	26398

[1] The compounds labeled as HPROM are a subset of the compounds labeled as PROM.

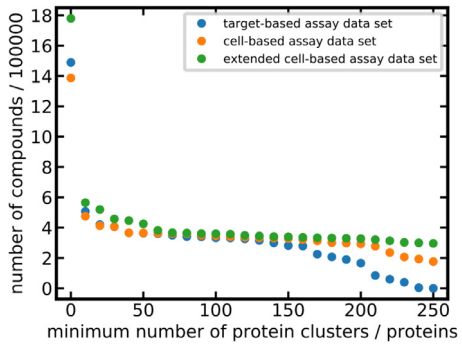


Fig. 2. Data set size (number of compounds) as a function of the minimum number of protein clusters (in the case of target-based assays) or proteins (in the case of cell-based assays) for which measured data are available.

Analysis of compound hit rates and assignment of promiscuity class labels

The ATR (Eq. (1)) can be used to assign categorical promiscuity values to compounds, such as “non-promiscuous”, “promiscuous” or “highly promiscuous”. The significance and robustness of the ATR depends on the quality and quantity of the underlying data: the higher the value of T (i.e. the total number of assays a compound was tested in), the more robust the ATR. The main advantage of the ATR over alternative metrics is its interpretability as it reflects the hit rate of a compound.

In this work, we set the minimum threshold of T for a compound to be included in the data sets used for model development to 100, which represents a good balance between ATR quality and coverage (Fig. 2). This filtering procedure resulted in a set of 332,653 compounds measured in target-based assays, 345,743 compounds measured in cell-based assays designed to measure a specific bioactivity, and 360,094 compounds measured in an extended set of cell-based assays.

Based on the ATR thresholds reported in Table 5, all compounds were assigned a promiscuity label: highly promiscuous (HPROM), promiscuous (PROM) or non-promiscuous (NPROM). According to these definitions, roughly 2% of the compounds are labeled HPROM across the three assay data sets. Likewise, the percentages of compounds labeled PROM were around 9% across the three assay data sets (note that all HPROM compounds are also part of the PROM subset). The percentages of compounds labeled NPROM are approximately 90% across the three assay data sets (Table 5 and Fig. 3).

To obtain a training set and a test set (separately for all three data sets), a stratified random split was performed to obtain 90% training data and 10% test (hold out) data. Following a fingerprint-based data merging procedure (i.e. merging of instances having identical fingerprints and identical class labels, and removal of any instances having identical fingerprints but conflicting class labels; see Materials and Methods for details) the target-based, cell-based and extended cell-based training sets contain 243,949, 256,873 and 265,049 compounds, respectively (Table 5).

As shown in Table 5, the average ATR across the extended set of cell-based assays is higher than for the cell-based and the target-based assay sets, suggesting that non-specific interactions are likely to play an important role in the assays exclusive to the extended set of cell-based assays (i.e. cell-based assays not designed to measure specific biological processes but to capture properties such as cell-viability and cytotoxicity).

Analysis of the chemical space covered by the training sets

The chemical space covered by the training set is a decisive factor for the applicability domain of a model. In order to obtain an understanding of the relevance of our three training sets to early drug discovery we run a pairwise comparison of the molecular structures included in these training sets and all molecular structures included in the ChEMBL database. Fig. 4 shows the distributions of the pairwise, maximum Tanimoto coefficients based on Morgan2 fingerprints (with a length of 1024 bits) for the three data sets vs. the ChEMBL database. The distributions are similar for the three data sets, with approximately

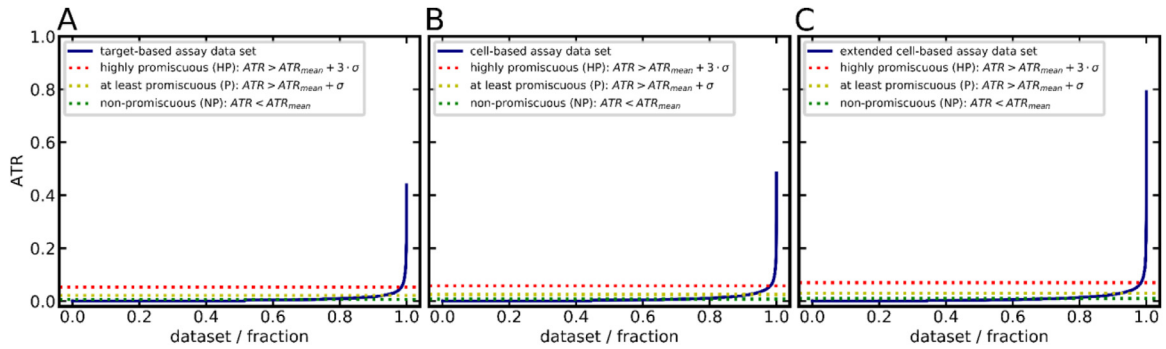


Fig. 3. ATR distribution among compounds of the (A) target-based assay data set, (B) cell-based assay data set, and (C) extended cell-based assay data set.

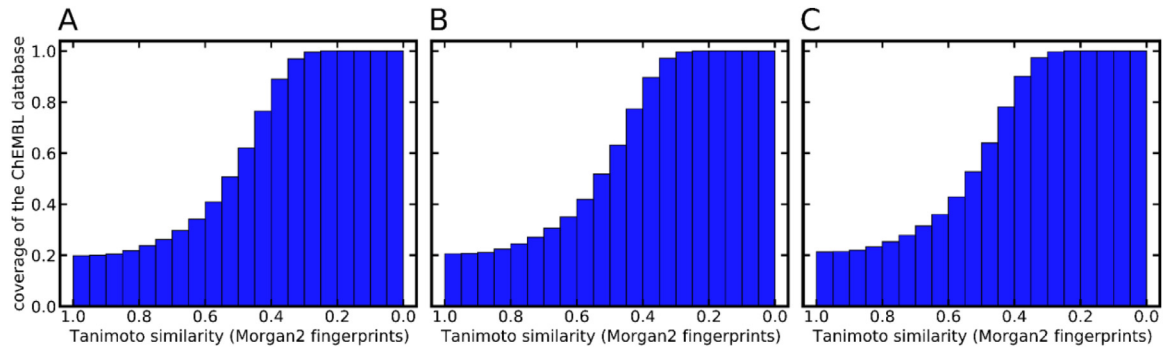


Fig. 4. Cumulative coverage of the compounds included in the ChEMBL database by the compounds included in the (A) target-based assay data set (B) cell-based assay data set, and (C) extended cell-based assay data set.

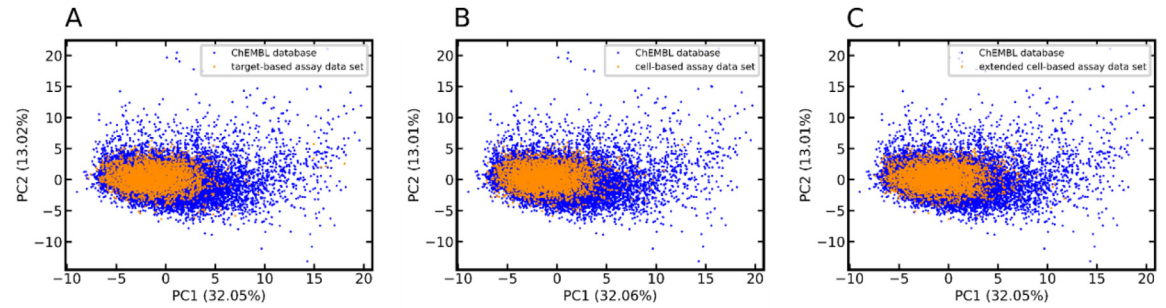


Fig. 5. PCA of the ChEMBL database and the (A) target-based assay set, (B) cell-based assay set, and (C) extended cell-based assay set. The PCA is derived from the 44 2D molecular property descriptors (see Table S1 in Ref. [46]) implemented in MOE. For the sake of clarity, only 1% of the data points (randomly selected) are visualized. The numbers in parentheses report the variance explained by the respective principal component (PC).

50% of the compounds in the ChEMBL database represented by at least one compound in the respective training set with a Tanimoto coefficient of 0.5 or higher.

The Principal Component Analysis (PCA) scatter plots presented in Fig. 5 show that the areas in chemical space that are most densely populated with the compounds from the ChEMBL23 database are also well represented by the assay data sets used for model training. However, there are a significant number of compounds included in the ChEMBL database that are chemically distinct from those represented by the training sets. These are in particular compounds with PC1 values greater than 10, which account for 2.5% of the total number of compounds of the ChEMBL database. Visual inspection of these compounds reveals that they are unusually large, with molecular weight between 575 and 900 Da.

The target-based and the cell-based assay data sets (training data only) have an overlap of 180,278 compounds (representing 75% of the target-based and 72% of the cell-based assay data set, respectively). Only 13,045 (7%) of these compounds have contradicting promiscuity labels (with HPROM treated as a subset of PROM). At first sight the level of agreement between readouts from target-based and cell-based assays seems surprisingly high. However, a closer look reveals that the agreement stems primarily from compounds consistently labeled as NPROM. Among the 20,481 compounds present in both data sets and labeled as PROM in at least one of them, only 6616 (32%) have identical class labels. This indicates that target-based and cell-based assays perform indeed differently and that they should be represented by dedicated models.

Development of machine learning models for compound promiscuity prediction

Two types of classifiers were generated for the target-based, cell-based and extended cell-based assay data sets: classifiers discriminating HPROM from NPROM compounds, and classifiers discriminating PROM from NPROM compounds.

Identification of the best setup for model generation

In order to identify the best setup for model generation we tested all possible combinations of four machine learning algorithms (i.e. KNN,

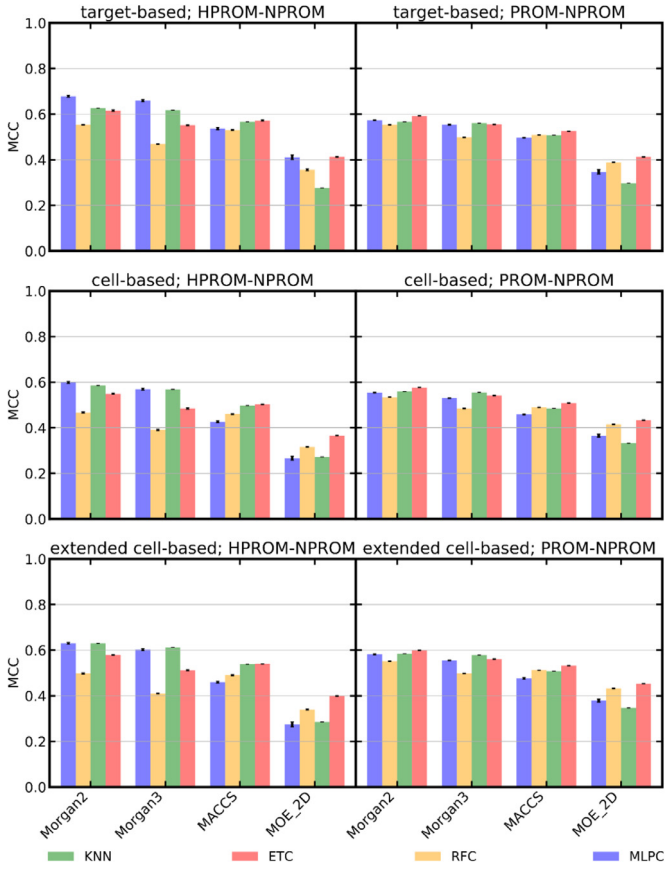


Fig. 6. Performance (quantified as MCC) of machine learning models trained on different types of descriptors. The variance of the ten experiments (each using a distinct random seed between 42 to 51; see Materials and Methods for details) is indicated by error bars.

ET, RF, MLP) and four sets of descriptors (i.e. Morgan2 and Morgan3 fingerprints, each of 1024 bits in length, MACCS keys, and the complete set of 206 2D physicochemical property descriptors implemented in MOE, referred to as "MOE_2D") within a 10-fold cross-validation framework. For each setup ten of these cross-validation experiments were performed using distinct random seeds. This allowed, for each setup, the calculation of a standard deviation that is independent of the cross-validation.

As shown in Fig. 6, the task of discriminating HPROM from NPROM compounds (MCCs of up to 0.679) is simpler than that of discriminating PROM from NPROM compounds across the three assay data sets (the MCC of the best HPROM-NPROM classifier, 0.679, is significantly higher than that of the best PROM-NPROM classifier, 0.599; p-value 2.48×10^{-12}). This is expected because of the larger ATR margin between the HPROM and the NPROM class (margin of 3σ) than between the PROM and the NPROM class (margin of 1σ). No substantial differences in model performance were observed with respect to the type of assay modeled: the best setups yielded comparable MCCs for the target-based assay set (MCCs 0.679 and 0.592 for HPROM-NPROM and PROM-NPROM classification, respectively), cell-based assay set (MCCs 0.602 and 0.577, respectively), and extended cell-based assay set (MCCs 0.631 and 0.599, respectively).

The differences in model performance that can be attributed to the model algorithms are rather small, on average 0.104 in MCC. The maximum difference in MCC observed for any model trained on identical input (i.e. same data set and same descriptor set) was 0.224. Overall, the MLP classifiers performed best in HPROM-NPROM classification (the MCC of the best MLP classifier, 0.679, is significantly higher than that of the second-best model, a KNN model that obtained an MCC of 0.630; p-value of 8.81×10^{-11}), and the ET classifiers performed best in PROM-NPROM classification (the MCC of the best ET classifier, 0.599, is significantly higher than that of the second-based model, a KNN model that obtained an MCC of 0.585; p-value of 7.79×10^{-11}). Interestingly, in this cross-validation scenario the simple one-nearest neighbor approach performed almost as well as the more complex machine learning algorithms (MCCs of up to 0.587; p-value of 7.79×10^{-11} against the best MLP classifier).

In contrast to what we observed for the machine learning algorithms, the differences in model performance that can be attributed to the molecular descriptors were, in part, substantial. On average, the best performance was obtained by models trained on Morgan2 fingerprints (MCC averaged over all models trained on Morgan2 fingerprints: 0.679). They were closely followed by the models based on Morgan3 fingerprints (MCC averaged over all models trained on Morgan3 fingerprints: 0.659; the difference in the average MCC of models trained on Morgan2 and Morgan3 fingerprints is significant, with a p-value of 1.10×10^{-79}). The MOE_2D physicochemical property descriptors and the MACCS keys yielded models that are clearly inferior, with MCCs not exceeding 0.453 and 0.572, respectively.

The highest MCC during model optimization (0.679) was obtained by the HPROM-NPROM MLP classifier in combination with Morgan2 fingerprints (the MCC of the second-best classifier, which is the respective model trained on Morgan3 fingerprints, was 0.659; the difference in the MCCs is significant, with a p-value of 3.98×10^{-5}).

Hyperparameter optimization

Focusing now on Morgan2 fingerprints, in the next phase of model development we optimized the hyperparameters of the individual algorithms (i.e. KNN, ET, RF and MLP). More specifically, we conducted a grid search within a 10-fold cross-validation framework to identify the hyperparameters yielding the best performing models for a particular combination of machine learning algorithm and descriptors in terms of MCC (averaged over the respective HPROM-NPROM and PROM-NPROM classifiers for the three assay data sets). An overview of the explored hyperparameters and value ranges, as well as the selected hyperparameter values, is provided in Table 6.

The impact of individual hyperparameter settings on model performance is generally small (Table SI_3). The largest improvement in MCC observed during hyperparameter optimization was 0.037 (for the PROM-NPROM MLP classifier trained on the cell-based assay data set; the optimized classifier performed significantly better than the classifier using default hyperparameters; p-value of 1.01×10^{-10}). The AUC values improved consistently with the MCCs (Table SI_3), except for KNN, for which the MCCs increased with fewer numbers of neighbors while the AUC values decreased. In the case of the RF and ET classifiers, gains in performance beyond 200 estimators were marginal and do not justify the additional demands in computational power and memory. The same is true for the MLP classifier, for which we identified 250 as the most suitable number of perceptrons for our purposes.

The best of all models (an HPROM-NPROM MLP classifier for target-based assays; single hidden layer with 250 perceptrons; activation function relu) yielded an MCC of 0.686 (the optimized classifier performed significantly better than the classifier using default hyperparameters; p-value of 3.79×10^{-3}). The models chosen from hyperparameter optimization are listed in Table 7.

Model performance as a function of the size of the training set

In order to determine the impact of the size of the training set on model performance we trained and tested the optimized HPROM-NPROM and PROM-NPROM MLP classifiers on fractions of 0.01 to 1.00 of the full training sets (within a 10-fold cross-validation framework). From Fig. 7 it is observed that models built on just 20% of the data already achieve good performance (MCCs between 0.434 and 0.524). Larger data sets may add significant value but primarily if they cover

Table 6
Overview of hyperparameters optimized during grid search within a 10-fold cross-validation framework.[1]

Classifier	Parameter	Values
KNN	n_neighbors (number of neighbors considered)	1, **3**, 5, 10
RF,	n_estimators (number of trees)	50, 100, **200**, 300, 400, 500
ET	max_features (features taken into account for best split search)	'sqrt', 'none', **'0.2'**, '0.4', '0.6', '0.8'
MLP	hidden_layer_sizes (number of perceptrons per layer)[2]	50, 100, **250**, 500
	hidden_layer_sizes (number hidden layer)[2]	**1**, 2, 3, 4, 5
	activation (activation function)	**'relu'**, 'tanh', 'logistic'

[1] The hyperparameter values indicated in bold are those we identified as most suitable for model building. These values were used for the generation of the final models.
[2] hidden_layer_sizes accepts two values: one for the number of perceptrons per layer and one for the number of hidden layers.

Table 7
Cross-validation and test set performance of the best models of different types.

Data	Classification	Machine learning algorithm	Cross-validation performance[1]					Test set performance				
			MCC[2]	AUC[2]	Balanced accuracy	Sensitivity	Specificity	MCC	AUC	Balanced accuracy	Sensitivity	Specificity
target-based assay data set	HPROM-NPROM	KNN	0.624	0.843	0.733	0.469	0.998	0.376	0.909	0.871	0.818	0.924
		ET	0.630	0.964	0.734	0.469	0.998	0.508	0.966	0.677	0.357	0.997
		RF	0.588	0.964	0.695	0.392	0.999	0.484	0.965	0.677	0.358	0.996
		MLP	0.686	0.946	0.796	0.595	0.997	0.648	0.949	0.798	0.601	0.995
	PROM-NPROM	KNN	0.587	0.844	0.745	0.506	0.984	0.412	0.864	0.816	0.822	0.809
		ET	0.597	0.928	0.746	0.504	0.986	0.518	0.910	0.721	0.464	0.977
		RF	0.578	0.929	0.718	0.445	0.991	0.518	0.908	0.731	0.489	0.973
		MLP	0.599	0.907	0.777	0.578	0.975	0.580	0.899	0.768	0.562	0.974
cell-based assay data set	HPROM-NPROM	KNN	0.571	0.827	0.704	0.41	0.998	0.338	0.899	0.857	0.812	0.902
		ET	0.572	0.950	0.697	0.395	0.998	0.531	0.940	0.692	0.387	0.997
		RF	0.514	0.947	0.651	0.303	0.999	0.520	0.932	0.692	0.387	0.996
		MLP	0.611	0.929	0.754	0.512	0.996	0.576	0.915	0.767	0.541	0.992
	PROM-NPROM	KNN	0.566	0.845	0.74	0.501	0.979	0.413	0.860	0.809	0.834	0.783
		ET	0.593	0.925	0.747	0.511	0.983	0.551	0.911	0.743	0.513	0.973
		RF	0.572	0.925	0.717	0.445	0.989	0.543	0.910	0.747	0.525	0.968
		MLP	0.579	0.910	0.77	0.571	0.969	0.561	0.901	0.764	0.562	0.965
extended cell-based assay data set	HPROM-NPROM	KNN	0.600	0.842	0.721	0.443	0.998	0.340	0.895	0.858	0.798	0.919
		ET	0.599	0.956	0.708	0.417	0.999	0.527	0.944	0.683	0.368	0.998
		RF	0.537	0.956	0.662	0.325	0.999	0.519	0.943	0.686	0.374	0.997
		MLP	0.639	0.939	0.764	0.532	0.997	0.567	0.921	0.753	0.511	0.994
	PROM-NPROM	KNN	0.586	0.854	0.749	0.516	0.981	0.429	0.871	0.819	0.835	0.804
		ET	0.618	0.935	0.757	0.529	0.985	0.565	0.923	0.742	0.506	0.978
		RF	0.590	0.934	0.725	0.459	0.990	0.554	0.920	0.748	0.523	0.973
		MLP	0.607	0.921	0.783	0.593	0.972	0.587	0.910	0.781	0.596	0.967

[1] The optimized hyperparameters are reported in Table 6.
[2] The variance is reported in Table SI_3.

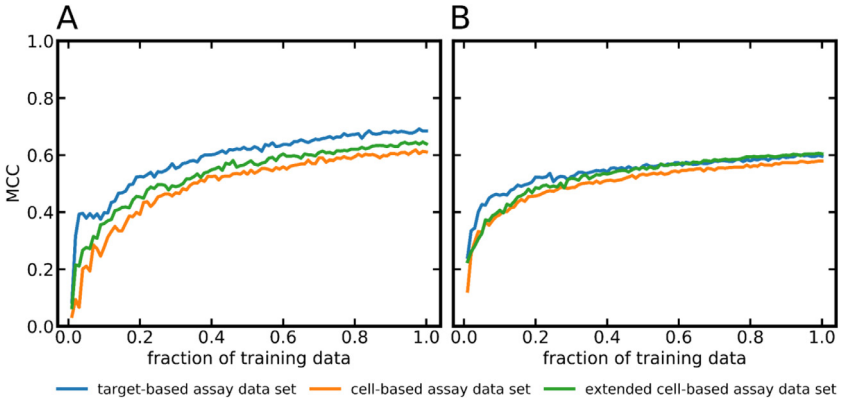


Fig. 7. Performance (quantified as MCC) of the hyperparameter-optimized MLP classifiers trained on the target-based, cell-based and extended cell-based assay data set as a function of training set size. (A) HPROM-NPROM classifiers, (B) PROM-NPROM classifiers. For each data point the variance was calculated from ten calculations (with different random seeds; see Materials and Methods for details). Because the variance values were within the range of 1.4×10^{-6} to 5.9×10^{-4} they are not visualized in these graphs.

distinct areas in the chemical space and hence contribute to the extension of the applicability domain of the model.

Evaluation of the final machine learning models

A total of 24 final models of different types (i.e. models trained on the full training set, balanced with SMOTE; see Materials and Methods for details) were tested on the holdout data (i.e. 10% of the data that was set aside prior to model building). The 24 models result from the combination of three different training sets (i.e. target-based, cell-based and extended cell-based assay data set), four machine learning algorithms (KNN, ET, RF, MLP), and two different types of classification (i.e. HPROM-NPROM and PROM-NPROM). All of these models are built on Morgan2 fingerprints and utilize the

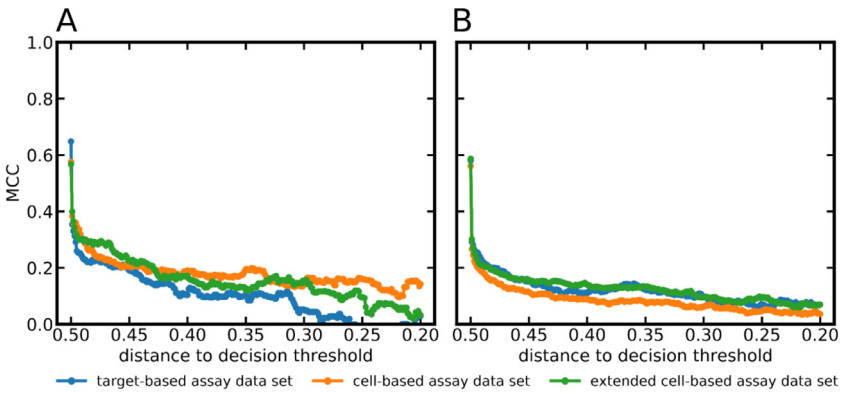


Fig. 8. Performance (quantified as MCC) of (A) the HPROM-NPROM MLP classifiers and (B) the PROM-NPROM MLP classifiers as a function of the distance of the predicted class probability to the decision threshold (the decision threshold applied to all models in this study is 0.5). For each data point the variance was estimated by running the models on ten randomly selected subsets of the test data (see Materials and Methods for details). Because the variance values were within the range of 9.6×10^{-6} to 4.5×10^{-3} they are not visualized in these graphs.

hyperparameters sets optimized during the previous cross-validation experiments.

Model performance on the test set

The average MCC obtained by the 24 models on the respective test sets was 0.507, which is 0.087 lower than in the cross-validation scenario (Table 7). Overall, the decrease in performance (on the test set compared to cross-validation) was more pronounced for the HPROM-NPROM classifiers (average decline in MCC 0.103) than the PROM-NPROM classifiers (average decline 0.070). The steeper drop in performance observed for the HPROM-NPROM classifiers is likely related to the fact that the number of compounds representing the active class is much lower for the HPROM-NPROM training set (approximately 5000 compounds) than for the PROM-NPROM training set (approximately 23,000 compounds). The best MCC among all HPROM-NPROM classifiers was obtained by the MLP classifier trained on the target-based assay data set (MCC 0.648). The best-performing PROM-NPROM classifier was the MLP classifier trained on the extended cell-based assay data set (MCC 0.587).

Importantly, a substantial decrease in performance was observed for the HPROM-NPROM KNN classifiers for all three assay data sets (for example, the KNN classifier of the target-based assay data set; three nearest neighbors; cross-validation MCC 0.624; test set MCC 0.376) and also the PROM-NPROM KNN classifiers for all three assay data sets (for example, the KNN classifier of the target-based assay data set; three nearest neighbors; cross-validation MCC 0.587; test set MCC 0.421). This decline in the performance may be related to model overfitting.

In contrast to the observations made for the KNN, the MCC values obtained by the RF and ET classifiers remained stable. The maximum decline in MCC observed for these models was 0.122. The MLP classifiers showed the most robust performance across the three data sets and the two types of classifications (i.e. HPROM-NPROM and PROM-NPROM), with a maximum decline in MCC of 0.072. For this reason these six MLP classifiers were selected to form the Hit Dexter 3 set of machine learning models and they were investigated further regarding their applicability domains.

Prediction success as a function of the distance of the predicted class probability from the decision threshold

Commonly, a directly proportional relationship is observed between the reliability of class assignments and the distances between the predicted class probabilities and the decision threshold. This holds true also for the Hit Dexter 3 models. Fig. 8 shows that class assignments based on predicted probabilities close to 0 or close to 1.0 (this corresponds to a distance to the decision threshold of approximately 0.5 as we apply a decision threshold of 0.5 in all cases) are particularly reliable (MCC values of up to 0.648) for the Hit Dexter 3 models. The MLP classifiers differentiating PROM and NPROM compounds for the three data sets report predicted class probabilities greater than 0.95 or smaller than 0.05 for on average 97% of the compounds in the test set.

Prediction success as a function of the distance of the test compounds to the training set

It is expected that test compounds that are structurally dissimilar from those represented by the training data pose greater challenges to the model than those that are structurally related. Fig. 9 shows that the Hit Dexter 3 models perform well for compounds represented by at least one molecule in the training set that is structurally related (i.e. having at least one compound in the training set for which the pairwise Tanimoto coefficient based on Morgan2 fingerprints is at least 0.7). Predictions for compounds that are more dissimilar to those represented in the training data are less reliable and should be considered with the necessary caution.

Prediction success as a function of the applied decision threshold

In the current context, the decision threshold applied to a classifier decides on when a compound is classified as a frequent hitter or as a non-promiscuous compound. The default value for the decision threshold is 0.5. There are some use cases where a different decision threshold may be preferred. For example, in cases where the detection of frequent hitters is a priority (i.e. prioritization of sensitivity over specificity), a lower decision threshold may result in better predictions. Fig. 10 visualizes the effects of changes in the decision threshold on the MCC, balanced accuracy, sensitivity and specificity. The fact that the curves remain fairly stable until the decision threshold approaches extreme values (i.e. values close to 0.0 or 1.0) indicates that the classifiers produce clear predictions for most compounds. In cases where sensitivity is of primary importance, users are advised to consider any compounds with predicted probabilities greater than 0.0 as potential frequent hitters.

Predicting frequent hitters in cell-based assays with models trained on data from target-based assays and vice versa

Compounds may behave differently in target-based and cell-based assays, in particular also with regard to their assay interference and frequent hitter behavior. In order to obtain a better understanding of the relevance and value of dedicated models for the prediction of frequent hitters in target-based and cell-based assays, we compared the performance of the MLP classifiers on test data of the same assay domain to their performance on test data of the other assay domain (i.e. classifiers trained on target-based assay data were tested on cell-based assay test set and vice versa).

As shown in Fig. 11, the PROM-NPROM MLP classifiers trained and tested on data from the same assay domain clearly outperformed those trained on the other domain. The graphs also indicate that the difference in performance is not the result of differences in the chemical space covered by the individual data sets: even for test compounds that are structurally closely related to those represented by the training set, models trained on target-based assay data do not perform well on cell-based assay data and vice versa.

For the cell-based assay test data, the MCC of the PROM-NPROM MLP classifier trained on target-based assay data was just 0.189 (vs.

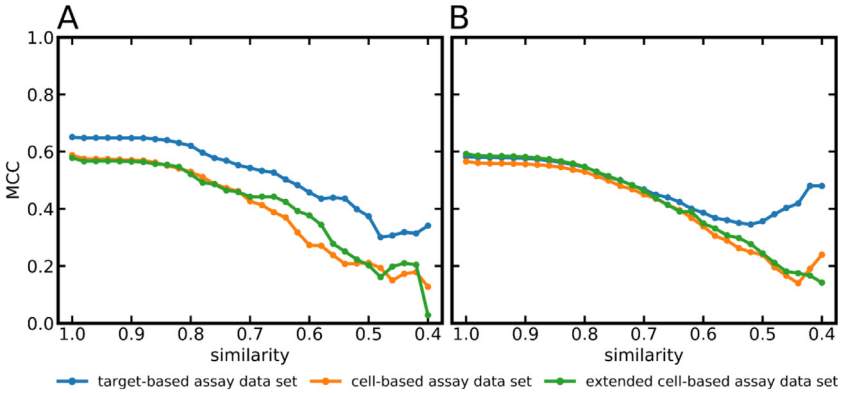


Fig. 9. Performance (quantified as MCC) of (A) the HPROM-NPROM MLP classifiers and (B) the PROM-NPROM MLP classifiers as a function of the structural similarity (measured as Tanimoto similarity on Morgan2 fingerprints) between the compounds in the test and the training sets. For each data point the variance was estimated by running the models on ten randomly selected subsets of the test data (see Materials and Methods for details). Because the variance values were within the range of 1.9×10^{-6} to 2.9×10^{-3} they are not visualized in these graphs.

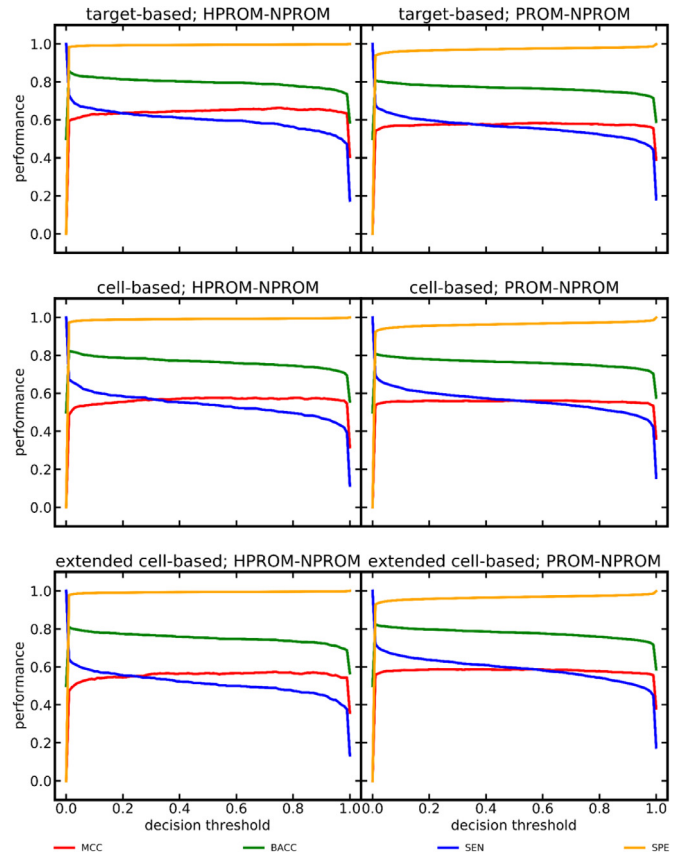


Fig. 10. Performance of the Hit Dexter 3 models as a function of the selected decision threshold.

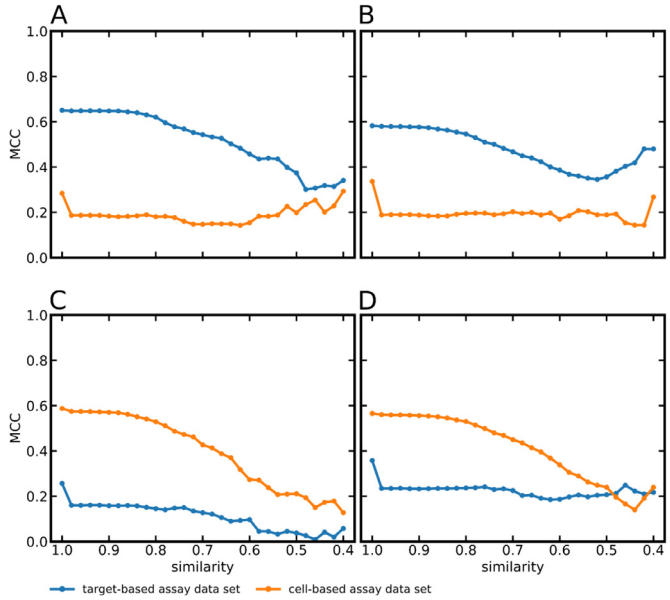


Fig. 11. Performance (quantified as MCC) of MLP classifiers as a function of the pairwise similarity between the test compound and its nearest neighbor in the training set (measured as Tanimoto coefficient derived from Morgan2 fingerprints). (A) HPROM-NPROM MLP classifier trained on the target-based assay data set, (B) PROM-NPROM MLP classifier trained on the target-based assay data set, (C) HPROM-NPROM MLP classifier trained on the cell-based assay data set, (D) PROM-NPROM classifier trained on the cell-based assay data set. For each data point the variance was estimated by running the models on ten randomly selected subsets of the test data (see Materials and Methods for details). Because the variance values were within the range of 9.2×10^{-6} to 6.2×10^{-3} they are not visualized in these graphs.

0.561 for the classifier trained on the cell-based assay data; any compounds present in the training and in the test set are disregarded in the calculation of these MCC values). Likewise, for the target-based assay test data the MCC for the PROM-NPROM MLP classifier trained on cell-based assay data was just 0.235 (vs. 0.580 for the classifier trained on the target-based assay data). These results show that target-based and cell-based assays clearly behave differently and that dedicated models are required to adequately predict their behavior.

Model performance on dark chemical matter

We tested the Hit Dexter 3 models also on the dark chemical matter (DCM) data set compiled by Wassermann et al. The DCM data set consists of 135,489 compounds which have been tested in at least 100 target-based and cell-based assays without a single positive assay outcome. These compounds are not necessarily without activity on any protein but they are unlikely frequent hitters.

In the test of the Hit Dexter 3 models on the DCM data set, any test compounds also present in the training set of the respective models were disregarded (leaving 24,111 to 37,711 DCM compounds for testing, depending on the individual training set). The target-based, cell-based and extended cell-based HPROM-NPROM MLP models correctly assigned 99.0%, 98.6% and 98.7% of the DCM compounds to the NPROM class. In comparison, the percentage of correct assignments of the PROM-NPROM models were 95.4%, 93.7% and 93.6%, respectively. This result corroborates the validity (in particular the specificity) of the models.

Model performance on known bad actors

To test the capacity of the six Hit Dexter 3 models to identify bad actors in biological assays, we ran the models on two recently published

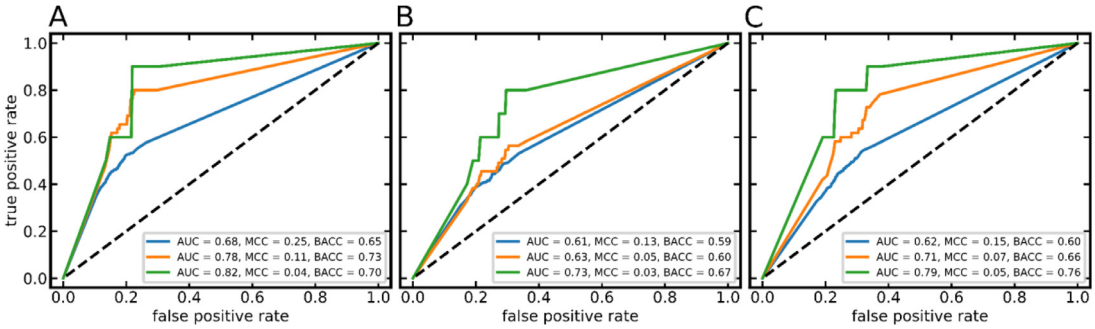


Fig. 12. ROC curves obtained with the Hit Dexter 3 PROM-NPROM classifiers trained on (A) target-based assay data, (B) cell-based assay data, and (C) extended cell-based assay data, and tested on the data set of Borrel et al. The compounds of the test set were annotated as frequent hitters according to Definition 1 (blue curves), Definition 2 (orange curves) and Definition 3 (green curves).

data sets containing experimentally confirmed bad actors (cave: bad actors are not necessarily frequent hitters; Hit Dexter 3 is designed to identify frequent hitters). As in all previous experiments, we disregarded all compounds present in these test sets that are also part of the training data of the individual models.

The first data set is from the work of Dahlin et al. [37]. This data set consists of 1139 compounds that are known to cause false readouts in various types of biological assays. For the 891 to 1002 test compounds not represented in the training set of the individual models, the target-based, cell-based and extended cell-based HPROM-NPROM MLP classifiers assigned 24.1%, 25.5% and 23.0% of all compounds to the HPROM class. The models distinguishing PROM and NPROM compounds flagged 40.3%, 39.3% and 40.3% as promiscuous, respectively. Because bad actors are not necessarily frequent hitters (and vice versa), the percentages of compounds reported by our models as PROM or HPROM are within the expected range.

The second data set is from the work of Borrel et al. [22]. This data set contains 8947 compounds, 891 of which have been observed to cause false positive readouts in bioluminescence assays due to luciferase inhibition (in one out of one assay) or autofluorescence (in one or several out of 24 assays), and 8056 compounds that are confirmed to behave benign in these assays. We explored three ways of translating the measurements recorded with these interference assays into “frequent hitter data"”: Compounds were labeled as frequent hitter if they produced

Definition 1. a false-positive signal in at least one assay (luciferase assay or assay to test for autofluorescence).

Definition 2. a false-positive signal in the luciferase assay AND at least one of the (24) assay setups to test for autofluorescence.

Definition 3. a false-positive signal in the luciferase assay AND at least nine of the (24) assay setups to test for autofluorescence.

All other compounds were labeled as non-promiscuous.

As shown in Fig. 12, the Hit Dexter 3 models reached AUC values of up to 0.82 (PROM-NPROM MLP classifier trained on the target-based assay data set, in combination with Definition 3), which confirms the ability of the models to rank bad actors early in a rank-ordered list of compounds. The MCC and balanced accuracy indicate moderate perfor-

Table 8
Performance of the Hit Dexter 2.0 and Hit Dexter 3 machine learning models on the DCM data sets and the known bad actors data set of Dhalin et al.

Hit Dexter version	training set	test set	classification	number of compounds in test set[1]	number of compounds correctly classified as DCM or bad actors[4]	fraction of compounds correctly classified as DCM or bad actors[4]
Hit Dexter 2.0	PSA data[2]	DCM	HPROM-NPROM	20,894	20,806	0.996
Hit Dexter 2.0	CDRA data[3]	DCM	HPROM-NPROM	42,567	42,341	0.995
Hit Dexter 3	target-based assay data	DCM	HPROM-NPROM	37,711	37,317	0.990
Hit Dexter 3	cell-based assay data	DCM	HPROM-NPROM	30,967	30,529	0.986
Hit Dexter 3	extended cell-based assay data	DCM	HPROM-NPROM	24,327	24,015	0.987
Hit Dexter 2.0	PSA data[2]	DCM	PROM-NPROM	20,872	20,472	0.981
Hit Dexter 2.0	CDRA data[3]	DCM	PROM-NPROM	41,587	40,080	0.964
Hit Dexter 3	target-based assay data	DCM	PROM-NPROM	37,421	35,695	0.954
Hit Dexter 3	cell-based assay data	DCM	PROM-NPROM	30,875	28,942	0.937
Hit Dexter 3	extended cell-based assay data	DCM	PROM-NPROM	24,111	22,572	0.936
Hit Dexter 2.0	PSA data[2]	Known Bad Actors [37]	HPROM-NPROM	974	140	0.144
Hit Dexter 2.0	CDRA data[3]	Known Bad Actors [37]	HPROM-NPROM	963	140	0.145
Hit Dexter 3	target-based assay data	Known Bad Actors [37]	HPROM-NPROM	1002	241	0.241
Hit Dexter 3	cell-based assay data	Known Bad Actors [37]	HPROM-NPROM	987	252	0.255
Hit Dexter 3	extended cell-based assay data	Known Bad Actors [37]	HPROM-NPROM	965	222	0.230
Hit Dexter 2.0	PSA data[2]	Known Bad Actors [37]	PROM-NPROM	910	304	0.334
Hit Dexter 2.0	CDRA data[3]	Known Bad Actors [37]	PROM-NPROM	896	330	0.368
Hit Dexter 3	target-based assay data	Known Bad Actors [37]	PROM-NPROM	941	379	0.403
Hit Dexter 3	cell-based assay data	Known Bad Actors [37]	PROM-NPROM	906	356	0.393
Hit Dexter 3	extended cell-based assay data	Known Bad Actors [37]	PROM-NPROM	891	359	0.403

[1] Any compounds present in the training set were removed from the test set.
[2] Primary screening assay data.
[3] Confirmatory dose-response assay data.
[4] Note that Hit Dexter models are not designed to identify all different kinds of bad actors but rather to identify frequent hitters (of which a significant portion are in fact bad actors).

mance but, again, Hit Dexter 3 is designed to predict frequent hitters, and it can be expected that a substantial proportion of the compounds observed to cause false-positive readouts in these interference assays will behave benign in other assay types and setups.

Model performance compared to Hit Dexter 2.0

The set of machine learning models developed in this work to form Hit Dexter 3 differ from the Hit Dexter 2.0 models in several ways. For Hit Dexter 3,

- dedicated models for target-based and cell-based assays were developed whereas the previous set of models only cover target-based assays.
- four different machine learning algorithms (KNN, ET, RF and MLP) instead of just ET were explored. This led to the finding that MLP classifiers perform best.
- the minimum number of data points required to calculate the ATR has been increased from 50 to 100. This results in more robust ATRs (based on which the class labels, i.e. HPROM, PROM and NPROM, are assigned).

Given the fact that the training and test sets utilized for the development and validation of the Hit Dexter 3 and Hit Dexter 2.0 models differ, a 1:1 comparison of model performance is difficult. For models of the same type (e.g. HPROM-NPROM classifier), differences in MCCs on the test data were in the range of -0.035 to +0.015 (cell-based models not included as they are not available in Hit Dexter 2.0). Also on the DCM data sets (the DCM data sets used for testing differ in their composition because of the removal of any compounds that are also present in the training set of the respective model), the models behave similarly, with the Hit Dexter 3 PROM-NPROM classifier (for target-based assays) assigning 5% to the PROM class and the respective Hit Dexter 2.0 models assigning 2% to 4% of the DCM compounds to the PROM class (Table 8). On the set of known bad actors [37], the percentage of compounds predicted as frequent hitters is 40% for Hit Dexter 3 (PROM-NPROM classifier trained on target-based assay data) and 33% to 37% for Hit Dexter 2.0 (PROM-NPROM classifiers; Table 8).

Overall, these results indicate that the performance of the Hit Dexter 3 and Hit Dexter 2.0 machine learning models is comparable. The Hit Dexter 3 models perform a bit better on the set of known bad actors. Finally, the addition of dedicated models for predicting a compound's behavior in cell-based assays is an important advantage of Hit Dexter 3 over Hit Dexter 2.0.

Conclusions

In this work we present the development, refinement and validation of new models for the prediction of frequent hitters in biological assays. The models are trained on a manually curated assay data set that we extracted from the PubChem Bioassay database and, for the first time, these models cover cell-based assays in addition to target-based assays. Further additions include the exploration of four sets of descriptors with additional machine learning algorithms such as KNN and MLP, and the use of more robust ATRs (calculated now on a minimum of 100 distinct assays compared to 50 previously).

The MLP classifiers turned out to obtain the best classification performance and robustness in most cases, with MCCs of up to 0.648 in discriminating HPROM from NPROM compounds, and MCCs of up to 0.580 in discriminating PROM from NPROM compounds. Use cases that require models with high sensitivity or high specificity can be approached by adjusting the decision threshold applied in classification.

Tests of the MLP classifiers on DCM compounds and sets of known bad actors corroborate good performance of the models: the models correctly identified 94 to 99% of all compounds of the DCM data set as non-promiscuous and flagged up to 40% of the known bad actors as frequent hitters (because bad actors are not necessarily frequent hitters this number is in line with the expectations for a good model).

We found that it is indeed important to use dedicated models for predicting the behavior of compounds in target-based and cell-based assays as assays from the different domains can behave very differently. At the same time it is clear that for the further development of this and similar computational methods it will be important to consider assay types and conditions, which poses fundamental challenges related to the scarcity and heterogeneity of the available data.

The best models presented in this work are available via a refined, free web service at https://nerdd.univie.ac.at/hitdexter3/. This web service offers many additional features, encrypted communication via HTTPS and the possibility for users to immediately and permanently delete their data from the web server.

We hope that the new Hit Dexter models, in particular the new models for cell-based assays, will be of high value to the scientific community to tackle the challenge of hit prioritization and the identification of problematic compounds in biological screens.

Funding

C.S. and J.K. are supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) project number KI 2085/1-1. N.M. and J.K. are supported by the Trond Mohn Foundation (BFS2017TMT01).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ailsci.2021.100007.

References

[1] Macarron R, Banks MN, Bojanic D, Burns DJ, Cirovic DA, Garyantes T, et al. Impact of high-throughput screening in biomedical research. Nat Rev Drug Discov 2011;10:188–95.
[2] Bajorath J. Evolution of assay interference concepts in drug discovery. Expert Opin Drug Discov 2021:1–3.
[3] Evans BE, Rittle KE, Bock MG, DiPardo RM, Freidinger RM, Whitter WL, et al. Methods for drug discovery: development of potent, selective, orally effective cholecystokinin antagonists. J Med Chem 1988;31:2235–46.
[4] In: Auld DS, Inglese J. Interferences with Luciferase reporter enzymes. Assay Guidance Manual. Markossian S, Grossman A, Brimacombe K, Arkin M, Auld D, Austin CP, editors. et al., editors. Eli Lilly & Company and the National Center for Advancing Translational Sciences, Bethesda (MD); 2016.
[5] Reker D, Bernardes GJL, Rodrigues T. Computational advances in combating colloidal aggregation in drug discovery. Nat Chem 2019;11:402–18.
[6] Yang Z-Y, He J-H, Lu A-P, Hou T-J, Cao D-S. Frequent hitters: nuisance artifacts in high-throughput screening. Drug Discov Today 2020;25:657–67.
[7] Dantas RF, Evangelista TCS, Neves BJ, Senger MR, Andrade CH, Ferreira SB, et al. Dealing with frequent hitters in drug discovery: a multidisciplinary view on the issue of filtering compounds on biological screenings. Expert Opin Drug Discov 2019;14:1269–82.
[8] Ferreira LLG, Andricopulo AD. ADMET modeling approaches in drug discovery. Drug Discov Today 2019;24:1157–65. doi:10.1016/j.drudis.2019.03.015.
[9] Kar S, Leszczynski J. Open access in silico tools to predict the ADMET profiling of drug candidates. Expert Opin Drug Discov 2020;15:1473–87.
[10] Feldmann C, Bajorath J. Machine learning reveals that structural features distinguishing promiscuous and non-promiscuous compounds depend on target combinations. Sci Rep 2021;11:7863.
[11] Irwin JJ, Duan D, Torosyan H, Doak AK, Ziebart KT, Sterling T, et al. An aggregation advisor for ligand discovery. J Med Chem 2015;58:7076–87.
[12] Yang Z-Y, Yang Z-J, Dong J, Wang L-L, Zhang L-X, Ding J-J, et al. Structural analysis and identification of colloidal aggregators in drug discovery. J Chem Inf Model 2019;59:3714–26.
[13] Alves VM, Capuzzi SJ, Braga RC, Korn D, Hochuli JE, Bowler KH, et al. SCAM detective: accurate predictor of small, colloidally aggregating molecules. J Chem Inf Model 2020;60:4056–63.
[14] Reactive compounds and in vitro false positives in HTS. Drug Discov Today 1997;2:382–4.
[15] Hann M, Hudson B, Lewell X, Lifely R, Miller L, Ramsden N. Strategic pooling of compounds for high-throughput screening. J Chem Inf Comput Sci 1999;39:897–902.

[16] Baell JB, Holloway GA. New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. J Med Chem 2010;53:2719–40.
[17] Baell J, Walters MA. Chemistry: chemical con artists foil drug discovery. Nature 2014;513:481–3.
[18] Baell JB, Nissink JWM. Seven year itch: pan-assay interference compounds (PAINS) in 2017-utility and limitations. ACS Chem Biol 2018;13:36–44.
[19] M. Koptelov, A. Zimmermann, P. Bonnet, R. Bureau, B. Crémilleux. PrePeP: a tool for the identification and characterization of pan assay interference compounds. Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining, p. 462–71.
[20] Ghosh D, Koch U, Hadian K, Sattler M, Tetko IV. Luciferase advisor: high-accuracy model to flag false positive hits in Luciferase HTS assays. J Chem Inf Model 2018;58:933–42.
[21] Yang Z-Y, Dong J, Yang Z-J, Lu A-P, Hou T-J, Cao D-S. Structural analysis and identification of false positive hits in Luciferase-based assays. J Chem Inf Model 2020;60:2031–43.
[22] Borrel A, Huang R, Sakamuru S, Xia M, Simeonov A, Mansouri K, et al. High-throughput screening to predict chemical-assay interference. Sci Rep 2020;10:3986.
[23] M Nissink JW, Blackburn S. Quantification of frequent-hitter behavior based on historical high-throughput screening data. Future Med Chem 2014;6:1113–26.
[24] Yang JJ, Ursu O, Lipinski CA, Sklar LA, Oprea TI, Bologa CG. Badapple: promiscuity patterns from noisy evidence. J Cheminform 2016;8:29.
[25] Stork C, Chen Y, Šícho M, Kirchmair J. Hit Dexter 2.0: machine-learning models for the prediction of frequent hitters. J Chem Inf Model 2019;59:1030–43.
[26] Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, et al. PubChem in 2021: new data content and improved web interfaces. Nucleic Acids Res 2021;49:D1388–95.
[27] Feldmann C, Yonchev D, Stumpfe D, Bajorath J. Systematic data analysis and diagnostic machine learning reveal differences between compounds with single- and multitarget activity. Mol Pharm 2020;17:4652–66.
[28] Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, et al. PubChem substance and compound databases. Nucleic Acids Res 2016;44:D1202–13.
[29] Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, et al. PubChem 2019 update: improved access to chemical data. Nucleic Acids Res 2019;47:D1102–9.
[30] PubChem. https://pubchem.ncbi.nlm.nih.gov/ (accessed June 3, 2021).
[31] NCBI Resource CoordinatorsDatabase resources of the National Center for Biotechnology Information. Nucleic Acids Res 2016;44:D7–19.
[32] Kim S, Thiessen PA, Cheng T, Yu B, Bolton EE. An update on PUG-REST: RESTful interface for programmatic access to PubChem. Nucleic Acids Res 2018;46:W563–70.
[33] Patrícia Bento A, Hersey A, Félix E, Landrum G, Gaulton A, Atkinson F, et al. An open source chemical structure curation pipeline using RDKit. J Cheminform 2020;12:1–16.
[34] RDKit: Open-source cheminformatics; http://www.rdkit.org/ (accessed June 3, 2021).
[35] ChEMBL 23. http://www.ebi.ac.uk/chembl/ (accessed June 3, 2021).
[36] Wassermann AM, Lounkine E, Hoepfner D, Le Goff G, King FJ, Studer C, et al. Dark chemical matter as a promising starting point for drug lead discovery. Nat Chem Biol 2015;11:958–66.
[37] Dahlin JL, Auld DS, Rothenaigner I, Haney S, Sexton JZ, Nissink JWM, et al. Nuisance compounds in cellular assays. Cell Chem Biol 2021;28:356–70.
[38] NCICADD Group, National Cancer Institute. Chemical Identifier Resolver. https://cactus.nci.nih.gov/chemical/structure (accessed July 26, 2021).
[39] Rogers D, Hahn M. Extended-connectivity fingerprints. J Chem Inf Model 2010;50:742–54.
[40] Morgan HL. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. J Doc 1965;5:107–13. doi:10.1021/c160017a018.
[41] Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 2009;25:1422–3.
[42] Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 2012;28:3150–2.
[43] Garreta R, Moncecchi G. Learning scikit-learn: machine learning in Python. Packt Publishing Ltd; 2013.
[44] https://www.chemcomp.com/MOE-Molecular_Operating_Environment.htm.
[45] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority oversampling technique. J Artif Intell Res 2002;16:321–57. doi:10.1613/jair.953.
[46] Stork C, Wagner J, Friedrich N-O, de Bruyn Kops C, Šícho M, Kirchmair J. Hit Dexter: a machine-learning model for the prediction of frequent hitters. ChemMedChem 2018;13:564–71.

5.4 New e-resource for early drug discovery (NERDD) for the availability of early drug discovery tools

In order to make the developed machine learning models available to the public, a web server was developed, called NERDD. The advantage of a web server is that for users no software installation is required. Hence, the aim of this part of this Ph.D. study was to develop an easily extensible and maintainable web server to support early drug discovery. Eight in silico tools are currently available via NERDD. Besides the Hit Dexter models, which were developed as part of this Ph.D. study, NP-Scout for the prediction of natural product likeness and Skin Doctor CP for the prediction of skin sensitization potential of molecules are accessible. Metabolism-related tools, including CYPstrate for the prediction of cytochrome P450 substrates and CYPlebrity for the prediction of Cytochrome P450 inhibitors are also available. FAME3, GLORY and GLORYx, which are tools for the prediction of sites-of-metabolism and the likely metabolites of small compounds, can also be accessed via NERDD, at `https://nerdd.univie.ac.at`.

[D4] NERDD: a web portal providing access to in silico tools for drug discovery

Conrad Stork, Gerd Embruch, Martin Šícho, Christina de Bruyn Kops, Ya Chen, Daniel Svozil and Johannes Kirchmair

Bioinformatics, 2019

Available at `https://doi.org/10.1093/bioinformatics/btz695`.

Contribution:
C. Stork and J. Kirchmair conceptualized the research. C. Stork, M. Šícho, C. de Bruyn Kops and Y. Chen developed the individual computational tools as documented in the scientific publications accompanying these tools. C. Stork implemented the web server, with G. Embruch providing technical support. C. Stork wrote the manuscript, with contributions from G. Embruch, M. Šícho, Y. Chen, C. de Bruyn Kops, D. Svozil and J. Kirchmair. J. Kirchmair and D. Svozil supervised the work.

The following article was reprinted with permission from:
Stork, C.; Embruch, G.; Šícho, M.; de Bruyn Kops, C; Chen, Y. and Kirchmair, J. NERDD: a web portal providing access to in silico tools for drug discovery, *Bioinformatics* **2019**, *36*, 1291–1292.
Copyright The Author(s) 2019. Published by Oxford University Press

Bioinformatics, 36(4), 2020, 1291–1292
doi: 10.1093/bioinformatics/btz695
Advance Access Publication Date: 6 September 2019
Applications Note

OXFORD

Structural bioinformatics

NERDD: a web portal providing access to *in silico* tools for drug discovery

Conrad Stork[1], Gerd Embruch[1], Martin Šícho[2], Christina de Bruyn Kops[1], Ya Chen[1], Daniel Svozil[2] and Johannes Kirchmair [1,3,4,*]

[1]Department of Informatics, Universität Hamburg, Faculty of Mathematics, Informatics and Natural Sciences, Center for Bioinformatics (ZBH), Hamburg 20146, Germany, [2]Department of Informatics and Chemistry, CZ-OPENSCREEN: National Infrastructure for Chemical Biology, University of Chemistry and Technology Prague, Faculty of Chemical Technology, 166 28 Prague 6, Czech Republic, [3]Department of Chemistry, University of Bergen, Bergen N-5020, Norway and [4]Computational Biology Unit (CBU), Bergen N-5020, Norway

*To whom correspondence should be addressed.

Associate Editor: Yann Ponty

Received on July 15, 2019; revised on August 16, 2019; editorial decision on September 2, 2019; accepted on September 3, 2019

Abstract

Summary: The New E-Resource for Drug Discovery (NERDD) is a quickly expanding web portal focused on the provision of peer-reviewed in silico tools for drug discovery. NERDD currently hosts tools for predicting the sites of metabolism (FAME) and metabolites (GLORY) of small organic molecules, for flagging compounds that are likely to interfere with biological assays (Hit Dexter), and for identifying natural products and natural product derivatives in large compound collections (NP-Scout). Several additional models and components are currently in development.

Availability and implementation: The NERDD web server is available at https://nerdd.zbh.uni-hamburg.de. Most tools are also available as software packages for local installation.

Contact: kirchmair@zbh.uni-hamburg.de

1 Introduction

Modern computational approaches make a substantial contribution to the development of safe and efficacious drugs. Our laboratories specialize in the development of new in silico approaches for drug discovery, including methods for the prediction of bioactivity, drug metabolism, natural product-likeness and interference of small molecules with biological assays. In an effort to make our software and models available to the scientific community we have developed the New E-Resource for Drug Discovery (NERDD), available at https://nerdd.zbh.uni-hamburg.de.

2 The NERDD web server

NERDD is built on the Django web framework (https://www.djangoproject.com) deployed with the Apache HTTP server (https://www.apache.org). The web service is designed to be maintainable and scalable. NERDD meets modern security standards and supports encrypted communication via HTTPS. The web service is linked to an in-house high-performance computing facility that can handle large numbers of concurrent requests.

When visiting NERDD, users are presented a homepage that lists all the available tools. The individual start pages of the tools offer different options to provide molecular structures as input (including bulk data upload). Users can also change some settings concerning the calculation, visualization and reporting of results. Following the submission of a calculation, users are presented a status page with an estimate of the remaining calculation time. The waiting time for individual queries will usually not be longer than a few seconds. During peak load and when calculations for large sets take more time, users may prefer to return to the website at a later point in time to collect their results. The results page provides a tabular or table-like overview of all results. Tools generating more complex reports offer options to show, hide and expand specific types of information. All tools offer options to export all the results in standard file formats. The results are stored for a specified period of time for users to retrieve their data. An option for the immediate deletion of the user's data is also provided. NERDD currently features four tools, which are briefly discussed in the following sections.

2.1 FAME 3 for site of metabolism prediction

FAME 3 is the third generation of machine learning models for the prediction of sites of metabolism (SoMs), the atom positions in a molecule at which metabolic reactions are initiated (Šícho *et al.*, 2019). Several advances distinguish this model from its predecessors and other existing models. First, the extremely randomized trees classifiers are trained on a large, expert-curated dataset covering phase 1 and 2 metabolism (Pedretti *et al.*, 2018). Second, the

© The Author(s) 2019. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

Downloaded from https://academic.oup.com/bioinformatics/article/36/4/1291/5564116 by Bibliothekssystem Universität Hamburg user on 10 January 2021

models, based on specifically designed atom descriptors, generalize well, making them applicable not only to synthetic compounds but also to natural products. Third, a newly introduced, atom-based distance measure ('FAMEscore') allows the estimation of the reliability of predictions individually for each atom in a molecule.

On holdout data, a global model for phase 1 and phase 2 metabolism reached competitive performance, with a Matthews correlation coefficient (MCC) of 0.50 and an area under the receiver operating characteristic curve (AUC) of 0.90. Focused models performed even better; a model for phase 2 metabolism achieved an MCC of 0.71 and AUC of 0.97.

2.2 GLORY for metabolite structure prediction

GLORY (de Bruyn Kops *et al.*, 2019) uses the results obtained from FAME to apply 73 reaction rules and generate the molecular structures of likely metabolites formed by CYPs. GLORY features two operation modes: MaxEfficiency and MaxCoverage. While MaxEfficiency gives the most relevant metabolites, MaxCoverage does a better job of covering all possible metabolites and is recommended. One important feature that sets GLORY apart from many of the existing metabolite structure predictors is its capability to rank the predicted metabolites according to their likelihood. The model ranked at least one known metabolite within the top three positions for 76% of the molecules of an independent test set.

2.3 Hit Dexter 2.0 for assay interference prediction

Hit Dexter 2.0 (Stork *et al.*, 2019) is designed as a one-stop shop for identifying small molecules that are likely to interfere with biological assays or show promiscuous behaviour. 'Badly behaving compounds', 'bad actors' or 'nuisance compounds' are abundantly present in screening libraries as well as the chemical biology and medicinal chemistry literature (Baell and Holloway, 2010), and have been setting researchers on the wrong track all too often. The interference of nuisance compounds with biochemical assays is based on various physical and chemical processes.

In Hit Dexter 2.0, we implemented several available in silico approaches (Stork and Kirchmair, 2018), including the well-known set of 480 substructures observed in pan-assay interference (PAINS; Baell and Holloway, 2010) and a large variety of published rule sets encoding substructures regarded as undesirable in the context of drug discovery. Hit Dexter 2.0 also provides results of Aggregator Advisor (Irwin *et al.*, 2015), a similarity-based approach for comparing compounds of interest with known aggregators.

The core of Hit Dexter 2.0 is a set of extremely randomized trees classifiers for identifying frequent hitters (i.e. small molecules for which a higher than expected hit rate is observed in biological assays). The Hit Dexter 2.0 machine learning models are trained on large sets of experimental data (up to approximately 250 000 compounds). Separate classifiers are available for primary screening assays and confirmatory dose-response assays. On holdout data, the classifiers obtained MCC values of up to 0.64 and AUC values of up to 0.96 in discriminating (highly) promiscuous from non-promiscuous compounds. Hit Dexter 2.0 should not be used as a hard filter to eliminate compounds but as a tool for hit prioritization.

2.4 NP-Scout for the identification of natural products

NP-Scout (Chen *et al.*, 2019) is a machine learning approach that allows the identification of natural products and natural product-like molecules in large compound collections. Natural products are the most prolific resource of inspiration for the development of modern small-molecule drugs. However, only an estimated 10% of known natural products are readily obtainable from commercial and other sources, and despite their value they are often concealed in mixed compound collections and not labelled as natural products (Chen *et al.*, 2017). In response to this situation we developed NP-Scout. The classifier is trained on approximately 265 000 natural products and synthetic molecules and obtained an MCC of 0.960 and AUC of 0.997 on holdout data. NP-Scout utilizes similarity maps (Riniker and Landrum, 2013) to visualize areas in a molecule that are characteristic of natural products or synthetic compounds.

3 Conclusions

NERDD is a quickly expanding web portal offering a variety of tools for drug discovery efforts. One of the most important features is the flexible linkage to an in-house high-performance computing facility, which enables the handling of large numbers of concurrent requests. All tools currently offered with NERDD are free for non-commercial and academic research. With the exception of FAME 3, all the tools are also free for commercial research. FAME 3 may be used by for-profit institutions for testing purposes; a fee applies for commercial use. Upon request to the authors, these tools are also available as software packages for local installation.

Funding

This work was supported by the DFG, German Research Foundation [KI 2085/1-1], Bergen Research Foundation (BFS) [BFS2017TMT01], Ministry of Education, Youth and Sports of the Czech Republic [LM2015063, RVO 68378050-KAV-NPUI, 21-SVV/2018] and China Scholarship Council [201606010345].

Conflict of Interest: The authors declare a potential financial interest in the event that FAME 3 or other components of the web portal will be licensed for a fee to for-profit institutions in the future.

References

Baell,J.B. and Holloway,G.A. (2010) New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *J. Med. Chem.*, **53**, 2719–2740.

de Bruyn Kops,C. *et al.* (2019) GLORY: generator of the structures of likely cytochrome P450 metabolites based on predicted sites of metabolism. *Front. Chem*, **7**, 402.

Chen,Y. *et al.* (2017) Data resources for the computer-guided discovery of bioactive natural products. *J. Chem. Inf. Model*, **57**, 2099–2111.

Chen,Y. *et al.* (2019) NP-Scout: machine learning approach for the quantification and visualization of the natural product-likeness of small molecules. *Biomolecules*, **9**, 43.

Irwin,J.J. *et al.* (2015) An aggregation advisor for ligand discovery. *J. Med. Chem.*, **58**, 7076–7087.

Pedretti,A. *et al.* (2018) MetaQSAR: an integrated database engine to manage and analyze metabolic data. *J. Med. Chem*, **61**, 1019–1030.

Riniker,S. and Landrum,G.A. (2013) Similarity maps – a visualization strategy for molecular fingerprints and machine-learning methods. *J. Cheminf.*, **5**, 43.

Šícho,M. *et al.* (2019) FAME 3: predicting the sites of metabolism in synthetic compounds and natural products for phase 1 and phase 2 metabolic enzymes. *J. Chem. Inf. Model*, **59**, 3400.

Stork,C. *et al.* (2019) Hit Dexter 2.0: machine-learning models for the prediction of frequent hitters. *J. Chem. Inf. Model*, **59**, 1030–1043.

Stork,C. and Kirchmair,J. (2018) PAIN(S) relievers for medicinal chemists: how computational methods can assist in hit evaluation. *Future Med. Chem.*, **10**, 1533–1535.

Downloaded from https://academic.oup.com/bioinformatics/article/36/4/1291/5564116 by Bibliothekssystem Universität Hamburg user on 10 January 2021

6. Conclusions and future directions

A major challenge in high-throughput screening (HTS) is the identification of compounds that are likely to trigger false positive assay readouts. The deprioritization of these badly behaving compounds is of utmost importance as the publication or the usage within a company of such compounds would lead to expensive and time consuming follow-up studies which would waste a significant amount of resources. The existing computational models for the identification of bad actors and nuisance compounds are primarily using similarity-based or rule-based approaches. Less often are machine learning models that use recently developed algorithms for more accurate predictions of badly behaving compounds. These algorithms are more complex than, for example, rule-based approaches (which often only match substructures) and hence need to be made easily accessible and well explained. Further limitations of the existing approaches are the often narrow applicability domain of the models and a clear constraint to target-based assay technologies.

During this Ph.D. study three generations of machine learning models for the prediction of frequent hitters in biochemical and biological assays were developed (Hit Dexter, Hit Dexter 2.0 and Hit Dexter 3). Frequent hitters are compounds that show higher-than-expected hit rates and are therefore likely to trigger false positive assay readouts and are also known as bad actors, nuisance compounds and badly behaving compounds. The high hit rates are often a result of undesirable, nonspecific interactions of the compound with the target protein or the assay ingredients. In order to make the Hit Dexter models freely and easily accessible a web server was developed called New E-Resource for Drug Discovery (NERDD), which has become an established platform for several bioinformatic tools developed in our research group. Additionally, it was co-authored to several peer-reviewed publications, which can be found in detail in Refs. [A1–A11].

The first generation of machine learning models for the prediction of frequent hitters in biochemical and biological assays are based on a large data set which was extracted from PubChem Bioassay database and only contains target-based assay data. Compounds within this large data set (around 311 000 compounds), were divided into three classes according to their hit rate, using the average plus standard deviation approach: non-promiscuous compounds (hit rates not larger that the average), promiscuous compounds (hit rates above the averages plus

one standard deviation) and highly promiscuous compounds (hit rates above the average plus three standard deviations). Two models were built: one for the differentiation of non-promiscuous and promiscuous compounds and another for the differentiation of non-promiscuous and highly promiscuous compounds. In order to decrease the statistical fluctuation of compounds that were tested less often, only compounds that were tested against at least 50 distinct protein targets were used for machine learning model building. Machine learning models based on extra tree classifiers and Morgan2 fingerprints, called Hit Dexter, reached Matthews correlation coefficient (MCCs) and area under the receiver operating characteristic curve (AUC) values of up to 0.67 and 0.96, respectively, and were made publicly available as part of the web server framework NERDD (for details see below).

The second generation of machine learning models for frequent hitter prediction in biochemical and biological assay screening was extended to primary screen assay (PSA) data. Initial PSA experiments often have lower accuracy and different compounds will be detected as frequent hitters compared to follow-up confirmatory dose-response assay (CDRA) screenings due to the different assay screening setups. Large data sets were extracted from PubChem Bioassay database for PSA data and CDRA data and dedicated models for both assay setups were built. Another important refinement of the newly developed models is the exclusion of compounds that show activity on structural related proteins such as kinases. Therefore the available protein target primary structures were clustered and compounds were only labeled as frequent hitters if they showed activity on multiple protein target clusters. The best-performing models (again based on extra tree classifiers and Morgan2 fingerprints) reached MCCs and AUC values of up to 0.64 and 0.96, respectively. Further, the performance of the developed machine learning models were evaluated on different data sets, including collections of approved drugs, known aggregation forming compounds, potential PAINS (compounds that were detected by the well-known PAINS filter set), natural products, consistently inactive compounds from target-based and cell-based assay screening (also known as dark chemical matter) and several screening libraries. Models for the differentiation of non-promiscuous and promiscuous compounds as well as for the differentiation of non-promiscuous and highly promiscuous compounds for both PSAs and CDRAs were made available within NERDD as Hit Dexter 2.0 models substituting the first generation of Hit Dexter models. In addition to the frequent hitter predictions, the Hit Dexter web server was extended with several establish rule-based approaches (like the PAINS filter set) and similarity-based approaches (like AggregatorAdvisor) to develop the Hit Dexter web server into a one-stop-shop for HTS hit (de-)prioritization.

To further extend the applicability domain (AD) of the Hit Dexter models, a differentiation of frequent hitters into cell-based and target-based assay screenings, respectively, was achieved in the third generation of machine learning models. Three data sets were compiled from the PubChem Bioassay database, which were divided into a target-based assay data set, a cell-based assays data set (excluding assays measuring nonspecific interactions like cytotoxicity) and an extended cell-based assay data set (including all types of cell-based assays). An increased number of available data were used to have a more significant hit rate for each compound. The minimum number of proteins (and protein clusters) a compound had to be tested against was increased to 100, which adds to the robustness of the models. Dedicated models were built for each of the data sets and multiple machine learning algorithms, including k-nearest neighbors (KNN), random forest (RF), extra tree (ET) and multilayer perceptron (MLP) classifiers were investigated. The best performing models (based on MLP classifiers combined with Morgan2 fingerprints) reached MCCs of up to 0.69, 0.61 and 0.64 for the target-based, cell-based and extended cell-based assay data sets for the differentiation of non-promiscuous and highly promiscuous compounds, respectivly. These models were released as part of the Hit Dexter 3 models, which also include models for the differentiation of non-promiscuous and promiscuous compounds, and are available at the web server NERDD. It was shown that models trained on the target-based assay data set do not perform well on test sets based on cell-based assay data and vice versa which underlines the need of dedicated models for target-based and cell-based assay screenings.

In order to make the Hit Dexter models publically available, a web server was developed called NERDD. NERDD is developed in a modular way which makes it easy to implement newly developed tools within the web server framework. The server uses HTTPS encryption, which makes it secure and all uploaded data can be deleted by the user immediately after downloading the results which allows the users also to work with confidential data. NERDD is continuously growing and has become an established platform in the scientific community. The web server accommodates, besides Hit Dexter, seven additional tools which are of high value to (early) drug discovery, including NP-Scout for the prediction of natural product-likeness and Skin Doctor CP for the prediction of skin sensitization potential. Addinally, five tools for metabolism predictions are available, including FAME3 for the sites of metabolism prediction of Phase 1 and Phase 2 metabolism. Based on FAME3, GLORYx can predict the likely metabolic structures which are formed during Phase 1 and Phase 2 metabolism whereas GLORY can predict the metabolic structures during Cytochrome P450 metabolism. Further, CYPstrate and CYPlebrity are machine learning tools for the prediction of Cytochrome P450 substrates and inhibitors, respectively.

During this Ph.D. study the understanding of frequent hitters in biochemical and biological assays increased and the prediction of frequent hitter compounds

by machine learning models was achieved. However, many ideas and follow-up studies could be implemented and performed to further increase our understanding of frequent hitters in biochemical and biological assay screenings.

One use case of the Hit Dexter models is, for example, the support of researchers that perform biological and biochemical assay screenings. A follow-up study, which would need the support of an experimental cooperation partner could investigate the following scenario: Supposing a researcher wants to find privileged scaffolds to start a multi-target drug discovery campaign. Hit Dexter could already support the building of the screening library as such a library should contain frequent hitter compounds. A diverse subset of the Hit Dexter selected compounds could be the starting point for such a multi-target drug campaign. However, an experimental workflow using Hit Dexter needs to be validated and established to make researchers familiar with the advantages of machine learning models during drug development. The screening outcome could serve as an external and experimental validation of the Hit Dexter models and an optimization of the models could be performed.

A major limitation of the Hit Dexter models is that the machine learning models cannot distinguish between frequent hitters that act as "bad actors" or "nuisance compounds", and frequent hitters that are valuable hits because of their true promiscuous behaviors due to privileged scaffolds. Such an extension of the Hit Dexter model would achieve a large impact on modern HTS campaigns as the identification of bad actors is still a major problem during assay screenings. Approaches exist that predict true promiscuous compounds based on data sets that were filtered with the most important and relevant existing models that can detect bad actors. However, models for bad actor detection are often incomplete and have a small applicability domain which may lead to a high estimated number of unreported cases of bad actors in the training data.

In the present work only ligand-based machine learning approaches were investigated and developed. One possibility for an improvement of the models could be the inclusion of structure-based approaches. The current version of the Hit Dexter models uses the protein target sequence only for clustering structurally similar proteins to avoid that compounds are classified as frequent hitters whereas they are active on one target family (e.g. kinases). In an extended model the protein structure could be used to extend the descriptors of the compounds to take also, for example, the binding mode of a compound with a protein target into account. One could imagine that especially the prediction of true promiscuous compounds could be boosted as privileged scaffolds might have similar binding modes on multiple proteins. But also the reaction patterns of bad actors could be revealed (for example covalent binding) taking the binding mode into account.

The analysis of the chemical structures of frequent hitters is another important step toward a better understanding of frequent hitter behavior. The present study revealed that the separation of non-promiscuous and (highly) promiscuous compounds is possible with machine learning models based on Morgan2 fingerprints. However, the analysis of the chemical structure of frequent hitter compounds has not yet been performed. In a next step, the used Morgan2 descriptors could be analysed in more detail within a backpropagation process. For example, it could be analysed, which chemical groups and chemical core fragments are enriched in frequent hitter compounds. A deeper understanding of the chemical structure of frequent hitters would make interpretations by experimentalists easier and would further increase the awareness of the challenges caused by frequent hitters.

Another important step regarding the identification of compounds triggering false positive assay results in biochemical and biological assays would be the development of models for specific assay reading technology setups. Each assay setup has different reaction mechanisms and assay reading technologies which makes some compound classes problematic only for particular setups. For example, autofluorescence compounds are mainly problematic in assay screening setups in which the detection method is based on electromagnetic radiation as the detector could accidentally measure the radiation of the autofluorescence compound instead of the radiation that is emitted during a positive assay outcome event. Few models for predicting compounds likely to trigger false positive assay outcomes for specific assay types already exist, mainly for luciferase (bioluminescence) assays. However, the amount of publicly available data for specific assay reading technologies is low and a reproducible assay ontologie is still in their infancy. These facts makes it difficult to easily and quickly develop machine learning models in the public domain.

The combination of several specific assay reading technology data sets containing bad actors for each of the available assay setups could present an opportunity to maximize the use of the limited amount of available data. A possible solution comes with multi-class neural networks, which could be used on multiple combined, small data sets to build a larger data set with different endpoints. The neural network could learn from each of the different data sets to cross link information of the different data sets to make predictions of bad actors and/or frequent hitters more accurate for specific assay setups. Structural analysis of the cross linked information would lead to an even deeper understanding of nuisance compounds.

The same multi-class approach could work for the Hit Dexter models. A multi-classification model can make use of the information of the target-based data set to learn for the cell-based assay and vice versa. Here a much more performance orientated way of programming would be necessary as the large data sets com-

bined with large neural networks need an increased amount of computational resources and the use of GPUs could be essential.

An important part of this Ph.D. study was the development of the web server NERDD which was implemented in a modular way which makes it easy to add newly developed tools within the web server framework, which will be done for newly developed tools. However, there is room for improvement in the implementation of some components of NERDD. For example, a performance-oriented implementation could speed up the calculation time thereby reducing the time a user needs to wait for the results. At the moment all calculations are sent to a high performance cluster and for each calculation large models and data sets have to be loaded which produces a large overhead. Simple workers running directly on the server itself having the models already loaded could be used to speed up calculations, especially calculation with only a single molecule. A well-tested environment is needed here to avoid memory leaks and server crashes which are producing downtimes of the server. Following this, an automated down time recognition and an automated bug report could be extremely useful to detect and solve server problems immediately. Several tests should be implemented that check if all tools are running and working as expected to avoid manual testing after a server crash or after new tools are migrated. Another problem that exists since the move from Hamburg to Vienna is the absence of an internal test server where new features can be implemented without interfering with the productive online web server. While setting such an environment up, the possibilities of automated building, testing and deploying pipelines which come along with git repository hosting services could be used for an automated updating of the productive server as this is done manually at the moment. This also includes the manual duplication of updated models, data sets and other larger files. For the tracking of larger files programs like data version control (DVC) should be established inside the web server framework. Taking together the web server is stably running, but a lot of space for improvement exits within the web server framework.

7. Bibliography

[1] DiMasi, J. A.; Grabowski, H. G.; Hansen, R. W. Innovation in the pharmaceutical industry: new estimates of R&D costs, *Journal of Health Economics* **2016**, *47*, 20–33.

[2] Gleeson, M. P.; Modi, S.; Bender, A.; L Marchese Robinson, R.; Kirchmair, J.; Promkatkaew, M.; Hannongbua, S.; Glen, R. C. The challenges involved in modeling toxicity data in silico: a review, *Current Pharmaceutical Design* **2012**, *18*, 1266–1291.

[3] Schneider, G. Automating drug discovery, *Nature Reviews Drug Discovery* **2018**, *17*, 97–113.

[4] Vamathevan, J.; Clark, D.; Czodrowski, P.; Dunham, I.; Ferran, E.; Lee, G.; Li, B.; Madabhushi, A.; Shah, P.; Spitzer, M.; Zhao, S. Applications of machine learning in drug discovery and development, *Nature Reviews Drug Discovery* **2019**, *18*, 463–477.

[5] Macarron, R.; Banks, M. N.; Bojanic, D.; Burns, D. J.; Cirovic, D. A.; Garyantes, T.; Green, D. V. S.; Hertzberg, R. P.; Janzen, W. P.; Paslay, J. W.; Schopfer, U.; Sittampalam, G. S. Impact of high-throughput screening in biomedical research, *Nature Reviews Drug Discovery* **2011**, *10*, 188–195.

[6] Szymański, P.; Markowicz, M.; Mikiciuk-Olasik, E. Adaptation of high-throughput screening in drug discovery—toxicological screening tests, *International Journal of Molecular Sciences* **2012**, *13*, 427–452.

[7] Janzen, W. P. Screening technologies for small molecule discovery: the state of the art, *Chemistry & Biology* **2014**, *21*, 1162–1170.

[8] Markossian, S.; Sittampalam, G. S.; Grossman, A., et al. *Assay guidance manual [internet]*, Bethesda (MD): Eli Lilly & Company and the National Center for Advancing Translational Sciences, available from: https://www.ncbi.nlm.nih.gov/books/NBK53196/, **2004**–.

[9] Singh, G.; Dahlin, J. L.; Walters, M. A. Risk management in early discovery medicinal chemistry, *Methods in Enzymology* **2018**, *610*, 1–25.

[10] Aldrich, C.; Bertozzi, C.; Georg, G. I.; Kiessling, L.; Lindsley, C.; Liotta, D.; Merz Jr, K. M.; Schepartz, A.; Wang, S. The ecstasy and agony of assay interference compounds, *ACS Central Science* **2017**, *8*, 379–382.

[11] Dahlin, J. L.; Walters, M. A. The essential roles of chemistry in high-throughput screening triage, *Future Medicinal Chemistry* **2014**, *6*, 1265–1290.

[12] Roche, O.; Schneider, P.; Zuegge, J.; Guba, W.; Kansy, M.; Alanine, A.; Bleicher, K.; Danel, F.; Gutknecht, E.-M.; Rogers-Evans, M.; Neidhart, W.; Stalder, H.; Dillon, M.; Sjögren, E.; Fotouhi, N.; Gillespie, P.; Goodnow, R.; Harris, W.; Jones, P.; Taniguchi, M.; Tsujii, S.; von der Saal, W.; Zimmermann, G.; Schneider, G. Development of a virtual screening method for identification of "frequent hitters" in compound libraries, *Journal of Medicinal Chemistry* **2002**, *45*, 137–142.

[13] Irwin, J. J.; Duan, D.; Torosyan, H.; Doak, A. K.; Ziebart, K. T.; Sterling, T.; Tumanian, G.; Shoichet, B. K. An aggregation advisor for ligand discovery, *Journal of Medicinal Chemistry* **2015**, *58*, 7076–7087.

[14] Baell, J. B.; Holloway, G. A. New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays, *Journal of Medicinal Chemistry* **2010**, *53*, 2719–2740.

[15] Rishton, G. M. Reactive compounds and in vitro false positives in HTS, *Drug Discovery Today* **1997**, *2*, 382–384.

[16] Nelson, K. M.; Dahlin, J. L.; Bisson, J.; Graham, J.; Pauli, G. F.; Walters, M. A. The essential medicinal chemistry of curcumin: miniperspective, *Journal of Medicinal Chemistry* **2017**, *60*, 1620–1637.

[17] Schmidt, R.; Ehmki, E. S.; Ohm, F.; Ehrlich, H.-C.; Mashychev, A.; Rarey, M. Comparing molecular patterns using the example of SMARTS: theory and algorithms, *Journal of Chemical Information and Modeling* **2019**, *59*, 2560–2571.

[18] Ehmki, E. S.; Schmidt, R.; Ohm, F.; Rarey, M. Comparing molecular patterns using the example of SMARTS: applications and filter collection analysis, *Journal of Chemical Information and Modeling* **2019**, *59*, 2572–2586.

[19] Pearce, B. C.; Sofia, M. J.; Good, A. C.; Drexler, D. M.; Stock, D. A. An empirical process for the design of high-throughput screening deck filters, *Journal of Chemical Information and Modeling* **2006**, *46*, 1060–1068.

[20] Nissink, J. W. M.; Blackburn, S. Quantification of frequent-hitter behavior based on historical high-throughput screening data, *Future Medicinal Chemistry* **2014**, *6*, 1113–1126.

[21] Chakravorty, S. J.; Chan, J.; Greenwood, M. N.; Popa-Burke, I.; Remlinger, K. S.; Pickett, S. D.; Green, D. V. S.; Fillmore, M. C.; Dean, T. W.; Luengo, J. I.; Macarrón, R. Nuisance compounds, PAINS filters, and dark chemical matter in the GSK HTS collection, *SLAS DISCOVERY: Advancing Life Sciences R&D* **2018**, *23*, 532–545.

[22] Wilkens, S. J.; Janes, J.; Su, A. I. HierS: hierarchical scaffold clustering using topological chemical graphs, *Journal of Medicinal Chemistry* **2005**, *48*, 3182–3193.

[23] Yang, J. J.; Ursu, O.; Lipinski, C. A.; Sklar, L. A.; Oprea, T. I.; Bologa, C. G. Badapple: promiscuity patterns from noisy evidence, *Journal of Cheminformatics* **2016**, *8*, 1–14.

[24] Couronne, C.; Koptelov, M.; Zimmermann, A. *PrePeP: A light-weight, extensible tool for predicting frequent hitters*, **2020**.

[25] Koptelov, M.; Zimmermann, A.; Bonnet, P.; Bureau, R.; Crémilleux, B. PrePeP: a tool for the identification and characterization of pan assay interference compounds, **2018**, DOI `10.1145/3219819.3219849`.

[26] McGovern, S. L.; Caselli, E.; Grigorieff, N.; Shoichet, B. K. A common mechanism underlying promiscuous inhibitors from virtual and high-throughput screening, *Journal of Medicinal Chemistry* **2002**, *45*, 1712–1722.

[27] Reker, D.; Bernardes, G. J.; Rodrigues, T. Computational advances in combating colloidal aggregation in drug discovery, *Nature Chemistry* **2019**, *11*, 402–418.

[28] Sassano, M. F.; Doak, A. K.; Roth, B. L.; Shoichet, B. K. Colloidal aggregation causes inhibition of G protein-coupled receptors, *Journal of Medicinal Chemistry* **2013**, *56*, 2406–2414.

[29] Jadhav, A.; Ferreira, R. S.; Klumpp, C.; Mott, B. T.; Austin, C. P.; Inglese, J.; Thomas, C. J.; Maloney, D. J.; Shoichet, B. K.; Simeonov, A. Quantitative analyses of aggregation, autofluorescence, and reactivity artifacts in a screen for inhibitors of a thiol protease, *Journal of Medicinal Chemistry* **2010**, *53*, 37–51.

[30] Feng, B. Y.; Shelat, A.; Doman, T. N.; Guy, R. K.; Shoichet, B. K. High-throughput assays for promiscuous inhibitors, *Nature chemical biology* **2005**, *1*, 146–148.

[31] Rao, H.; Li, Z.; Li, X.; Ma, X.; Ung, C.; Li, H.; Liu, X.; Chen, Y. Identification of small molecule aggregators from large compound libraries by support vector machines, *Journal of Computational Chemistry* **2010**, *31*, 752–763.

[32] Yang, Z.-Y.; Yang, Z.-J.; Dong, J.; Wang, L.-L.; Zhang, L.-X.; Ding, J.-J.; Ding, X.-Q.; Lu, A.-P.; Hou, T.-J.; Cao, D.-S. Structural analysis and identification of colloidal aggregators in drug discovery, *Journal of chemical information and modeling* **2019**, *59*, 3714–3726.

[33] Alves, V. M.; Capuzzi, S. J.; Braga, R. C.; Korn, D.; Hochuli, J. E.; Bowler, K. H.; Yasgar, A.; Rai, G.; Simeonov, A.; Muratov, E. N.; Zakharov, A. V.; Tropsha, A. SCAM detective: accurate predictor of small, colloidally aggregating molecules, *Journal of Chemical Information and Modeling* **2020**, *60*, 4056–4063.

[34] Ash, S.; Cline, M. A.; Homer, R. W.; Hurst, T.; Smith, G. B. SYBYL line notation (SLN): A versatile language for chemical structure representation, *Journal of Chemical Information and Computer Sciences* **1997**, *37*, 71–79.

[35] DAYLIGHT. SMARTS – A Language for Describing Molecular Patterns, `https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html` (visited on 08/29/2021).

[36] Baell, J. B.; Walters, M. A. Chemistry: Chemical con artists foil drug discovery, *Nature News* **2014**, *513*, 481–483.

[37] Vidler, L. R.; Watson, I. A.; Margolis, B. J.; Cummins, D. J.; Brunavs, M. Investigating the behavior of published PAINS alerts using a pharmaceutical company data set, *ACS Medicinal Chemistry Letters* **2018**, *9*, 792–796.

[38] Gilberg, E.; Gütschow, M.; Bajorath, J. X-ray structures of target–ligand complexes containing compounds with assay interference potential, *Journal of Medicinal Chemistry* **2018**, *61*, 1276–1284.

[39] David, L.; Walsh, J.; Sturm, N.; Feierberg, I.; Nissink, J. W. M.; Chen, H.; Bajorath, J.; Engkvist, O. Identification of compounds that interfere with high-throughput screening assay technologies, *ChemMedChem* **2019**, *14*, 1795–1802.

[40] Siramshetty, V. B.; Preissner, R.; Gohlke, B.-O. Exploring activity profiles of PAINS and their structural context in target–ligand complexes, *Journal of Chemical Information and Modeling* **2018**, *58*, 1847–1857.

[41] Schorpp, K.; Rothenaigner, I.; Salmina, E.; Reinshagen, J.; Low, T.; Brenke, J. K.; Gopalakrishnan, J.; Tetko, I. V.; Gul, S.; Hadian, K. Identification of small-molecule frequent hitters from AlphaScreen high-throughput screens, *Journal of Biomolecular Screening* **2014**, *19*, 715–726.

[42] Senger, M. R.; Fraga, C. A.; Dantas, R. F.; Silva Jr, F. P. Filtering promiscuous compounds in early drug discovery: is it a good idea?, *Drug Discovery Today* **2016**, *21*, 868–872.

[43] Dahlin, J. L.; Auld, D. S.; Rothenaigner, I.; Haney, S.; Sexton, J. Z.; Nissink, J. W. M.; Walsh, J.; Lee, J. A.; Strelow, J. M.; Willard, F. S., et al. Nuisance compounds in cellular assays, *Cell Chemical Biology* **2021**, *28*, 356–370.

[44] Lagorce, D.; Oliveira, N.; Miteva, M. A.; Villoutreix, B. O. Pan-assay interference compounds (PAINS) that may not be too painful for chemical biology projects, *Drug Discovery Today* **2017**, *22*, 1131–1133.

[45] Bajorath, J. Evolution of assay interference concepts in drug discovery, *Expert Opinion on Drug Discovery* **2021**, *16*, 719–721.

[46] Sun, J.; Zhong, H.; Wang, K.; Li, N.; Chen, L. Gains from no real PAINS: Where 'Fair Trial Strategy' stands in the development of multi-target ligands, *Acta Pharmaceutica Sinica B* **2021**, DOI `https://doi.org/10.1016/j.apsb.2021.02.023`.

[47] Capuzzi, S. J.; Muratov, E. N.; Tropsha, A. Phantom PAINS: problems with the utility of alerts for pan-assay interference compounds, *Journal of Chemical Information and Modeling* **2017**, *57*, 417–427.

[48] Baell, J. B.; Nissink, J. W. M. Seven year itch: pan-assay interference compounds (PAINS) in 2017 utility and limitations, *ACS Chemical Biology* **2018**, *13*, 36–44.

[49] Jasial, S.; Gilberg, E.; Blaschke, T.; Bajorath, J. Machine learning distinguishes with high accuracy between pan-assay interference compounds that are promiscuous or represent dark chemical matter, *Journal of Medicinal Chemistry* **2018**, *61*, 10255–10264.

[50] Jöst, C.; Nitsche, C.; Scholz, T.; Roux, L.; Klein, C. D. Promiscuity and selectivity in covalent enzyme inhibition: a systematic study of electrophilic fragments, *Journal of Medicinal Chemistry* **2014**, *57*, 7590–7599.

[51] Šink, R.; Gobec, S.; Pečar, S.; Zega, A. False positives in the early stages of drug discovery, *Current Medicinal Chemistry* **2010**, *17*, 4231–4255.

[52] Curpăn, R.; Avram, S.; Vianello, R.; Bologa, C. Exploring the biological promiscuity of high-throughput screening hits through DFT calculations, *Bioorganic & Medicinal Chemistry* **2014**, *22*, 2461–2468.

[53] Metz, J. T.; Huth, J. R.; Hajduk, P. J. Enhancement of chemical rules for predicting compound reactivity towards protein thiol groups, *Journal of computer-aided molecular design* **2007**, *21*, 139–144.

[54] Zega, A. NMR methods for identification of false positives in biochemical screens: miniperspective, *Journal of Medicinal Chemistry* **2017**, *60*, 9437–9447.

[55] Hann, M.; Hudson, B.; Lewell, X.; Lifely, R.; Miller, L.; Ramsden, N. Strategic pooling of compounds for high-throughput screening, *Journal of Chemical Information and Computer Sciences* **1999**, *39*, 897–902.

[56] Matlock, M. K.; Hughes, T. B.; Dahlin, J. L.; Swamidass, S. J. Modeling small-molecule reactivity identifies promiscuous bioactive compounds, *Journal of Chemical Information and Modeling* **2018**, *58*, 1483–1500.

[57] Wassermann, A. M.; Lounkine, E.; Hoepfner, D.; Le Goff, G.; King, F. J.; Studer, C.; Peltier, J. M.; Grippo, M. L.; Prindle, V.; Tao, J.; Schuffenhauer, A.; Wallace, I. M.; Chen, S.; Krastel, P.; Cobos-Correa, A.; Parker, C. N.; Davies, J. W.; Glick, M. Dark chemical matter as a promising starting point for drug lead discovery, *Nature Chemical Biology* **2015**, *11*, 958–966.

[58] Muegge, I.; Mukherjee, P. Performance of dark chemical matter in high throughput screening, *Journal of Medicinal Chemistry* **2016**, *59*, 9806–9813.

[59] Ballante, F.; Rudling, A.; Zeifman, A.; Luttens, A.; Vo, D. D.; Irwin, J. J.; Kihlberg, J.; Brea, J.; Loza, M. I.; Carlsson, J. Docking finds GPCR ligands in dark chemical matter, *Journal of Medicinal Chemistry* **2019**, *63*, 613–620.

[60] Medina-Franco, J. L.; Giulianotti, M. A.; Welmaker, G. S.; Houghten, R. A. Shifting from the single to the multitarget paradigm in drug discovery, *Drug Discovery Today* **2013**, *18*, 495–501.

[61] Schneider, P.; Schneider, G. Privileged structures revisited, *Angewandte Chemie International Edition* **2017**, *56*, 7971–7974.

[62] Evans, B. E.; Rittle, K. E.; Bock, M. G.; DiPardo, R. M.; Freidinger, R. M.; Whitter, W. L.; Lundell, G. F.; Veber, D. F.; Anderson, P. S.; Chang, R. S. L.; Lotti, V. J.; Cerino, D. J.; Chen, T. B.; Kling, P. J.; Kunkel, K. A.; Springer, J. P.; Hirshfield, J. Methods for drug discovery: development of potent, selective, orally effective cholecystokinin antagonists, *Journal of Medicinal Chemistry* **1988**, *31*, 2235–2246.

[63] Gilberg, E.; Jasial, S.; Stumpfe, D.; Dimova, D.; Bajorath, J. Highly promiscuous small molecules from biological screening assays include many pan-assay interference compounds but also candidates for polypharmacology, *Journal of Medicinal Chemistry* **2016**, *59*, 10285–10290.

[64] Hu, Y.; Bajorath, J. High-resolution view of compound promiscuity, *F1000Research* **2013**, *2*, 1–10.

[65] Ferreira, L. L.; Andricopulo, A. D. ADMET modeling approaches in drug discovery, *Drug Discovery Today* **2019**, *24*, 1157–1165.

[66] Kar, S.; Leszczynski, J. Open access in silico tools to predict the ADMET profiling of drug candidates, *Expert Opinion on Drug Discovery* **2020**, *15*, 1473–1487.

[67] Lounkine, E.; Keiser, M. J.; Whitebread, S.; Mikhailov, D.; Hamon, J.; Jenkins, J. L.; Lavan, P.; Weber, E.; Doak, A. K.; Côté, S.; Shoichet, B. K.; Urban, L. Large-scale prediction and testing of drug activity on side-effect targets, *Nature* **2012**, *486*, 361–367.

[68] Simeonov, A.; Jadhav, A.; Thomas, C. J.; Wang, Y.; Huang, R.; Southall, N. T.; Shinn, P.; Smith, J.; Austin, C. P.; Auld, D. S.; Inglese, J. Fluorescence spectroscopic profiling of compound libraries, *Journal of Medicinal Chemistry* **2008**, *51*, 2363–2371.

[69] Auld, D. S.; Southall, N. T.; Jadhav, A.; Johnson, R. L.; Diller, D. J.; Simeonov, A.; Austin, C. P.; Inglese, J. Characterization of chemical libraries for luciferase inhibitory activity, *Journal of Medicinal Chemistry* **2008**, *51*, 2372–2386.

[70] Ghosh, D.; Koch, U.; Hadian, K.; Sattler, M.; Tetko, I. V. Luciferase advisor: high-accuracy model to flag false positive hits in luciferase HTS assays, *Journal of Chemical Information and Modeling* **2018**, *58*, 933–942.

[71] Yang, Z.-Y.; Dong, J.; Yang, Z.-J.; Lu, A.-P.; Hou, T.-J.; Cao, D.-S. Structural analysis and identification of false positive hits in luciferase-based assays, *Journal of Chemical Information and Modeling* **2020**, *60*, 2031–2043.

[72] Borrel, A.; Huang, R.; Sakamuru, S.; Xia, M.; Simeonov, A.; Mansouri, K.; Houck, K. A.; Judson, R. S.; Kleinstreuer, N. C. High-throughput screening to predict chemical-assay interference, *Scientific Reports* **2020**, *10*, 1–20.

[73] Bruns, R. F.; Watson, I. A. Rules for identifying potentially reactive or promiscuous compounds, *Journal of Medicinal Chemistry* **2012**, *55*, 9763–9772.

[74] Brenke, J. K.; Salmina, E. S.; Ringelstetter, L.; Dornauer, S.; Kuzikov, M.; Rothenaigner, I.; Schorpp, K.; Giehler, F.; Gopalakrishnan, J.; Kieser, A.; Gul, S.; Tetko, I. V.; Hadian, K. Identification of Small-Molecule Frequent Hitters of Glutathione S-Transferase–Glutathione Interaction, *Journal of Biomolecular Screening* **2016**, *21*, 596–607.

[75] Lor, L. A.; Schneck, J.; Mcnulty, D. E.; Diaz, E.; Brandt, M.; Thrall, S. H.; Schwartz, B. A simple assay for detection of small-molecule redox activity, *Journal of Biomolecular Screening* **2007**, *12*, 881–890.

[76] Tomohara, K.; Adachi, I.; Horino, Y.; Kesamaru, H.; Abe, H.; Suyama, K.; Nose, T. DMSO-Perturbing assay for identifying promiscuous enzyme inhibitors, *ACS Medicinal Chemistry Letters* **2019**, *10*, 923–928.

[77] Crisman, T. J.; Parker, C. N.; Jenkins, J. L.; Scheiber, J.; Thoma, M.; Kang, Z. B.; Kim, R.; Bender, A.; Nettles, J. H.; Davies, J. W.; Glick, M. Understanding false positives in reporter gene assays: in silico chemogenomics approaches to prioritize cell-based HTS data, *Journal of Chemical Information and Modeling* **2007**, *47*, 1319–1327.

[78] Ekins, S.; Kaneko, T.; Lipinski, C. A.; Bradford, J.; Dole, K.; Spektor, A.; Gregory, K.; Blondeau, D.; Ernst, S.; Yang, J.; Goncharoff, N.; Hohmana, M. M.; Bunina, B. A. Analysis and hit filtering of a very large library of compounds screened against Mycobacterium tuberculosis, *Molecular BioSystems* **2010**, *6*, 2316–2324.

[79] Alves, V.; Muratov, E.; Capuzzi, S.; Politi, R.; Low, Y.; Braga, R.; Zakharov, A. V.; Sedykh, A.; Mokshyna, E.; Farag, S.; Andrade, C.; Kuz'min, V.; Fourches, D.; Tropsha, A. Alarms about structural alerts, *Green Chemistry* **2016**, *18*, 4348–4360.

[80] Kenny, P. W. Comment on the ecstasy and agony of assay interference compounds, *Journal of Chemical Information and Modeling* **2017**, *57*, 2640–2645.

[81] Plemper, R. K.; Cox, R. M. Biology must develop herd immunity against bad-actor molecules, *PLoS Pathogens* **2018**, *14*, 1–6.

[82] Baell, J. B. Screening-based translation of public research encounters painful problems, *ACS Medicinal Chemistry Letters* **2015**, *6*, 229–234.

[83] Blake, J. F. Identification and evaluation of molecular properties related to preclinical optimization and clinical fate, *Medicinal Chemistry* **2005**, *1*, 649–655.

[84] NIH Molecular Libraries Small Molecule Repository. `https://grants.nih.gov/grants/guide/notice-files/not-rm-07-005.html` (visited on 08/29/2021).

[85] Brenk, R.; Schipani, A.; James, D.; Krasowski, A.; Gilbert, I. H.; Frearson, J.; Wyatt, P. G. Lessons learnt from assembling screening libraries for drug discovery for neglected diseases, *ChemMedChem* **2008**, *3*, 435–444.

[86] Sushko, I.; Salmina, E.; Potemkin, V.; Poda, G.; Tetko, I. ToxAlerts: a web server of structural alerts for toxic chemicals and compounds with potential adverse reactions, *Journal of Chemical Information and Modeling* **2012**, *52*, 2310–2316.

[87] Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H. PubChem substance and compound databases, *Nucleic Acids Research* **2016**, *44*, D1202–D1213.

[88] Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. PubChem 2019 update: improved access to chemical data, *Nucleic Acids Research* **2019**, *47*, D1102–D1109.

[89] PubChem Database, `https://pubchem.ncbi.nlm.nih.gov/` (visited on 08/29/2021).

[90] Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magariños, M. P.; Overington, J. P.; Papadatos, G.; Smit, I.; Leach, A. R. The ChEMBL database in 2017, *Nucleic Acids Research* **2017**, *45*, D945–D954.

[91] Avram, S.; Curpan, R.; Bora, A.; Neanu, C.; Halip, L. Enhancing molecular promiscuity evaluation through assay profiles, *Pharmaceutical Research* **2018**, *35*, 1–10.

[92] Hu, Y.; Bajorath, J. How promiscuous are pharmaceutically relevant compounds? A data-driven assessment, *The AAPS Journal* **2013**, *15*, 104–111.

[93] 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.; Maciejewski, A.; Gale, N.; Wilson, A.; Chin, L.; Cummings, R.; Le, D.; Pon, A.; Knox, C.; Wilson, M. DrugBank 5.0: a major update to the DrugBank database for 2018, *Nucleic acids research* **2018**, *46*, D1074–D1082.

[94] Bisson, J.; McAlpine, J. B.; Friesen, J. B.; Chen, S.-N.; Graham, J.; Pauli, G. F. Can invalid bioactives undermine natural product-based drug discovery?, *Journal of Medicinal Chemistry* **2016**, *59*, 1671–1690.

[95] Chen, Y.; Garcia de Lomana, M.; Friedrich, N.-O.; Kirchmair, J. Characterization of the chemical space of known and readily obtainable natural products, *Journal of Chemical Information and Modeling* **2018**, *58*, 1518–1532.

[96] Hert, J.; Irwin, J. J.; Laggner, C.; Keiser, M. J.; Shoichet, B. K. Quantifying biogenic bias in screening libraries, *Nature Chemical Biology* **2009**, *5*, 479–483.

[97] Hu, Y.; Bajorath, J. Compound promiscuity: what can we learn from current data?, *Drug Discovery Today* **2013**, *18*, 644–650.

[98] Hu, Y.; Gupta-Ostermann, D.; Bajorath, J. Exploring compound promiscuity patterns and multi-target activity spaces, *Computational and Structural Biotechnology Journal* **2014**, *9*, e201401003.

[99] Jasial, S.; Hu, Y.; Bajorath, J. Determining the degree of promiscuity of extensively assayed compounds, *PLoS One* **2016**, *11*, 1–15.

[100] Hu, Y.; Jasial, S.; Bajorath, J. Promiscuity progression of bioactive compounds over time, *F1000Research* **2015**, *4*, 1–18.

[101] Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-learn: machine learning in Python, *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.

[102] Buitinck, L.; Louppe, G.; Blondel, M.; Pedregosa, F.; Mueller, A.; Grisel, O.; Niculae, V.; Prettenhofer, P.; Gramfort, A.; Grobler, J.; Layton, R.; VanderPlas, J.; Joly, A.; Holt, B.; Varoquaux, G. in ECML PKDD Workshop: Languages for Data Mining and Machine Learning, **2013**, pp. 108–122.

[103] Breiman, L. Random forests, *Machine learning* **2001**, *45*, 5–32.

[104] Geurts, P.; Ernst, D.; Wehenkel, L. Extremely randomized trees, *Machine Learning* **2006**, *63*, 3–42.

[105] Hinton, G. E. Connectionist learning procedures, *Artificial Intelligence* **1989**, *40*, 185–234.

[106] Glorot, X.; Bengio, Y. in Proceedings of the thirteenth international conference on artificial intelligence and statistics, JMLR Workshop and Conference Proceedings, **2010**, pp. 249–256.

[107] He, K.; Zhang, X.; Ren, S.; Sun, J. Delving deep into rectifiers: surpassing human-level performance on imagenet classification, *arXiv preprint arXiv:1502.01852* **2015**.

[108] Kingma, D. P.; Ba, J. Adam: A method for stochastic optimization, *arXiv preprint arXiv:1412.6980* **2014**.

[109] Bento, A. P.; Hersey, A.; Félix, E.; Landrum, G.; Gaulton, A.; Atkinson, F.; Bellis, L. J.; De Veij, M.; Leach, A. R. An open source chemical structure curation pipeline using RDKit, *Journal of Cheminformatics* **2020**, *12*, 1–16.

[110] RDKit: Open-Source Cheminformatics Software, `https://www.rdkit.org/` (visited on 08/29/2021).

[111] Cock, P. J. A.; Antao, T.; Chang, J. T.; Chapman, B. A.; Cox, C. J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; de Hoon, M. J. L. Biopython: freely available Python tools for computational molecular biology and bioinformatics, *Bioinformatics* **2009**, *25*, 1422–1423.

[112] ULC, C. C. G. Molecular Operating Environment (MOE), `https://www.chemcomp.com/Products.htm` (visited on 08/29/2021).

[113] Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL keys for use in drug discovery, *Journal of Chemical Information and Computer Sciences* **2002**, *42*, 1273–1280.

[114] Morgan, H. L. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service., *Journal of Chemical Documentation* **1965**, *5*, 107–113.

[115] Rogers, D.; Hahn, M. Extended-connectivity fingerprints, *Journal of Chemical Information and Modeling* **2010**, *50*, 742–754.

[116] Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme, *Biochimica et Biophysica Acta (BBA)-Protein Structure* **1975**, *405*, 442–451.

[117] Django (Version 1.5) [Computer Software], **2013**, `https://djangoproject.com` (visited on 08/29/2021).

[118] Bienfait, B.; Ertl, P. JSME: a free molecule editor in JavaScript, *Journal of Cheminformatics* **2013**, *5*, 1–6.

[119] Riniker, S.; Landrum, G. A. Similarity maps-a visualization strategy for molecular fingerprints and machine-learning methods, *Journal of Cheminformatics* **2013**, *5*, 1–7.

8. Appendix

A Gefahrstoffe nach GHS

In this work no hazardous compounds according to the GHS (Globally Harmonized System Of Classification and Labeling of Chemicals) were used.

B Supporting information for [D2]

Supporting Information for the following publication:

Stork, C.; Wagner, J.; Friedrich N.-O.; de Bruyn Kops, C.; Šícho, M. and Kirchmair, J. Hit Dexter: A Machine-Learning Model for the Prediction of Frequent Hitters, *ChemMedChem* **2018**, *13*, 564–571.

CHEMMEDCHEM
Chemistry & Drug Discovery

Supporting Information

Hit Dexter: A Machine-Learning Model for the Prediction of Frequent Hitters

Conrad Stork,[a] Johannes Wagner,[a] Nils-Ole Friedrich,[a] Christina de Bruyn Kops,[a] Martin Šícho,[a, b] and Johannes Kirchmair*[a]

cmdc_201700673_sm_miscellaneous_information.pdf

Supporting Information

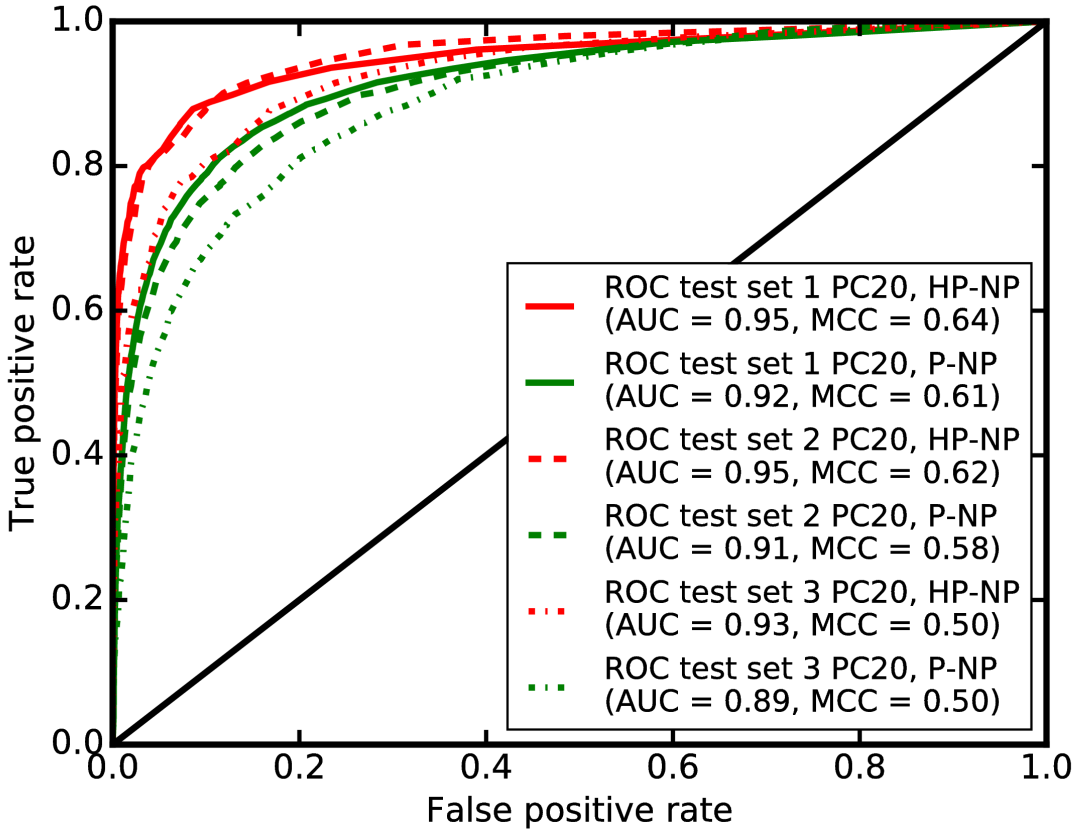


Figure S1. ROC curves obtained for the three independent test sets, which are independent test sets created from the PC20 data set.

Table S1. Molecular descriptors for principal component analysis (calculated with MOE).

Code	Class	Description
a_acc	2D	Number of H-bond acceptor atoms
a_acid	2D	Number of acidic atoms
a_aro	2D	Number of aromatic atoms
a_base	2D	Number of basic atoms
a_don	2D	Number of H-bond donor atoms
a_heavy	2D	Number of heavy atoms
a_hyd	2D	Number of hydrophobic atoms
a_nB	2D	Number of boron atoms
a_nBr	2D	Number of bromine atoms
a_nC	2D	Number of carbon atoms
a_nCl	2D	Number of chlorine atoms
a_nF	2D	Number of fluorine atoms
a_nH	2D	Number of hydrogen atoms
a_nI	2D	Number of iodine atoms
a_nN	2D	Number of nitrogen atoms
a_nO	2D	Number of oxygen atoms
a_nP	2D	Number of phosphorus atoms
a_nS	2D	Number of sulfur atoms
b_ar	2D	Number of aromatic bonds
b_count	2D	Number of bonds
b_double	2D	Number of double bonds
b_rotN	2D	Number of rotatable bonds
b_rotR	2D	Fraction of rotatable bonds
b_single	2D	Number of single bonds
b_triple	2D	Number of triple bonds
chiral	2D	Number of chiral centers
FCharge	2D	Sum of formal charges

logP(o/w)	2D	Log octanol/water partition coefficient
logS	2D	Log Solubility in Water
mr	2D	Molar refractivity
PC+	2D	Total positive partial charge
PC-	2D	Total negative partial charge
rings	2D	Number of rings
TPSA	2D	Topological Polar Surface Area (A**2)
vdw_area	2D	Van der Waals surface area (A**2)
vdw_vol	2D	Van der Waals volume (A**3)
vsa_acc	2D	VDW acceptor surface area (A**2)
vsa_acid	2D	VDW acidic surface area (A**2)
vsa_base	2D	VDW basic surface area (A**2)
vsa_don	2D	VDW donor surface area (A**2)
vsa_hyd	2D	VDW hydrophobe surface area (A**2)
vsa_other	2D	VDW other surface area (A**2)
vsa_pol	2D	VDW polar surface area (A**2)
Weight	2D	Molecular weight (CRC)

Table S2. Grid Search Results for P-NP Classifiers Trained on the PC20 Dataset.

max_features \ number of estimators	Metric	10	50	100	150	200	250	300	400	500	600
sqrt	MCC	0.55	0.59	0.59	0.59	0.59	0.59	0.59	0.59	0.59	0.59
sqrt	AUC	0.88	0.91	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92
0.2	MCC	0.57	0.59	0.60	0.60	0.60	0.60	0.60	0.60	0.60	0.60
0.2	AUC	0.88	0.91	0.91	0.92	0.92	0.92	0.92	0.92	0.92	0.92
0.4	MCC	0.56	0.59	0.59	0.59	0.59	0.59	0.59	0.59	0.59	0.59
0.4	AUC	0.88	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.91
0.6	MCC	0.56	0.58	0.58	0.58	0.58	0.58	0.58	0.58	0.58	0.58
0.6	AUC	0.88	0.90	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.91
0.8	MCC	0.54	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.56
0.8	AUC	0.87	0.89	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90
None (1.0)	MCC	0.45	0.46	0.46	0.46	0.46	0.46	0.46	0.46	0.46	0.46
None (1.0)	AUC	0.77	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78

Table S3. Grid Search Results for HP-NP Classifiers Trained on the PC20 Dataset.

max_features \ number of estimators	Metric	10	50	100	150	200	250	300	400	500	600
sqrt	MCC	0.56	0.59	0.60	0.60	0.60	0.60	0.60	0.60	0.60	0.60
sqrt	AUC	0.89	0.95	0.95	0.96	0.96	0.96	0.96	0.96	0.96	0.96
0.2	MCC	0.58	0.61	0.62	0.62	0.61	0.62	0.62	0.62	0.62	0.62
0.2	AUC	0.90	0.94	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.96
0.4	MCC	0.58	0.60	0.60	0.60	0.60	0.60	0.60	0.61	0.61	0.60
0.4	AUC	0.89	0.94	0.94	0.95	0.95	0.95	0.95	0.95	0.95	0.95
0.6	MCC	0.57	0.59	0.59	0.59	0.59	0.59	0.59	0.59	0.59	0.59
0.6	AUC	0.88	0.93	0.94	0.94	0.94	0.95	0.95	0.95	0.95	0.95
0.8	MCC	0.54	0.56	0.56	0.55	0.56	0.56	0.56	0.56	0.56	0.56
0.8	AUC	0.87	0.91	0.92	0.93	0.93	0.93	0.93	0.93	0.94	0.94
None (1.0)	MCC	0.44	0.44	0.44	0.44	0.44	0.44	0.44	0.44	0.44	0.44
None (1.0)	AUC	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77

Table S4. Grid Search Results for P-NP Classifiers Trained on the PC50 Dataset.

max_features \ number of estimators	Metric	10	50	100	150	200	250	300	400	500	600
sqrt	MCC	0.53	0.57	0.58	0.58	0.58	0.58	0.58	0.58	0.58	0.58
sqrt	AUC	0.88	0.91	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92
0.2	MCC	0.55	0.58	0.58	0.58	0.58	0.58	0.58	0.58	0.58	0.58
0.2	AUC	0.88	0.91	0.91	0.91	0.91	0.91	0.91	0.92	0.92	0.92
0.4	MCC	0.54	0.57	0.57	0.57	0.58	0.58	0.57	0.58	0.58	0.58
0.4	AUC	0.87	0.90	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.91
0.6	MCC	0.53	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.57
0.6	AUC	0.87	0.90	0.90	0.90	0.91	0.91	0.91	0.91	0.91	0.91
0.8	MCC	0.52	0.54	0.54	0.55	0.55	0.55	0.55	0.55	0.55	0.55
0.8	AUC	0.86	0.89	0.89	0.90	0.90	0.90	0.90	0.90	0.90	0.90
None (1.0)	MCC	0.44	0.44	0.44	0.44	0.44	0.44	0.44	0.44	0.44	0.44
None (1.0)	AUC	0.76	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77

Table S5. Grid Search Results for HP-NP Classifiers Trained on the PC50 Dataset.

max_features \ number of estimators	Metric	10	50	100	150	200	250	300	400	500	600
sqrt	MCC	0.54	0.59	0.59	0.59	0.59	0.59	0.59	0.59	0.59	0.59
sqrt	AUC	0.90	0.95	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96
0.2	MCC	0.57	0.60	0.61	0.61	0.61	0.61	0.61	0.61	0.61	0.61
0.2	AUC	0.90	0.94	0.95	0.96	0.96	0.96	0.96	0.96	0.96	0.96
0.4	MCC	0.56	0.60	0.60	0.60	0.60	0.60	0.60	0.60	0.60	0.60
0.4	AUC	0.89	0.94	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
0.6	MCC	0.56	0.58	0.58	0.59	0.59	0.58	0.58	0.58	0.58	0.59
0.6	AUC	0.88	0.93	0.94	0.94	0.94	0.94	0.95	0.95	0.95	0.95
0.8	MCC	0.53	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.55
0.8	AUC	0.87	0.91	0.92	0.93	0.93	0.93	0.93	0.93	0.94	0.94
None (1.0)	MCC	0.43	0.44	0.44	0.44	0.44	0.44	0.44	0.44	0.44	0.44
None (1.0)	AUC	0.76	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77

C Supporting information for [D3]

Supporting Information for the following publication:

Stork, C.; Chen, Y.; Šícho, M. and Kirchmair, J. Hit Dexter 2.0: Machine-Learning Models for the Prediction of Frequent Hitters, *J. Chem. Inf. Model.* **2018**, *59*, 1030–1043.

Supporting Information

Hit Dexter 2.0: Machine-Learning Models for the Prediction of Frequent Hitters

Conrad Stork,[1] Ya Chen,[1] Martin Šícho,[1,2] Johannes Kirchmair[1,3,4]*

[1] Center for Bioinformatics (ZBH), Department of Computer Science, Faculty of Mathematics, Informatics and Natural Sciences, Universität Hamburg, Hamburg, 20146, Germany

[2] CZ-OPENSCREEN: National Infrastructure for Chemical Biology, Laboratory of Informatics and Chemistry, Faculty of Chemical Technology, University of Chemistry and Technology Prague, 166 28 Prague 6, Czech Republic

[3] Department of Chemistry, University of Bergen, N-5020 Bergen, Norway

[4] Computational Biology Unit (CBU), University of Bergen, N-5020 Bergen, Norway

[*]J. Kirchmair. E-mail: johannes.kirchmair@uib.no. Tel.: +47 55 58 34 64.

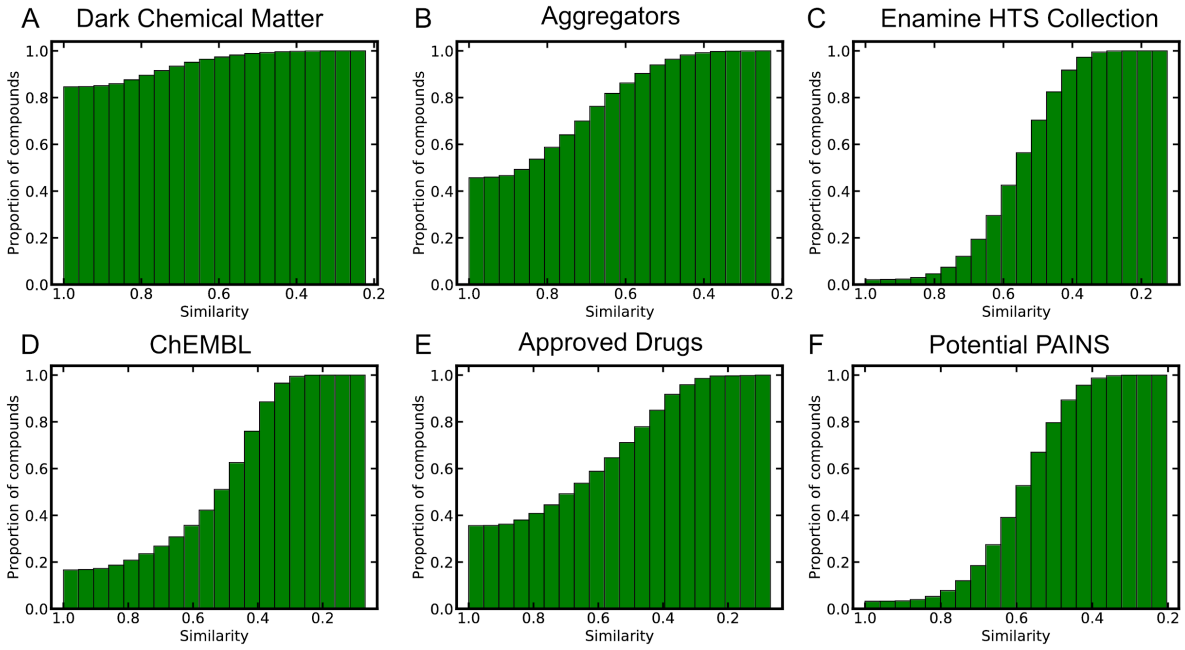


Figure S1. Proportion of (A) dark chemical matter compounds, (B) known aggregators, (C) screening compounds from the Enamine HTS Collection, (D) compounds of the ChEMBL database, (E) approved drugs from DrugBank and (F) compounds of the Enamine HTS collection that match at least one PAINS pattern, represented by the PSA50 training set of the P-NP classifier at a given minimum similarity (Tanimoto coefficient calculated from Morgan2 fingerprints).

D Supporting information for [D5]

Supporting Information for the following publication:

Stork, C.; Mathai N. and Kirchmair, J. Computational prediction of frequent hitters in target-based and cell-based assays, *Artificial Intelligence in the Life Sciences* **2021**, *1*, 100007.

Supporting Information

Computational prediction of frequent hitters in target-based and cell-based assays

Conrad Stork,[1] Neann Mathai[2] and Johannes Kirchmair[1,2,3]*

[1] Universität Hamburg, Faculty of Mathematics, Informatics and Natural Sciences, Department of Informatics, Center for Bioinformatics, 20146 Hamburg, Germany.

[2] Department of Chemistry and Computational Biology Unit (CBU), University of Bergen, N-5020 Bergen, Norway.

[3] Department of Pharmaceutical Sciences, Division of Pharmaceutical Chemistry, Faculty of Life Sciences, University of Vienna, 1090 Vienna, Austria.

[*] Corresponding author email: johannes.kirchmair@univie.ac.at

Table SI_1: Number of Assay Data Sets Removed During Data Preprocessing.

	Number of assays			
Assay annotation	**annotated**	**removed because they have not exactly one GI annotated**	**removed due to unusually high hit rates[2]**	**after removing all assays with unusually high hit rates and assays without a single GI**
target-based	425	66[1]	7	352
cell-based	369	-	4	365
extended cell-based	619	-	7	612

[1] These assays were not considered during the calculation of the average assay hit rate and σ.

[2] Assays with a hit rate greater than the average plus 3σ for the respective assay data set were removed.

E Scientific contribution

[A1] Šícho, M.; de Bruyn Kops, C.; Stork, C.; Svozil, D.; Kirchmair, J. FAME 2: simple and effective machine learning model of cytochrome P450 regioselectivity, *Journal of Chemical Information and Modeling* **2017**, *57*, 1832–1846.

[A2] Chen, Y.; Stork, C.; Hirte, S.; Kirchmair, J. NP-Scout: machine learning approach for the quantification and visualization of the natural product-likeness of small molecules, *Biomolecules* **2019**, *9*, 43.

[A3] De Bruyn Kops, C.; Stork, C.; Šícho, M.; Kochev, N.; Svozil, D.; Jeliazkova, N.; Kirchmair, J. GLORY: generator of the structures of likely cytochrome P450 metabolites based on predicted sites of metabolism, *Frontiers in Chemistry* **2019**, *7*, 402.

[A4] Šícho, M.; Stork, C.; Mazzolari, A.; de Bruyn Kops, C.; Pedretti, A.; Testa, B.; Vistoli, G.; Svozil, D.; Kirchmair, J. FAME 3: predicting the sites of metabolism in synthetic compounds and natural products for phase 1 and phase 2 metabolic enzymes, *Journal of Chemical Information and Modeling* **2019**, *59*, 3400–3412.

[A5] Wilm, A.; Stork, C.; Bauer, C.; Schepky, A.; Kühnl, J.; Kirchmair, J. Skin Doctor: machine learning models for skin sensitization prediction that provide estimates and indicators of prediction reliability, *International Journal of Molecular Sciences* **2019**, *20*, 4833.

[A6] Fan, N.; Bauer, C. A.; Stork, C.; de Bruyn Kops, C.; Kirchmair, J. ALADDIN: docking approach augmented by machine learning for protein structure selection yields superior virtual screening performance, *Molecular Informatics* **2020**, *39*, 1900103.

[A7] Wilm, A.; Norinder, U.; Agea, M. I.; de Bruyn Kops, C.; Stork, C.; Kühnl, J.; Kirchmair, J. Skin Doctor CP: conformal prediction of the skin sensitization potential of small organic molecules, *Chemical Research in Toxicology* **2020**, *34*, 330–344.

[A8] Holmer, M.; de Bruyn Kops, C.; Stork, C.; Kirchmair, J. CYPstrate: a Set of machine learning models for the accurate classification of cytochrome P450 enzyme substrates and non-substrates, *Molecules* **2021**, *26*, 1–20.

[A9] Mathai, N.; Stork, C.; Kirchmair, J. BonMOLière: small-sized libraries of readily purchasable compounds, optimized to produce genuine hits in biological screens across the protein space, *International Journal of Molecular Sciences* **2021**, *22*, 1–20.

[A10] Plonka, W.; Stork, C.; Šícho, M.; Kirchmair, J. CYPlebrity: machine learning models for the prediction of inhibitors of cytochrome P450 enzymes, *Bioorganic & Medicinal Chemistry* **2021**, 116388.

[A11] Wilm, A.; Garcia de Lomana, M.; Stork, C.; Mathai, N.; Hirte, S.; Norinder, U.; Kühnl, J.; Kirchmair, J. Predicting the skin sensitization potential of small molecules with machine learning models trained on biologically meaningful descriptors, *Pharmaceuticals* **2021**, *14*, 1–21.

9. Danksagung

An erster Stelle möchte ich meinem Doktorvater Johannes Kirchmair danken, der diese Arbeit betreut und ermöglicht hat. Ich bedanke mich für seine Hilfe und Unterstützung während der gesamten Arbeit und dafür, dass er immer wieder neue Ideen und Visionen mit mir entwickelt und umgesetzt hat. Johannes hat mich im richtigen Maß gefordert und gefördert, mir aber auch die Zeit gegeben, mich weiterzuentwickeln.

Als nächstes möchte ich meinen Zweitgutachter Herr Prof. Dr. Andrew Torda für seine Zeit und sein Interesse danken, diese Arbeit zu lesen und zu bewerten.

Ich möchte mich bei der gesamten ACM/COMP3D Gruppe bedanken, besonders bei Nils und Christina, die mir zu Beginn meiner Arbeit viel beigebracht haben und viele lange Diskussionen mit mir geführt haben. Es war immer schön mit euch in einem Büro zu arbeiten. Anya, Ningning, Martin und Meliné danke ich für interessante Diskussionen, Kollaborationen, tolle Konferenzen und Ausflüge. Zudem bedanke ich mich bei Marina, Neann und Christina für das Korrekturlesen von Papern und dieser Arbeit.

Großer Dank geht an Anke und ihre Familie. Unsere Familien sind gute Freunde auch über die Arbeit hinaus geworden.

Ich möchte den großen technischen Support, den ich im ZBH von Gerd und Jörn erhalten habe, hervorheben. Es ist nicht selbstverständlich mit wie viel Geduld und Ausdauer meine vielen Fragen und Anliegen behandelt wurden. Ihr habt mir sehr viel beigebracht.

Matthias Rarey und seiner Gruppe danke ich für viele Diskussionen und Seminare, die wir zusammen hatten. Es war immer interessant bei euch. Besonderer Dank geht an Raini, Flaxi und den Ehmki.

Ein großer Dank geht auch die guten Freunde neben der Arbeit Robert, Alex und Timo. Robert für seine intensiven Ausflüge ins Waldstadion. Alex für die vielen und langen Diskussionen über Alles beim Mittagessen und Timo für die guten Spieletage die wir mit dir verbracht haben.

Zum Schluss danke ich meiner tollen Familie für die großartige Unterstützung.

10. Eidesstattliche Versicherung

Declaration on oath
I hereby declare, on oath, that I have written the present dissertation by my own and have not used other than the acknowledged resources and aids.

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.

Conrad Stork
Hamburg, 6th of September, 2021