# Development of a Web-based System for Geometric Mining in Protein Structure Collections

Kumulative Dissertation

zur Erlangung des akademischen Grades

*Dr. rer. nat.*

an der Fakultät
für Mathematik, Informatik und Naturwissenschaften der
Universität Hamburg

eingereicht beim Fach-Promotionsausschuss Informatik

von

**Konrad Diedrich**

geboren in Henstedt-Ulzburg

Hamburg, November 2024

Erstgutachter: Prof. Dr. Matthias Rarey
Zweitgutachter: Prof. Dr. Andrew Torda

Tag der Disputation: 04.04.2025

# Kurzfassung

Proteine sind für das Leben essenziell und folglich auch für lebenswissenschaftliche Fragestellungen wie der Wirkstoffentwicklung. Die Bindung von Liganden an Proteine aktiviert, deaktiviert oder reguliert biologische Prozesse in Organismen und kontrolliert deshalb deren Entwicklung, Fähigkeiten und Erhaltung. Dreidimensionale Strukturen von Protein-Ligand-Komplexen und Proteinen werden in Datenbanken wie der Protein Data Bank (PDB) gesammelt und stellen eine kontinuierlich wachsende und immer wichtigere Datenquelle für den computergestützten Wirkstoffentwurf dar. Für die Forschung sind Softwareanwendungen erforderlich, um die Ligandenbindung ermöglichenden Eigenschaften von Protein-Ligand-Bindestellen im Rahmen des Wirkstoffentwurfs zu analysieren. Insbesondere die Visualisierung und 3D-Suche dieser Daten generiert wertvolle Erkenntnisse. Die Visualisierung einer Protein-Ligand-Bindestelle in Berichten, Präsentationen und Artikeln unterstützt Wissenschaftler dabei, die Affinität, Selektivität und daraus resultierende biologische Aktivität eines Liganden darzustellen, zu analysieren und zu modifizieren. Die 3D-Suche in Protein-Ligand-Schnittstellen hilft Forschende dabei, ähnliche Bindungsstellen zu identifizieren, um beispielsweise *Off Target*-bezogene Forschungsfragen wie *Drug Repurposing* oder Nebenwirkungen zu beantworten. Teilweise basierend auf neu implementierten, überarbeiteten und umfassend weiterentwickelten Grundlagen wurden zwei Methoden und Tools entwickelt, um den jeweiligen Stand der Wissenschaft voranzutreiben.

Neben der Anwendbarkeit der Tools lag das Hauptaugenmerk auf der Entwicklung von intuitiv nutz- und bedienbaren Methoden, um die komplexen Daten und die entsprechende funktionale Komplexität der beiden Tools zu adressieren. Beide Tools sind über einen kostenlosen Webdienst und einen Software-Container verfügbar. Für das erste Tool wurde ein Visualisierungskonzept für Protein-Ligand-Schnittstellen entwickelt, das interaktive und synchronisierte Text-, 2D- und 3D-Darstellungen synergetisch integriert. Die innovativste Komponente des Tools ist der 2D-Editor, der hochgradig anpassbare 2D-Diagramme von Ligandenbindungsmodi anzeigt. Das Visualisierungskonzept wurde dann für das zweite Tool genutzt, um die Benutzerfreundlichkeit der Suchoberfläche angesichts der hohen Komplexität der Suchanfrage zu adressieren. Eine Abfrage kann einfache textbasierte Schlüsselwörter und benutzerdefinierte 3D-Objekte umfassen, die relative räumliche Anordnungen verschiedener chemischer Komponenten wie Atome darstellen. Der Suchraum besteht aus algorithmisch vorhergesagten ligandengebundenen und ligandenungebundenen Bindungsstellen der gesamten PDB. Beide Anwendungen erwiesen sich als einzigartige Alternativen zu anderen entwickelten Tools und ermöglichen eine umfassende und benutzerfreundliche Analyse von Protein-Ligand-Bindestellen zur Beantwortung biologischer funktionsbezogener Fragestellungen.

# Abstract

Proteins are essential to life and, therefore, to life science-related topics like drug development. The binding of ligands to proteins activates, deactivates, or regulates biological processes in organisms, thereby controlling their development, capabilities, and maintenance. Three-dimensional structures of proteins and protein-ligand complexes are collected in databases like the Protein Data Bank (PDB) and represent a continuously growing and increasingly important data source for computer-aided drug design. Scientific research requires software applications to study the ligand binding-enabling characteristics of protein-ligand interfaces in the context of drug development. Notably, the visualization and 3D searching of that data generates valuable knowledge. The visualization of a protein-ligand interface in reports, presentations, and articles supports scientists to present, analyze, and modify a ligand's affinity, selectivity, and consequential biological activity. 3D searching in protein-ligand interfaces, on the other hand, helps scientists identifying similar binding sites to, for example, address off target-related research questions like drug repurposing or drug side effects. Partly based on reimplemented, refactored and extensively further developed groundwork, two methods and tools were developed to advance the respective state of the art.

In addition to the applicability of the two tools, a major focus was centered on the development of usability-enhancing methods to address the complex data and the corresponding functional complexity of the tools. Both tools are available via a free-to-use web service and a standalone software container, ensuring their accessibility to all researchers in the field. A visualization concept for protein-ligand interfaces was developed for the first tool, synergistically integrating interactive and synchronized textual, 2D, and 3D representations. The tool's most innovative component is the 2D editor, which generates and displays highly user-customizable 2D diagrams of ligand binding modes. The visualization concept was then exploited for the second tool to address the search interface's usability in view of the query's high complexity. A query can include basic text-based keywords and user-customized 3D objects describing relative spatial arrangements of different chemical features like atoms. The search space is composed of algorithmically predicted ligand-bound and ligand-unbound binding sites of the complete PDB. Both applications were demonstrated to be unique alternatives to other existing tools, enabling comprehensive and user-friendly analyses of protein-ligand interfaces to answer biological function-related research questions.

# Acknowledgments

I want to thank Matthias Rarey for the interesting doctoral topic and the excellent supervision. I would also like to thank Christiane Ehrt, Joel Graef, and Martin Poppinga for their helpful collaboration on GeoMine and several scientific publications. I would also like to thank the BMBF for funding this work. I would especially like to thank all members of the AMD group for the relaxed working environment and cooperative work ethic that enables the group's extensive technical infrastructure. Lastly, I would like to thank Joel Graef and Christiane Ehrt for proofreading my doctoral thesis.

# Contents

# Chapter 1

# Introduction

The existence of all life forms is based on the information encoded in their genetic material. Its decoding generates a large variety of macromolecular chemical structures called proteins. According to the Human Proteome Project, 19,778 proteins are predicted to be encoded by the human genome, of which 18,397 have been identified as of April 1, 2023, resulting in 93.01% human proteome coverage [1].

The high number of unique proteins is based on the combinatorial diversity of their building blocks, the 20 proteinogenic $\alpha$-amino acids. Each encoding gene generates a different sequence of these building blocks, which folds into a unique three-dimensional (3D) protein structure [2]. Proteins carry out specific functions within organisms to control their development, maintenance, and abilities. From hormones that promote the organism's growth to antibodies that protect it against a virus, from enzymes that catalyze the digestion of nutrients to receptors that enable pain perception, the biological functions of proteins are highly diverse. The functional diversity of proteins is primarily based on their differentiating $\alpha$-amino acid sequences and resulting 3D structures. These varying properties lead to the formation of unique regions on a protein's surface that can bind to other small chemical structures, which in this context are called ligands. The spatial shape of a binding site and the electrical charges of the amino acid atoms exposed on its surface determine the affinity for ligands that are complementary in shape and atomic charges. During the ligand binding process, the ligand spatially fits into the binding site, and attractive forces occur between oppositely charged binding site and ligand atoms. The subsequent binding of a ligand to the binding site can alter the protein's functional state by either enhancing or inhibiting its activity. Online resources like the Catalytic Site Atlas (CSA) [3] and PROSITE [4] provide descriptions of important amino acid motifs for ligand recognition in binding sites.

Due to their ubiquitous presence in all life forms and essential functional importance to life, proteins are also an important research topic in life sciences, especially in the context of drug development to treat human diseases. This process exploits the previously described molecular control mechanism of protein activity. The drug development process includes the visualization, analysis, and identification of proteins with disease-related activity, of their activity-associated binding sites, and of already existing ligands or newly designed ones that bind these binding sites with high affinity. The ultimate goal is to obtain a ligand that manipulates a target protein's disease-related biological activity in such a way that a therapeutic effect is achieved in the patient [5–7]. To ensure the effectiveness of an administered drug, it is important also to investigate and improve other drug-related effects in the human body that may lead to complications such as side effects due to low target protein selectivity, insufficient bioavailability, or short metabolic half-life.

Due to the significant complexity of a drug development process and the resulting timeframe, costs, and risk of failure in clinical phases [8], computational approaches have become increasingly important as an efficiency-enhancing support to experimental ones, particularly in the early stages of development. Both approaches are used for the early discovery of compounds with more favorable absorption, distribution, metabolism, excretion, toxicological, and affinity profiles [9–12]. For example, the identification of ligands that bind to a protein target can be approached through an in silico or in vitro high-throughput screening, in which screening libraries of potential drug candidates are tested in computational experiments or in the laboratory using a computational 3D model or a biological assay of a protein target respectively [13–15]. Subsequently, the visualization and analysis of a computational 3D model of a ligand-bound binding site and its calculated intermolecular interactions can provide medicinal chemists insights into how a drug and its intermolecular interactions might be further optimized to improve its affinity [16, 17].

As the example described previously also illustrates, the knowledge about the functionally relevant 3D structures of proteins, ligand-unbound binding sites, ligands, and ligand-bound binding sites can particularly support the visualization, analysis, and identification tasks in computer-aided drug design. The determination of molecular 3D structures can be accomplished through X-ray crystallography [18] and nuclear magnetic resonance (NMR) spectroscopy [19]. In addition, the importance of electron microscopy (EM) is growing as it improves speed, costs, and resolvable structure size [20, 21]. X-ray crystallography is still a frequently used technology for 3D structure determination (Figure 1.1 taken from [22]). An X-ray crystallographic experiment generates an electron

density map, which is processed to build and refine a 3D structural model in agreement with the experimental data. In addition to the experimental technologies, a computational approach in this context is the artificial intelligence(AI)-based tool AlphaFold [23], which predicts 3D structures of proteins based on 3D reference structures and sequence alignments as training data.

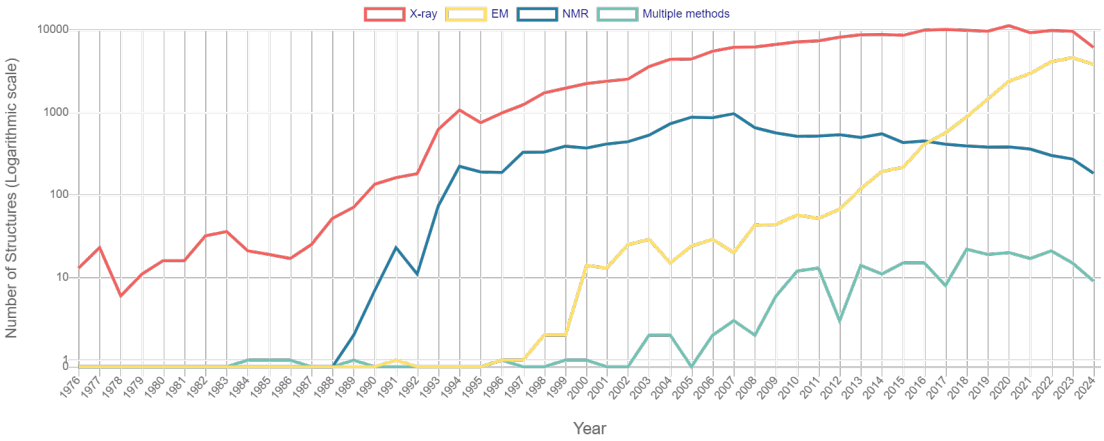



**Figure 1.1:** Number of released PDB structures per year. Red: x-ray diffraction, fiber diffraction, or powder diffraction. Blue: solution NMR or solid-state NMR. Yellow: electron microscopy, electron crystallography, or electron tomography. Turquoise: Multiple experimental methods. Taken from [22]

The experimental and computational advancements in 3D determination technologies have led to a significant increase in 3D structures, which is collected by databases such as the Cambridge Structural Database (CSD) [24], the above-mentioned AlphaFold database, the Protein Data Bank (PDB) [25], and the sc-PDB [26]. The CSD contains experimental 3D structures of organic, metal-organic, and organometallic molecules. As previously highlighted, the AlphaFold database includes 3D structures of proteins predicted by artificial intelligence. The PDB contains experimentally determined 3D structures of proteins, nucleic acids, and corresponding hybrids and ligand complexes. In addition, the PDB provides access to a subselection of 3D structures of the AlphaFold database. The sc-PDB contains a subselection of PDB entries for which binding sites that were predicted to be druggable are annotated. The content of the PDB is freely accessible on a web page and updated weekly with 3D structures submitted by scientists from all over the world. The PDB has established itself as a central exchange platform for molecular 3D structures and, thus, as an important resource for life scientists to visualize, analyze, and identify proteins, binding sites, and binding ligands for the development of new drugs.

The PDB itself provides several internal tools for the therefore necessary data exploration. For example, the search options include keyword-based searches, sequence searches utilizing the BLAST algorithm [27], chemical substructure searches, and 3D structural searches. The 3D structural data of the PDB entries can also be downloaded in various file formats (PDB, PDBx/mmCIF, XML, BinaryCIF) [28]. The PDBx/mmCIF is an improvement to the PDB file format supporting data of large 3D structures, complex chemistry, and new and hybrid experimental methods [29]. Although the PDB file format is no further developed and will become outdated due to the PDBx/mmCIF format's evolvement [30], it is still widely used to exchange 3D structure data of proteins and protein-ligand complexes. External visualization and analysis tools can also use PDB files as input. The origin of PDB files is not limited to the PDB website and other publicly accessible internet sources. For example, external sources like molecular docking tools that use PDB files as input might also generate PDB files as output [31]. A PDB file contains atomic 3D coordinates and other atomic properties such as the element and amino acid. In addition, the PDB file content includes information about the source organism, the protein, its amino acid sequence, and secondary structure elements, as well as experimental metadata, for example, regarding the 3D structure's quality. The 3D structural data in the PDB files might be parsed and further processed by computational tools to complement and represent it regarding user-specific requirements. For example, ligand-unbound binding sites [32], polar hydrogen atoms [33], and intermolecular ligand interactions [34] could be predicted computationally. As previously described, proteins, their binding sites, and binding ligands are of essential importance for life and, consequently, for life sciences and drug development in particular. Furthermore, the corresponding 3D structural data, such as the one in the PDB, is a significant source for computer-aided drug design due to its functional relevance.

Based on this context, this doctoral project resulted in two mutually dependent new tools to explore structural data called PoseEdit [D1] and GeoMine [D2–D5]. PoseEdit takes a 3D structure in PDB format of a protein in complex with one or multiple ligands as input and allows the automatic depiction of ligand-specific and interactive two-dimensional (2D) interaction diagrams. A diagram illustrates a ligand's binding mode, i.e., the ligand in its binding site and its intermolecular interactions. GeoMine, which is the main tool of this doctoral project, is a search engine for querying predicted ligand-bound and ligand-unbound binding sites in large collections of 3D structures. The comprehensive 3D structural data of the PDB was selected as a searchable data basis for GeoMine. GeoMine's search functionality includes a text-based and a 3D structural query type, which describes relative spatial arrangements of binding site characteristics

like atoms. The query design process of GeoMine can partially be done using PoseEdit diagrams, while the interaction detection done by GeoMine is the basis for their generation. The development of both tools is highly focused on the structural data's complexity and the associated usability of their graphical user interfaces, i.e., user-centric design problems and concepts. Both tools are available as software containers and are embedded in the publically available Proteins*Plus* web server (https://proteins.plus) [35], which was further developed for this doctoral project [36, D6, D7]. In addition, software container-based and publically available standalone web servers were developed (https://poseedit.proteins.plus, https://geomine.proteins.plus).

In sections 1.1 and 1.2, the chemical and physical principles of protein-ligand interactions and the corresponding structural data quality are described in detail. PoseEdit and GeoMine are then addressed in the following two chapters (Chapter 2 and Chapter 3). In both chapters, the general motivations for tools like PoseEdit and GeoMine are presented by highlighting the exploitability of the 3D structures in the PDB by such tools and the research questions they can answer. Subsequently, the specific motivations for developing PoseEdit and GeoMine are outlined based on the respective tool-specific state of the art. The strengths and limitations of existing tools are highlighted regarding application examples of interest to identify the perspectives and objectives for the development of PoseEdit and GeoMine. Next, the technical groundwork on which both tools are based and the corresponding technical challenges are summarized, followed by a description of the tools and potential applications. Lastly, a respective outlook section proposes further developments, followed by a conclusion (Chapter 4) about the final results of this doctoral project.

## 1.1 Protein-Ligand Interactions

A sequence of the 20 $\alpha$-amino acids forms the primary structure of a protein in which the amino acids are linked by covalent peptide bonds. The backbone of a protein consists of the peptide bonds, the $\alpha$-carbons of the amino acids connected by the peptide bonds, and the amino group and carboxyl group of the terminal amino acids. The variable side chains of the amino acids are connected to the $\alpha$-carbon atoms of the backbone. Intramolecular interactions between the backbone atoms lead to the secondary structure of a protein, causing local spatial conformations of the backbone. Secondary structure elements include $\alpha$-helices, $\beta$-sheets, and turns. The 3D structure of a protein chain, commonly referred to as tertiary structure, is determined by intramolecular interactions between the side chain atoms. Furthermore, proteins may also have a quaternary structure. This structure classification level refers to large protein complexes,

which are themselves composed of multiple identical or different protein chains through intermolecular interactions. For the formation of a protein-ligand complex, a ligand must first fit sterically into an often hollow-shaped binding site on the protein's surface. Opposite transient, partial, and formal atomic charges are then the basis for forming non-covalent and attractive electrostatic forces between binding site and ligand atoms, such as hydrogen bonds, ionic interactions, and van der Waals forces. The 20 $\alpha$-amino acids have side chains with different functional groups and consequently varying atomic charges. However, in this context, some amino acids are more similar than others. Based on the properties of the side chains, amino acids can be classified as acidic if negatively charged, basic if positively charged, polar if they have polar atoms, hydrophobic if composed of non-polar atoms, and aromatic if aromatic rings are present.

The binding affinity depends, besides other factors, on the extent to which the charge distribution on the binding site's surface complements the charge distribution on the ligand's surface. Furthermore, geometric properties of intermolecular interactions, such as the distance between two interacting atoms, determine the binding strength. Hydrogen bonds are based on the opposite partial charges of a hydrogen atom donor and a heavy atom acceptor. Ionic interactions occur between a cationic and anionic atom due to their opposite formal charges. Van der Waals forces are formed between two atoms by the spontaneous occurrence of transient atomic charges. Protein-ligand interactions can occur directly or be mediated through other interaction partners in the binding site, such as the solvent molecules or cofactors.

From the thermodynamic perspective, protein-ligand binding is based on the Gibbs free energy, which is described by the formula $\Delta G = \Delta H - T\Delta S$. $\Delta H$ describes the total enthalpy change caused by the disruption and formation of intermolecular interactions between the protein, ligand, and solvent. $\Delta S$ is the total entropy change resulting from solvent release and the conformational, translational, and rotational degrees of freedom before and after protein-ligand complex formation. A negative change in Gibbs free energy expresses that the final bound state has a lower energy than the initial unbound state at a fixed temperature. A low Gibbs free energy is the prerequisite for a high binding affinity and consequential formation of a stable protein-ligand complex. Consequently, the protein-ligand binding benefits from a negative total enthalpy change and a positive total entropy change.

## 1.2 Structural Data Quality

The quality of visualization and analysis results of 3D structures depends on the quality of the 3D structures themselves. When using experimental or artificial intelligence-predicted 3D structures, the associated limitations of the underlying data and subsequential 3D structural model generation must be considered. For example, certain regions of a 3D structure may not be sufficiently resolved due to a lack of experimental or AI training data. A local lack of experimental data for a specific region could, for example, be caused by its high structural flexibility. Furthermore, the 3D structural model building and refinement process can introduce errors that result in atomic collisions and distortions of atomic geometry, such as bond length deviations. The PDB provides several reports and quality measures for 3D structures.

The validation of experimental 3D structures is based on experts' recommendations for X-ray crystallography, NMR, and EM [37–39]. For example, the overall quality of an X-ray-determined 3D structure is highlighted by its resolution. A low resolution indicates that two adjacent atoms are well distinguishable. The local quality of regions of 3D structures predicted by artificial intelligence is commonly validated using the pLDDT (Predicted Local Distance Difference Test) score. A high pLDDT value for a region indicates that the region is predicted with high confidence based on the 3D reference structure and sequence alignment training data. Consequently, the quality of a 3D structure determines its reliability and should be assessed before its visualization and analysis by computational tools. Another specific limitation to consider is the technological capability to resolve hydrogen atoms. In PDB files of structures solved by X-ray crystallography, hydrogen atoms are usually absent. In contrast, NMR analysis and theoretical models typically determine hydrogen atom positions [40]. Consequently, without hydrogen atoms, the exact position of oxygen and nitrogen atoms in the sidechains of, for example, asparagine and glutamine is difficult to determine due to the similar electron numbers of the atoms.

# Chapter 2

# 2D Visualization of Ligand Binding Modes

The general motivation for the 2D visualization of ligand binding modes and, consequently, the specific motivation for the development of PoseEdit is presented in the following sections. The motivations are based on the scientific applicability, the exploitability of the 3D structures in the PDB (section 2.1), and the strengths and limitations of respective state-of-the-art tools (section 2.2). The consequential perspectives and objectives for developing PoseEdit are summarized in section 2.3. Subsequently, the technical groundwork and challenges of the development of PoseEdit (section 2.4), the tool's description (section 2.5), and its application (section 2.6) are presented. In the outlook (section 2.7), extensions of PoseEdit are suggested. [D1] references the scientific publication about PoseEdit. The second tool, GeoMine, which was also developed for this doctoral project (Chapter 3), integrates PoseEdit's functionality [D5]. [D2–D4] reference additional GeoMine-related scientific publications. The PoseEdit-hosting Proteins*Plus* web server and its associated development performed for this doctoral project are described in [36, D6, D7].

## 2.1 Relevance for Scientific Research

Experimental and computational approaches like in vitro or in silico high-throughput screening provide large numbers of candidate compounds in the context of drug development. Despite all the related workflow automatizations, the visual examination of chemical and spatial aspects of a ligand binding mode by scientists is still an indispensable task. It can help to answer research questions such as the following:

- Which residues form intermolecular interactions with the ligand?
- Which intermolecular interaction types occur?
- Which functional groups and atoms of the protein and ligand are interacting?
- What is the 3D conformation of the ligand?
- What is the 3D shape of the binding site?

The 3D structure of a protein-ligand complex is the basis for the computational generation of corresponding 2D and 3D representations of a ligand binding mode [41]. Both visualization concepts provide a familiar environment for scientists to visualize chemical structures. Furthermore, such graphical representations enable scientists to concisely communicate a ligand's binding mode to themselves as well as to others through reports, presentations, and scientific articles. Therefore, a suitable 2D or 3D visualization of a ligand binding mode must contain a sufficient amount of relevant chemical and spatial information that is presented clearly and attractively. Subsequently, the visual examination of a ligand's binding mode enables scientists to analyze it in the context of the above-described research questions and subsequently improve a ligand's affinity, selectivity, and consequential biological activity based on their expert knowledge. The chemical and spatial complementarity of the ligand and the binding site and, consequently, the strength of intermolecular interactions has, therefore, to be optimized to this end.

Dimensionality reduction from 3D to 2D may seem counterintuitive. However, both representation concepts have specific advantages and disadvantages depending on what type of information scientists want to focus on. A 3D representation is a suitable choice to comprehend spatial information relevant to the ligand binding mode, for example, the ligand's conformation or the binding site's shape. However, the large amount of chemical and spatial information of a 3D representation requires a time-consuming navigation through the 3D scene, which complicates the ligand binding mode's examination. Furthermore, a 2D screenshot of a 3D scene prevents a clear and immediate communication of a ligand binding mode due to the large amount of spatially overlapping chemical information. A 2D scene provides a decreased amount of chemical and spatial information due to its planarity, limiting boundaries, and consequential risk of graphical collisions. However, by excluding unnecessary chemical information, a 2D representation can provide a faster communication of chemical key aspects of a ligand's binding mode while maintaining a collision-free diagram layout. For example, a 2D representation might not depict all amino acids of a binding site but only those that directly interact with the

ligand. The precise and quickly communicated chemical information of a 2D representation is valuable to present and compare ligand binding modes in reports, presentations, and scientific publications. Furthermore, when a large number of ligand binding modes needs to be examined sequentially, the speedup of a 2D representation is particularly helpful.

## 2.2 State of the Art

Compared to other computational tools for structure-based modeling, only a few published tools exist for automatically generating 2D diagrams of ligand binding modes based on 3D structures of protein-ligand complexes in PDB format: PoseView [34, 42, 43], LeView [44], LigPlot+ [45, 46], and MOE [47]. All tools were published more than ten years ago. In particular, LigPlot+ and PoseView became popular. As of June 18, 2024, the scientific publications of LigPlot+ and its predecessor LIGPLOT counted on Google Scholar amount to 11,290 citations [48, 49] and of the more recent PoseView [50–55] in 985 citations.

Figure 2.1, Table 2.1, and the following paragraphs summarize the differences of the diagrams generated by these tools in terms of chemical information content and graphical styles. Furthermore, the following paragraphs highlight the corresponding tool-specific strengths and limitations. The above-mentioned tools and their interaction models specify which chemical structures are considered ligands and potential interaction partners, as well as which and how intermolecular interaction types are calculated. All tools allow the generation of diagrams based on the 3D structure of a protein in PDB format and one of its ligands as input. These tools also support diagrams for DNA-ligand and RNA-ligand complexes. A ligand can be any small molecule. In addition, diagrams for simple ions like metal ions can be generated using PoseView, LeView, and LigPlot+.

All tools calculate and display direct intermolecular interactions of a ligand to protein and nucleic acid residues. PoseView also draws intermolecular ligand interactions with metal ions and LeView with water molecules. LigPlot+ and MOE support both mentioned interaction types. Furthermore, LigPlot+ also visualizes intermolecular ligand interactions with other small molecules, for example, organic cofactors. In LeView and MOE, all residues within a tool-specific cut-off radius of the ligand's heavy atoms are displayed, thereby including also those residues that do not form intermolecular interactions with the ligand. While LeView only considers hydrogen bonds, MOE and PoseView visualize a large variety of intermolecular interactions. The chemical and spatial criteria that the tools apply to calculate the supported intermolecular interaction types are

based on different scientific studies and are, therefore, highly tool-specific. For example, PoseView defines a hydrogen bond by a specific donor hydrogen (hydrogen atoms of nitrogen, oxygen, or sulfur atoms) and specific acceptor atom (solvent-exposed uncharged or negatively charged nitrogen, oxygen, or sulfur atoms) with an optimal distance of 1.9 Å, 0.5 Å tolerance, and an acceptor-hydrogen-donor angle above 120°. LigPlot+ and MOE detect covalent bonds between the ligand and residues. Explicit polar hydrogens and atom charges are displayed in the diagrams of PoseView and MOE.

All tools show the ligand in its skeletal representation and calculated intermolecular interactions as connecting lines. The visualization of the interaction partners differs between the tools. While LigPlot+ and PoseView use the skeletal representation for residues, LeView and MOE visualize them as circles or text labels. LigPlot+ displays the complete residue. In contrast, PoseView shows its backbone, its sidechain, or both, depending on which residue parts form intermolecular interactions. Only PoseView strictly follows the Union of Pure and Applied Chemistry (IUPAC) [56] guidelines for drawing chemical structures in skeletal representation. The IUPAC constraints-based depiction style ensures drawings of chemical structures with, for example, consistent bond lengths and angle sizes. With PoseView, hydrophobic contacts are represented by the labels of contacting residues that annotate specific spline segments at the corresponding hydrophobic ligand parts. LigPlot+ represents hydrophobic contacts through spiked arcs, which are individually labeled by contacting residues and oriented towards the corresponding hydrophobic ligand parts.

**Table 2.1:** Comparison of the diagrams of PoseView, LeView, LigPlot+, and MOE regarding chemical information content and graphical styles. Taken and adapted from [D1]

| | **PoseView** | **LeView** | **LigPlot+** | **MOE** |
|---|---|---|---|---|
| **chemical structures** | | | | |
| **ligands** | small molecules, metal ions | small molecules, metal ions | small molecules, metal ions | small molecules |
| **interaction partners** | amino acids, nucleic acid residues, metal ions | amino acids, nucleic acid residues, water | amino acids, nucleic acid residues, water, metal ions, small molecules | amino acids, nucleic acid residues, water, metal ions |

| **intermolecular interactions** | | | | |
|---|---|---|---|---|
| **covalent ligand bonds to the protein** | - | - | x | x |
| **hydrogen bonds** | x | x (+ $H_2O$-mediated) | x (+ $H_2O$-mediated) | x (+ $H_2O$-mediated) |
| **ionic interactions** | - | - | - | x |
| **metal coordination** | x (direct) | - | x (direct and coordinated) | x (direct and coordinated) |
| **pi-pi interactions** | x | - | - | x |
| **cation-pi interactions** | x | - | - | x |
| **pi-H interactions** | - | - | - | x |
| **hydrophobic contacts** | x | - (near residues) | - | - (near residues) |
| **explicit hydrogens** | x | - | - | x |
| **charges** | x | - | - | x |
| **representation styles** | | | | |
| **ligand** | skeletal | skeletal | skeletal | skeletal |
| **interaction partners** | skeletal (side chain only, backbone only, or both) | circle | skeletal | circle |
| **IUPAC** | x | - | - | - |

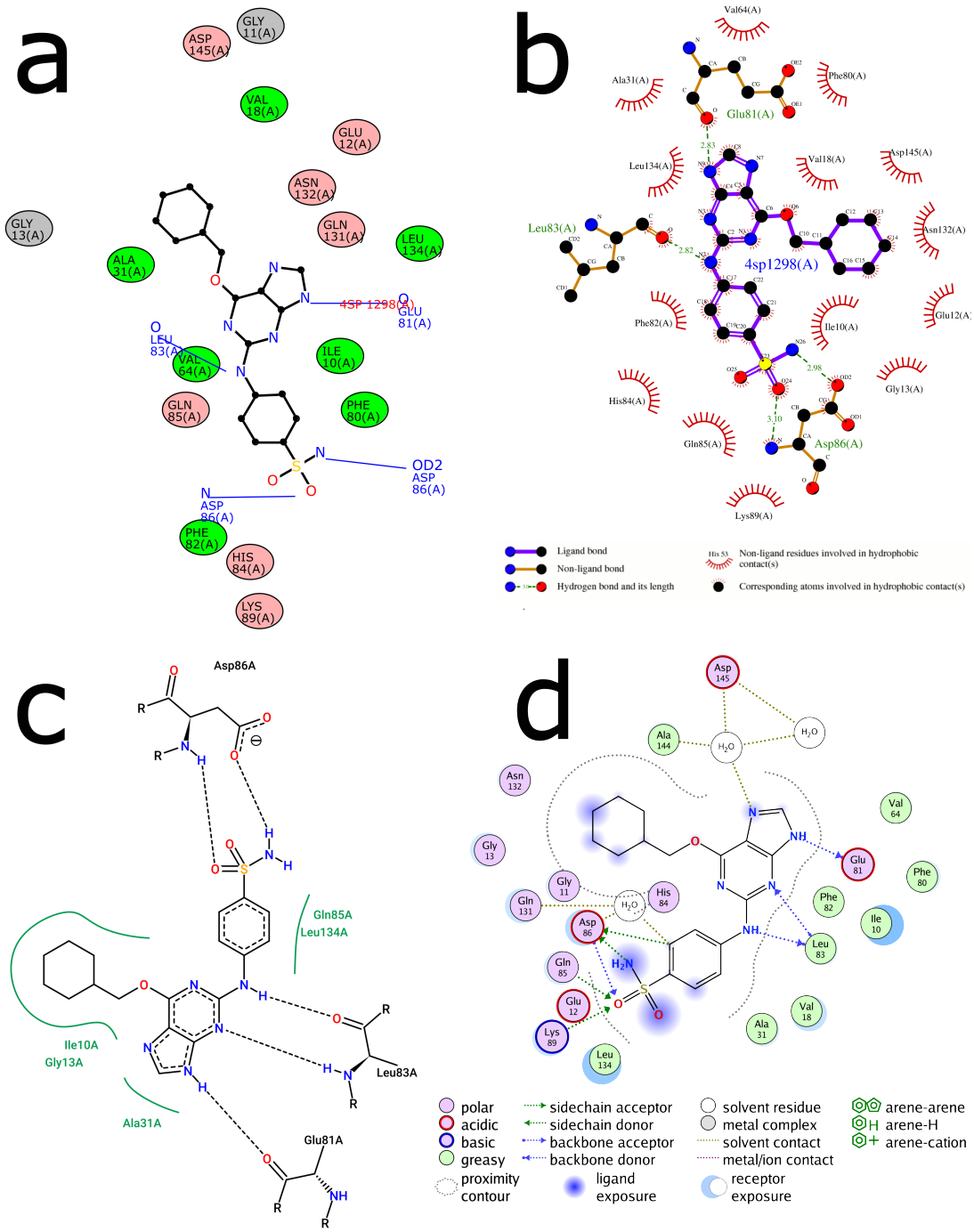



**Figure 2.1:** Diagrams of different tools of the inhibitor 4-[[6-(cyclohexylmethoxy)-7H-purin-2-yl]amino]benzenesulfonamide in complex with a cyclin-dependent kinase [57] (PDB identifier: 1H1S). **a** LeView. **b** LigPlot+. **c** PoseView. **d** MOE. Taken from [D1]

Depending on the complexity of the ligand binding mode, for example caused by a large number of interacting residues and intermolecular interactions, a 2D diagram layout can show graphical deficiencies, such as:

- overcrowded diagrams due to a large number of closely located graphical objects like chemical structures, text labels, etc.
- overlapping graphical objects, like covalent bond lines and atom labels of chemical structures
- intersecting intermolecular interaction lines
- intermolecular interaction lines crossing other graphical objects like chemical structures

Therefore, the algorithmic challenge of these tools is to create a clear and collision-free 2D diagram layout while maintaining a sufficient amount of relevant chemical information and attractive graphical styles. The algorithmic performance might reach its limits for specific diagrams. For example, a large-scale application of PoseView in 2010 on 201,245 ligand-bound binding sites from the PDB resulted in 155,612 (77.3%) calculatable diagrams, including 123,535 (79.4%) diagrams with a good layout and 32,077 (20.6%) diagrams with improvable or insufficient layout quality. The remaining 45,633 (22.7%) diagrams could not be calculated due to technical reasons like a computing timeout (>450 s, 1,038) or because PoseView's interaction model did not detect any (32,549) or too many (897) intermolecular interactions (>18) or interaction partners (>14) causing the diagram's omission. Both limitations are internally set. While graphical collisions in the diagram are objective layout quality deficiencies, the diagram's chemical information content and graphical styles both depend on the user's subjective preferences regarding, for example:

- the presence or absence of specific intermolecular interactions or interaction partners
- the protonation states of specific atoms
- the ligand's structural 2D layout
- the overall 2D arrangement of chemical structures
- the colors of intermolecular interactions

The relevance of some of the previously described objective and subjective criteria for diagram quality is further illustrated in Figure 2.2. Figure 2.2 shows a PoseView diagram with objectively and subjectively suboptimal quality. The corresponding PDB structure is lysine-specific histone demethylase 1A in complex with a cofactor and inhibitor in

the same binding site [58] (PDB identifier: 5LGT; ligand identifiers: FAD_A_901 and 6W3_A_902). The diagram is generated using the cofactor FAD_A_901 as input ligand.

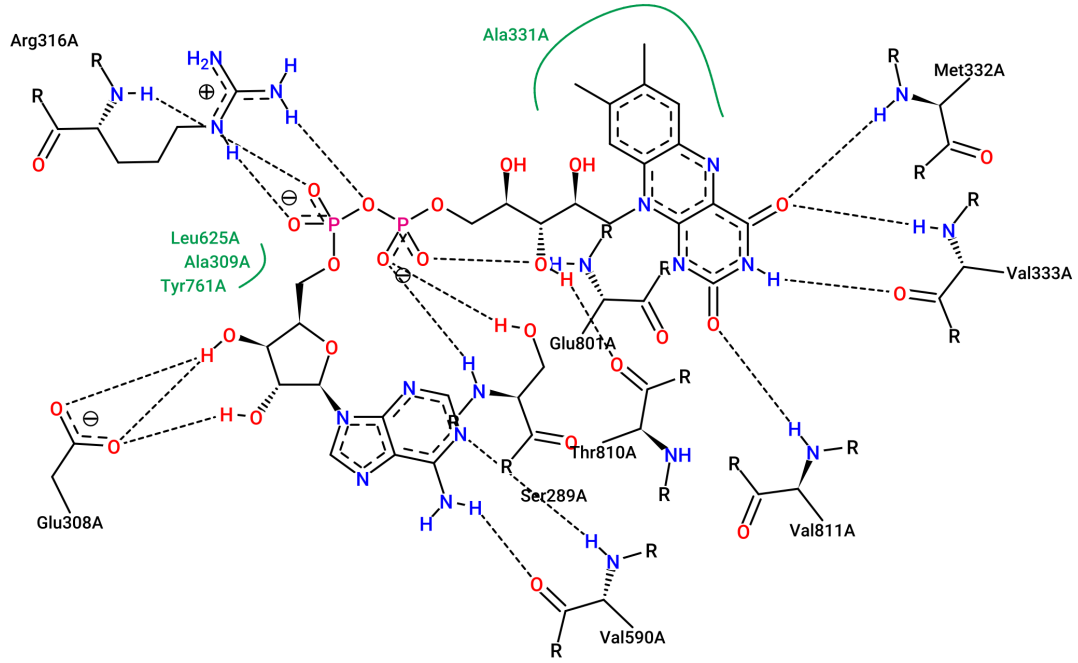



**Figure 2.2:** Example of an improvable PoseView diagram showing lysine-specific histone demethylase 1A in complex with a cofactor [58] (PDB identifier: 5LGT; ligand identifier: FAD_A_-901)

Due to the highly complex ligand binding mode, the algorithmically generated diagram shows numerous intersections and overlaps between intermolecular interactions, chemical structures, and text labels, for example:

- the hydrogen bond of Arg316A's backbone intersects its side chain
- the double bonds of the ligand's two diphosphate groups intersect
- the two hydrogen bonds of Ser289A intersect the ligand's charge annotation
- the structure diagram of Ser289A and the ligand overlap
- the hydrogen bond of Val590A intersects the text label of Ser289A

Furthermore, graphical aspects like the bent structural 2D layout of the cofactor or the black color of intermolecular interactions could be subjectively unattractive style

choices. In addition, the chemical information content of the diagram excludes the inhibitor 4-methyl-N-[2-[[4-(1-methylpiperidin-4-yl)oxyphenoxy]methyl]phenyl]thieno[3,2-b]pyrrole-5-carboxamide and therefore an important interaction partner.

The tool-specific algorithmic approaches and diagram contents provide users with different solutions and trade-offs to address such problems, some of which are highlighted in the following:

- LeView and MOE visualize all residues as simple circles or text labels and are, therefore, convenient workarounds for users that want to reduce graphical collisions
- PoseView and LigPlot+ display all chemical structures in the atomic skeletal representation and are consequently appropriate options for users that want ligand binding modes displayed with a high level of chemical detail
- PoseView applies the IUPAC guidelines, making it a suitable choice for users who want chemical drawings of high aesthetic quality
- LigPlot+ does not exclude any specific small molecule types as interaction partners like organic cofactors or water, making it a good solution for users interested in ligand binding modes of high complexity

However, for a specific ligand binding mode, none of the tools might sufficiently address all objective graphical issues or subjective user preferences regarding the diagram's chemical content and graphical styles, forcing users to accept dissatisfactory compromises. For example, a ligand binding mode might be too complex to be algorithmically drawn without graphical collisions by any tool or users might need PoseView's IUPAC-based depiction style but do not favor its interaction model. Therefore, some tools provide interactive graphical user interfaces to support users to manually adapt diagrams according to objective and subjective criteria. Table 2.2 and the following paragraphs provide a comparative summary of the graphical user interfaces of these tools regarding their accessibility, and the distribution, comprehension, exploration, and especially editing functionalities of diagrams. Furthermore, the following paragraphs highlight the corresponding tool-specific strengths and limitations.

LigPlot+, LeView, MOE, and PoseView are desktop tools. In addition to its desktop version [59], PoseView is also available via a web server. Web-based tools offer users high accessibility independent of device, operating system, software installation, and manual updates. In contrast, the accessibility of desktop tools is relatively limited. However, desktop tools can provide users with more data privacy and are accessible

without an internet connection. LeView and the web version of PoseView are free to use and, therefore, accessible to the general public. LigPlot+ requires an academic license. MOE and the desktop version of PoseView are commercially available tools, with the latter being free for academic usage. All tools allow saving diagrams in various image file formats for their integration in reports, presentations, or scientific articles. The comprehension and exploration of diagrams are further addressed by LeView, LigPlot+, and MOE. LigPlot+ and MOE annotate the 2D diagram with a legend facilitating the comprehension of drawn chemical features like the types of intermolecular interactions. A text file with information about the depicted ligand binding mode, e.g., the intermolecular interactions, is provided by LigPlot+, MOE, and LeView. Therefore, users do not need to extract related information from the diagram into text themselves. They can directly analyze the text file's content and integrate it into reports, presentations, or scientific publications. LigPlot+ and MOE provide a corresponding 3D visualization for each diagram. Both visualization concepts can synergistically complement each other regarding the analysis of a ligand binding mode's chemical and spatial aspects. Multiple diagrams, for example, of the same ligand bound to different proteins, can be displayed together for comparison using the 2D editor of LigPlot+.

While the diagrams of PoseView are static, LeView, LigPlot+, and MOE offer 2D editors for modifying a diagram's content. The relative arrangement of chemical structures and text labels in the 2D scene can be changed using LigPlot+ and, to a limited extent, using LeView. Furthermore, chemical structures can be mirrored at covalent bonds with LigPlot+. Graphical collisions and personal layout preferences of the automatically generated diagrams can, therefore, be manually addressed by users.

The chemical information content of diagrams can only be customized with LeView by removing chemical structures. Graphical styles in the diagram, like the coloring of chemical structures, can be set with LeView, LigPlot+, and MOE and thereby adapted to personal preferences. The export and import of diagrams via a specific file type is possible using LigPlot+. Consequently, LigPlot+ enables users to reload and subsequently modify previously generated diagrams in its interactive 2D editor. Furthermore, LigPlot+ provides additional features that assist in modifying diagrams. The complete diagram can be zoomed, translated, rotated, and recentered, supporting its interactive editing by users. Furthermore, LigPlot+ offers a 2D editor history to undo the last ten tracked changes or to reset the diagram to its initial unmodified state.

**Table 2.2:** Comparison of the graphical user interfaces of PoseView, LeView, LigPlot+, and MOE regarding accessibility, and the distribution, comprehension, exploration, and editing of diagrams. Taken and adapted from [D1]

| | **PoseView** | **LeView** | **LigPlot+** | **MOE** |
|---|---|---|---|---|
| **accessibility** | free web application/commercial desktop application (free with academic license) | free desktop application | commercial desktop application (free with academic license) | commercial desktop application |
| **diagram comprehension/exploration/distribution** | | | | |
| **diagram export** | PDF, PNG, SVG | PNG, JPG, GIF, PDF, SVG, EPS, TXT | PS, DRW | PNG, JPG, EPS, PS, BMP, TIF, EMF, SVG |
| **diagram import** | - | - | DRW | - |
| **interactions list export** | - | x | x | x |
| **3D visualization** | - | - | x (only interaction partners) | x (complete binding site) |
| **multiple diagrams** | - | - | x | - |
| **diagram legend** | - | - | x | x |
| **diagram editing** | | | | |
| **graphical styles (sizes, colors, ...)** | - | x | x | x |

| | | | | |
|---|---|---|---|---|
| **interactive objects** | - | chemical structures | chemical structures, atoms, text labels | - |
| **object translation** | - | x | x | - |
| **object rotation** | - | - | x | - |
| **object removal** | - | x | - | - |
| **mirror structure at bond** | - | - | x | - |
| **diagram reset** | - | x | - | x |
| **editing history** | - | - | x (undo of the last ten structural movements) | - |

## 2.3 Perspectives and Objectives

This doctoral project addresses two goals in the context of the 2D visualization of ligand binding modes. The first goal is the development of a new tool, called PoseEdit, to generate 2D diagrams of ligand binding modes. The low number, limited accessibility, and advanced age of the existing state-of-the-art tools, in contrast to their high and continuously growing number of citations, suggest potential for further development. The objective of PoseEdit is to ensure that quality issues of automatically generated diagrams no longer present an insuperable obstacle to their usability. This issue can be addressed by enabling the manual refinement of diagrams. Since a tool's layout algorithm and its chemical and graphical default settings might generate insufficient compromises, manually modifying diagrams is a helpful approach for resolving

objective graphical deficiencies and satisfying subjective preferences regarding chemical information content and graphical styles. PoseEdit is based on the tool PoseView and provides diagrams with improved chemical information content, graphical styles, and, most importantly, a high level of interactivity via a freely accessible web-based graphical user interface. PoseEdit's feature list, i.e., its improvements relative to PoseView, are summarized in Tables 2.3 and 2.4. The list is inspired by the previously described research questions of interest and related strengths and limitations of the state-of-the-art tools. A central aspect in this context was to build the 2D visualization functionality on top of a corresponding 3D visualization concept to exploit the strengths of both.

Another primary focus was developing an easily accessible and user-friendly frontend solution to provide PoseEdit's comprehensive functionality. The second and foremost goal in the context of the 2D visualization of ligand binding modes is the integration of PoseEdit's functionality into the tool GeoMine, the second tool developed for this doctoral project, see Chapter 3.

**Table 2.3:** Comparison of the diagrams of PoseView and PoseEdit regarding chemical information content and graphical styles

| | **PoseView** | **PoseEdit** |
|---|---|---|
| **ligands** | small molecules, metal ions | small molecules, metal ions |
| **interaction partners** | amino acids, nucleic acid residues, metal ions | amino acids, nucleic acid residues, metal ions |
| **intermolecular interactions** | | |
| **covalent ligand bonds to the protein** | - | x |
| **hydrogen bonds** | x | x |
| **ionic interactions** | - | x |
| **metal coordination** | x (direct) | x (direct) |
| **pi-pi interactions** | x (complete ring system) | x (single rings) |
| **cation-pi interactions** | x (complete ring system) | x (single rings) |

| | | |
|---|---|---|
| **hydrophobic contacts** | x | x |
| **explicit hydrogens** | x | x |
| **charges** | x | x |
| **diagram import** | - | x |
| **representation styles** | | |
| **ligand** | skeletal | skeletal/circle |
| **interaction partners** | skeletal (side chains up to $\beta$-carbon) | skeletal/circle (side chains up to $\alpha$-carbon) |
| **IUPAC** | x | x |
| **intermolecular interactions/bonds** | | color gradients, consistent length of lines/line dashes, minimal atom radius against collisions |

**Table 2.4:** Comparison of the graphical user interfaces of PoseView and PoseEdit regarding accessibility, and the distribution, comprehension, exploration, and editing of diagrams

| | **PoseView** | **PoseEdit** |
|---|---|---|
| **accessibility** | free web application/commercial desktop application (free with academic license) | free web application/standalone container application |
| **diagram comprehension/exploration/distribution** | | |
| **diagram export** | PDF, PNG, SVG | SVG, TXT, JSON |
| **diagram import** | - | JSON |
| **interactions list export** | - | x |

| | | |
|---|---|---|
| **3D visualization** | - | x (complete binding site, 2D-3D synchronization of structure highlighting triggered by mouse pointer hover and selection) |
| **diagram legend** | - | x (SVG export) |
| **info section** | - | x |
| **diagram editing** | | |
| **diagram rotation** | - | x |
| **diagram translation** | - | x |
| **zoom** | - | x |
| **diagram recentering** | - | x |
| **graphical styles (sizes, colors, . . . )** | - | x |
| **merge multiple diagrams** | - | x |
| **interactive objects** | - | chemical structures, atoms, bonds, rings, intermolecular interactions, hydrophobic contact splines and spline control points, text labels |
| **object selection, grouping, highlighting** | - | x |
| **object translation** | - | x |
| **object rotation** | - | x |
| **object adding** | - | x |
| **object removal** | - | x |

| | | |
|---|---|---|
| **object editing (atom charge, bond type, ...)** | - | x |
| **mirror structure at bond** | - | x |
| **mirror structure at line** | - | x |
| **diagram reset** | - | x |
| **editing history** | - | x (all changes undo & redo) |

## 2.4 Technical Groundwork and Challenges

The technical groundwork and challenges to address the development of the key concepts and components of PoseEdit described in Tables 2.3 and 2.4 are summarized in this section. To this end, this section presents a brief overview of the tool's primary development steps. Corresponding general aspects regarding, for example, the technical implementation, are described in more detail in the Appendix.

PoseEdit is based on the 2D layout algorithm of PoseView. The first technical challenge in this context was to enable the building and execution of the tool PoseView. This is because the build scripts of the tool were no longer available. Furthermore, the archived C and C++ code of PoseView and corresponding legacy libraries are uncommented and have not been maintained for more than ten years, requiring time-consuming code revision, bug fixing, and reimplementation.

PoseView's layout algorithm had to be decoupled from the tool's diagram preprocessing steps, which include the calculation of the chemical content of a diagram based on a PDB file and a corresponding MOL2 ligand file. The chemical diagram content, e.g., the interacting chemical structures and intermolecular interactions, is now externally calculated by the tool GeoMine and directly parsed by the command line interface of PoseView via its adapted input functionality that accepts files with chemical data in MOL2 format. The calculated intermolecular interaction types and corresponding calculation criteria are described in B.5.4. The MOL2 file content, which was further

adapted for this doctoral project, is described in B.1.2. Consequently, PoseView applies a new and shared interaction model to generate the chemical information content for PoseEdit, additionally detecting ionic interactions, covalent bonds between the ligand and the protein, cation-pi and pi-pi interactions to single aromatic rings instead of complete aromatic ring systems, and a side chain depiction including not only the $\beta$-carbon but also the $\alpha$-carbon. These changes address additional user preferences and support GeoMine's query design process.

PoseView's layout algorithm had to be decoupled from the tool's image drawing functionality and desktop-based graphical user interface. The output of PoseView's command line interface was extended by a new output file type that stores textual diagram data directly generated by the algorithm in JSON format. The JSON file content is described in B.2.3. The new command line interface version of PoseView was integrated into the backend of the Proteins*Plus* web server. The general usage and technical implementation of the Proteins*Plus* web server are described in B.3.1 and B.3.2, respectively. The usage of the newly implemented and web server-integrated functionality of GeoMine's and PoseView's command line interfaces are described in B.1.1 and B.5.1, respectively. The frontend of the Proteins*Plus* web server was extended by a corresponding new graphical user interface.

The graphical user interface is based on the newly implemented InteractionDrawer library and the further developed Proteins*Plus* web server's source code and frontend components like its molecular 3D viewer. The previously mentioned JSON file generated by PoseView is usable as data input by the InteractionDrawer library, which draws interactive 2D diagrams of ligand binding modes in a web-based graphical user interface. The new drawing engine enables addressing specific graphical aspects like color gradients for intermolecular interactions and bonds, bonds of constant length, intermolecular interaction lines with dashes of constant length, and a minimal atom radius beyond which bonds and intermolecular interactions cannot extend to avoid collisions. The resulting web-based tool is called PoseEdit. The previously described technical workflow of PoseEdit is graphically illustrated in Figure 2.3.

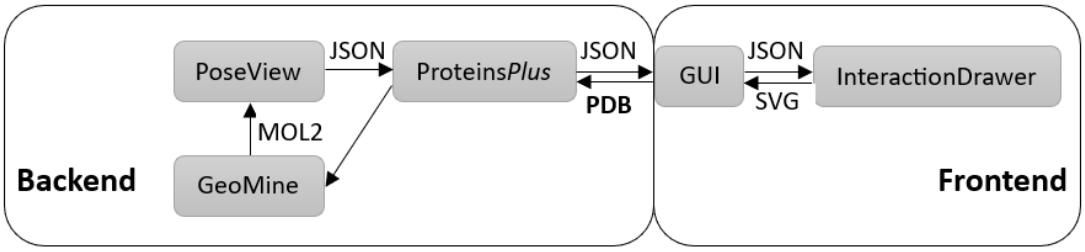



**Figure 2.3:** Technical workflow of PoseEdit

The InteractionDrawer library's JavaScript code is freely available for reuse and further development at [60]. The usage, configuration, and technical implementation of the InteractionDrawer library are documented in B.2.1, B.2.2, and B.2.4, respectively. The InteractionDrawer library is based on a prototype implemented in the context of the bachelor thesis of Lennart Weihs [61]. The prototype was then integrated into the frontend of the Proteins*Plus* web server for the master thesis of Bennet Krause [62]. The prototype's code quality required a high level of additional documentation, refactoring, bug fixing, and testing to develop new features. The prototype's untested and monolithic JavaScript code structure provided 21 files containing 31 classes with up to 10,000 lines and correspondingly long methods, making the library challenging to read and develop. The prototype code and the refactored, fully tested, and further developed code of the resulting InteractionDrawer library are described in B.2.5. The consequential reusability and extendability of the InteractionDrawer library are of particular importance to implement the tool GeoMine, the second tool developed during this doctoral project (Chapter 3). The additional feature development regarding the prototype of the InteractionDrawer library and its usability-improving Proteins*Plus* web server integration to implement PoseEdit's feature list (Tables 2.3 and 2.4) is summarized in the following and includes:

- adaptions to represent the new chemical information content and graphical styles specific to PoseEdit

- new *Edit* mode to edit visualized chemical and general graphical properties of specific atoms, bonds, text labels, and chemical structures

- extended *Add* mode for also adding text labels, atoms, covalent bonds, ionic interactions, and various chemical structures like specific amino acids or nucleic acid residues via a list and the Simplified Molecular Input Line Entry Specification (SMILES) language at the user-specified position in the drawing area

- extended *Move* mode and *Remove* mode that also affect single atoms, bonds, and rings
- extended *Rotation* mode that also affects complete hydrophobic contact splines
- extended *Select* mode to also affect text labels and intermolecular interactions, and that allows modifying, e.g., rotating or removing, selected graphical objects together as a group. Furthermore, the mode was further extended to enable GeoMine's 2D query interface (Chapter 3)
- various graphical configuration options for general styles of the 2D viewer and graphical objects, i.e., color themes, bond line width, atom label colors, circle representation for individual structures
- extended editing history that also tracks all graphical configuration changes and modifications of extended and new editing modes
- the complete ligand-bound binding site that corresponds to the diagram is visualized as a 3D representation in the 3D viewer of the Proteins*Plus* web server. Unlike the diagram, the 3D representation's chemical content is not limited, thereby showing the complete binding site, including all residues and additional intermolecular interactions like hydrogen bonds with water from both the binding site residues and the ligand. Various chemical features, such as the intermolecular interactions, are highlighted in the 3D binding site. 3D binding sites and their chemical features are calculated by the tool GeoMine, which also provides related chemical information as input to PoseView for generating corresponding 2D diagrams
- 2D viewer-3D viewer synchronization that allows the highlighting and focusing of mouse pointer-hovered and mouse pointer-selected atoms, bonds, and chemical structures simultaneously in both scenes
- user-friendly handling of the 2D viewer's functionality through, for example, user-activatable modes in a button toolbar instead of numerous and complex mouse and keyboard combinations
- a diagram legend illustrating all intermolecular interaction types and an info section that textually highlights focused scene objects
- export of a diagram in JSON format, the diagram and its legend in SVG format, and textual data about the ligand binding mode (intermolecular interactions, the interacting structures, and atoms) in TXT format

- import of a diagram in JSON format to reload a modified diagram in JSON or to load and simultaneously visualize multiple diagrams in the 2D viewer

- usage of PoseEdit via its integration in the Representational State Transfer Application Programming Interface (REST API) of the Proteins*Plus* web server, see B.4.1 for its documentation. Since the JavaScript-based InteractionDrawer library cannot be technically integrated into the Ruby-based Proteins*Plus* web server's backend, REST API-derived diagrams are directly generated by PoseView's own image drawing functionality. The graphical default styles of PoseView's images, for example, the colors of intermolecular interactions, have been adapted to match the ones set by PoseEdit

In addition, a distributable standalone version of PoseEdit was implemented. The in-house version of PoseEdit is based on a containerized version of the Proteins*Plus* web server, which was simplified to PoseEdit. The build process and usage of the PoseEdit container are documented in B.4.2.

## 2.5 Conceptual Summary

This section illustrates the newly developed graphical user interface of PoseEdit and its capabilities, which were previously summarized in Tables 2.3 and 2.4. The general description of the PoseEdit-hosting Proteins*Plus* web server and its usage are available in B.3.1. To start a PoseEdit calculation, a ligand from the central *Ligands* list (see B.3.1) must be selected by clicking on the corresponding 2D structure diagram. Figures 2.4 and 2.5 illustrate the key concepts and components used by PoseEdit in more detail: the 2D viewer and the 3D viewer. When a 2D ligand interaction diagram is calculated and subsequently visualized in the 2D viewer, the corresponding 3D binding site is also automatically displayed in the 3D viewer. The visual correspondence of chemical structures between both dimensions can be obtained by mouse hover highlighting or selection highlighting of displayed structural objects like atoms, bonds, and text labels representing structures. To provide an example, all 2D diagram structures were highlighted in transparent green via the *Select* mode in Figure 2.4. Due to the 2D-3D viewer synchronization, these structures are also highlighted in the corresponding complete 3D binding site.

The section at the top of the 2D editor (Figure 2.4a) shows the names of all diagram structures. Name-associated checkboxes can be marked to display or hide specific structures in the diagram. The two button toolbars below provide several interactive

diagram editing modes (Figure 2.4b), like modes for the adding or removing of structures and general viewer controls (Figure 2.4c), such as the diagram's download as an image file or the scene's zooming. Furthermore, the second toolbar provides a toggle list of general viewer settings to directly modify graphical styles in the diagram, such as the text size or colors of intermolecular interactions. All buttons are labeled with text and icons to highlight their functionality. A diagram editing mode is activated by a button click, which changes the button's color to blue. Buttons with an additional icon represent diagram editing modes whose activation requires additional input by users. For example, the *Add* mode requires the user-specification of the object type that shall be added. Table 2.5 describes the diagram editing modes from left to right through the toolbar's icon buttons. The complete list of general viewer settings is given in B.4.3. The general viewer controls in the second toolbar are listed in the following and include:

- upload and loading of a new diagram via a JSON file
- download of diagram data and the general viewer settings as JSON file
- download of a diagram image file in SVG format
- download of a legend image file in SVG format
- download of interaction data as TXT file
- upload and loading of other additional diagrams that are placed next to the already loaded ones via a JSON file
- *Undo* and *Redo* buttons to move stepwise through the editing history, which tracks all changes of diagram editing modes and changes of general viewer settings
- *Center* button, which centers the scene's content to fill the drawing area
- *Zoom* buttons to zoom in and out of the scene
- *Reset* button to revert all scene changes
- *Opts* button to access and modify a list of general viewer settings

Usage information about corresponding control elements of the 2D editor, like the button of a diagram editing mode, can be obtained via tooltips that are triggered by mouse hovering. The diagram is displayed in the drawing area below the two toolbars (Figure 2.4d). The section below the drawing area displays textual information about atoms, bonds, and structures hovered by the mouse pointer in either the 2D diagram or the

complete 3D binding site (Figure 2.4e). The last section shows a legend illustrating the supported intermolecular interaction types (Figure 2.4f). A new diagram calculation for another ligand can be started by clicking the red button below the legend section (Figure 2.4g).

**Table 2.5:** Diagram editing modes of PoseEdit

| **mode** | **options** | **function** |
|---|---|---|
| **Move** | Structure freedom level | translate the complete scene, a structure, structure circle, hydrophobic contact spline, spline control point, or annotation. Individual atoms, bonds, and rings can be translated by setting the option *Structure freedom level* to *Atoms and bonds* or *Rings*, respectively. Intermolecular interactions, hydrophobic contact splines, spline control points, and annotations are linked to a specific structure and specific structure atoms. Translating a structure, structure circle, ring, or atom, all linked intermolecular interactions, hydrophobic contact splines, spline control points, and annotations follow that translation |
| **Rotation** | - | rotate the complete scene, a structure, structure circle, or hydrophobic contact spline regarding their midpoints. Rotating a structure or structure circle, all linked intermolecular interactions, hydrophobic contact splines, spline control points, and annotations follow that rotation |

| | | |
|---|---|---|
| **Select** | Click, Lasso, Rectangle | select one or multiple diagram objects via mouse click or multiple ones at once via a rectangular or lasso selection tool. Deselect a specific diagram object with the *Click* selection mode by mouse clicking on it again and deselect everything by clicking in the blank of the 2D scene. Selected objects are highlighted in transparent green. The selection highlighting is also visualized in the complete 3D binding site in the 3D viewer. Furthermore, the selection highlighting is also shown in the downloadable SVG diagram. Atoms, bonds, structures, and structure circles can be moved, rotated, and removed together when grouped by selecting them |
| **Mirror** | Bond, Line | mirror a structure at a specific bond or a structure or hydrophobic contact spline at a user-adjustable line that intersects the diagram object's midpoint. Mirroring a structure, all linked intermolecular interactions, hydrophobic contact splines, spline control points, and annotations are also mirrored |

| | | |
|---|---|---|
| **Add** | Annotation, Atom with covalent bond, Atom-atom interaction, Cation-pi interaction Pi-pi interaction, Explicit H with covalent bond, Hydrophobic contact, Structure | specify an object type from a list to draw a new object of that type at the mouse pointer's position. Several properties can be set via a popup window to specify atoms (text label, element, charge, number of implicit hydrogens), annotations (text label, color, linkage to nearest atom for synchronizing repositionings), and structures (text label, SMILES, or structure from list). The complete list of structures that can be added from a list is given in B.4.3 |
| **Remove** | Structure freedom level | remove a structure, structure circle, hydrophobic contact spline, spline control point, annotation, or intermolecular interaction. Individual atoms, bonds, and rings can be removed, setting the option *Structure freedom level* to *Atoms and bonds* or *Rings*, respectively. Removing a structure or structure circle, all linked intermolecular interactions, hydrophobic contact splines, spline control points, and annotations are also removed. Removing a ring or atom, all linked intermolecular interactions are also removed |
| **Edit** | Annotation, Atom, Bond, Structure | specify from a list an object type to edit several properties of a specific object of this type via a popup window: atom (text label, element, charge, number of implicit hydrogens), bond (type), annotations (text label), and structures (text label, skeletal or circle representation) |

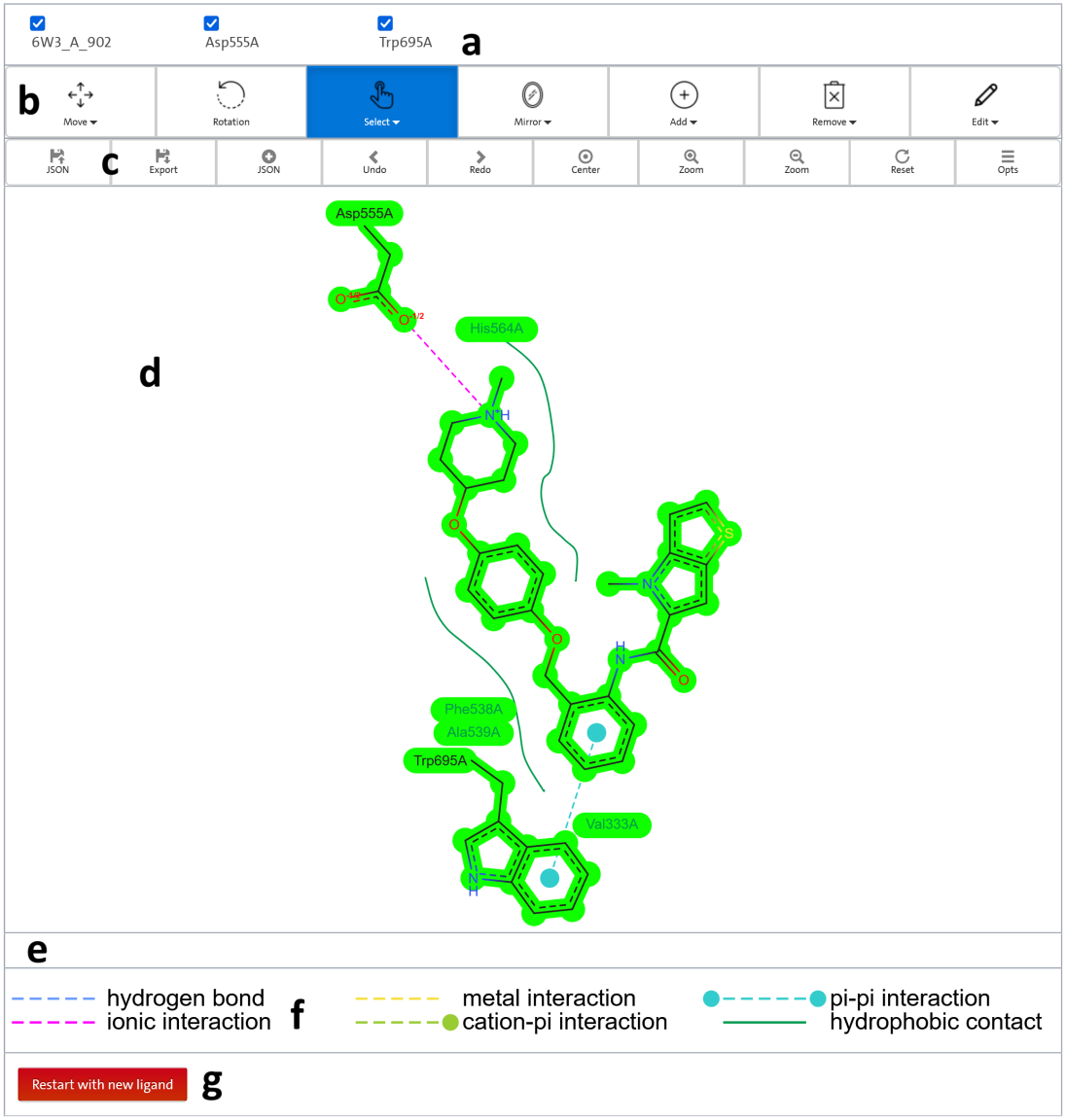



**Figure 2.4:** 2D viewer showing a PoseEdit diagram of lysine-specific histone demethylase 1A in complex with an inhibitor [58] (PDB identifier: 5LGT; Proteins*Plus* ligand identifier: 6W3_A_902). All structures are highlighted in transparent green. **a** Names of all diagram structures with checkboxes to toggle their visualization status. **b** Button toolbar with diagram editing modes. **c** Button toolbar for general viewer controls. **d** Drawing area showing the 2D ligand pose diagram. **e** Info section displaying information of atoms, bonds, and structures on mouse hover in the 2D viewer and 3D viewer. **f** Diagram legend showing the supported intermolecular interaction types. **g** Button to restart PoseEdit with another ligand

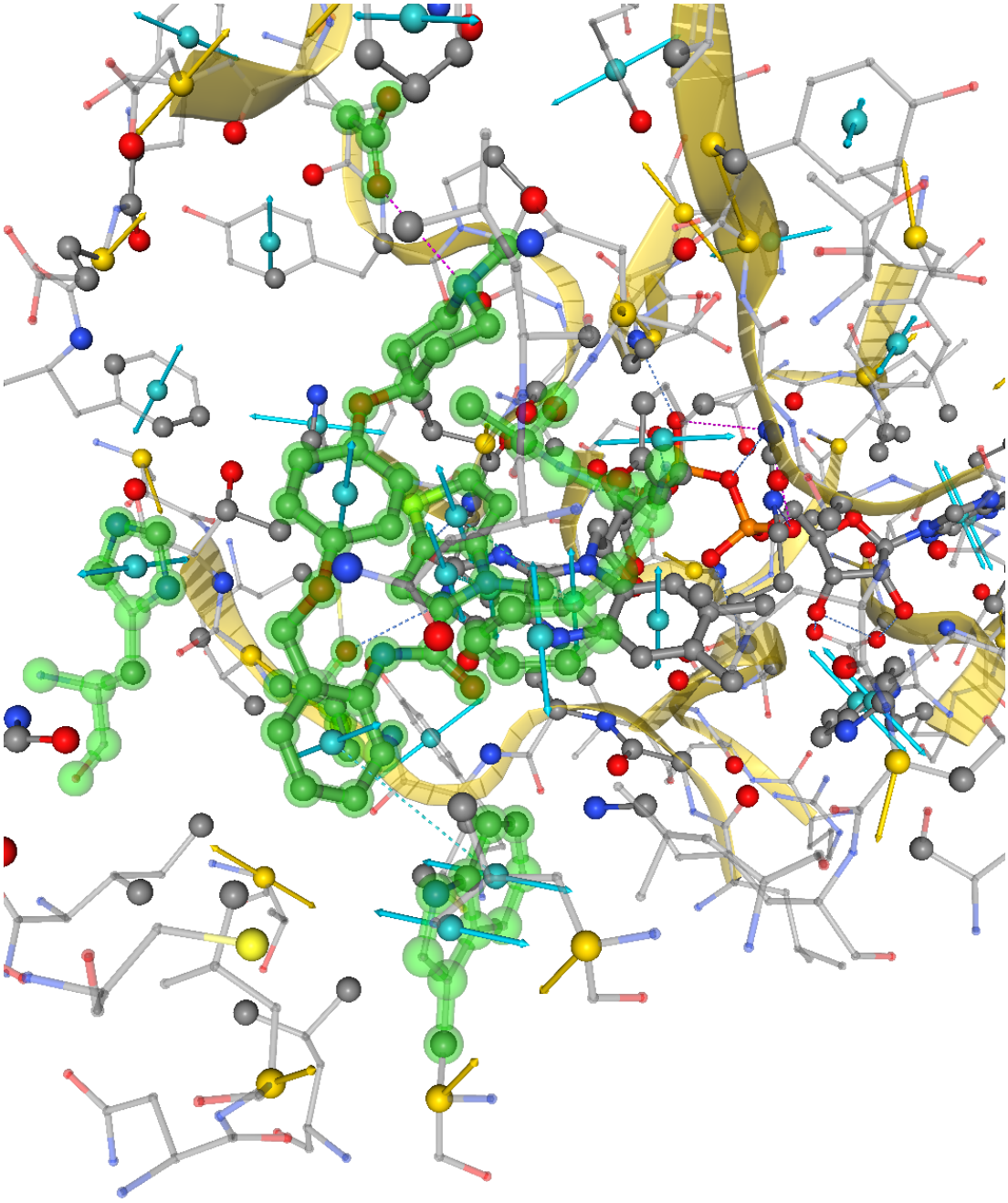


**Figure 2.5:** 3D viewer showing the complete binding site of lysine-specific histone demethylase 1A and an inhibitor [58] (PDB identifier: 5LGT; Proteins*Plus* ligand identifier: 6W3_A_902). All structures that are displayed in the 2D diagram are highlighted in transparent green

## 2.6 Application

Scientific publications [63] about PoseEdit's development or that cite the corresponding scientific article [D1] provide various application examples. The following application example taken and adapted from [D1] comprehensively demonstrates the capabilities of PoseEdit to interactively solve diagram issues, which have been identified as problematic for scientists (sections 2.1-2.2). The two diagrams in Figures 2.6, 2.7, and 2.8 are based on the previously mentioned PDB entry 5LGT, which represents lysine-specific histone demethylase 1A with a binding site containing flavin adenine dinucleotide (FAD) and the inhibitor 4-methyl-N-[2-[[4-(1-methylpiperidin-4-yl)oxyphenoxy]methyl]phenyl]thieno[3,2-b]pyrrole-5-carboxamide. Figure 2.6 and Figure 2.7 show the cofactor's and inhibitor's PoseEdit diagram.

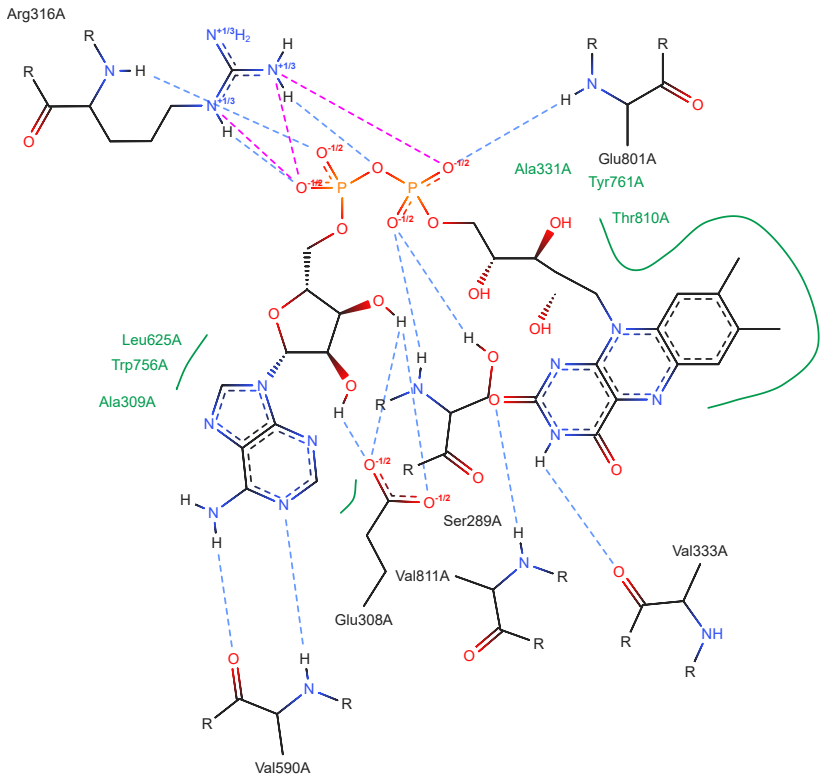



**Figure 2.6:** Unmodified PoseEdit diagram showing lysine-specific histone demethylase 1A in complex with a cofactor [58] (PDB identifier: 5LGT; Proteins*Plus* ligand identifier: FAD_A_901)

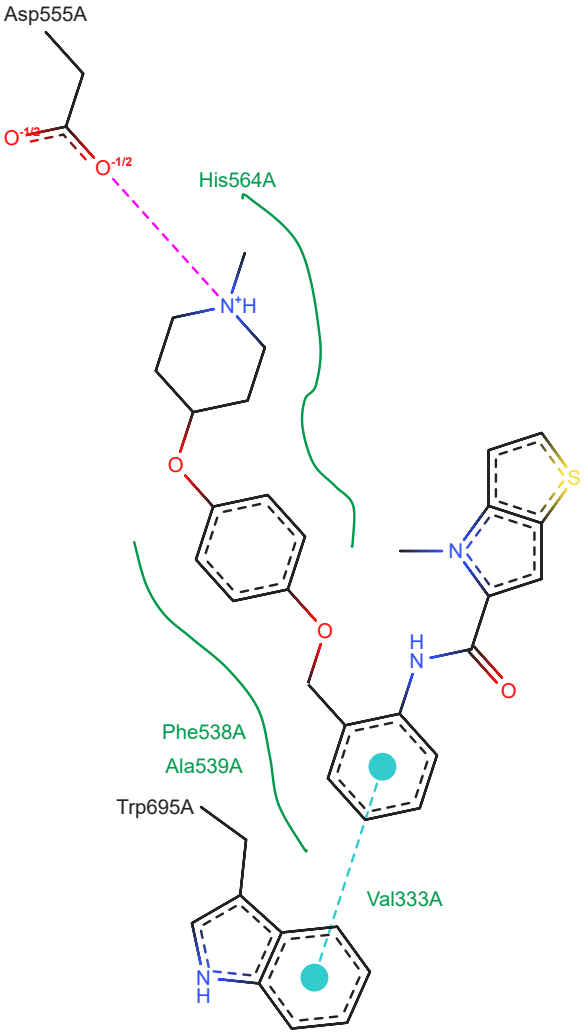



**Figure 2.7:** Unmodified PoseEdit diagram showing lysine-specific histone demethylase 1A in complex with an inhibitor [58] (PDB identifier: 5LGT; Proteins*Plus* ligand identifier: 6W3_A_-902)

Figure 2.8 shows a corresponding diagram after its manual modification with PoseEdit based on objective and exemplary subjective criteria. The general viewer controls, like the editing history or the zooming functionality, support users in applying the diagram editing modes precisely and via a trial-and-error approach. The 2D editor's functionality to hide and show specific chemical structures in the diagram helps users to focus on specific issues and subsequently solve them with diagram editing modes. The following diagram changes addressing objective graphical issues and subjective preferences regarding chemical information content and graphical styles were performed:

- removal of intersections and overlaps of intermolecular interactions, chemical structures, and text labels with the *Move*, *Rotate*, and *Mirror* modes
- relative rearrangements of individual chemical structures, hydrophobic contact splines, and text labels with the *Move*, *Rotate*, and *Mirror* modes to generate sufficient space and consequently improve the diagram's comprehensibility
- elongation of the cofactor's bent structural 2D layout to obtain a chemically more attractive representation
- exploring the corresponding complete 3D binding site supported by the 2D-3D synchronization of the mouse hover highlighting and the *Select* mode highlighting, users can verify additional chemical information content, like residues or water molecules that are relevant for the ligand binding mode's representation but not included by default. For this application example, the inhibitor 6W3_A_902 and its intermolecular interactions with the cofactor and with the binding site residues are manually added, which can be achieved either with the JSON file upload functionality that directly adds the diagram content of 6W3_A_902 or by individually adding all chemical structures, intermolecular interactions, hydrophobic contact splines, and text labels displayed in the diagram of 6W3_A_902 with the *Add* mode. Using the first approach, the complete diagram content of 6W3_A_902 can be aesthetically arranged relative to the cofactor FAD_A_901 with the *Select*, *Move*, and *Rotate* modes. Subsequently, the missing pi-stacking interactions displayed between the inhibitor and cofactor in the complete 3D binding site can be added to the diagram with the *Add* mode.
- removal of the hydrogen bond and corresponding explicit hydrogen atom of the N$\epsilon$ nitrogen atom of the Arg316A side chain with the *Remove* mode. Labeling of the N$\epsilon$ atom with one implicit hydrogen atom with the *Edit* mode. Consequently, only the stronger intermolecular interaction between the N$\epsilon$ atom and the ligand, the ionic one, is included in the diagram
- visualization of chemical structures by the circle representation with the *Edit* mode to reduce the visual overload and focus on interactions partners of interest, like the residues that interact with the cofactor
- highlighting important chemical features like intermolecular interactions, interacting atoms, and functional groups by transparent green color with the *Select* mode

- recoloring the diagram using the *Oldschool* theme via the general viewer settings to reduce the visual overload by too many colors
- recoloring the hydrophobic contact splines and corresponding text labels to an eyestrain-reducing darker color

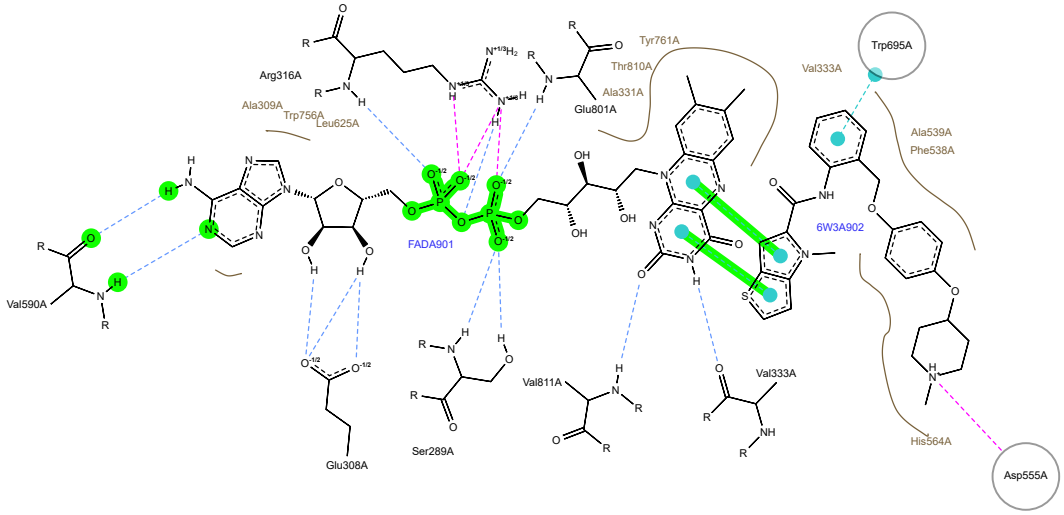



**Figure 2.8:** Modified PoseEdit diagram showing lysine-specific histone demethylase 1A in complex with a cofactor and inhibitor [58] (PDB identifier: 5LGT; Proteins*Plus* ligand identifier: FAD_A_901 and 6W3_A_902). Taken and adapted from [D1]

## 2.7 Outlook

Some additional improvements of PoseEdit might enhance the tool's usability and the diagram's information content:

- users might already have a custom interaction model that they want to apply to calculate 2D diagrams. A corresponding input option to specify the preferred intermolecular interaction types and their parametrization might reduce the necessity of manually adjusting the chemical information content of the diagram
- a chemical structure can be added via a SMILES string, a preselected list of commonly appearing chemical structures, or a previously exported JSON file of a diagram. However, chemical structures of interest may be present in the complete 3D binding site but missing in the 2D diagram. Therefore, an alternative approach might involve selecting chemical structures of interest directly in the complete 3D binding site, which consequently adds them to the 2D diagram

- the visualization of additional information, like the annotation of intermolecular interactions by geometric properties or hydrophobic contacts by the involved residue atoms, might simplify communicating and analyzing ligand binding modes. However, to address the layout quality-associated trade-off between the simplicity and the level of detail in the 2D scene, the visualization of such additional information might be kept optional, initially deactivated, and manually activatable

Furthermore, a new version of the Proteins*Plus* web server will be available in the near future. PoseEdit will not be included anymore due to its complex graphical user interface. Consequently, the tool is made independently available via its containerized web server version on `https://poseedit.proteins.plus` [64].

# Chapter 3

# 3D Structural Searching of Large Binding Site Collections

The general motivation for a 3D structural search engine for large 3D structure collections like the PDB and, consequently, the specific motivation for developing GeoMine is presented in the following sections. The motivation is based on the scientific applicability (section 3.1), the exploitability of the 3D structures in the PDB (section 3.2), and the strengths and limitations of respective state-of-the-art tools (section 3.3). The consequential perspectives and objectives regarding GeoMine's development are highlighted in section 3.4. The following sections present the technical groundwork and challenges of the development of GeoMine (section 3.5), the tool's underlying methodology (section 3.6), and its application (section 3.7). Section 3.8 gives an outlook on GeoMine's limitations and corresponding further tool development. [D2–D4] provide the tool's scientific publications. The tool PoseEdit [D1], which was also implemented for this doctoral project and is addressed in Chapter 2, is a building block of GeoMine's functionality [D5]. The Proteins*Plus* web server and its further development to integrate GeoMine are described in [36, D6, D7].

## 3.1 Relevance for Scientific Research

Similarity searches in large 3D structure collections like the PDB provide valuable insights in the context of drug development. There are many research questions that life scientists could ask. For illustrative purposes, two application examples are described below:

- side effect analysis: a drug candidate resulting from an in vitro high-throughput screening binds with high affinity to a functionally relevant binding site of a target protein. Life scientists want to investigate the drug candidate's selectivity based on that ligand-bound binding site. Does a similar relative spatial arrangement of interacting residues also occur in ligand-unbound binding sites of other proteins or protein classes? Resulting proteins might be off-targets of the drug candidate and, therefore, relevant for side effect analyses or selectivity optimization.

- drug repurposing: life scientists predict a ligand-unbound binding site of a protein drug target. The potential binding site is located between two $\alpha$-helices of distinct protein chains. It is assumed that the disruption of the protein complex through a ligand's binding could be functionally relevant. The binding site is characterized by a relative spatial arrangement of conserved residues that might be relevant for protein-protein complex formation and consequently exploitable for ligand recognition [65]. Do ligand-bound binding sites of different proteins contain ligands that interact with a similar spatial arrangement of residues? Resulting ligands might be potential candidates for a drug repurposing endeavor.

The common aspect of these research questions is the search for user-specified relative spatial arrangements of chemical features based on and found in ligand-bound or predicted ligand-unbound binding sites. Binding sites are the basis for a protein's function and, consequently, are commonly addressed by drugs, leading to a pharmaceutical effect. Therefore, searchable chemical features and relative spatial relationships of interest may include those that are important for protein-ligand complex formation. For example, solvent-exposed atoms of amino acids are the ones that undergo intermolecular interactions with ligands. Relative spatial arrangements of chemical features might then be specified through various geometric relationships such as distances, directions, and relative orientations. As illustrated, through 3D structural searches, life scientists can uncover valuable scientific knowledge and similarity relationships hidden in large 3D structure collections based on chemical and spatial similarity, thereby enabling a comprehensive functional analysis of proteins, ligands, and protein-ligand complexes in the context of drug discovery.

## 3.2 Relevance of the Data in the PDB

Due to the significant size and growth of the PDB, a purely visual approach to capture common chemical and spatial aspects of binding sites in the PDB is infeasible. As of January 1, 2024, the PDB has released 214,226 experimental entries, of which 185,697

are proteins, 12,476 are protein-nucleic acid complexes, and 16,053 are nucleic acids [66]. Since a PDB entry represents a 3D structure at a specific point in time under specific experimental conditions, the same protein can be represented by multiple entries. For example, structures of the same protein in the presence or absence of ligands have separate PDB entries and individual structural characteristics. On January 1, 2024, the PDB contains over 800,000 single protein chains in the asymmetric unit. Since each protein chain is assumed to have four functionally relevant binding sites on average [67], it is predicted that the PDB contains at this date more than 3.2 million binding sites [68]. The number of experimental 3D structure submissions to the PDB has increased exponentially in the past, with 10,959 new 3D structures before 2000, 17,726 from 2000 to 2004, 33,064 from 2005 to 2009, 43,311 from 2010 to 2014, 53,742 from 2015 to 2019, and 55,424 from 2020 to September 2, 2024 [69], see Figure 3.1.

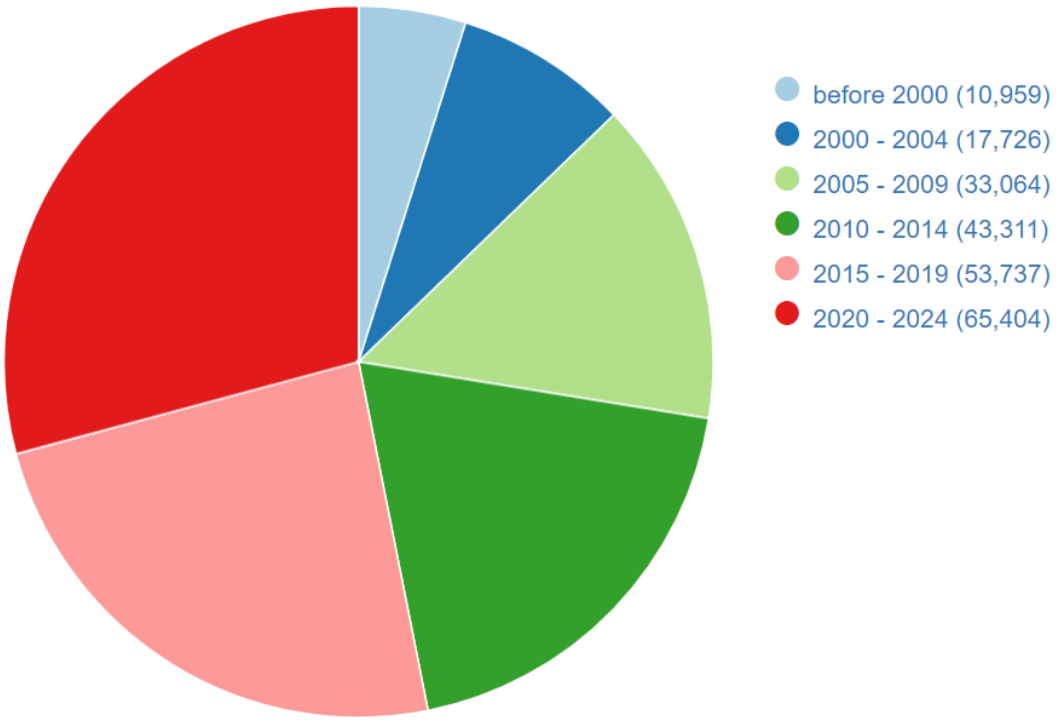



**Figure 3.1:** PDB structures by release date till September 2, 2024. Taken from [69]

An extrapolation based on that growth rate suggests that by 2030, 294,000 3D structures, 1.38 protein chains, and 5.52 million binding sites will be stored in the PDB [68]. In addition to the experimentally determined 3D structures, as of October 23, 2024, the PDB provides access to 999,251 artificial intelligence-predicted 3D structures of proteins from the AlphaFold database.

The current size of the PDB is sufficiently large enough to be considered for a 3D structural search engine that aims to extract meaningful chemical and spatial similarities to a query, answering research questions like the above-described ones. The more data is available, the more likely such queries can detect relevant matches. Therefore, the exponential growth of the PDB also indicates an increasingly attractive opportunity to exploit this data source with a 3D structural search engine even more successfully in the future.

## 3.3 State of the Art

Algorithmic advancements for the 3D querying of protein structures range back to 1991 [70]. 3DinSight [71] and SPASM [72], which were released in 1998 and 1999, respectively, were the first tools to feature such search functionalities supporting queries of $\alpha$-carbons and pairwise distance ranges on complete protein structures. Both tools are not available anymore. A decade later, various tools for the 3D structural querying of the continuously growing PDB were published, including the search functionality of the PDB itself [73], CSD-CrossMiner [74], PRDB [75], PROLIX [76], Relibase+ [77], PDBeMotif [78], PELIKAN [79], and GSP4PDB [80]. To provide a complete overview of features, this section also includes tools that are no longer available.

Table 3.1, Table 3.2, and the following paragraphs highlight the applicability of the state-of-the-art tools in comparison, considering the supported PDB search space and query content. Furthermore, the following paragraphs highlight the corresponding tool-specific strengths and limitations.

The PDB enables querying complete protein structures. Consequently, it is not possible to limit the search space to protein binding sites. However, as highlighted in the two application examples described above, protein binding sites are the target of important research questions like off-target prediction for side effect analyses and drug repurposing. All other tools search in ligand-bound binding sites, which is necessary for drug repurposing. However, none of the tools allows searching in predicted ligand-unbound binding sites, which is required for off-target-based side effect analyses. Ligand-bound binding sites are defined by a reference ligand and amino acids located within a tool-specific distance to one of the reference ligand's heavy atoms. For example, CSD-CrossMiner and PELIKAN consider all ligands with more than five atoms and fewer than 100 atoms as reference ligands and include amino acids within a radius of 6 Å and 6.5 Å, respectively. The binding site definitions of CSD-CrossMiner, Relibase+, and PELIKAN also include all simple ions like metal ions and additional small molecules within that radius, such as cofactors or water molecules. Their extended search spaces enable additional queries of

interest. For example, PELIKAN allows the search of water-mediated hydrogen bonds between a reference ligand and amino acids. In contrast to such a radius-based definition of ligand-bound binding sites, a predictive approach that analyzes the protein's surface might lead to a more comprehensive determination of a binding site's shape and boundaries [32].

A query of the PDB web service can have two to ten amino acids, which can have up to four alternative amino acid types. The spatial arrangement of the amino acids is automatically defined by distance ranges with a tolerance of 1 Å between their $\alpha$-carbons and $\beta$-carbons. All other tools provide users more control over the spatial query content, for example, to incorporate the known spatial flexibility of specific amino acid positions into the query. The query of CSD-CrossMiner is based on the spatial arrangement of pharmacophore spheres, which represent predefined chemical features, including specific atom types and aromatic ring centers. The radius of a pharmacophore sphere can be individually adjusted to increase or decrease the spatial tolerance regarding the chemical feature's position.

With Relibase+ and PELIKAN, arbitrary pairs of atoms and aromatic ring centers of the reference ligand and other small molecules, amino acids, and simple ions can be connected by distance ranges. Using PRDB, distance ranges can be specified between atoms of the reference ligand and amino acids. Distance ranges between amino acids can also be added, but only regarding their $\alpha$-carbons. GSP4PDB supports distance ranges between the reference ligand and amino acids and between amino acids considering the reference ligand's center of mass and the $\alpha$-carbons of amino acids.

PROLIX and PDBeMotif only support distance ranges between $\alpha$-carbons of amino acids. Consequently, these tools cannot define the relative position of a reference ligand. However, this could be useful, for example, regarding the application example about side effect analysis. The corresponding query could alternatively be searched in ligand-bound binding sites. Consequently, the additional incorporation of the ligand structure into the query could further support off-target detection based on the similarity of the ligand binding modes.

PROLIX, PDBeMotif, and PELIKAN allow to specify intermolecular interactions. With PROLIX, the exact interacting atoms of amino acids cannot be set. This chemical feature is useful, for example, in the context of the application example about side effect analysis. The relative spatial arrangement of intermolecular ligand interactions of the ligand could be alternatively searched in ligand-bound binding sites to identify potential off-targets.

PRDB, Relibase+, and PELIKAN allow the specification of angle ranges between pairs of distance ranges. With Relibase+ and PELIKAN, aromatic ring normal can also be used to define angle ranges. Furthermore, PELIKAN supports angle ranges with intermolecular interactions. Angle ranges can be useful, for example, to specify the relative orientation of two interacting functional groups like two aromatic rings that interact via pi-stacking interactions.

Tools that offer a 3D structural query with full atomic precision are CSD-CrossMiner, Relibase+, and PELIKAN. Since intermolecular interactions occur at the atomic level, a query with full atomic precision is useful to describe where and how a ligand and a binding site interact or could interact with each other. For example, regarding the application example about side effect analysis, specifying the relative spatial arrangement of the interacting atoms of amino acids instead of their $\alpha$-carbons could return more relevant results.

The atomic precision of the 3D structural query also depends on the supported chemical atom properties. The queries of CSD-CrossMiner, Relibase+, and PELIKAN can include various properties, which are highlighted in Table 3.2. In PELIKAN, all properties that are not atom-specific can also be set for aromatic ring centers. Chemical specifications of atoms can be used to further refine the results. For example, in the two previously described application examples, users might be interested in specifying the class of amino acids instead of their exact types, which may be too restrictive to screen for potential off-targets.

PELIKAN offers the largest number of chemical atom properties. However, several important ones are missing. Atoms or aromatic ring centers belonging to protein residues or nucleic acids cannot be differentiated, which prevents a targeted search in protein structures. Furthermore, PELIKAN does not allow to quantify the solvent exposure of atoms. However, only solvent-accessible atoms can undergo molecular interactions. Also, the position, length, and orientation of secondary structure elements like $\alpha$-helices and strands of $\beta$-sheets cannot be specified. Secondary structure specifications could be useful, for example, to detect reactive cysteine residues as potential targets for covalent inhibitors [81].

**Table 3.1:** Comparison of the search capabilities of tools for the 3D structural querying of the PDB regarding search space and query content

| | **PDB search space** | **user-specified distances** | **interactions** | **user-specified angles** | **atomic precision** | **PDB filter** |
|---|---|---|---|---|---|---|
| **PDB** | entire proteins (amino acids) | - | - | - | - | x |
| **CSD-CrossMiner** | ligand-bound binding sites (6 Å, amino acids + all other small molecules and simple ions) | - | - | - | x | x |
| **PRDB** | ligand-bound binding sites (8 Å, amino acids) | x | - | x | x (only for ligand-amino acid distances) | x |
| **PROLIX** | ligand-bound binding sites (4.5 Å, amino acids) | x (only for amino acids) | x | - | x (only for intermolecular interactions and the reference ligand) | - |

| | | | | | | |
|---|---|---|---|---|---|---|
| **Relibase+** | ligand-bound binding sites (7 Å, amino acids + all other small molecules and simple ions) | x | - | x | x | x |
| **PDBeMotif** | ligand-bound binding sites (16 Å, amino acids) | x (only for amino acids) | x | - | x (only for intermolecular interactions) | - |
| **PELIKAN** | ligand-bound binding sites (6.5 Å, amino acids + all other small molecules and simple ions) | x | x | x | x | x |
| **GSP4PDB** | ligand-bound binding sites (7 Å, amino acids) | x | - | - | - | x |

**Table 3.2:** Comparison of the supported atom properties of tools that offer 3D structural queries with full atomic precision

| | **atom properties** |
|---|---|
| **CSD-CrossMiner** | halogen, bromine, chlorine, fluorine, heavy atom, acceptor, donor, hydrophobic, metal, water, amino acid type from list (e.g., alanine), nucleic acid type from list (e.g., adenosine) |
| **Relibase+** | atom defined by its element and user-depicted chemical substructure to which it is covalently bound to |
| **PELIKAN** | atom element from list, acceptor, donor, anion, cation, hydrophobic, metal, water, reference ligand, ligand, residue, functional group from list, residue class and type from list (e.g., hydrophobic or alanine), backbone, sidechain, helix, sheet, no secondary structure, SMARTS environment, all properties can be set to match all properties |

An additional text-based query type for filtering PDB entries based on keywords and scalar values like the PDB identifier or the experimental resolution is offered by the PDB, CSD-CrossMiner, PRDB, Relibase+, PELIKAN, and GSP4PDB. This feature helps to limit the 3D structural search. For example, a filter for a low experimental resolution reduces the experimental uncertainty of resulting structures that might prevent a precise matching with atomistic detail. PELIKAN offers the largest quantity of textual, numerical, and chemical filters, including 48 searchable properties regarding the ligand, binding site, protein, and experimental methodology. However, corresponding data is not displayed for query template binding sites, making it difficult to filter efficiently for properties like a binding site's depth or volume. PELIKAN provides a SMARTS-based search for ligands [82]. However, PELIKAN does not support fingerprint-based similarity searches, which are widely used in drug discovery to detect similar ligands [83]. Furthermore, PELIKAN lacks several other useful filters that, for example, return only binding site matches that satisfy a specific RMSD range to the 3D structural query or are composed of multiple protein chains.

The differences of the graphical user interfaces of state-of-the-art tools regarding accessibility, query specification, and results presentation are summarized in Table 3.3

and in the following paragraphs. In addition, the following paragraphs highlights the corresponding tool-specific strengths and limitations.

PRDB, PROLIX, Relibase+, and PDBeMotif are not available anymore. Of the tools still available, the PDB search functionality and GSP4PDB are freely accessible web applications, and CSD-CrossMiner and PELIKAN are desktop applications. CSD-CrossMiner is a commercial tool and PELIKAN is freely accessible to academic users only, limiting their accessibility. While a desktop application like PELIKAN offer more data privacy and its usage without an internet connection, a web-based platform like the PDB is independent of the device, operating system, software dependencies as well as installation and update obstacles.

The PDB, CSD-CrossMiner, and PELIKAN provide a 3D editor for generating queries. The query formulation process via the 3D editor requires a structural template from the PDB. Furthermore, CSD-CrossMiner and PELIKAN allow to load a custom 3D structure file as a query template. While the PDB web service allows query selection anywhere in the template structure, CSD-CrossMiner and PELIKAN provide corresponding 3D visualizations of ligand-bound binding sites. The chemical content of the visualizable query template binding sites of CSD-CrossMiner and PELIKAN is the same as the one described above regarding their search spaces.

In addition, CSD-CrossMiner supports query generation from scratch with the 3D editor by placing the query directly in the 3D space. PROLIX, Relibase+, PDBeMotif, and GSP4PDB provide a 2D editor for drawing queries and no query templates. PRDB requires a purely textual query design in SQL format. A PELIKAN query can also be formulated textually via a file in Extensible Markup Language (XML) format and by tables, which are synchronized with the 3D editor.

In contrast to enabling text-based queries on the PDB, e.g., with keywords or amino acid sequences, realizing a user-friendly interface for 3D structural queries is a challenging task. The purely textual formulation of 3D structural queries, as in PRDB, can be difficult. However, it could enable a tool's integration via a command line interface (CLI) in automated computational workflows. 3D and 2D scenes are familiar graphical environments for life scientists to visualize chemical structures. Particularly, 2D depictions of molecules as IUPAC style-based chemical structure diagrams are widely used. Therefore, constructing a query in 2D or 3D space is a reasonable approach. Both visualization concepts have advantages and disadvantages, which are outlined in Chapter 2.

Unlike a 2D scene, a 3D scene provides precise 3D spatial information, which supports the generation of 3D structural queries. A 2D scene shows only distorted 3D spatial

information, for example, through the explicit depiction of distance constraints between chemical features. However, depending on the level of 3D spatial information required for formulating a specific query, a 2D scene can enable easier query generation with and without template structure and a faster overview of its content. In contrast to a 2D scene, a 3D scene requires complex and time-consuming user operations in 3D space to formulate and inspect a query. Due to the 3D spatial information, it might be necessary to intensively zoom, rotate, and translate the 3D scene to place and rearrange the query's components. The query generation with the 3D editor is also complicated using a query template structure. In contrast, a 2D scene offers a simplified visualization of a query template through chemical structure diagrams. While a 3D scene displays a query template structure without limitations, its chemical information content is restricted in a 2D scene to avoid layout issues like graphical overlaps caused by dimensionality reduction. Consequently, although a 2D scene only displays a limited amount of chemical information of a template structure, it has the advantage of highlighting those chemical features that are most important to search for.

A template-based query generation approach in 2D and 3D can significantly support users in translating a research question into a query. For example, as in the two previously described application examples about off-target prediction, users might want to use a protein binding site as a starting point for generating queries. None of the tools provides 3D visualizations of predicted ligand-unbound binding sites as templates preventing query formulation for the application example about drug repurposing. In contrast, 3D visualizations of ligand-bound binding sites are provided by CSD-CrossMiner and PELIKAN, which enables query formulation for the application example about side effect analysis. Directly providing 3D visualizations of binding sites as query templates directs the focus on these important aspects, thereby simplifying the query generation process. In contrast, the complete 3D structure as a query template provides the advantage to select queries beyond the limitations of tool-specific binding site definitions. Although 3D structures of the AlphaFold database can be loaded with CSD-CrossMiner and PELIKAN via corresponding PDB files, the direct loading of database entries via the corresponding UniProt accession numbers is not supported.

Template-free query editing in 2D or 3D can be useful when the query template structure is insufficient to describe the complete query, e.g., due to a missing ligand, which can be included by placing a point in the part of the binding site where a ligand might be situated. This can be the case, for example, regarding the application example about drug repurposing to annotate binding sites with potential ligands. A relative spatial arrangement of amino acids is selected in a ligand-unbound binding site. The query

is then searched in ligand-bound binding sites to detect potential ligands. However, the matching amino acids in the ligand-bound binding sites may not be in a sufficient proximity of ligands. Therefore, to obtain more relevant results, users might also be required to incorporate the ligand structure into the query via a template-free query editing approach.

The additional textual editing of the query's elements, for example, via its tabular representation in PELIKAN, can be useful to examine their chemical and spatial properties and adapt them to user-specific requirements, for example, the element of a hydrogen bond acceptor or the width of a distance range.

All tools provide a list of search results with information about matching PDB entries, such as the PDB identifier code or the title of the PDB entry. Depending on the specificity of the query, the number of results can vary significantly, from a few to several thousand. Therefore, an intuitive navigation through the result set is required. The result lists of the PDB and CSD-CrossMiner are ranked by the root-mean-square deviation (RMSD) between the 3D structural query points and the corresponding matching structural parts. The RMSD ranking helps to browse through results more quickly and helps to filter the best-matching results.

The 3D superimposition of results onto the query and the export of results statistics are supported by the PDB, CSD-CrossMiner, and PELIKAN. The 3D visualization of individual query-superimposed results is important to visually inspect how they match the query's chemical and spatial constraints. 3D viewer options to change the representation styles and displayed chemical content of results support result inspection by simplifying unnecessarily overloaded visualizations or by improving the visibility of specific aspects. For example, PELIKAN provides the option to display the complete binding sites for all results or only those residues or ligand parts that were directly matched by the query. Statistics and query-based 3D superimpositions of multiple results can be helpful in elucidating common or varying chemical and spatial aspects. The applicability of these two approaches depends on the number of results and if the query is kept generic, which leads to more different results. Furthermore, both approaches can be of additional benefit if they include not only the structural parts of the template and the results that were matched by the query and consequently superimposed onto it. For example, regarding the application example about off-target-based side effect analysis, statistics and query-based 3D superimpositions of multiple results might reveal alternative interacting amino acid types or additional ones, which could be relevant for the off-target's affinity. Furthermore, clashes between the ligand and a query site are not reported and these can only be found by visual inspection.

A query might not return results of interest. Therefore, users may want to modify it. CSD-CrossMiner and PELIKAN support query refinements. Users can interactively modify an already-used query and continue to search in the results of the previous query. For example, the search for relative spatial arrangements of amino acids, like in the two previously described application examples, might return more relevant results with a lower spatial query flexibility.

Only the PDB provides a history functionality, which tracks all performed searches. Furthermore, a query file can be imported and exported with CSD-CrossMiner and PELIKAN. Both features help users to archive, share, and later modify a query and to reproduce or update its results. None of the tools offer the export of results in PDB format, which complicates their postprocessing with other tools.

**Table 3.3:** Comparison of graphical user interfaces of tools for the 3D structural querying of the PDB regarding accessibility, query specification, and results presentation

| | **accessibility** | **query specification** | **results presentation** |
|---|---|---|---|
| **PDB** | web application, free | 3D editor, 3D template-based anywhere in the complete structure of a PDB entry | RMSD-ranked list of results, 3D, query-based 3D superimpositions, results history |
| **CSD-CrossMiner** | desktop application, commercial | 3D editor, 3D template-based in ligand-bound binding sites (6 Å, amino acids + all other small molecules and simple ions) of a PDB entry or custom PDB file, 3D template-free, refinement, import/export | RMSD-ranked list of results, 3D, query-based 3D superimpositions |
| **PRDB** | desktop application, discontinued | SQL | list of results |

| | | | |
|---|---|---|---|
| **PROLIX** | desktop application, discontinued | 2D editor, 2D template-free | list of results, 3D, statistics |
| **Relibase+** | web application, discontinued | 2D editor, 2D template-free | list of results, 3D |
| **PDBeMotif** | web application, discontinued | 2D editor, 2D template-free | list of results, 3D, statistics |
| **PELIKAN** | desktop application, commercial, free for academic use | 3D editor, 3D template-based in ligand-bound binding sites (6.5 Å, amino acids + all other small molecules and simple ions) of a PDB entry or custom PDB file, text-based via tables or XML file, refinement, import/export | list of results, 3D, statistics, query-based 3D superimpositions |
| **GSP4PDB** | web application, free | 2D editor, 2D template-free | list of results, 2D, 3D |

## 3.4 Perspectives and Objectives

As indicated by the presentation and analysis of state-of-the-art tools, the development of a 3D structural search engine is based on three initial key problems:

- Usability: due to the large amount of data and the high complexity of the three-dimensional data type, the development of a user-friendly search engine represents a significant challenge. Search engines that handle such data must offer easily accessible, comprehensively functional, yet reasonably intuitive graphical user interfaces for query generation and results presentation. Therefore, the development of a 3D structural search engine is highly focused on usability-centric design aspects.

- Applicability: the supported search space and 3D structural query must be relevant to research questions of life scientists, including searchable regions, chemical features, and spatial relationships of interest. In addition, the 3D structural search engine must be supplemented by a text-based query type for the specification of relevant search terms, such as the molecular weight of a ligand, the volume of a binding site, the organism the protein is derived from, or the experimental resolution of the PDB entry. Both query types can then be used synergistically to limit and explore the search space.
- Efficiency: due to the high complexity and large amount of data, the search engine's performance represents an additional challenge. Runtime and memory requirements of the search algorithm and database must be optimized and scalable with a rapidly growing search space.

The primary aim of this doctoral project is to develop a search engine for the 3D structural querying of chemical features in large molecular 3D structure collections like the PDB, focusing on protein-ligand interfaces. The resulting tool is called GeoMine. The starting point of GeoMine's development is the tool PELIKAN and the data of the PDB. PELIKAN has been improved in terms of usability, applicability, and efficiency. GeoMine's usability is the primary focus of this doctoral project. Its efficiency is addressed in the doctoral projects of Joel Graef and Martin Poppinga and in the corresponding scientific publications [D3, D4]. Therefore, the primary objective of this doctoral project is to enable life scientists to intuitively apply GeoMine's search algorithm. All mentioned doctoral projects deal with applicability-related issues since the search space and query content must be integrated from the search algorithm's and graphical user interface's perspective. GeoMine's development is based on application examples like those previously described, i.e., off-target prediction for side effect analysis and drug repurposing, as well as the strengths and limitations of state-of-the-art tools. Table 3.4 highlights GeoMine's development goals with PELIKAN as a starting point.

**Table 3.4:** Development of the graphical user interface of GeoMine regarding accessibility, query specification, and results presentation, and of GeoMine's search capabilities regarding search space and query content with PELIKAN as starting point. GeoMine provides all features of PELIKAN. The right column only lists those GeoMine features that are new or were further developed with respect to PELIKAN

| | **PELIKAN** | **GeoMine** |
|---|---|---|
| **accessibility** | desktop application, commercial, free for academic use | free web application, standalone version |
| **query specification** | 3D editor, 3D template-based in ligand-bound binding sites (6.5 Å, amino acids + all other small molecules and simple ions) of a PDB entry or custom PDB file, text-based via tables or an XML file, refinement, import/export | 3D template-based anywhere in the complete structure and in predicted ligand-bound and ligand-unbound binding sites (amino acids + all other small molecules and simple ions) of a PDB entry, AlphaFold entry, or custom PDB file, 3D template-free, 2D template-free, 2D template-based in interactive 2D ligand interaction diagrams, XML (REST API), JSON file (GUI), textual data for template binding sites |
| **results presentation** | list of results, 3D, statistics, query-based 3D superimposition | RMSD-ranked list of results, download of resulting binding sites in PDB file format, results history, extended 3D visualization options |
| **PDB search space** | ligand-bound binding sites (6.5 Å, amino acids + all other small molecules and simple ions) | predicted ligand-bound and ligand-unbound binding sites (amino acids + all other small molecules and simple ions) |

| | | |
|---|---|---|
| **chemical properties of atoms** | atom element from list, acceptor, donor, anion, cation, hydrophobic, metal, water, reference ligand, ligand, residue, functional group from list, residue class and type from list (e.g., hydrophobic or alanine), backbone, sidechain, helix, sheet, no secondary structure, SMARTS environment, all properties can be set to match all values | amino acid, amino acid type and class (e.g., hydrophobic or alanine) from list, nucleic acid, nucleic acid type (e.g., adenosione) from list, solvent exposure, midpoints and endpoints ($\alpha$-carbons) of $\alpha$-helices and $\beta$-sheet strands |
| **PDB filter** | | fingerprint-based ligand similarity, RMSD range, binding sites consisting of single/multiple chains, ligand-bound/ligand-unbound binding sites, no symmetrical results, maximum number of results per pocket |

## 3.5 Technical Groundwork and Challenges

This section summarizes the technical groundwork and challenges to develop the key concepts and components of GeoMine, which are described in Table 3.4. To this end, the main development steps of GeoMine are briefly introduced. The Appendix explains corresponding general aspects like the technical implementation in more detail. GeoMine is based on the desktop tool PELIKAN, which was further developed for this doctoral project, resulting in the new tool GeoMine. The primary aim and technical challenge was rebuilding and further developing PELIKAN's functionality on a freely available and easily accessible web-based platform. Therefore, PELIKAN's graphical user interface and functionality had to be completely decoupled from its desktop-based

graphical user interface and rebuilt on a web server. The tool code for communicating with the database and its API regarding query delivery and retrieval of search results was refactored and reimplemented. The input and output functionalities of the command line interface of PELIKAN were further specialized to develop GeoMine. The tool's main input, i.e., the query, is provided via an XML file. The XML file content is described in B.5.5. The capability to read query files in XML format already existed and has been further developed in this doctoral project to integrate new features. The tool's main output, i.e., the search results, is saved in a JSON file. Furthermore, two additional JSON file-based command line interface outputs were implemented. The first JSON returns the number of stored PDB entries and binding sites. The second JSON contains chemical and spatial data of binding sites for a specific PDB entry, which is required for a template-based query generation. The content of the respective JSON files is described in B.5.6. The usage of GeoMine's command line interface is described in B.5.1. The command line interface version of GeoMine was integrated into the backend of the Proteins*Plus* web server. B.3.1 and B.3.2 describe the general usage and technical implementation of the Proteins*Plus* web server. Based on the web server's source code and frontend components like the molecular 3D viewer, a new graphical user interface was built for GeoMine, which requires the JSON output of the tool's command line interface as input. Details about the technical implementation of GeoMine's backend and frontend are given in B.5.7.

The graphical user interface is based on a prototype, which was developed by the author of this doctoral thesis for his master's thesis [84] and further developed regarding code quality and features. Furthermore, the InteractionDrawer library, which was implemented in the context of this doctoral project, see Chapter 2, is reused and integrated into GeoMine's frontend to develop one of its primary components, the 2D query editor. The library's editing modes were extended to enable the selection and placement of query components in 2D. Furthermore, the functionality of the 2D interface was synchronized with other frontend components, including the 3D query editor and the tabular query representation. The previously described technical workflow of GeoMine is graphically illustrated in Figure 3.2. The additional feature development of the prototype to implement GeoMine's features list (see Table 3.4) is summarized in the following and includes:

- the searching and loading of AlphaFold entries via the Proteins*Plus* web server
- the on-the-fly calculation and visualization of predicted ligand-bound and ligand-unbound binding sites in the 3D viewer for an entry of the PDB database, AlphaFold database, or a custom PDB file. The calculation criteria for binding sites

and the displayed additional chemical binding site content like solvent-exposed atoms or aromatic ring centers are described in B.5.4

- the calculation and visualization of a 2D ligand interaction diagram for a specific ligand of the input structure in the 2D viewer
- the template-based design of queries in predicted ligand-bound and ligand-unbound 3D binding sites or 2D ligand interaction diagrams by selecting visualized chemical features
- the template-free query generation from scratch in the 3D viewer and 2D viewer by placing new hypothetical chemical features
- the support of the query design process through 2D-3D synchronization and instant visual correspondence of the query and chemical features between both dimensions and the query tables through a synchronized mouse-hover and color highlighting
- the download and upload of a query as JSON file
- an optional full-screen mode of the 3D viewer to design queries
- the structural template and query visualization can be switched on and off, guiding a query generation process that mixes a template-based and template-free design approach
- the visualization of corresponding textual, numerical, and chemical properties of template binding sites that can be set as values for corresponding PDB filters
- extension of GeoMine's search capabilities by new textual, numerical, and chemical PDB filters, such as the fingerprint-based ligand similarity filter, and by PELIKAN-derived and completely new 3D query content like aromatic ring normals and secondary structure points. As highlighted in section 3.4, the development and implementation of the search algorithm is addressed by two other doctoral projects. The new filters that include or exclude results based on the RMSD similarity to the 3D query and a maximal number of matches per binding site are post-processing steps on the results that were completely implemented in this doctoral project
- a results refinement functionality and browsable results history, which adds one history step for each executed search step

- a results table with a sortable, searchable, and RMSD-ranked list of query matches
- additional 3D visualization options that can also be set for individual results instead of all results
- the download of resulting superposed binding sites as PDB files
- integration of GeoMine in the REST API of the Proteins*Plus* web server, see B.5.2 for its documentation

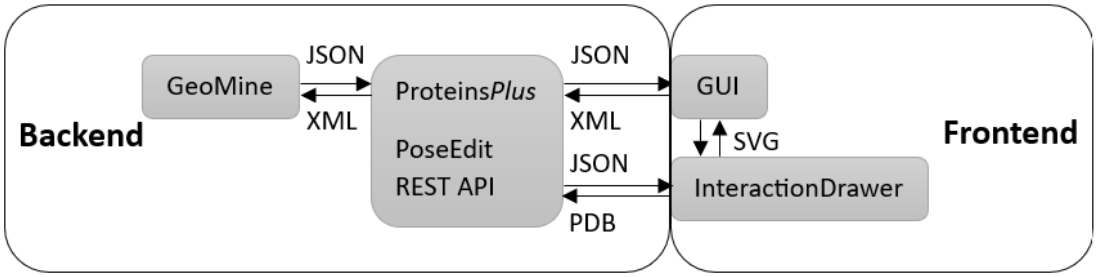



**Figure 3.2:** Technical workflow of GeoMine

In addition to the Proteins*Plus* web server version of GeoMine, a distributable standalone version of GeoMine was created. A containerized version of the Proteins*Plus* web server reduced to GeoMine was the basis for developing an in-house version of the tool. A description of the GeoMine container's build process and usage is given in B.5.3.

## 3.6 Conceptual Summary

The new graphical user interface of GeoMine and its features (see Table 3.4) are outlined in this section. B.3.1 provides the general description of the GeoMine-hosting Proteins*Plus* web server and the usage of its graphical user interface. The workflow concept of the user interface of GeoMine can be subdivided into three parts: query generation, results analysis, and results refinement. A cyclin-dependent kinase structure bound to the inhibitor 4-[[6-(cyclohexylmethoxy)-7H-purin-2-yl]amino]benzenesulfonamide [57] (PDB identifier: 1H1S; Proteins*Plus* ligand identifier: 4SP_A_1298) is used as query template to illustrate the workflow.

### 3.6.1 Query Generation Workflow

The query generation workflow of GeoMine is subdivided into two steps: template selection and query design. Figure 3.4 illustrates the associated concepts and synchronized

user interface components of GeoMine in more detail: the 2D viewer (Figure 3.4a) and 3D viewer (Figure 3.4b), and the 3D query tables (Figure 3.4c). Additional graphical user interface components include the textual, numerical, and chemical PDB filter table and the PDB subselection list. The 2D viewer, 3D query tables, PDB filter table, and PDB subselection list are GeoMine-specific and, therefore, separately accessible via the *3D query*, *PDB filter*, and *PDB subselection* tabs in the tool-specific graphical user interface on the right of the main page. GeoMine employs the 3D viewer on the left of the main page, whose functionality is shared by all the Proteins*Plus* web server tools.

**Template selection**:
Loading a chemical structure entry in PDB format from the AlpahFold database, PDB database, or a custom PDB file on the Proteins*Plus* landing page, GeoMine automatically calculates ligand-bound and ligand-unbound binding sites of the input structure. The calculated binding sites are loaded into the central *Pockets* list of the main page, while all ligands of the input structure are loaded in its central *Ligands* list. A predicted ligand-bound or ligand-unbound binding site of the input structure can be loaded into the 3D viewer from the *Pockets* list. A 2D ligand interaction diagram can be displayed for all ligands of the input structure in the 2D viewer by selecting a ligand from the *Ligands* list. The 2D or 3D scene can be rotated with a left mouse click and translated with a right mouse click. Alternatively, to directly load a diagram from a PoseEdit session, a diagram file in JSON format can be uploaded into the 2D viewer of GeoMine. This approach is useful, for example, if users want to manually optimize the diagram's layout with PoseEdit before designing a 3D query. Binding sites can be independently visualized from the complete input structure in the 3D viewer. Consequently, the complete input structure, except the binding site, can be set invisible to enable the binding site's visual focus for 3D query generation. Alternatively, the complete input structure can be displayed, for example, in the *Cartoon* representation, to better understand the secondary structure elements beyond the binding site boundaries. The loading of a 2D ligand interaction diagram for a ligand of the *Ligands* list automatically displays the corresponding ligand-bound binding site of the *Pockets* list in the 3D viewer while hiding everything else, accelerating the template-based 3D query generation.

Figure 3.3 illustrates and compares the selectable content of a 3D binding site and a corresponding 2D diagram. In the 2D and 3D viewer, displayed chemical features that are selectable for 3D query design include buried atoms represented by small colored spheres, solvent-exposed atoms, aromatic ring centers, secondary structure points represented by big colored spheres, and intermolecular interactions highlighted by dashed colored lines. Furthermore, 3D binding sites display arrows representing ring normals

of aromatic ring centers and directions of secondary structure elements for corresponding secondary structure points. Buried atoms and covalent bonds are transparently visualized to set a focus on chemical features that are particularly important for ligand recognition, like solvent-exposed atoms. Furthermore, a grid that visually highlights spatial aspects like the binding site's surface, volume, and depth can be loaded into the 3D viewer. Each *Pockets* list entry provides textual, numerical, and chemical properties of the represented binding site, like its surface, volume, and depth, which facilitate the specification of corresponding PDB filter values.

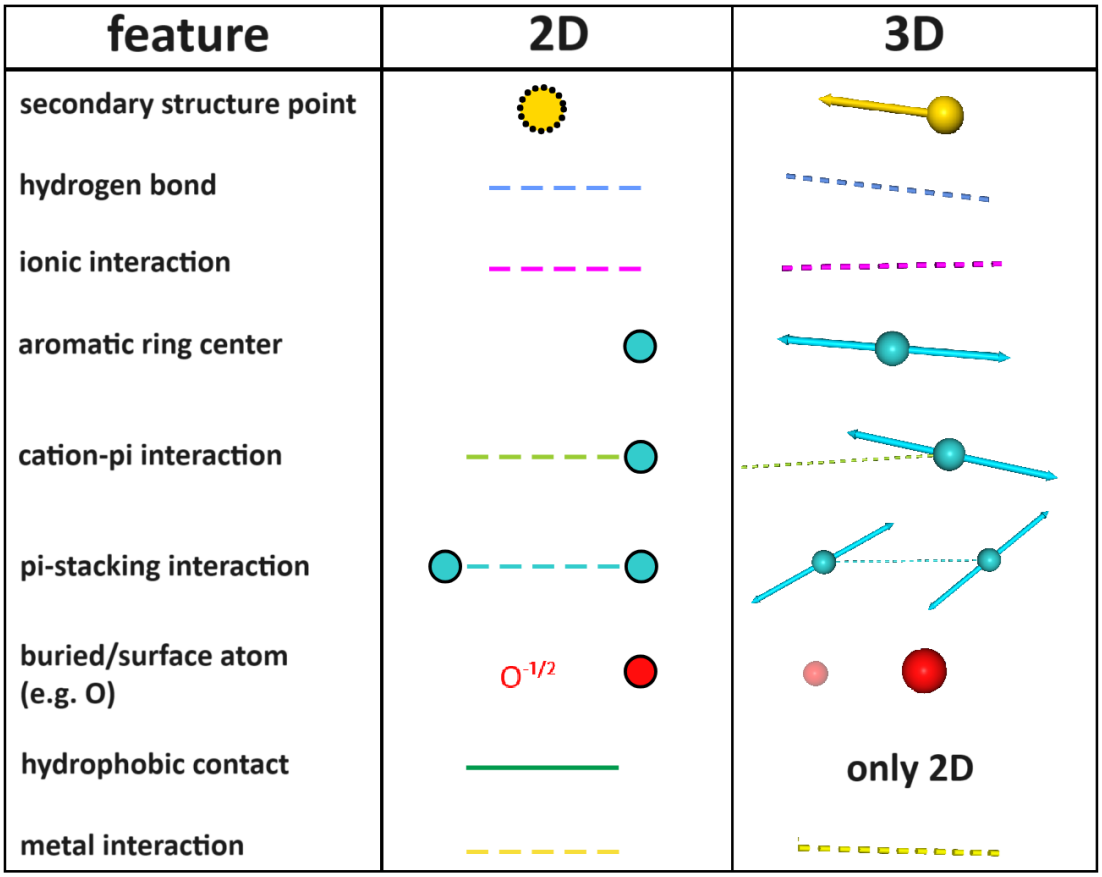



**Figure 3.3:** Selectable content of 2D diagrams and 3D binding sites displayed in the left and right column respectively

**Query definition**:
A selection mode can be set via a drop-down list on the right of GeoMine's tool-specific graphical user interface to add specific 3D query objects in the 2D diagram and 3D binding site with a left mouse click. 3D query objects include so-called points represented by big transparent colored spheres, and distances and interactions displayed

by continuous and dashed transparently colored cylinders between points, respectively. Figure 3.4 shows an example query selected in a 2D diagram (Figure 3.4a) and a corresponding 3D binding site (Figure 3.4b). All 2D and 3D selected points, distances, and interactions are automatically added to the respective 3D query tables (Figure 3.4c) to examine and adjust chemical and spatial properties, e.g., residue type and distance ranges. Figure 3.4 illustrates the table data of all query object types, showing one row for each 3D query table. The 3D query design is synchronized between the 2D and 3D spaces and the 3D query tables. The 2D-3D-3D query tables correspondences regarding 3D query and structure visualization are easily detectable by the unique coloring and synchronized mouse hover highlighting of 3D query objects and chemical binding site features like atoms or intermolecular interactions. The 3D query table section is scrollable to allow the simultaneous visualization of the 2D viewer, 3D viewer, and the 3D query table entries of interest. Furthermore, the 3D viewer can be switched to full screen, facilitating query generation in 3D.

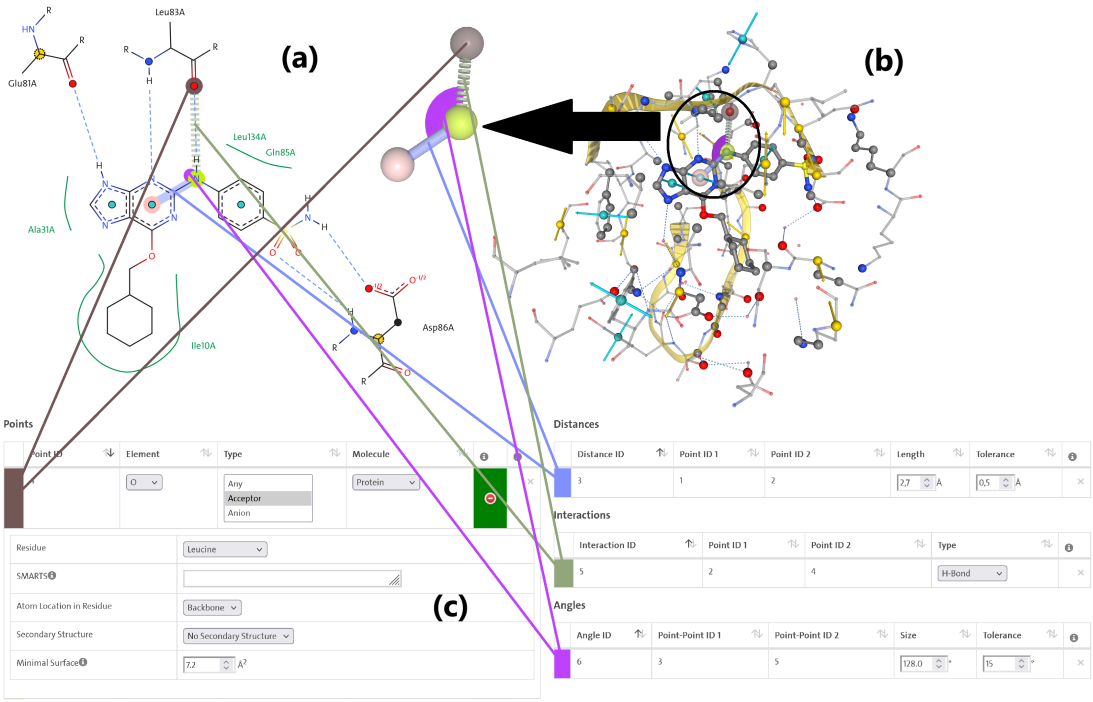



**Figure 3.4:** Example 3D query selected in a cyclin-dependent kinase structure bound to the inhibitor 4-[[6-(cyclohexylmethoxy)-7H-purin-2-yl]amino]benzenesulfonamide [57] (PDB identifier: 1H1S; Proteins*Plus* ligand identifier: 4SP_A_1298) and visualized in GeoMine's primary and synchronized graphical user interface components. **a** 2D viewer showing a 2D pose diagram. **b** 3D viewer displaying the corresponding complete binding site. A zoomed-in view highlights the 3D query. **c** 3D query tables listening chemical and spatial data of the 3D query objects

In the *Point* mode, buried and solvent-exposed atoms, aromatic ring centers, and secondary structure points can be left mouse-clicked to select points. Furthermore, undefined points can be directly placed and moved in the 2D or 3D scene. Placing a hypothetical point in a 2D diagram automatically adds it to the center of the corresponding ligand-bound 3D binding site. Consequently, the point's relative position must be manually adapted by its translation in 3D and its subsequential connection to other points via distances and interactions via the 2D or 3D viewer. Alternatively, the relative positioning of the point can be purely specified in 2D, adjusting only the range sizes of connecting distances in the 3D query tables.

Using the *Distance* or *Interaction* mode, any pair of atoms, aromatic ring centers, secondary structure points, or selected points can be left mouse-clicked to connect them by a distance or interaction. The missing points, for example, when directly specifying a distance constraint between two atoms, are automatically added. Non-hypothetical intermolecular interactions can be defined by directly selecting them in the 3D binding site or 2D diagram with a left mouse click. The corresponding points are also automatically added if missing. With the *Angle* mode, angle constraints represented by colored arcs can be set between any pair of connected distances, interactions, aromatic ring normals, and secondary structure point directions by left mouse click.

A checkbox allows to automatically load all template binding site-derived chemical and spatial properties of a 3D query object into the corresponding 3D query table row. Otherwise, only the primary properties are automatically set, including the element, interaction type and molecule type for atoms, the interaction type and molecule type for aromatic ring centers, and the element, molecule type, residue location (backbone or side chain), and secondary structure type for secondary structure points. All other properties are set to generic default values. A 3D query object can be removed from the 2D viewer, 3D viewer, and respective 3D query table by activating the corresponding selection mode and clicking on it again either in one of the two scenes or on the cross icon in its respective 3D query table row. If a 3D query object is removed, for example, a point, all dependent 3D query objects, in this case, connected distances, interactions, and angles between those, are also deleted.

Various PDB filters can be added to the PDB filter table (Figure 3.5), restricting the search by textual, numerical, and chemical properties of the ligand, binding site, protein, PDB entry, and results. This search functionality can also be used without a 3D query, for example, to limit the 3D search space in advance. Figure 3.5 highlights one exemplary row of the PDB filter table.

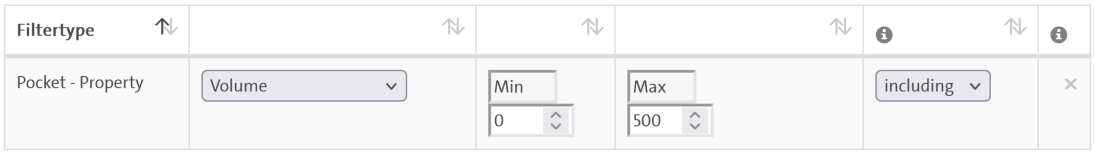



**Figure 3.5:** Example table row of the PDB filter table limiting the search space to all binding sites with a volume between 0 $Å^3$ and 500 $Å^3$

Finally, a list of PDB identifiers can be provided in the PDB subselection list (Figure 3.6) to avoid searching the complete GeoMine database, but only specific entries of interest. A GeoMine query can be downloaded and re-uploaded in JSON file format. Furthermore, a GeoMine query file can be exported in XML format. The XML format is usable by the REST API of GeoMine. Multiple XML queries could be automatically generated and used in a GeoMine search, for example, by subsequently increasing the sizes of the 3D query's distances before each request to the REST API. Clicking the *Search* button starts the GeoMine query search. If the query is invalid or unsuitable, for example, due to incorrect PDB filter values or a highly unspecific 3D query, the search is stopped and the user gets a corresponding feedback.

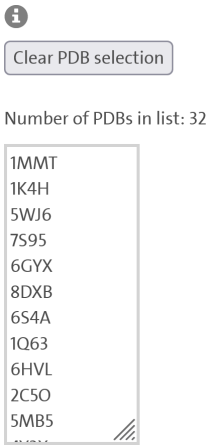



**Figure 3.6:** PDB subselection list limiting the search space to 32 PDB entries

### 3.6.2 Results Analysis and Refinement Workflow

After the search is performed, GeoMine's tool-specific graphical user interface on the right shows an additional *Results* tab (Figure 3.7) to inspect the database matches. The corresponding section provides the total number of matched PDB entries, matched binding sites of those PDB entries, and the total number of 3D query hits in those binding sites if the GeoMine search comprehends a 3D query (Figure 3.7a). The results

are ranked by RMSD. The RMSD is calculated between the 3D query's points, like atoms and aromatic ring centers, and the corresponding matching chemical features in the detected binding sites. The 150 hits with the lowest RMSD are listed in a searchable and sortable table (Figure 3.7b). The 3D visualizable result set is limited to 150 to improve the search's runtime and because a larger result set is difficult to thoroughly inspect in 3D. A larger set might contain results that are too unspecific regarding the associated application scenarios. The size of the result set and the query's specificity can be iteratively adapted through the refinement functionality described below. Each table entry represents a 3D query hit and is annotated by the corresponding RMSD value, the title, identifier, and protein class of the matched PDB, the name of the detected binding site based on the concatenated names of all contained ligands, and the binding site-specific unique identifier of the match. The table allows to visualize results in the 3D viewer individually or as superimpositions onto the 3D query together with the template binding site (Figure 3.7c). If only PDB filters are specified, the first 150 results are listed in the table. With such a GeoMine query, the table entries are not annotated by RMSD values and resulting binding sites are overlayed in the 3D viewer on their centers. Various toggleable options can be set for individual 3D-visualized results (Figure 3.7d), including:

- visualization of a result via the complete corresponding PDB structure, the matching binding site, the relevant residue parts, the relevant ligand parts, or the relevant residue and ligand parts. Chemical structures, structural parts like functional groups, and intermolecular interactions that were directly matched by the 3D query are considered relevant to limit and clarify the result's visualization
- hiding of all residues or ligands or their visualization in the *Licorice*, *Line*, *Ball+Stick*, or *Spacefill* representation
- highlighting of the 3D query match and the result's visualized chemical structure by a unique color
- visualization of intermolecular interactions

The refinement functionality (Figure 3.7e) permits to continue to search in a result set after modifying the corresponding GeoMine query. The PDB identifiers of the result set are loaded into the PDB subselection list and the corresponding GeoMine query is loaded into the 2D viewer, 3D viewer, 3D query tables, and PDB filter table as a new starting point for the query's further refinement. In addition, with the results history functionality (Figure 3.7f), any set of results and corresponding queries is reloaded to

continue the refinement process. The following files can be downloaded (Figure 3.7g) on the results page:

- a statistics file
- the transformed and superimposed binding sites of the 150 matches with the lowest RMSD or the 150 first matches in PDB files format
- results table data in CSV and JSON file format

The statistics file is generated not only for the visualizable results but for the complete search result set. The statistics file contains the following information about the results:

- list of PDB entries that match the query
- unique SMILES and names of all ligands in binding sites that match the query
- list of RMSD values for all 3D query matches
- list of query points annotated each with the number of respective detected hits and the percentages of atom elements, molecule types, amino acid types, secondary structure types, and atom types
- list of point-to-point constraints in the 3D query (distances and intermolecular interactions) annotated each with the number of respective detected hits, a list of matching distances in Å, and the distribution of matching intermolecular interaction types expressed in percentages
- list of 3D query angles annotated each with the number of respective detected hits and a list of matching angles in degree

Results are cached for 30 days and can be reaccessed through a link provided by the results page.

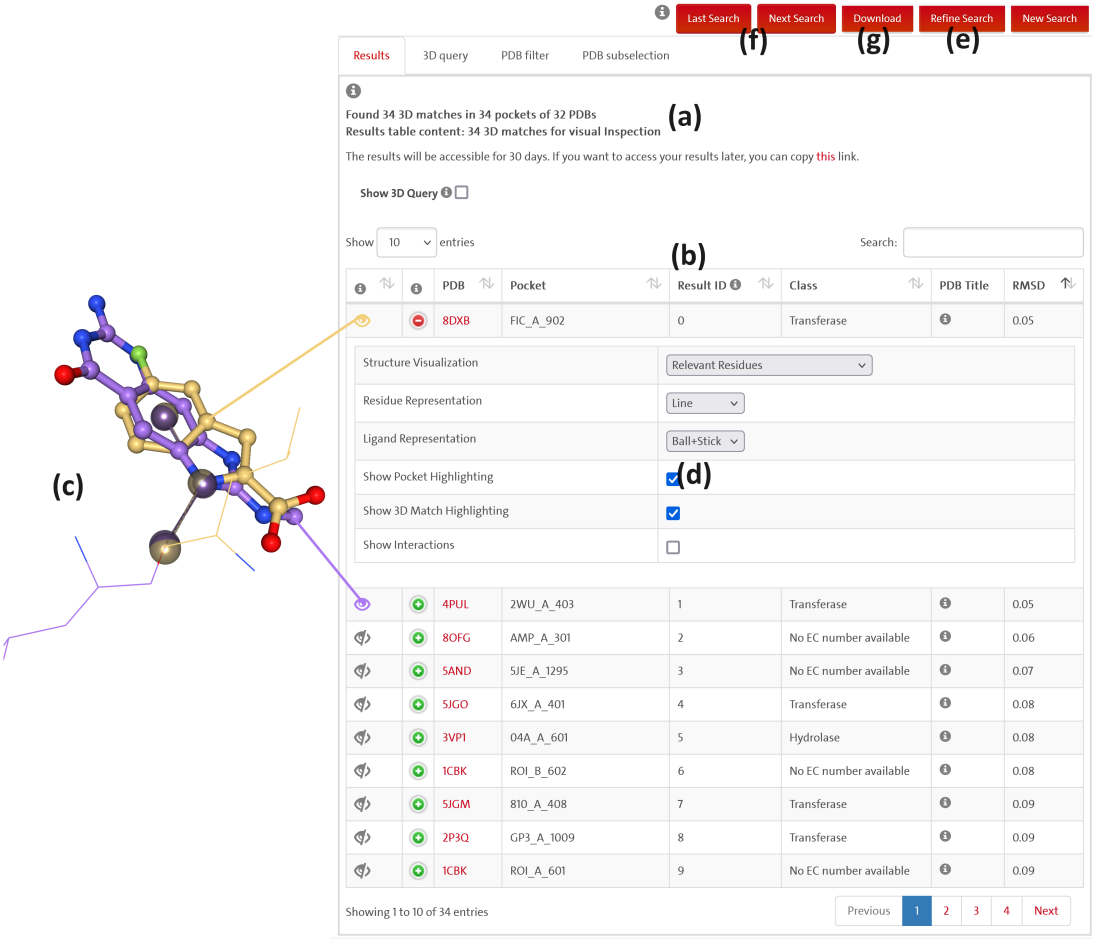



**Figure 3.7:** Results of an example GeoMine search designed in a cyclin-dependent kinase structure bound to the inhibitor 4-[[6-(cyclohexylmethoxy)-7H-purin-2-yl]amino]benzenesulfonamide [57] (PDB identifier: 1H1S; Proteins*Plus* ligand identifier: 4SP_A_1298). **a** Number of search results. **b** Results table with at most 150 hits. **c** 3D viewer superimposition of two query matches in the table. **d** Visualization options for the first match in the table. **e** Button for the results refinement functionality. **f** Buttons for the search history. **g** Button for the download of results-related files

## 3.7 Application

Examples of GeoMine's application can be found in the corresponding scientific publications [D2–D5]. In particular, application examples underscoring the benefits of new GeoMine features developed in this doctoral project, like AlphaFold structures as 3D query templates or the geometric query design-supporting 2D interface, are available in [D3] and [D5].

The following application example (Figure 3.8) taken and adapted from [D5] demonstrates GeoMine's applicability to find off-targets, a scientifically relevant research question that helps, for example, to explore the potential side effects of drugs. The query is designed in an S-adenosyl methionine (SAM)-bound site of the enzyme alpha N-terminal protein methyltransferase 1 of *Leishmania major* [85] (PDB identifier: 1XTP, Proteins*Plus* ligand identifier: SAI_A_401) and searched for in human protein structures in the PDB that are bound to SAM-related compounds. The aim is to predict potential off-targets for SAM-competitive inhibitors, i.e., compounds that competitively bind to the SAM binding site and thereby inhibit the enzyme's function. Solvent-exposed residue atoms and aromatic ring centers that interact with SAM are selected and connected by distance ranges with a tolerance of 1 Å. The selected search points were specified by their pharmacophoric properties (acceptor, donor, hydrophobic, or aromatic center). The interacting atoms of Thr167, Arg106A, and Gln165 were omitted to focus on the adenosyl moiety and to exclude weak hydrogen bond acceptors. The 45 results include several hits that indicate selectivity-related challenges for SAM-competitive inhibitors.

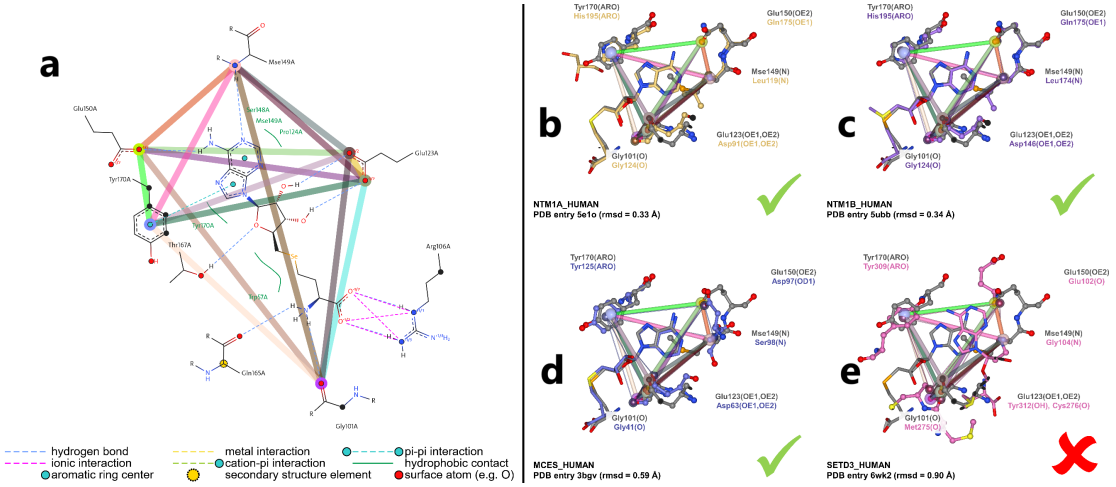



**Figure 3.8:** Example 3D query to search off-targets of SAM-competitive inhibitors based on a SAM-bound binding site of the enzyme alpha N-terminal protein methyltransferase 1 of *L. major* [85] (PDB identifier: 1XTP, Proteins*Plus* ligand identifier: SAI_A_401). Taken and adapted from [D5]. **a** A 2D editor-designed query, which models SAM-interacting atoms and aromatic ring centers of binding site residues. **b-e** Template binding site-aligned hits and the RMSD between the query and matched points after superposition. The alignments of the first three results (PDB identifier: 5E1O, 5UBB, 3BGV) are convincing, indicating potential off-targets. In contrast, the alignment with actin-histidine N-methyltransferase (PDB identifier: 6WK2) is suboptimal and presents structural clashes, indicating a different SAM binding mode

Besides GeoMine's applicability, the application example also highlights several advantageous frontend features that enhance GeoMine's usability. The 2D query editor (Figure 3.8a) enables a fast and intuitive selection to design complex queries, such as queries with many points and distances like in the described application example. In contrast to the binding site's 3D representation, the 2D pose diagram clearly focuses on the ligand binding mode, selectively highlighting information such as solvent exposure, interacting atoms, and intermolecular interactions. The RMSD-ranked result list provides easy prioritization of results of relevance. Figure 3.8 shows four matches with low RMSD values for further visual validation. The alignments of the 3D query template site and results regarding the query's points and the matching points of the results enable easy comparison of structural aspects. Additional fine-tuned visualization options presented in Figure 3.8 improve the visual analysis by highlighting relevant aspects and enabling a straightforward comparison:

- the individual coloring of a result's and 3D query template site's chemical structures helps to compare them structurally
- individual representation styles for residues (*Ball+Stick*) and ligands (*Licorice*) support their visual discrimination
- the visualization of the 3D query consisting of points and distances and the visualization of the matching points and distances in the results helps to understand how the result fulfilled the query's chemical and spatial constraints. In the application example, the 3D query match is colored like the result's chemical structures to emphasize their correspondence. Furthermore, the semi-transparent color of a 3D query match does not affect the visibility of chemical structures covered by it in the 3D scene, for example, spheres around atoms highlighting matches of points
- the visualization of only those residues of a result that were matched by the 3D query enables focusing on ligand binding mode-relevant information

Further application examples of the previously mentioned scientific articles include:

- off-target searches for acetazolamide and celecoxib
- the intermolecular interaction geometry analysis for halogen-aromatic interactions
- selective protein kinase inhibitor design
- binding site function prediction and off-target analyses for methyltransferases from parasites of the genus Leishmania

In some of the GeoMine-citing non-technical articles, GeoMine was applied in the context of interaction profiling of the eucalyptus essential oils targeting angiotensin-converting enzyme 2 and lipoxygenase to explore and verify its therapeutic effect on upper respiratory tract diseases [86] and of the chromobacterium violaceum-produced pigment violacein targeting low molecular weight protein tyrosine phosphatase to explore violaceum's effect on the energetic metabolism to inhibit tumor cells [87]. Additional application examples for 3D structural searches of some other related tools can be found in the corresponding scientific articles [74, 79, 88, 89], further highlighting the general relevance of similarity searches in 3D structural data.

## 3.8 Outlook

Some additional usability-related and applicability-related enhancements of GeoMine regarding query specification, results presentation, and search space include:

- AlphaFold database entries that are already utilizable for a template-based 3D query design could enhance GeoMine's search space currently encompassing the PDB database. For example, the search space could provide a significantly larger number of potential search results of interest to identify new targets in drug repurposing or side effect analyses. The database's and search algorithm's performance regarding runtime and memory requirements must be further improved to process the significantly increased amount of searchable data

- since important chemical information of results might be communicated more conveniently in 2D than in 3D, the 2D viewer utilized for query design could also be usable to display 2D diagrams of results representing ligand-bound binding sites. Multiple results could be simultaneously visualized in 2D by displaying and overlaying the 3D query-matching parts of the respective 2D diagrams highlighting chemical similarities and dissimilarities. Since the current 2D diagram implementation shows only structures interacting with the ligand, structures that do not interact but are included by the query must be visualized in addition

- a 2D visualization of ligand-unbound binding sites by PoseEdit could be implemented and integrated into GeoMine to facilitate the generation of 3D queries and the inspection of results. However, due to the unlimited maximal number of residues that a DoGSite3-predicted empty binding site can include, a comprehensible 2D visualization must reasonably simplify the represented chemical information content to be advantageous compared to the 3D binding site. For

example, all residues could be positioned in a circle and initially displayed by the simple circle representation type. Subsequently, the skeletal representation type might be set for specific relevant residues. Furthermore, visualizing only a preselection of potentially relevant residues can also lead to a helpful 2D simplification. Corresponding selection criteria might include the level of solvent exposure of side chain atoms or the sequence-based or spatial conservation levels of residues

- a GeoMine query can only include chemical features that are present at specific positions and that have specific chemical property values. A query object to define a space not occupied by chemical features like atoms and the option to exclude specific chemical properties for points could extend the application scenarios and quality of results. For example, structural clashes between results and the query template binding site could be reduced. A query object to define an unoccupied space could be graphically represented by a sphere with a user-specified radius and corresponding tolerance value

- without prior knowledge on which chemical features are important to select in a template binding site, the combinatorial space of generatable queries can be too large to be manually processed through multiple results refinement steps. The GeoMine search only returns complete results, i.e., all chemical features and their chemical properties of the query were matched. Query flexibility is integrable by setting chemical properties to "Any" and by specifying the relative spatial arrangement of chemical features through varying sizes of distance and angle ranges. Furthermore, query points can be specified by a residue class, e.g., polar or hydrophobic, rather than its precise residue type and by multiple interaction types, e.g., donor and acceptor. For example, in a search for ligands that interact with a ligand-unbound template binding site, the binding site might present a large variety of solvent-exposed donor and acceptor residue atoms. All atoms could be potential interaction partners for a ligand and are, therefore, query points of interest. Consequently, users have to manually include and exclude specific atoms and corresponding properties through numerous and time-consuming results refinement steps to cover every combination of atoms. Two solutions to this challenge are addressable. Defining optional chemical features like atoms or intermolecular interactions and optional values for corresponding chemical properties. The solutions allow to specify, for example, a query that matches the presence or absence of a residue atom, which belongs either to alanine or leucine. Both extensions of the query features can make a results refinement process obsolete since the combinatorial space of queries is automatically processed by only one search. Valid

results can be priorly defined in the graphical user interface by a minimum number of matching chemical features and subsequently ranked in the results table by the level of completeness. The extended search is of increased complexity due to the calculation of incomplete matches, therefore requiring additional optimizations of GeoMine's performance

- the on-the-fly loading of results can address searches with elevated runtimes and consequently improve the tool's usability. However, the algorithm returns all results at once after processing the complete query. Binding site clusters can be precalculated to independently search in those and return only cluster representatives as results on the fly. A careful evaluation of clustering results is important to verify if valuable results get lost by the initial focus on cluster representatives

- a prediction of a search's runtime would improve the user's acceptance of long waiting times, which can also occur with highly specific queries. For example, a query consisting of a large number of residue atoms connected by distance ranges and specified by generic chemical property values would only produce a low number of results but still cause a high runtime due to the low chemical specificity of the atoms. To predict a search's runtime, the associated impact of query object combinations must be specified by comprehensive statistical analyses of the database occurrences and extraction times of corresponding chemical features and their chemical properties under various geometric constraints

- a PDB filter for user-specified protein sequences, for example, implemented with the tool SIENA [90], could enhance GeoMine's search capabilities by providing a sequence-based preselection of interesting binding sites

In addition to general improvements, a new version of the Proteins*Plus* web server is currently in development. Due to its complex graphical user interface, GeoMine will not be included in the Proteins*Plus* web server anymore. Therefore, the tool is made separately available via its containerized web server version on `https://geomine.proteins.plus` [91].

# Chapter 4

# Summary and Conclusion

This doctoral project resulted in two new tools called PoseEdit and GeoMine and a respectively further developed version of the Proteins*Plus* web server hosting the two tools. In addition, a software container-based standalone web server was developed for each tool. Both standalone web server versions are also publicly available via webpages. The new developments resulted in several corresponding scientific publications [D1–D7]. Both tools are partly based on previous existing technical groundwork, which was reimplemented, refactored, and extensively further developed. The general goal of the tools is to advance the respective state of the art in exploring chemical and spatial data of protein binding sites in the context of drug development. In addition to the applicability of the two tools, another major aspect explored by this doctoral project was their usability. Well-designed graphical user interfaces significantly simplify the use of such highly complex tools for scientists with varying levels of experience regarding computational analyses. A corresponding focus is on facilitating complex structural data exploration by developing different and complementing graphical representations.

- PoseEdit offers a new way to display ligand binding modes of ligand-bound binding sites through a comprehensive visualization concept. Gained information helps scientists to communicate and explore reasons for a ligand's affinity, selectivity, and consequential biological activity, which is a common task in drug development

- GeoMine enables 3D query-based searches in ligand-bound and ligand-unbound binding sites of large binding site collections via its comprehensive and intuitive query design and results inspection interfaces. Various drug development-related applications for a 3D structural similarity search are thinkable, including protein function elucidation, drug repurposing, and drug side effect analyses

PoseEdit stands out relative to other state-of-the-art tools in several aspects. The high number of tool features to explore and modify diagrams down to the atomistic level distinguishes PoseEdit from other state-of-the-art tools, which provide significantly more static 2D representations. PoseEdit's capabilities support users to overcome limitations of the layout algorithm and the tool's chemical and graphical presets that could otherwise result in unsolvable diagram issues ranging from objective lacks to highly subjective deficiencies. Automatically generated diagrams with such shortcomings are rejected by scientists to communicate a ligand's binding mode. With PoseEdit, all diagrams can still be highly usable thanks to various editing options. Additional preset improvements to the PoseEdit diagram's represented information, like covalent bonds between ligands and residues and esthetic graphical styles, such as color gradients for covalent bonds, round off its revised content.

PoseEdit seamlessly integrates classic textual ligand binding mode data, chemical data-focused 2D diagrams, and algorithmically predicted complete 3D binding sites into one unique synchronized visualization concept. This approach offers a synergistic way to explore and verify chemical and spatial key features of binding sites based on the strengths and limitations of the three representation types. For example, it can help to generate ideas about how to modify the restricted chemical content of a 2D diagram by examining the corresponding complete 3D binding site.

Unlike other state-of-the-art tools, PoseEdit is accessible to a large user community through its free web interface. In particular, academic research will significantly benefit from its easy accessibility, which, as of October 25, 2024, is highlighted by 35 references [63] to the PoseEdit publication. In contrast, the distributable standalone PoseEdit container also makes the tool applicable to confidential in-house data, providing a high level of control over data privacy while keeping a flexible and dependency-free usage via a web browser. Furthermore, the reasonably compressed but highly functional graphical user interface and the required minimum usage of mouse and keyboard controls allow a quick overview, instant comprehension, and easy application of PoseEdit's features, rendering the tool a user-friendly solution compared to other state-of-the-art tools.

The code base of the interactive frontend is publicly available on GitHub and the corresponding JSON input is accessible via PoseEdit's REST API, facilitating the integration of PoseEdit's functionality in other scientific software projects, which is exemplarily demonstrated by GeoMine, the second tool of this doctoral project.

GeoMine is set apart from other state-of-the-art tools by multiple aspects. First, GeoMine's usability regarding query formulation and results representation is relatively enhanced through various user-centric design concepts. Second, the possibilities of

search space definition and query definition options are the most comprehensive ones of all available tools. GeoMine is the only tool that enables a comprehensive 3D query-building process, smoothly integrating text-based tabular, 2D, and 3D input types, enabling users to profit from their respective advantages. Furthermore, all three input types are synergistically synchronized and comprehensively providable through input structure-based query templates as well as from scratch. In particular, generating queries with the easy-to-use 2D editor in protein-ligand interactions-focused PoseEdit diagrams is a unique feature of GeoMine, which further supports scientists in obtaining and realizing query ideas. In contrast, other related state-of-the-art tools only allow drawing of 3D queries in 2D space without providing 2D templates.

Unlike other state-of-the-art tools, GeoMine's template binding sites are not ligand radius-based but computationally predicted, enabling a more comprehensive determination of a binding site's shape and structural boundaries. Furthermore, this approach provides ligand-bound but also ligand-unbound binding sites as query templates, enabling scientists to easily address query design for additional relevant research questions, such as drug repurposing. If the template binding sites of a specific input structure are insufficient, scientists can still go beyond the spatial limitation of query templates by verifying and performing query design anywhere in the complete 3D input structure. Finally, supporting not only PDB database entries and custom structures but also AlphaFold database entries as input structures provides scientists with a wide range of usability-enhancing starting points for generating queries of interest. GeoMine's result browser enables a comprehensive analyses of the retrieved query matches through the visual examination of ranked tabular data and various overlaid 3D representations, as well as further analyses with other tools through exportable statistical data. The overall usability of the GeoMine is rounded off by the import and export functionality for queries, the history and refinement functionality, and the RMSD-based results ranking. These features help scientists to accustom themselves to GeoMine's functionality and to explore different queries and results.

Like PoseEdit and unlike the state-of-the-art tools, GeoMine's free web service and its standalone container version make the tool highly accessible to a large user community with different requirements, including academia and industry. For example, the GeoMine container was integrated for in-house application by a large pharmaceutical company. Furthermore, GeoMine's functionality can be directly integrated into scientific workflows via its REST API and the XML query representation.

The calculated query templates of input structures display the same information content calculated for PDB entries stored in the searchable GeoMine database. In

comparison to other state-of-the-art tools, a relatively large variety of comprehensively visualized chemical features and corresponding properties are specifiable via query templates for 3D query formulation. GeoMine-specific chemical features of scientific interest include, for example, solvent-exposed atoms and secondary structure points. A distinguishing aspect of GeoMine is the large number of text-based filters, which can be set in addition to or without a 3D query to limit the search space and results based on various chemical and spatial properties.

In summary, PoseEdit and GeoMine provide new, unique approaches to a comprehensive and user-friendly visualization and searching of 3D structural data, enabling the analysis of intermolecular interfaces and the investigation of biological function-related research questions.

# Bibliography


[1] Human Proteome Organization (HUPO). *HPP progress to date*. URL: https://hupo.org/hpp-progress-to-date (visited on 08/24/2024).

[2] J. Berg, J. Tymoczko, and L. Stryer. *Biochemistry*. W. H. Freeman, 2010. ISBN: 9781429229364.

[3] A. J. M. Ribeiro, G. L. Holliday, N. Furnham, J. D. Tyzack, K. Ferris, and J. M. Thornton. "Mechanism and Catalytic Site Atlas (M-CSA): a database of enzyme reaction mechanisms and active sites". *Nucleic Acids Research* 46 (2017), pp. 618–623. DOI: https://doi.org/10.1093/nar/gkx1012.

[4] N. Hulo, A. Bairoch, V. Bulliard, L. Cerutti, E. De Castro, P. S. Langendijk-Genevaux, M. Pagni, and C. J. A. Sigrist. "The PROSITE database". *Nucleic Acids Research* 34 (2006), pp. 227–230. DOI: https://doi.org/10.1093/nar/gkj063.

[5] S. J. Dixon and B. R. Stockwell. "Identifying druggable disease-modifying gene products". *Current Opinion in Chemical Biology* 13 (2009), pp. 549–555. DOI: https://doi.org/10.1016/j.cbpa.2009.08.003.

[6] P. Imming, C. Sinning, and A. Meyer. "Drugs, their targets and the nature and number of drug targets". *Nature Reviews Drug Discovery* 5 (2006), pp. 821–834. DOI: https://doi.org/10.1038/nrd2132.

[7] A. C. Anderson. "The Process of Structure-Based Drug Design". *Chemistry & Biology* 10 (2003), pp. 787–797. DOI: https://doi.org/10.1016/j.chembiol.2003.09.002.

[8] B. Munos. "Lessons from 60 years of pharmaceutical innovation". *Nature Reviews Drug Discovery* 8 (2009), pp. 959–968. DOI: https://doi.org/10.1038/nrd2961.

[9] G. Schneider and U. Fechner. "Computer-based de novo design of drug-like molecules". *Nature Reviews Drug Discovery* 4 (2005), pp. 649–663. DOI: https://doi.org/10.1038/nrd1799.

[10] A. V. Sadybekov and V. Katritch. "Computational approaches streamlining drug discovery". *Nature* 616 (2023), pp. 673–685. DOI: https://doi.org/10.1038/s41586-023-05905-z.

[11] M. J. Waring, J. Arrowsmith, A. R. Leach, P. D. Leeson, S. Mandrell, R. M. Owen, G. Pairaudeau, W. D. Pennie, S. D. Pickett, J. Wang, O. Wallace, and A. Weir. "An analysis of the attrition of drug candidates from four major pharmaceutical companies". *Nature Reviews Drug Discovery* 14 (2015), pp. 475–486. DOI: https://doi.org/10.1038/nrd4609.

[12] H. Yu and A. Adedoyin. "ADME-Tox in drug discovery: integration of experimental and computational technologies". *Drug Discovery Today* 8 (2003), pp. 852–861. DOI: https://doi.org/10.1016/S1359-6446(03)02828-9.

[13] J. Bajorath. "Integration of virtual and high-throughput screening." *Nature Reviews Drug Discovery* 1 (2002), pp. 882–894. DOI: https://doi.org/10.1038/nrd941.

[14] J. Singh, C. E. Chuaqui, P. Boriack-Sjodin, W.-C. Lee, T. Pontz, M. J. Corbley, H.-K. Cheung, R. M. Arduini, J. N. Mead, M. N. Newman, J. L. Papadatos, S. Bowes, S. Josiah, and L. E. Ling. "Successful shape-Based virtual screening: The discovery of a potent inhibitor of the type I TGF$\beta$ receptor kinase (T$\beta$RI)". *Bioorganic & Medicinal Chemistry Letters* 13 (2003), pp. 4355–4359. DOI: https://doi.org/10.1016/j.bmcl.2003.09.028.

[15] G. Klebe. "Recent developments in structure-based drug design". *Journal of Molecular Medicine* 78 (2000), pp. 269–281. DOI: https://doi.org/10.1007/s001090000084.

[16] G. A. C. Rajamani R. "Ranking poses in structure-based lead discovery and optimization: current trends in scoring function development". *Current Opinion in Drug Discovery and Development* 10 (2007), pp. 308–315.

[17] F. W. de Azevedo Jr. and R. Dias. "Computational Methods for Calculation of Ligand-Binding Affinity". *Current Drug Targets* 9 (2008), pp. 1031–1039. DOI: https://doi.org/10.2174/138945008786949405.

[18] J. R. Deschamps. "X-ray crystallography of chemical compounds". *Life Sciences* 86.15 (2010), pp. 585–589. DOI: https://doi.org/10.1016/j.lfs.2009.02.028.

[19] Y. Hu, K. Cheng, L. He, X. Zhang, B. Jiang, L. Jiang, C. Li, G. Wang, Y. Yang, and M. Liu. "NMR-Based Methods for Protein Analysis". *Analytical Chemistry* 93 (2021), pp. 1866–1879. DOI: https://doi.org/10.1021/acs.analchem.0c03830.

[20] E. Callaway. "The revolution will not be crystallized: a new method sweeps through structural biology". *Nature* 525 (2015), pp. 172–174. DOI: https://doi.org/10.1038/525172a.

[21] E. Y. D. Chua, J. H. Mendez, M. Rapp, S. L. Ilca, Y. Z. Tan, K. Maruthi, H. Kuang, C. M. Zimanyi, A. Cheng, E. T. Eng, A. J. Noble, C. S. Potter, and B. Carragher. "Better, Faster, Cheaper: Recent Advances in Cryo-Electron Microscopy". *Annual Review of Biochemistry* 91 (2022), pp. 1–32. DOI: https://doi.org/10.1146/annurev-biochem-032620-110705.

[22] RCSB PDB. *Number of released PDB structures per year*. URL: https://www.rcsb.org/stats/all-released-structures (visited on 08/24/2024).

[23] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis. "Highly accurate protein structure prediction with AlphaFold". *Nature* 596 (2021), pp. 583–589. DOI: https://doi.org/10.1038/s41586-021-03819-2.

[24] C. R. Groom, I. J. Bruno, M. P. Lightfoot, and S. C. Ward. "The Cambridge Structural Database". *Acta Crystallographica Section B* 72 (2016), pp. 171–179. DOI: https://doi.org/10.1107/S2052520616003954.

[25] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. "The Protein Data Bank". *Nucleic Acids Research* 28 (2000), pp. 235–242. DOI: https://doi.org/10.1093/nar/28.1.235.

[26] J. Desaphy, G. Bret, D. Rognan, and E. Kellenberger. "sc-PDB: a 3D-database of ligandable binding sites-10 years on". *Nucleic Acids Research* 43 (2014), pp. 399–404. DOI: https://doi.org/10.1093/nar/gku928.

[27] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs". *Nucleic Acids Research* 25 (1997), pp. 3389–3402. DOI: https://doi.org/10.1093/nar/25.17.3389.

[28] Protein Data Bank. *File download services of the PDB*. URL: https://www.rcsb.org/docs/programmatic-access/file-download-services (visited on 08/24/2024).

[29] Protein Data Bank. *Beginner's Guide to PDB Structures and the PDBx/mmCIF Format*. URL: https://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/beginner%E2%80%99s-guide-to-pdbx-mmcif (visited on 08/24/2024).

[30] H. Berman, K. Henrick, and H. Nakamura. "Announcing the worldwide Protein Data Bank". *Nature Structural & Molecular Biology* 10 (2003). DOI: https://doi.org/10.1038/nsb1203-980.

[31] F. Flachsenberg, A. Meyder, K. Sommer, P. Penner, and M. Rarey. "A Consistent Scheme for Gradient-Based Optimization of Protein-Ligand Poses". *Journal of Chemical Information and Modeling* 60 (2020), pp. 6502–6522. DOI: https://doi.org/10.1021/acs.jcim.0c01095.

[32] J. Graef, C. Ehrt, and M. Rarey. "Binding Site Detection Remastered: Enabling Fast, Robust, and Reliable Binding Site Detection and Descriptor Calculation with DoGSite3". *Journal of Chemical Information and Modeling* 63 (2023), pp. 3128–3137. DOI: https://doi.org/10.1021/acs.jcim.3c00336.

[33] S. Bietz, S. Urbaczek, B. Schulz, and M. Rarey. "Protoss: a holistic approach to predict tautomers and protonation states in protein-ligand complexes". *Journal of Cheminformatics* 6 (2014). DOI: https://doi.org/10.1186/1758-2946-6-12.

[34] K. Stierand and M. Rarey. "Drawing the PDB: Protein-Ligand Complexes in Two Dimensions". *ACS Medicinal Chemistry Letters* 1 (2010), pp. 540–545. DOI: https://doi.org/10.1021/ml100164p.

[35] *ProteinsPlus web service*. URL: https://proteins.plus (visited on 08/24/2024).

[36] R. Fährrolfes, S. Bietz, F. Flachsenberg, A. Meyder, E. Nittinger, T. Otto, A. Volkamer, and M. Rarey. "ProteinsPlus: a web portal for structure analysis of macromolecules". *Nucleic Acids Research* 45 (2017), pp. 337–343. DOI: https://doi.org/10.1093/nar/gkx333.

[37] R. J. Read, P. D. Adams, I. Arendall W. Bryan, A. T. Brunger, P. Emsley, R. P. Joosten, G. J. Kleywegt, E. B. Krissinel, T. Lütteke, Z. Otwinowski, A. Perrakis, J. S. Richardson, W. H. Sheffler, J. L. Smith, I. J. Tickle, G. Vriend, and P. H. Zwart. "A New Generation of Crystallographic Validation Tools for the Protein Data Bank". *Structure* 19 (2011), pp. 1395–1412. DOI: https://doi.org/10.1016/j.str.2011.08.006.

[38] G. T. Montelione, M. Nilges, A. Bax, P. Güntert, T. Herrmann, J. S. Richardson, C. D. Schwieters, W. F. Vranken, G. W. Vuister, D. S. Wishart, H. M. Berman, G. J. Kleywegt, and J. L. Markley. "Recommendations of the wwPDB NMR Validation Task Force". *Structure* 21 (2013), pp. 1563–1570. DOI: https://doi.org/10.1016/j.str.2013.07.021.

[39] R. Henderson, A. Sali, M. L. Baker, B. Carragher, B. Devkota, K. H. Downing, E. H. Egelman, Z. Feng, J. Frank, N. Grigorieff, W. Jiang, S. J. Ludtke, O. Medalia, P. A. Penczek, P. B. Rosenthal, M. G. Rossmann, M. F. Schmid, G. F. Schröder, A. C. Steven, D. L. Stokes, J. D. Westbrook, W. Wriggers, H. Yang, J. Young, H. M. Berman, W. Chiu, G. J. Kleywegt, and C. L. Lawson. "Outcome of the First Electron Microscopy Validation Task Force Meeting". *Structure* 20 (2012), pp. 205–214. DOI: https://doi.org/10.1016/j.str.2011.12.014.

[40] Protein Data Bank. *Missing coordinates in the PDB*. URL: https://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/missing-coordinates (visited on 08/24/2024).

[41] S. I. O'Donoghue, D. S. Goodsell, Frangakis, F. A. S. Jossinet, R. A. Laskowski, M. Nilges, H. R. Saibil, A. Schafferhans, R. C. Wade, E. Westhof, and A. J. Olson. "Visualization of macromolecular structures". *Nature methods* 7 (2010), pp. 42–55. DOI: https://doi.org/10.1038/nmeth.1427.

[42] K. Stierand and M. Rarey. "From Modeling to Medicinal Chemistry: Automatic Generation of Two-Dimensional Complex Diagrams". *ChemMedChem* 2 (2007), pp. 853–860. DOI: https://doi.org/10.1002/cmdc.200700010.

[43] K. Stierand, P. C. Maaß, and M. Rarey. "Molecular complexes at a glance: automated generation of two-dimensional complex diagrams". *Bioinformatics* 22 (2006), pp. 1710–1716. DOI: https://doi.org/10.1093/bioinformatics/btl150.

[44] C. Ségolène. "LeView: automatic and interactive generation of 2D diagrams for biomacromolecule/ligand interactions". *Journal of Cheminformatics* 5 (2013). DOI: https://doi.org/10.1186/1758-2946-5-40.

[45] R. A. Laskowski and M. B. Swindells. "LigPlot+: Multiple Ligand-Protein Interaction Diagrams for Drug Discovery". *Journal of Chemical Information and Modeling* 51 (2011), pp. 2778–2786. DOI: https://doi.org/10.1021/ci200227u.

[46] A. C. Wallace, R. A. Laskowski, and J. M. Thornton. "LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions". *Protein Engineering, Design and Selection* 8 (1995), pp. 127–134. DOI: https://doi.org/10.1093/protein/8.2.127.

[47] A. M. Clark and P. Labute. "2D Depiction of Protein-Ligand Complexes". *Journal of Chemical Information and Modeling* 47 (2007), pp. 1933–1944. DOI: https://doi.org/10.1021/ci7001473.

[48] Google Scholar. *LIGPLOT citations*. URL: https://scholar.google.co.uk/citations?view_op=view_citation&hl=en&user=BzD6L54AAAAJ&citation_for_view=BzD6L54AAAAJ:d1gkVwhDpl0C (visited on 08/24/2024).

[49] Google Scholar. *LigPlot+ citations*. URL: https://scholar.google.co.uk/citations?view_op=view_citation&hl=en&user=BzD6L54AAAAJ&citation_for_view=BzD6L54AAAAJ:fPk4N6BV_jEC (visited on 08/24/2024).

[50] Google Scholar. *PoseView citations*. URL: https://scholar.google.com/citations?view_op=view_citation&hl=de&user=dZNhOjkAAAAJ&sortby=pubdate&citation_for_view=dZNhOjkAAAAJ:qjMakFHDy7sC (visited on 08/24/2024).

[51] Google Scholar. *PoseView citations*. URL: https://scholar.google.com/citations?view_op=view_citation&hl=de&user=dZNhOjkAAAAJ&sortby=pubdate&citation_for_view=dZNhOjkAAAAJ:d1gkVwhDpl0C (visited on 08/24/2024).

[52] Google Scholar. *PoseView citations*. URL: https://scholar.google.com/citations?view_op=view_citation&hl=de&user=dZNhOjkAAAAJ&sortby=pubdate&citation_for_view=dZNhOjkAAAAJ:zYLM7Y9cAGgC (visited on 08/24/2024).

[53] Google Scholar. *PoseView citations*. URL: https://scholar.google.com/citations?view_op=view_citation&hl=de&user=dZNhOjkAAAAJ&sortby=pubdate&citation_for_view=dZNhOjkAAAAJ:W7OEmFMy1HYC (visited on 08/24/2024).

[54] Google Scholar. *PoseView citations*. URL: https://scholar.google.com/citations?view_op=view_citation&hl=de&user=dZNhOjkAAAAJ&sortby=pubdate&citation_for_view=dZNhOjkAAAAJ:2osOgNQ5qMEC (visited on 08/24/2024).

[55] Google Scholar. *PoseView citations*. URL: https://scholar.google.com/citations?view_op=view_citation&hl=de&user=dZNhOjkAAAAJ&sortby=pubdate&citation_for_view=dZNhOjkAAAAJ:Tyk-4Ss8FVUC (visited on 08/24/2024).

[56] A. D. McNaught and A. Wilkinson. *IUPAC. Compendium of Chemical Terminology, 2nd ed. (the "Gold Book")*. Blackwell Scientific Publications, Oxford, 1997.

[57] T. G. Davies, J. Bentley, C. E. Arris, F. T. Boyle, N. J. Curtin, J. A. Endicott, A. E. Gibson, B. T. Golding, R. J. Griffin, I. R. Hardcastle, P. Jewsbury, L. N. Johnson, V. Mesguiche, D. R. Newell, M. E. M. Noble, J. A. Tucker, L. Wang, and H. J. Whitfield. "Structure-based design of a potent purine-based cyclin-dependent kinase inhibitor". *Nature Structural & Molecular Biology* 9 (2002), pp. 745–749. DOI: https://doi.org/10.1038/nsb842.

[58] P. Vianello, L. Sartori, F. Amigoni, A. Cappa, G. Fagá, R. Fattori, E. Legnaghi, G. Ciossani, A. Mattevi, G. Meroni, L. Moretti, V. Cecatiello, S. Pasqualato, A. Romussi, F. Thaler, P. Trifiró, M. Villa, O. A. Botrugno, P. Dessanti, S. Minucci, S. Vultaggio, E. Zagarrí, M. Varasi, and C. Mercurio. "Thieno[3,2-b]pyrrole-5-carboxamides as New Reversible Inhibitors of Histone Lysine Demethylase KDM1A/LSD1. Part 2: Structure-Based Drug Design and Structure-Activity Relationship". *Journal of Medicinal Chemistry* 60 (2017), pp. 1693–1715. DOI: https://doi.org/10.1021/acs.jmedchem.6b01019.

[59] BioSolveIT. *PoseView*. URL: https://www.biosolveit.de/products/ (visited on 08/24/2024).

[60] Github. *InteractionDrawer library*. URL: https://github.com/rareylab/InteractionDrawer (visited on 08/24/2024).

[61] L. Weihs. "Interactive Molecular Structure Visualization for Web Applications". *Master's thesis, Universität Hamburg* (2020).

[62] B. Krause. "Interaktive webbasierte Generierung von Protein-Ligand-Komplex-Diagrammen". *Master's thesis, Universität Hamburg* (2021).

[63] Google Scholar. *PoseEdit citations*. URL: `https://scholar.google.com/scholar?cluster=13024362095153361591&hl=de&scisbd=1&as_sdt=2005&sciodt=0,5` (visited on 08/24/2024).

[64] *PoseEdit standalone version*. URL: `https://poseedit.proteins.plus` (visited on 07/11/2024).

[65] Y. S. Choi, J. S. Yang, Y. Choi, S. H. Ryu, and S. Kim. "Evolutionary conservation in multiple faces of protein interaction". *Proteins* 77 (2009), pp. 14–25. DOI: `https://doi.org/10.1002/prot.22410`.

[66] RCSB PDB. *PDB structures by polymer entity type*. URL: `https://www.rcsb.org/stats/explore/polymer_entity_type` (visited on 08/24/2024).

[67] R. R. Thangudu, M. Tyagi, B. A. Shoemaker, S. H. Bryant, A. R. Panchenko, and T. Madej. "Knowledge-based annotation of small molecule binding sites in proteins". *BMC Bioinformatics* 11 (2010). DOI: `https://doi.org/10.1186/1471-2105-11-365`.

[68] ediss.sub.hamburg. *Dissertation - Graef, Joel*. URL: `https://ediss.sub.uni-hamburg.de/handle/ediss/10954?mode=full` (visited on 08/24/2024).

[69] RCSB PDB. *PDB structures By release date*. URL: `https://www.rcsb.org/stats/explore/release_date` (visited on 08/24/2024).

[70] R. Nussinov and H. J. Wolfson. "Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques". *Proceedings of the National Academy of Sciences of the United States of America* 88 (1991), pp. 10495–10499. DOI: `https://doi.org/10.1073/pnas.88.23.10495`.

[71] J. An, T. Nakama, Y. Kubota, and A. Sarai. "3DinSight: an integrated relational database and search tool for the structure, function and properties of biomolecules." *Bioinformatics* 14 (1998), pp. 188–195. DOI: `https://doi.org/10.1093/bioinformatics/14.2.188`.

[72] G. J. Kleywegt. "Recognition of spatial motifs in protein structures". *Journal of Molecular Biology* 285 (1999), pp. 1887–1897. DOI: https://doi.org/10.1006/jmbi.1998.2393.

[73] S. Bittrich, S. K. Burley, and A. S. Rose. "Real-time structural motif searching in proteins using an inverted index strategy". *PLOS Computational Biology* 16 (2020). DOI: 10.1371/journal.pcbi.1008502.

[74] O. Korb, B. Kuhn, J. Hert, N. Taylor, J. Cole, C. Groom, and M. Stahl. "Interactive and Versatile Navigation of Structural Databases". *Journal of Medicinal Chemistry* 59 (2016), pp. 4257–4266. DOI: https://doi.org/10.1021/acs.jmedchem.5b01756.

[75] D. Mobilio, G. Walker, N. Brooijmans, R. Nilakantan, R. A. Denny, J. DeJoannis, E. Feyfant, R. K. Kowticwar, J. Mankala, S. Palli, S. Punyamantula, M. Tatipally, R. K. John, and C. Humblet. "Protein relational database and protein family knowledge bases to facilitate structure-based design analyses". *Chemical Biology & Drug Design* 76 (2010), pp. 142–153. DOI: https://doi.org/10.1111/j.1747-0285.2010.00994.x.

[76] M. Weisel, H.-M. Bitter, F. Diederich, W. V. So, and R. Kondru. "PROLIX: Rapid Mining of Protein-Ligand Interactions in Large Crystal Structure Databases". *Journal of Chemical Information and Modeling* 52 (2012), pp. 1450–1461. DOI: https://doi.org/10.1021/ci300034x.

[77] M. Hendlich, A. Bergner, J. Günther, and G. Klebe. "Relibase: Design and Development of a Database for Comprehensive Analysis of Protein-Ligand Interactions††We dedicate this paper to Professor J. D. Dunitz." *Journal of Molecular Biology* 326 (2003), pp. 607–620. DOI: https://doi.org/10.1016/S0022-2836(02)01408-0.

[78] A. Golovin and K. Henrick. "MSDmotif: exploring protein sites and motifs". *BMC Bioinformatics* 9 (2008). DOI: https://doi.org/10.1186/1471-2105-9-312.

[79] T. Inhester, S. Bietz, M. Hilbig, R. Schmidt, and M. Rarey. "Index-Based Searching of Interaction Patterns in Large Collections of Protein-Ligand Interfaces". *Journal of Chemical Information and Modeling* 57 (2017), pp. 148–158. DOI: 10.1021/acs.jcim.6b00561.

[80] R. Angles, M. Arenas-Salinas, R. García, J. A. Reyes-Suarez, and P. E. "GSP4PDB: a web tool to visualize, search and explore protein-ligand structural patterns". *BMC Bioinformatics* 21 (2020). DOI: https://doi.org/10.1186/s12859-020-3352-x.

[81] A. J. Maurais and E. Weerapana. "Reactive-cysteine profiling for drug discovery". *Current Opinion in Chemical Biology* 50 (2019), pp. 29–36. DOI: https://doi.org/10.1016/j.cbpa.2019.02.010. URL: https://www.sciencedirect.com/science/article/pii/S1367593118301509.

[82] Daylight Chemical Information Systems, Inc. *SMARTS - A Language for Describing Molecular Patterns*. URL: https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html (visited on 08/24/2024).

[83] A. Cereto-Massagué, M. J. Ojeda, C. Valls, M. Mulero, S. Garcia-Vallvé, and G. Pujadas. "Molecular fingerprint similarity search in virtual screening". *Methods* 71 (2015), pp. 58–63. DOI: https://doi.org/10.1016/j.ymeth.2014.08.005.

[84] K. Diedrich. "Geometrische Suche in Protein-Ligand-Komplex-Datenbanken - Ein webbasierter Ansatz". *Master's thesis, Universität Hamburg* (2018).

[85] M. A. Robien and W. G. J. Hol. "Structural Analysis of Leishmania major LMAJ004091AAA, a SAM-dependent methyltransferase of the DUF858/Pfam05891 family". *Structural Genomics of Pathogenic Protozoa Consortium* (2004). DOI: https://doi.org/10.2210/pdb1XTP/pdb.

[86] E. Ak Sakall, K. Teral, A. E. Karadağ, S. N. Biltekin, M. Koşar, B. Demirci, K. Hüsnü Can Başer, and F. Demirci. "In vitro and in silico Evaluation of ACE2 and LOX Inhibitory Activity of Eucalyptus Essential Oils, 1,8-Cineole, and Citronellal". *Natural Product Communications* 17 (2022). DOI: https://doi.org/10.1177/1934578X221109409.

[87] A. V. Faria, E. M. Fonseca, P. de S. Fernandes-Oliveira, T. I. de Lima, S. P. Clerici, G. Z. Justo, L. R. Silveira, N. Durán, and C. V. Ferreira-Halder. "Violacein switches off low molecular weight tyrosine phosphatase and rewires mitochondria in colorectal cancer cells". *Bioorganic Chemistry* 127 (2022), p. 106000. DOI: https://doi.org/10.1016/j.bioorg.2022.106000.

[88] A. Bergner, J. Günther, M. Hendlich, G. Klebe, and M. Verdonk. "Use of Relibase for retrieving complex three-dimensional interaction patterns including crystallographic packing effects". *Biopolymers* 61 (2001), pp. 99–110. DOI: https://doi.org/10.1002/1097-0282(2001/2002)61:2<99::AID-BIP10075>3.0.CO;2-8.

[89] J. Günther, A. Bergner, M. Hendlich, and G. Klebe. "Utilising Structural Knowledge in Drug Design Strategies: Applications Using Relibase". *Journal of Molecular Biology* 326 (2003), pp. 621–636. DOI: https://doi.org/10.1016/S0022-2836(02)01409-2.

[90] S. Bietz and M. Rarey. "SIENA: Efficient Compilation of Selective Protein Binding Site Ensembles". *Journal of Chemical Information and Modeling* 56 (2016), pp. 248–259. DOI: https://doi.org/10.1021/acs.jcim.5b00588.

[91] *GeoMine standalone version*. URL: https://geomine.proteins.plus (visited on 11/07/2024).

[92] A. Weber, A. Casini, A. Heine, D. Kuhn, C. T. Supuran, A. Scozzafava, and G. Klebe. "Unexpected Nanomolar Inhibition of Carbonic Anhydrase by COX-2-Selective Celecoxib: New Pharmacological Opportunities Due to Related Binding Site Recognition". *Journal of Medicinal Chemistry* 47 (2004), pp. 550–557. DOI: https://doi.org/10.1021/jm030912m.

[93] J. L. Wang, D. Limburg, M. J. Graneto, J. Springer, J. R. B. Hamper, S. Liao, J. L. Pawlitz, R. G. Kurumbail, T. Maziasz, J. J. Talley, J. R. Kiefer, and J. Carter. "The novel benzopyran class of selective cyclooxygenase-2 inhibitors. Part 2: The second clinical candidate having a shorter and favorable human half-life". *Bioorganic & Medicinal Chemistry Letters* 20 (2010), pp. 7159–7163. DOI: https://doi.org/10.1016/j.bmcl.2010.07.054.

[94] C. A. Lesburg, C.-c. Huang, D. W. Christianson, and C. A. Fierke. "Histidine → Carboxamide Ligand Substitutions in the Zinc Binding Site of Carbonic Anhydrase II Alter Metal Coordination Geometry but Retain Catalytic Activity". *Biochemistry* 36 (1997), pp. 15780–15791. DOI: https://doi.org/10.1021/bi971296x.

[95] A. Meyder, E. Nittinger, G. Lange, R. Klein, and M. Rarey. "Estimating Electron Density Support for Individual Atoms and Molecular Fragments in X-ray Structures". *Journal of Chemical Information and Modeling* 57 (2017), pp. 2437–2447. DOI: https://doi.org/10.1021/acs.jcim.7b00391.

[96] J. T. Bolin, D. J. Filman, D. A. Matthews, R. C. Hamlin, and J. Kraut. "Crystal structures of Escherichia coli and Lactobacillus casei dihydrofolate reductase refined at 1.7 A resolution. I. General features and binding of methotrexate". *The Journal of Biological Chemistry* 257 (1982), pp. 13650–13662.

[97] *D3.js JavaScript library*. URL: `https://d3js.org/` (visited on 08/24/2024).

[98] *Fraction.js JavaScript library*. URL: `https://github.com/infusion/Fraction.js/` (visited on 08/24/2024).

[99] *SmilesDrawer JavaScript library*. URL: `https://github.com/reymond-group/smilesDrawer` (visited on 08/24/2024).

[100] D. Probst and J.-L. Reymond. "SmilesDrawer: Parsing and Drawing SMILES-Encoded Molecular Structures Using Client-Side JavaScript". *Journal of Chemical Information and Modeling* 58 (2018), pp. 1–7. DOI: `https://doi.org/10.1021/acs.jcim.7b00425`.

[101] *Jasmine Ruby on Rails library*. URL: `https://github.com/jasmine/jasmine-gem` (visited on 08/24/2024).

[102] *ProteinsPlus REST API (version 1)*. URL: `https://proteins.plus/api/v1` (visited on 08/24/2024).

[103] *Ruby on Rails web framework*. URL: `https://rubyonrails.org` (visited on 08/24/2024).

[104] *MySQL database*. URL: `https://www.mysql.com` (visited on 08/24/2024).

[105] *Delayed::Job*. URL: `https://github.com/collectiveidea/delayed_job` (visited on 08/24/2024).

[106] *rspec-rails testing framework*. URL: `https://github.com/rspec/rspec-rails` (visited on 08/24/2024).

[107] *Rack::Attack middleware*. URL: `https://github.com/rack/rack-attack` (visited on 08/24/2024).

[108] *Bootstrap frontend toolkit*. URL: `https://getbootstrap.com` (visited on 08/24/2024).

[109] *NGL library*. URL: `https://github.com/arose/ngl` (visited on 08/24/2024).

[110] A. S. Rose, A. R. Bradley, Y. Valasatava, J. M. Duarte, A. Prlić, and P. W. Rose. "NGL viewer: web-based molecular graphics for large complexes". *Bioinformatics* 34 (2018), pp. 3755–3758. DOI: https://doi.org/10.1093/bioinformatics/bty419.

[111] A. S. Rose and P. W. Hildebrand. "NGL Viewer: a web application for molecular visualization". *Nucleic Acids Research* 43 (2015), pp. 576–579. DOI: https://doi.org/10.1093/nar/gkv402.

[112] *Datatables library*. URL: https://datatables.net (visited on 08/24/2024).

[113] *Chart.js library*. URL: https://www.chartjs.org/ (visited on 08/24/2024).

[114] *Matomo web analytics platform*. URL: https://matomo.org/ (visited on 08/24/2024).

[115] *RCSB REST API*. URL: https://rcsb.org/ (visited on 08/24/2024).

[116] *Docker*. URL: https://www.docker.com/ (visited on 08/24/2024).

[117] *Podman*. URL: https://podman.io/ (visited on 08/24/2024).

[118] *QT*. URL: https://www.qt.io/ (visited on 08/24/2024).

[119] S. Bietz, T. Inhester, F. Lauck, K. Sommer, M. M. von Behren, R. Fährrolfes, F. Flachsenberg, A. Meyder, E. Nittinger, T. Otto, M. Hilbig, K. T. Schomburg, A. Volkamer, and M. Rarey. "From cheminformatics to structure-based design: Web services and desktop applications based on the NAOMI library". *Journal of Biotechnology* 261 (2017), pp. 207–214. DOI: https://doi.org/10.1016/j.jbiotec.2017.06.004.

[120] *PostgreSQL*. URL: https://www.postgresql.org/ (visited on 08/24/2024).

# Bibliography of the Cumulative Dissertation


[D1] K. Diedrich, B. Krause, O. Berg, and M. Rarey. "PoseEdit: enhanced ligand binding mode communication by interactive 2D diagrams". *Journal of Computer-Aided Molecular Design* 37 (2023), pp. 491–503. DOI: https://doi.org/10.1007/s10822-023-00522-4.

[D2] K. Diedrich, J. Graef, K. Schöning-Stierand, and M. Rarey. "GeoMine: interactive pattern mining of protein-ligand interfaces in the Protein Data Bank". *Bioinformatics* 37 (2020), pp. 424–425. DOI: https://doi.org/10.1093/bioinformatics/btaa693.

[D3] J. Graef, C. Ehrt, K. Diedrich, M. Poppinga, N. Ritter, and M. Rarey. "Searching Geometric Patterns in Protein Binding Sites and Their Application to Data Mining in Protein Kinase Structures". *Journal of Medicinal Chemistry* 65 (2022), pp. 1384–1395. DOI: https://doi.org/10.1021/acs.jmedchem.1c01046.

[D4] M. Poppinga, J. Graef, K. Diedrich, M. Rarey, and N. Ritter. "Database and Workflow Optimizations for Spatial-Geometric Queries in GeoMine". *Proceedings of the LWDA 2023 Workshops: FGDB, FGBIA, FGKDML, FGWM, and FGIR (LWDA 2023).* Maarburg, Deutschland: CEUR-WS.org. Available: https://ceur-ws.org/Vol-3630/LWDA2023-paper8.pdf.

[D5] K. Diedrich, C. Ehrt, J. Graef, M. Poppinga, N. Ritter, and M. Rarey. "User-centric design of a 3D search interface for protein-ligand complexes". *Journal of Computer-Aided Molecular Design* 38 (2024). DOI: https://doi.org/10.1007/s10822-024-00563-3.

[D6] K. Schöning-Stierand, K. Diedrich, R. Fährrolfes, F. Flachsenberg, A. Meyder, E. Nittinger, R. Steinegger, and M. Rarey. "ProteinsPlus: interactive analysis

of protein-ligand binding interfaces". *Nucleic Acids Research* 48 (2020), pp. 48–53. DOI: https://doi.org/10.1093/nar/gkaa235.

[D7] K. Schöning-Stierand, K. Diedrich, C. Ehrt, F. Flachsenberg, J. Graef, J. Sieg, P. Penner, M. Poppinga, A. Ungethüm, and M. Rarey. "ProteinsPlus: a comprehensive collection of web-based molecular modeling tools". *Nucleic Acids Research* 50 (2022), pp. 611–615. DOI: https://doi.org/10.1093/nar/gkac305.

# Appendix A

# Scientific Contributions

In this section, the author's scientific contributions related to the topics of the doctoral project are listed thematically in chronological order. M. Rarey supervised all scientific contributions mentioned below.

## A.1 Publications

**[D1]** K. Diedrich, B. Krause, O. Berg, and M. Rarey. "PoseEdit: enhanced ligand binding mode communication by interactive 2D diagrams". *Journal of Computer-Aided Molecular Design* 37 (2023), pp. 491–503. DOI: `https://doi.org/10.1007/s10822-023-00522-4`

This scientific publication introduces the tool PoseEdit. The concepts behind PoseEdit were developed by K. Diedrich and M. Rarey. A corresponding frontend-only prototype was implemented by B. Krause for his master's thesis to be revised and extended by K. Diedrich for this doctoral project. K. Diedrich wrote the manuscript draft. All authors read and improved the manuscript.

**[D2]** K. Diedrich, J. Graef, K. Schöning-Stierand, and M. Rarey. "GeoMine: interactive pattern mining of protein-ligand interfaces in the Protein Data Bank". *Bioinformatics* 37 (2020), pp. 424–425. DOI: `https://doi.org/10.1093/bioinformatics/btaa693`

This scientific publication introduces the tool GeoMine. The tool's method, graphical user interface, and application are described. The web-based graphical user interface was developed and implemented by K. Diedrich. The search algorithm and the underlying database were developed and implemented by J. Graef. K. Diedrich developed the application examples. K. Diedrich wrote the manuscript draft. All authors read

and improved the manuscript.

**[D3]** J. Graef, C. Ehrt, K. Diedrich, M. Poppinga, N. Ritter, and M. Rarey. "Searching Geometric Patterns in Protein Binding Sites and Their Application to Data Mining in Protein Kinase Structures". *Journal of Medicinal Chemistry* 65 (2022), pp. 1384–1395. DOI: https://doi.org/10.1021/acs.jmedchem.1c01046

This scientific publication focuses on the search algorithm, underlying database, new tool features, and corresponding application examples. C. Ehrt developed the application examples. J. Graef developed and implemented the search algorithm and database. K. Diedrich developed and implemented the web-based graphical user interface. The new features were developed and integrated into the backend and frontend by J. Graef and K. Diedrich, respectively. M. Poppinga and N. Ritter contributed to the database design. J. Graef and C. Ehrt wrote the manuscript draft. All authors read and improved the manuscript.

**[D4]** M. Poppinga, J. Graef, K. Diedrich, M. Rarey, and N. Ritter. "Database and Workflow Optimizations for Spatial-Geometric Queries in GeoMine". *Proceedings of the LWDA 2023 Workshops: FGDB, FGBIA, FGKDML, FGWM, and FGIR (LWDA 2023)*. Maarburg, Deutschland: CEUR-WS.org. Available: https://ceur-ws.org/Vol-3630/LWDA2023-paper8.pdf

This scientific publication describes database optimizations regarding the search's runtime. M. Poppinga developed and implemented the optimization of 3D searches. M. Poppinga and J. Graef developed and implemented the optimization of text-based searches. M. Poppinga performed the corresponding runtime experiments. All authors wrote and read the manuscript.

**[D5]** K. Diedrich, C. Ehrt, J. Graef, M. Poppinga, N. Ritter, and M. Rarey. "User-centric design of a 3D search interface for protein-ligand complexes". *Journal of Computer-Aided Molecular Design* 38 (2024). DOI: https://doi.org/10.1007/s10822-024-00563-3

This scientific publication describes the integration of PoseEdit's 2D interface into GeoMine to facilitate the 3D query generation process. Furthermore, application examples are provided to showcase the new functionalities. The 2D interface concept was developed and implemented by K. Diedrich. C. Ehrt developed the application examples. K. Diedrich and C. Ehrt wrote the manuscript draft. All authors read and improved the manuscript.

**[D6]** K. Schöning-Stierand, K. Diedrich, R. Fährrolfes, F. Flachsenberg, A. Meyder, E. Nittinger, R. Steinegger, and M. Rarey. "ProteinsPlus: interactive analysis of protein-ligand binding interfaces". *Nucleic Acids Research* 48 (2020), pp. 48–53. DOI: `https://doi.org/10.1093/nar/gkaa235`

This scientific publication describes new tools and features of the Proteins*Plus* web server, including StructureProfiler, WarPP, METALizer, the PDB keyword-search functionality, and REST API. The REST API, which is relevant for all existing tools as well as for the development of PoseEdit and GeoMine, was developed and implemented by K. Diedrich for a student project and extended for this doctoral project. The manuscript draft was written by K. Stierand. All authors read and improved the manuscript.

**[D7]** K. Schöning-Stierand, K. Diedrich, C. Ehrt, F. Flachsenberg, J. Graef, J. Sieg, P. Penner, M. Poppinga, A. Ungethüm, and M. Rarey. "ProteinsPlus: a comprehensive collection of web-based molecular modeling tools". *Nucleic Acids Research* 50 (2022), pp. 611–615. DOI: `https://doi.org/10.1093/nar/gkac305`

This scientific publication describes new tools and features of the Proteins*Plus* web server, including GeoMine and the integration of AlphaFold structures. Furthermore, one application example is provided, including the usage of all new tools and AlphaFold structures. The GeoMine and AlphaFold integration, which is relevant for all described tools, including GeoMine, was implemented by K. Diedrich. All authors read and improved the manuscript. The manuscript content regarding GeoMine was written by J. Graef and K. Diedrich.

## A.2 Conferences and Workshops

### A.2.1 Oral Presentations

**[V1]** Nittinger, E., **Diedrich, K.**, Rarey, M., Workshop: Protein Structure Fundamentals: Searching - Analyzing - Modelling. Universität Hamburg ZBH - Center for Bioinformatics. Hamburg, Germany, 2019

**[V2] Diedrich, K.**, J. Graef, J., Poppinga, M., Ehrt, C., Ritter, N., Rarey, M., GeoMine: Geometric Pattern Mining in PDB Binding Sites. 1st Nordic Conference on Computational Chemistry. Gothenburg, Sweden, 2022

**[V3]** Ehrt, C., **Diedrich, K.**, Graef, J., Poppinga, M., Ritter, N., Rarey, M., Workshop: GeoMine @ ProteinsPlus - Versatile Tools for Structural Investigations. 1st Nordic Conference on Computational Chemistry. Gothenburg, Sweden, 2022

### A.2.2 Poster Presentations

**[P1] Diedrich, K.**, Graef, J., Nittinger, E., Rarey, M., GeoMine: A Web-Based Tool for chemical 3D Searching of the PDB. German Conference on Cheminformatics (GCC). Mainz, Deutschland, 2019.

**[P2]** Graef, J., **Diedrich, K.**, Schöning-Stierand, K., Rarey, M., GeoMine: A Web-Based Tool for Chemical Three-Dimensional Searching of the PDB. International Society For Computational Biology (ISMB). Virtuelle Konferenz, 2020.

# Appendix B

# Software

The following sections and subsections explain the tools described in Chapter 2 and Chapter 3 in more detail.

## B.1 PoseView

### B.1.1 Usage

This subsection describes the command line interface of PoseView, highlighting the corresponding development done during this doctoral project. The following command executes PoseView's new functionality, creating a diagram in JSON and SVG format:

```
./bin/poseview -f ./input.mol2 -o ./output.svg -w ./output.json
```

The command line arguments are listed and further described in Table B.1.

**Table B.1:** Extended command-line interface options of PoseView used by PoseEdit

| **option** | **description** |
|---|---|
| -f <file path> | provide chemical input data for diagram calculation by other tools, for example, GeoMine, via a MOL2 file |
| -o <file path> | print the diagram in various image file formats (.png, .pdf, .svg) |
| -w <file path> | store textual diagram data in a JSON file |

### B.1.2 MOL2 Input

This subsection describes the MOL2 file content generated by GeoMine based on a PDB file and one of its ligands. MOL2 files are provided to PoseView's command line interface as input to specify the chemical information content of corresponding PoseEdit diagrams. The command line interface functionality to process MOL2 files already exists in PoseView. For this doctoral project, adjustments were made to the MOL2 content and its processing. Heavy atoms of molecular entities are listed by their PDB file-extracted serial numbers instead of newly created sequence indices. This adjustment enables later atomic assignment necessary, for example, regarding the 2D-3D synchronization of atom selection in PoseEdit and GeoMine. Hydrogen atoms were assigned incremental serial identifiers starting with the highest serial number of all heavy atoms. The MOL2 file content exemplary described below is based on the PDB entry 4DFR and one of its two ligands (ligand identifier: MTX_A_161) [96].

The first part is a comment block listing all interacting structures and corresponding intermolecular interactions, including hydrophobic contacts. The comment block is illustrated in the following text and Table B.2:

```
@<TRIPOS>COMMENT
MTX_A_161
%AMINO_ACIDS 6
# Format: <mol_id> <mol_nr> <chain_id> <name> <m_id>
2  4  A Ile5A 4
3 26 A Asp27A 26
4 30 A Phe31A 30
5 51 A Arg52A 51
6 56 A Arg57A 56
7 93 A Ile94A 93
%DIRECTED_BONDS 16
# Format: <aa_nr> <type> <pi_cat_type> <energy>
#         <nof_lig_ia_atm> <lig_ia_center> <lig_ia_atm1> ...
#         <nof_aa_ia_atm> <aa_ia_center> <aa_ia_atm1> ...
4 0 0 0.000000
1 39 13
1 2 2
26 0 0 0.000000
1 36 12
1 7 7
```

```
26 0 0 0.000000
1 34 6
1 8 8
30 4 0 0.000000
6 7 7 14 6 15 17 16
6 6 6 7 8 9 10 11
30 4 0 0.000000
6 9 9 21 19 8 17 15
6 6 6 7 8 9 10 11
51 1 0 0.000000
1 5 5
1 24 11
56 1 0 0.000000
1 3 3
1 21 10
56 1 0 0.000000
1 1 1
1 23 11
93 0 0 0.000000
1 38 13
1 2 2
26 4 0 0.000000
1 6 6
1 7 7
26 4 0 0.000000
1 6 6
1 8 8
51 4 0 0.000000
1 2 2
1 11 11
56 4 0 0.000000
1 1 1
1 10 10
56 4 0 0.000000
1 3 3
1 10 10
```

```
56 4 0 0.000000
1 1 1
1 11 11
56 4 0 0.000000
1 3 3
1 11 11
%HYDROPHOBIC_CONTACTS 3
# Format: <aa_name> <aa_nr> <aa_pdb_nr> <aa_chain_id><nof_cons> <con1> <con2> ... <conn>
LEU 28 28 A 5 9 20 21 25 31
PHE 31 31 A 8 7 8 14 15 16 17 23 25
ILE 50 50 A 3 18 22 23
```

**Table B.2:** MOL2 definition of interacting structures, intermolecular interactions, and hydrophobic contacts

| | |
|---|---|
| **interacting structures** | |
| mol_id | index of the interaction partner based on the order in which they are listed as molecular entries below in the MOL2 file. Since the ligand (ligand identifier: MTX_A_161) is internally always indexed with 1, its interaction partners start with index 2 |
| mol_nr | structure sequence number taken from the corresponding PDB file |
| chain_id | chain identifier the structure belongs to |
| name | name of the structure |
| m_id | additional structure identifier (here set same as mol_nr) |
| **intermolecular interactions** | |
| aa_nr | structure sequence number from the corresponding PDB file (same as mol_nr) |
| type | type of the intermolecular interaction. 0 if hydrogen bond and the donor atom belongs to the ligand, 1 if hydrogen bond and the donor atom belongs to the residue, 3 if it is a metal interaction, and 4 for ionic, cation-pi, and pi-stacking interactions |

| | |
|---|---|
| pi_cat_type | 1 if the interacting cationic atom belongs to the ligand and 2 if the interacting cationic atom belongs to the residue |
| energy | additional energy values of hydrogen bonds for their labeling. Energy values are not provided by GeoMine |
| nof_lig_ia_atm | number of ligand atoms participating in the intermolecular interaction |
| lig_ia_center | MOL2 file index of the interacting heavy atom if the atom is the acceptor or of the interacting hydrogen atom if the atom is the donor |
| lig_ia_atm1, lig_ia_atm2, ... | all ligand atoms participating in the intermolecular interaction. In the case of cation-pi and pi-stacking interactions, all atoms of the interacting aromatic ring are listed |
| nof_aa_ia_atm | number of atoms of the interaction partner that participate in the intermolecular interaction |
| aa_ia_center | MOL2 file index of the interacting heavy atom if the atom is the acceptor or of the interacting hydrogen atom if the atom is the donor |
| aa_ia_atm1, aa_ia_atm2, ... | all interaction partner atoms participating in the intermolecular interaction. In the case of cation-pi and pi-stacking interactions, all atoms of the interacting aromatic ring are listed |
| **hydrophobic contacts** | |
| aa_name | name of the residue with hydrophobic contact to the ligand |
| aa_nr | residue sequence number from the corresponding PDB file (same as mol_nr) |
| aa_pdb_nr | additional residue identifier, here set like aa_nr |
| aa_chain_id | chain identifier the residue belongs to |
| nof_cons | number of ligand atoms with hydrophobic contact to the residue |

| | |
|---|---|
| con1, con2, ... | indices of ligand atoms with hydrophobic contact to the residue. The indices are based on the order of the ligand atoms in the ligand's MOL2 entry located below |

Chemical and spatial information about the ligand and all its interaction partners are listed below the above-described comment block. All atoms are listed with their serial numbers taken from the corresponding PDB file. The corresponding MOL2 content is illustrated in the following text:

```
@<TRIPOS>MOLECULE
MTX_A_161
54 56
SMALL
FORMAL_CHARGES
@<TRIPOS>ATOM
2569 O2 19.921 68.75 23.149 O.co2 1 LIG -0.5
2574 OE2 19.441 69.469 27.489 O.co2 1 LIG -0.5
2568 O1 20.289 66.659 22.848 O.co2 1 LIG -0.5
...
@<TRIPOS>BOND
1 2569 2567 ar
2 2574 2572 ar
3 2568 2567 ar
...

@<TRIPOS>MOLECULE
Ile5A
21 20
SMALL
FORMAL_CHARGES
@<TRIPOS>ATOM
33 C 23.282 55.874 19.045 C.2 1 RES 0
34 O 22.074 56.043 19.096 O.2 1 RES 0
32 CA 24.107 56.528 17.919 C.3 1 RES 0
...
```

```
@<TRIPOS>BOND
70 32 33 1
74 33 34 2
78 33 39 am
...

...
```

## B.2 InteractionDrawer

### B.2.1 Usage

This subsection describes the usage of the InteractionDrawer library regarding test execution, drawer configuration, and diagram generation and its manipulation. The InteractionDrawer library is available on GitHub [60]. Files with information about the library and how to use it can be accessed in the README.md file and via files in the help folder of the repository:

- Setup.md: how to set up the library's drawing functionality
- Options.md: how to configure the library (styling, key bindings, etc.)
- JSON.md: how to provide the input to the drawing area (format of JSON input)

The library is compatible with Ruby on Rails applications whose assets pipelines automatically handle the library's code compression into one file. Integrating the library's source code and tests into a Ruby on Rails project, the library's tests can be executed as follows.
Start a Jasmine server:

```
rake jasmine
```

Start a browser and open:

```
http://localhost:8888/?spec=InteractionDrawer
```

These commands run all library tests in a web browser. Specific tests can be executed with a URL precision based on the library's given folder structure. For example, *http://localhost:8888/?spec=InteractionDrawer GeometryCalculation* runs all tests in the GeometryCalculation subfolder.

A diagram can be displayed for an SVG with id="draw-area" as described below. Users can provide various options during the initialization of an InteractionDrawer instance (see B.2.2) as part of the opts parameter. Any option that is not manually set is filled by a default value. In the following example, the general text size and the colors of dashed lines representing ionic interactions and text labels representing the element aluminum are set.

```
//set options as you like
const opts = {
 textSize: 6.5,
 colors: {
  ionicInteractions: '#ff00ff',
  AL: '#BFA6A6'
 }
};
//initialize the drawing area
const drawer = new InteractionDrawer.Drawer('draw-area', opts);
```

For each SVG drawing area users want to fill with the InteractionDrawer library, a separate InteractionDrawer instance is required, which can be provided with several types of input as illustrated below:

**JSON**
Via a json variable containing a JSON diagram string (see B.2.3).

```
drawer.addByJSON(json);
```

**PDB identifier and ligand name**
Via an id variable that can either be a string representing a four-letter PDB identifier or a Proteins*Plus* identifier of an uploaded PDB file and a ligandName variable that contains a ligand name as string in the format molecule_chain_number, e.g., 4SP_A_1298.

```
drawer.addById(id, ligandName);
```

**PDB file and ligand name**
Via a fileContent variable that contains the content of a PDB file as string and a ligandName variable that contains a ligand name as string in the format molecule_chain_number, e.g., 4SP_A_1298.

```
drawer.addByFile(fileContent, ligandName);
```

The InteractionDrawer library calls with this information the PoseEdit REST API of the Proteins*Plus* web server for calculating diagram data in JSON format, which is subsequently drawn by the InteractionDrawer library as interactive SVG images. The PoseEdit REST API documentation can be found in B.4.1. If users want to handle the communication with the REST API or the diagram calculation themselves, a scene in JSON format can be directly provided to an InteractionDrawer instance as input.

After the setup of the InteractionDrawer instance, several callbacks functions can be set with drawer.setCallbacks(), for example, a callback that is triggered after a scene element has been hit by mouse hover. Users can access and manipulate the diagram data in drawer.sceneData and the view in drawer.svgComponent directly or via class instances and methods exposed by drawer.userInteractionHandler and drawer.svgDrawer. In addition, the InteractionDrawer library also provides several classes and methods contained in files in the DataProcessing (functions for manipulating the data), Utils (generic helper functions), and GeometryCalculation (functions for geometric calculations) folders that can be used to process the scene. A diagram can be exported as an SVG file (drawer.getSvgBlob()), a JSON file with and without the current config (drawer.getJsonBlobWithConfig(), drawer.getJsonBlob()), or a text file containing information about all intermolecular interactions (drawer.getTxtBlob()).

### B.2.2 Config

This subsection describes the configuration options (see Options.md) of the InteractionDrawer library.

#### B.2.2.1 Debug Information

Debug text can be displayed on top of atom text/bond lines. This text can be completely disabled (so as not to take up any SVG elements). Otherwise, its current visibility can be controlled. See Table B.3 for the corresponding configuration options.

**Table B.3:** Options to display debug information

| **option** | **effect** | **default value** |
|---|---|---|
| debug.atoms | whether atom debug text can ever be shown | true |

| | | |
|---|---|---|
| debug.showAtoms | whether to immediately show atom debug text | false |
| debug.edges | whether bond debug text can ever be shown | true |
| debug.showEdges | whether to immediately show bond debug text | false |
| debug.textSize | size of debug text (in px) | 6.5 |

#### B.2.2.2 User Interaction

**Allowed modes:**
The drawer offers different user interaction modes which can be included or excluded. See Table B.4 for the corresponding configuration options.

**Table B.4:** Options to include or exclude specific user interactions

| **option** | **effect** | **default value** |
|---|---|---|
| allowInteraction | whether user interaction is allowed at all | true |
| allowedInteraction | JavaScript array of allowed user interactions; all others are disabled. By default, all interaction types are enabled. Users can use this parameter if they only want to specify a few interaction types | ["movement", "rotation", "scaledRotation", "rectSelect", "freeSelect", "lineMirror", "bondMirror", "zoomIn", "zoomOut", "advanceHistory", "revertHistory", "center", "structureReset", "remove", "addIntermolecular", "addAnnotation", "globalMovement", "doNothing", "addStructure", "clickSelect", "addAtom", "edit"] |

| | | |
|---|---|---|
| excludedInteraction | JavaScript array of explicitly not allowed interactions (to be removed from allowedInteraction). If users need to disable only a few interactions (but keep the rest), this parameter is advantageous instead of allowedInteraction | [] |

**Mouse bindings:**
Many interaction modes follow the same drag-and-drop pattern. These modes are all applicable with the mouse. Different modes are applicable depending on the pressed mouse button (field "key") and additionally pressed keyboard modifiers (field "modifiers"). Users can also add more than one binding to a specific interaction by expanding the JavaScript array. A single mode can also be set as the default mode (option defaultInteraction) if no other mode conditions are met. A preselection of modes is already set. See Table B.5 for the corresponding configuration options.

**Table B.5:** Options to bind user interaction modes

| **option** | **effect** | **default value** |
|---|---|---|
| buttons.mouse.rotation | mouse key bindings to trigger the rotation interaction mode | ['key': 1, 'modifiers': []] |
| buttons.mouse.addAnnotation | mouse key bindings to trigger the input popup window to add a new annotation | ['key': 1, 'modifiers': ['shift']] |

| | | |
|---|---|---|
| buttons.mouse.addIntermolecular | mouse key bindings to trigger the adding of new intermolecular interactions | ['key': 1, 'modifiers': ['ctrl']] |
| buttons.mouse.addStructure | mouse key bindings to trigger the input popup window to add a new structure | ['key': 1, 'modifiers': ['alt']] |
| buttons.mouse.movement | mouse key bindings to trigger the movement interaction mode | ['key': 2, 'modifiers': []] |
| buttons.mouse.bondMirror | mouse key bindings to trigger the mirror interaction mode on a bond of a structure | ['key': 2, 'modifiers': ['alt']] |
| buttons.mouse.lineMirror | mouse key bindings to trigger the mirror interaction mode on a free line | ['key': 2, 'modifiers': ['ctrl']] |
| buttons.mouse.scaledRotation | mouse key bindings to trigger the scaled rotation interaction mode | |

| | | |
|---|---|---|
| buttons.mouse.addAtom | mouse key bindings to trigger the input popup window for drawing a new atom | |
| buttons.mouse.edit | mouse key bindings to trigger the input popup window to edit a specific annotation, atom, bond, or structure | |
| buttons.mouse.doNothing | no interaction mode is triggered | |

**Other bindings:**
Other interactions can be bound to different keyboard buttons ("type": "key", with "button" set to the desired key) or the mouse wheel (type "wheel", with "button" set to "up" or "down"). Multiple buttons can be bound for one interaction type. See Table B.6 for the corresponding configuration options.

**Table B.6:** Options to bind additional user interaction modes

| **option** | **effect** | **default value** |
|---|---|---|
| buttons.zoomIn | key bindings to zoom into the scene | ["type": "wheel", "button": "up", "type": "key", "button": "+"] |
| buttons.zoomOut | key bindings to zoom out of the scene | ["type": "wheel", "button": "down", "type": "key", "button": "-"] |

| | | |
|---|---|---|
| buttons.advanceHistory | key bindings to advance the InteractionDrawer's history by one step | ["type": "key", "button": "ArrowRight"] |
| buttons.revertHistory | key bindings to revert the InteractionDrawer 's history by one step | ["type": "key", "button": "ArrowLeft"] |
| buttons.center | key bindings to center the scene in its drawing area | ["type": "key", "button": "c"] |
| buttons.structureReset | key bindings to reset the current scene to its initial (after processing JSON input) state | ["type": "key", "button": "r"] |
| buttons.remove | key bindings to remove hovered/selected objects | ["type": "key", "button": "Delete"] |

**Mode behavior:**
The behavior of some interaction modes can be further specified by setting certain properties. See Table B.7 for the corresponding configuration options.

**Table B.7:** Additional options to bind user interaction modes

| **option** | **effect** | **default value** |
|---|---|---|
| moveFreedomLevel | size of molecular units to be moved at once in the movement interaction mode:<br>1. "structures": can only move full structures<br>2. "free": can freely move atoms and bonds<br>3. "rings": can freely move atoms and bonds outside of ring systems, but move all atoms and bonds of ring systems together | "structures" |
| moveAllSelection | whether to move all selected elements at once when a selected element is moved during the movement interaction mode (while moveFreedomLevel is set to structures) | true |
| hoverAfterDeselection | keeps hover highlighting of an element active after its deselection | true |

| | | |
|---|---|---|
| selectionGrace | how far the mouse pointer must be moved from its initial click position before a movement is committed in the movement interaction mode (to allow selection for minor mouse slips, in px) | 0.4 |
| scaledRotationThreshold | threshold after which rotation is committed during the scaledRotation interaction mode (in degree) | 5 |
| zoomStrength | relative strength of zoom operations | 7 |
| sceneMaxScale | maximum scaling factor, which can be applied to the scene (can be set to null to allow infinite zoom-out) | 3 |
| sceneMinScale | minimum scaling factor, which can be applied to the scene (can be set to null to allow infinite zoom-in) | 0.5 |

| | | |
|---|---|---|
| handleCollisionWith | how collision detection on scene elements should be done:<br>1. "selector": checks collision against the selection shape(s) around draw elements<br>2. "drawingOnly": checks collision against the drawn shapes of draw elements only (so against much smaller elements than the selection shapes) | "selector" |
| historyCanClearScene | whether reverting the history can set the scene back to its empty state before adding elements (if set to false, the latest state that can be reverted back to the state in which elements were just added) | true |
| addIntermolecularSnapDist | the distance from the nearest atom, ring, or hydrophobic control point to the mouse pointer position where to snap the intermolecular interaction to the object during addIntermolecular interaction mode | 15 |

| | | |
|---|---|---|
| resetMode | to what state to reset during reset interaction (initial state off; 0 = all present elements; 1 = only the first loaded JSON, discard other elements; 2 = only all loaded JSONs which contain at least one structure) | 0 |

#### B.2.2.3 Representations

There are currently two different representations of structures.

- *default* is the skeletal representation. Internally, this is always present because the drawer relies on that information
- *circle* visualizes the structures as a circle

The allowed representations can be specified for each moleculeType defined in the loaded JSON. As stated above, the default representation must always be present. See Table B.8 for the corresponding configuration options.

**Table B.8:** Options to enable representation types for structures

| option | effect | default value |
|---|---|---|
| allowedStructureRepresentations | can contain a key with a JavaScript array of allowed representations for every moleculeType of loaded structures. The first item in the JavaScript array defines the initial representation. The JavaScript array at key default is applied when moleculeType of a loaded structure does not match any other key. | default: ["default", "circle"] |

**Circle options:**
Additional styling and functional options can be applied to the structure when in circle representation. The label of the circle shown inside the circle (JSON field "moleculeLabel") can be split into a maximum of three rows so as not to extend outside the circle. There are two split methods available:

- **Automatic**: splits the label so that all rows have the same number of characters. Note that this mode does not guarantee rows of the same width because the width of characters differ
- **Manual**: splits the label at the first occurrences of characters in a defined array

Further options on how splitting should be handled are described in the table below. See Table B.9 for the corresponding configuration options.

**Table B.9:** Options to style and functionally adapt the circle representation

| option | effect | default value |
|---|---|---|
| structureCircleOpts | can contain a key that holds a JavaScript object of additional options (described below) for every moleculeType of loaded structures. The object at key default is applied when moleculeType of a loaded structure does not match any other key | "default": "rad": 20, "textColor": "000000", "circleCss": "stroke": "000000", "fill": "808080", "opacity": "0.4" , "labelAutoSplit": true, "labelAutoSplitMinCharsPerLine": 3, "labelManualLineBreakChars": [" ", "-", "_"], "labelManualIncludeSplitChar": [false, false], "labelMaxLines": 3, "labelSameFontSize": true |
| structureCircleOpts .moleculeType.circleCss | JavaScript object containing all relevant styling that is directly applied to the circle SVG element. Every style attribute that fits an SVG circle element can be set. (Hint: use "stroke": undefined for no border) | "stroke": "000000", "fill": "808080", "opacity": "0.4" |

| | | |
|---|---|---|
| structureCircleOpts .moleculeType.rad | radius of the circle | 20 |
| structureCircleOpts .moleculeType.textColor | Cascading Style Sheets (CSS) color of the label in the middle of the circle | "000000" |
| structureCircleOpts .moleculeType.labelMaxLines | maximal number of lines the label string should be split into. Only 1, 2, and 3 are supported currently | 3 |
| structureCircleOpts .moleculeType.labelSameFontSize | whether each line of the label should fill the available space and may produce a different font size for each row (false) or each line should have the same font size (true). The max font size will always be textSize | true |
| structureCircleOpts .moleculeType.labelAutoSplit | the split mode. true: automatic mode; false: manual mode | true |
| structureCircleOpts .moleculeType .labelAutoSplitMinCharsPerLine | minimal number of characters that should be present in each row of the label. Only applies if labelAutoSplit is true | 3 |

| | | |
|---|---|---|
| structureCircleOpts .moleculeType .labelManualLineBreakChars | JavaScript array containing characters at which to split the label. Only applies if labelAutoSplit is false | [" ", "-", "_"] |
| structureCircleOpts .moleculeType .labelManualIncludeSplitChar | if the character on which the label string was split should be included in the final label. First value for first split, second value for second split. Only applies if labelAutoSplit is false | [false, false] |

#### B.2.2.4 Styling

The configuration options directly control the styling of scene elements. Parameters usually correspond directly with CSS properties to set. For styling of the structure circle representation, see the previous subsection. See Table B.10-14 for the corresponding configuration options.

**Text**:

**Table B.10:** Options to style SVG element representing atoms and text labels

| **option** | **effect** | **default value** |
|---|---|---|
| atomMode | how atoms are depicted. Currently only supports "name" mode, which displays atoms by a textual label | "name" |

| | | |
|---|---|---|
| textSelector | kind of selection shape to draw around text selectors: "circle": only draw a simple circle; "full": draw a more sophisticated shape which fully flows around the text | "full" |
| fontFamily | font family to use for text labels | "arial" |
| textSize | text size to use for text labels (in px) | 6.5 |
| textCrop | estimated extra space below/above glyphs in the current font (to get hydrogen text above/below atom text closer) | "0.285em" |
| textBorderCorrection | estimated excess space on border boxes of larger text elements (in px) | 1 |
| labelSideCorrection | correction to apply to anchor points (as selectors around labels are very large, in px) | 1 |
| hOffset | distance between atom text and hydrogen text | "0.2em" |
| chargeFontSize | percentage of regular font size to apply to the charge text font size | 0.775 |
| chargeOffset | percentage of regular font size to set as y offset between atom text and charge text | 0.3 |
| hNumberFontSize | percentage of regular font size to apply to the font size of hydrogen number text (the small number as subscript of hydrogen text) | 0.6 |

| | | |
|---|---|---|
| hNumberOffset | percentage of regular font size to set as y offset between the middle line of hydrogen text and the upper line of bounding box of hydrogen number text | 0.235 |
| atomRadius | radius around atoms in which bonds are not allowed to be drawn (in px) | 3.75 |
| atomSelectorRadius | radius around atoms/distance from atoms to base selection shape | 5 |
| smallestBboxWidth | when first drawing a scene, scale the scene such that a text of 'I' is rendered with a bbox at least of this width (in px) to avoid a too small drawing where rounding errors may distort the scene | 20 |

**Edges**:

**Table B.11:** Options to style SVG elements representing bonds and intermolecular interactions

| **option** | **effect** | **default value** |
|---|---|---|
| lineWidth | width of drawn lines (in px) | 0.65 |
| wedgeBaseWidth | width of stereo bonds at their smallest (in px) | 0.3 |
| wedgeFullWidth | width of stereo bonds at their widest (in px) | 2.5 |
| wedgeSpacing | space between individual segments of stereo bonds (in px) | 1 |

| | | |
|---|---|---|
| spaceBetweenDouble | space between individual lines of double bonds (in px) | 1.3 |
| spaceBetweenTriple | space between individual lines of triple bonds (in px) | 0.75 |
| spaceToRing | space between bonds and inner bonds of aromatic rings (in px) | 1.3 |
| cutoffAngleDouble | for double bonds where not both bonds are drawn full length: in which angle does the endpoint of the smaller bond lies relative to the endpoint of the larger bond (in degree) | 60 |
| edgeSelectorOffset | distance from lines representing bonds and the surrounding selection shape (in px) | 2.25 |
| lineDashDrawn | length of individual segments of dashed lines (in px) | 2 |
| lineDashGap | gap between segments of dashed lines (in px) | 2 |

**Coloring**:

**Table B.12:** Options to color SVG elements

| option | effect | default value |
|---|---|---|
| colors.DEFAULT | color of annotations that do not belong to a hydrophobic contact or with no given color. Color for atoms that have an unknown element. Color of undefined interactions and bonds | "222" |
| colors.BACKGROUND | background color of the drawing area | "fff" |
| colors.HOVER | color of hovered elements | "8eff7d" |
| colors.SELECTION | color of selected elements | "10ff00" |
| colors.MIRROR | color of the line to mirror during lineMirror interaction mode | "ff6600" |
| colors.multiSelectionToolBorder | border color of lasso and rectangle selection tool line | "ff33ff" |
| colors.multiSelectionToolFill | color of lasso and rectangle selection tool line | "ffe6ff" |
| colors.cationPiStackings | color of cation-pi stacking interactions | "99cc33" |
| colors.piStackings | color of pi-stacking interactions | "33cccc" |
| colors.atomPairInteractions | color of atom pair interactions | "6699ff" |

| | | |
|---|---|---|
| colors.ionicInteractions | color of ionic interactions | "ff00ff" |
| colors.metalInteractions | color of metal interactions | "f7de3a" |
| colors.hydrophobicContacts | color of hydrophobic contacts | "019a4d" |

Atom colors can be set as hex values for each element. The property of colors is defined by the element. There is also the possibility of adding colors for custom elements (e.g., "R" for a side chain)

**Table B.13:** Options to color atoms by type

| **option** | **default value** |
|---|---|
| colors.C | "222" |
| colors.N | "3050F8" |
| colors.O | "FF0D0D" |
| colors.H | "222" |
| colors.HE | "D9FFFF" |
| colors.LI | "CC80FF" |
| colors.BE | "C2FF00" |
| colors.B | "FFB5B5" |
| colors.F | "90E050" |
| colors.NE | "B3E3F5" |
| colors.NA | "AB5CF2" |
| colors.MG | "7CE500" |
| colors.AL | "BFA6A6" |
| colors.SI | "F0C8A0" |
| colors.P | "FF8000" |
| colors.S | "F7DE3A" |

| | |
|---|---|
| colors.CL | "1FF01F" |
| colors.AR | "80D1E3" |
| colors.K | "8F40D4" |
| colors.CA | "33D800" |
| colors.SC | "E6E6E6" |
| colors.TI | "BFC2C7" |
| colors.V | "A6A6AB" |
| colors.CR | "8A99C7" |
| colors.MN | "9C7AC7" |
| colors.FE | "E06633" |
| colors.CO | "F090A0" |
| colors.NI | "50D050" |
| colors.CU | "C88033" |
| colors.ZN | "7D80B0" |
| colors.GA | "C28F8F" |
| colors.GE | "668F8F" |
| colors.AS | "BD80E3" |
| colors.SE | "FFA100" |
| colors.BR | "A62929" |
| colors.KR | "5CB8D1" |
| colors.RB | "702EB0" |
| colors.SR | "00FF00" |
| colors.Y | "94FFFF" |
| colors.ZR | "94E0E0" |
| colors.NB | "73C2C9" |
| colors.MO | "54B5B5" |
| colors.TC | "3B9E9E" |
| colors.RU | "248F8F" |
| colors.RH | "0A7D8C" |

| | |
|---|---|
| colors.PD | "006985" |
| colors.AG | "C0C0C0" |
| colors.CD | "FFD98F" |
| colors.IN | "A67573" |
| colors.SN | "668080" |
| colors.SB | "9E63B5" |
| colors.TE | "D47A00" |
| colors.I | "940094" |
| colors.XE | "940094" |
| colors.CS | "57178F" |
| colors.BA | "00C900" |
| colors.LA | "70D4FF" |
| colors.CE | "FFFFC7" |
| colors.PR | "D9FFC7" |
| colors.ND | "C7FFC7" |
| colors.PM | "A3FFC7" |
| colors.SM | "8FFFC7" |
| colors.EU | "61FFC7" |
| colors.GD | "45FFC7" |
| colors.TB | "30FFC7" |
| colors.DY | "1FFFC7" |
| colors.HO | "00FF9C" |
| colors.ER | "00E675" |
| colors.TM | "00D452" |
| colors.YB | "00BF38" |
| colors.LU | "00AB24" |
| colors.HF | "4DC2FF" |
| colors.TA | "4DA6FF" |
| colors.W | "2194D6" |

| | |
|---|---|
| colors.RE | "267DAB" |
| colors.OS | "266696" |
| colors.IR | "175487" |
| colors.PT | "D0D0E0" |
| colors.AU | "FFD123" |
| colors.HG | "B8B8D0" |
| colors.TL | "A6544D" |
| colors.PB | "575961" |
| colors.BI | "9E4FB5" |
| colors.PO | "AB5C00" |
| colors.AT | "754F45" |
| colors.RN | "428296" |
| colors.FR | "420066" |
| colors.RA | "007D00" |
| colors.AC | "70ABFA" |
| colors.TH | "00BAFF" |
| colors.PA | "00A1FF" |
| colors.U | "008FFF" |
| colors.NP | "0080FF" |
| colors.PU | "006BFF" |
| colors.AM | "545CF2" |
| colors.CM | "785CE3" |
| colors.BK | "8A4FE3" |
| colors.CF | "A136D4" |
| colors.ES | "B31FD4" |
| colors.FM | "B31FBA" |
| colors.MD | "B30DA6" |
| colors.NO | "BD0D87" |
| colors.LR | "C70066" |

| | |
|---|---|
| colors.RF | "CC0059" |
| colors.DB | "D1004F" |
| colors.SG | "D90045" |
| colors.BH | "E00038" |
| colors.HS | "E6002E" |
| colors.MT | "EB0026" |
| colors.DS | "FFFFFF" |
| colors.RG | "FFFFFF" |
| colors.CN | "FFFFFF" |
| colors.UUT | "FFFFFF" |
| colors.FL | "FFFFFF" |
| colors.UUP | "FFFFFF" |
| colors.LV | "FFFFFF" |
| colors.UUH | "FFFFFF" |
| colors.D | "FFFFC0" |
| colors.T | "FFFFA0" |

**Other**:

**Table B.14:** Additional options to influence the rendering of the scene and specific SVG elements

| **option** | **effect** | **default value** |
|---|---|---|
| decimalPrecision | decimal precision to which SVG attributes are rounded | 8 |

| | | |
|---|---|---|
| svgElementOrder | order of elements within the SVG drawing area (later elements are drawn on top of earlier elements) | ["hydrophobicContactSelectors", "cationPiStackingsSelectors", "piStackingsSelectors", "atomPairInteractionsSelectors", "hydrophobicContacts", "cationPiStackings", "piStackings", "atomPairInteractions", "distancesSelectors", "interactionsSelectors", "bondSelectors", "atomSelectors", "annotationSelectors", "structureCirclesSelectors", "bonds", "atoms", "annotations", "structureCircles", "bondDebugTexts", "atomDebugTexts"] |
| selectorDashArray | value of "stroke-dasharray" for lines of rectangle/lasso selector (in px) | 4 |
| selectorDashWidth | value of "stroke-width" for lines of rectangle/lasso selector (in px) | 3 |
| mirrorLineWidth | line width of the line used in the mirror line mode (in px) | 0.65 |
| piPiRadius | radius of circles as part of pi-stacking interaction representations (in px) | 3.75 |

| | | |
|---|---|---|
| drawAreaPadding | space to leave between elements of the scene and the border of the drawing area (in any direction, in px) | 15 |

### B.2.3 JSON Input

This subsection describes the JSON file content (see JSON.md) generated by PoseView based on a MOL2 file of GeoMine. JSON files are provided as input to the InteractionDrawer library to draw interactive SVGs of ligand binding modes.

The JSON contains a single JSON object with just one key: "scene". Mapped to this key are several JSON arrays detailing specific diagram element types and the corresponding elements a scene is composed of.

```
{
 "scene": {
  "structures": [...],
  "atomPairInteractions": [...],
  "piStackings": [...],
  "cationPiStackings": [...],
  "hydrophobicContacts": [...],
  "annotations": [...],
 }
}
```

**Structures**:
Different structures of the scene are given as JSON objects inside a JSON array provided in the "structures" field (Table B.15). Each such JSON object has to contain a unique identifier. The structure JSON object can further contain the fields "atoms" (Table B.16), "bonds" (Table B.17), "rings" (Table B.18), and "ringsystems" (Table B.19), all optional, but potentially dependent on each other).

```
"structures": [
{
 "id": 0,
```

```
"structureName": "xyz",
"structureType": "residue",
"structureLabel": "abc",
"representation": 1,
"additionalInformation": {},
"atoms": [
{
 "id": 0,
 "element": "C",
 "label": "C",
 "coordinates": {
  "x": 5,
  "y": 5,
 },
 "color": "black",
 "charge": 0,
 "hydrogenCount": 0,
 "aromatic": false,
 "stereoCenter": false,
 "additionalInformation": {},
},
...
],
"bonds": [
{
 "id": 0,
 "from": 0,
 "to": 1,
 "type": "single",
 "aromatic": "false"
},
...
],
"rings": [
{
 "id": 0,
```

```
  "atoms": [0,1,2,3,4,5]
 },
 ...
 ],
 "ringsystems": [
 {
  "id": 0,
  "atoms": [0,1,2,3,4,5,6,7,8,9]
 },
 ...
 ]
},
...
]
```

**Table B.15:** JSON definition for structures

| **field** | **type** | **description** |
|---|---|---|
| "id" | (Number\|mandatory) | unique identifier of the structure within this scene |
| "structureName" | (String\|optional) | name of the structure |
| "structureType" | (String\|optional) | type of the structure, for example residue. This information is used to set allowed structure representations and graphical styles specific to that structure type |

| | | |
|---|---|---|
| "structureLabel" | (String\|optional) | label of the structure, which is displayed inside the structure circle of the circle representation. "structureName" is used as an alternative in case the full name is too long |
| "representation" | (Number\|optional) | initial representation of the structure (1 for skeletal representation, 2 for circle representation) |
| "additionalInformation" | (Object\|optional) | any additional information that is not required by the InteractionDrawer library directly and may be relevant to external sources during runtime |

**Atoms**:
Given as JSON array of JSON objects, each JSON object describes an individual atom.

**Table B.16:** JSON definition for atoms

| **field** | **type** | **description** |
|---|---|---|
| "id" | (Number\|mandatory) | unique identifier of the atom within this structure |
| "element" | (String\|mandatory) | element by periodic table letter code |

| | | |
|---|---|---|
| "label" | (String\|mandatory) | atom text to draw |
| "coordinates" | (Object\|mandatory) | x- and y-coordinates to place atom at |
| "color" | (String\|optional) | valid CSS color of the drawn text (otherwise deduced from atom's element) |
| "charge" | (Number\|optional) | charge of the atom to appear as text (otherwise deduced from the atom's neighbors) |
| "hydrogenCount" | (Number\|optional) | number of hydrogens bound to this atom (otherwise deduced from the atom's neighbors and charge) |
| "aromatic" | (Boolean\|optional) | whether the atom is part of an aromatic system |
| "stereoCenter" | (Boolean\|optional) | whether the atom is a stereo center (can be inferred from bond types of neighbors) |
| "additionalInformation" | (Object\|optional) | any additional information that is not required by the InteractionDrawer library directly and may be relevant to external sources during runtime |

**Bonds**:
Given as a JSON array of JSON objects, each JSON object describes an individual

bond. Provides information on which atoms are connected. Positions and colors are derived from the connected atoms.

**Table B.17:** JSON definition for bonds

| **field** | **type** | **description** |
|---|---|---|
| "id" | (Number\|mandatory) | unique identifier of the bond within this structure |
| "from" | (Number\|mandatory) | identifier of first atom the bond connects |
| "to" | (Number\|mandatory) | identifier of second atom the bond connects |

| | | |
|---|---|---|
| "type" | (String\|mandatory) | type of bond - can take either of the following values: "single", "double", "triple" (referring to the chemical bond types), "stereoFront", "stereoBack", " stereoFrontReverse", "stereoBackReverse" (for front and back facing stereo bonds, either from atom referenced in the field "to" to atom referenced in the field "from" or in backward direction), "up", " down" (for front and back facing stereo bonds with unspecified direction, which is deduced by surrounding stereo center atoms) |
| "aromatic" | (Boolean\|optional) | whether the bond is part of an aromatic system |

**Rings**:
Given as a JSON array of JSON objects, each JSON object describes an individual ring. Rings must be provided for cation-pi/pi-stacking interactions.

**Table B.18:** JSON definition for rings

| field | type | description |
|---|---|---|
| "id" | (Number\|mandatory) | unique identifier of the ring within this structure |
| "atoms" | (Array\|mandatory) | identifiers of atoms of this ring |

**Ring systems**:
Given as a JSON array of JSON objects, each JSON object describes an individual ring system. Describes the cyclic regions of the structure.

**Table B.19:** JSON definition for ring systems

| field | type | description |
|---|---|---|
| "id" | (Number\|mandatory) | unique identifier of the ring system within this structure |
| "atoms" | (Array\|mandatory) | identifiers of atoms of this ring system |

**Hydrogen bonds, metal interactions, ionic interactions ("atomPairInteractions")**:
Given as a JSON array of JSON objects, each JSON object describes an individual atom pair interaction (Table B.20). It provides information on which atoms of which structures shall be connected.

```
"atomPairInteractions": [
{
 "id": 0,
 "fromStructure": 0,
 "toStructure": 1,
 "from": 0,
```

```
 "to": 1
},
...
]
```

**Table B.20:** JSON definition for hydrogen bonds, metal interactions, and ionic interactions

| field | type | description |
|---|---|---|
| "id" | (Number\|mandatory) | unique identifier of the atom pair interaction within this scene |
| "fromStructure" | (Number\|mandatory) | identifier of the first structure to connect |
| "toStructure" | (Number\|mandatory) | identifier of the second structure to connect |
| "from" | (Number\|mandatory) | identifier of the atom to connect in the first structure |
| "to" | (Number\|mandatory) | identifier of the atom to connect in the second structure |
| "additionalInformation" | (Object\|optional) | any additional information that is not required by the InteractionDrawer library directly and may be relevant to external sources during runtime |

**Pi-Stacking Interactions ("piStackings")**:
Given as a JSON array of JSON objects, each JSON object describes an individual pi-stacking interaction (Table B.21). Provides information on which ring of which structures shall be connected.

```
"piStackings": [
{
 "id": 0,
 "fromStructure": 0,
 "toStructure": 1,
 "from": 0,
 "to": 1,
},
...
]
```

**Table B.21:** JSON definition for pi-stacking interactions

| **field** | **type** | **description** |
|---|---|---|
| "id" | (Number\|mandatory) | unique identifier of the pi-stacking interaction within this scene |
| "fromStructure" | (Number\|mandatory) | identifier of the first structure to connect |
| "toStructure" | (Number\|mandatory) | identifier of the second structure to connect |
| "from" | (Number\|mandatory) | identifier of the ring to connect in the first structure |
| "to" | (Number\|mandatory) | identifier of the ring to connect in the second structure |
| "additionalInformation" | (Object\|optional) | any additional information that is not required by the InteractionDrawer library directly and may be relevant to external sources during runtime |

**Cation-Pi Interactions ("cationPiStackings")**:
Given as a JSON array of JSON objects, each JSON object describes an individual cation-pi interaction (Table B.22). It provides information on which ring to connect with which atom.

```
"cationPiStackings": [
{
 "id": 0,
 "fromStructure": 0,
 "toStructure": 1,
 "from": 0,
 "to": 0
},
  ...
]
```

**Table B.22:** JSON definition for cation-pi interactions

| field | type | description |
|---|---|---|
| "id" | (Number\|mandatory) | unique identifier of the cation-pi interaction within this scene |
| "fromStructure" | (Number\|mandatory) | identifier of the structure containing the ring to connect |
| "toStructure" | (Number\|mandatory) | identifier of the structure containing the atom to connect |
| "from" | (Number\|mandatory) | identifier of the ring to connect |
| "to" | (Number\|mandatory) | identifier of the atom to connect |

| | | |
|---|---|---|
| "additionalInformation" | (Object\|optional) | any additional information that is not required by the InteractionDrawer library directly and may be relevant to external sources during runtime |

**Hydrophobic contacts ("hydrophobicContacts")**:
Given as a JSON array of objects, each JSON object describes an individual hydrophobic contact (Table B.23). Hydrophobic contacts are rendered as splines, which are defined by a series of control points.

```
"hydrophobicContacts": [
{
 "id": 0,
 "belongsTo": 0,
 "controlPoints": [
 {
  "x": 5,
  "y": 5,
  "atomLinks": [0, 1]
 },
 ...
 ],
 "controlPointsInsertId": 0
},
...
]
```

**Table B.23:** JSON definition for hydrophobic contacts

| field | type | description |
|---|---|---|
| "id" | (Number\|mandatory) | unique identifier of the hydrophobic contact within this scene |
| "belongsTo" | (Number\|mandatory) | identifier of the structure the hydrophobic contact interacts with |
| "controlPoints" | (Array\|mandatory) | JSON objects of control points to define the spline, which represents the hydrophobic contact in the drawing area. Each control point is defined by its position of x- and y-coordinates. Each control point can also optionally be provided with atom identifiers ("atomLinks"): if one of the atoms referenced in this JSON array is moved, the control point follows the movement. If not set, the nearest atom is automatically linked to the control point |

| | | |
|---|---|---|
| "controlPointsInsertId" | (Number\|optional) | if a hydrophobic contact with "id" already exists, the control points are added to that hydrophobic contact. This information defines the position in the hydrophobic contact where the given control points are inserted. Control points will be inserted at the end if undefined |

**Annotations**:
Given as a JSON array of JSON objects, each JSON object describes an individual annotation. Annotations represent text labels in the scene (Table B.24). They can be bound to structures/hydrophobic contacts to mimic movement applied to the associated element.

```
"annotations": [
{
 "id": 0,
 "label": "Asp86A",
 "coordinates": {
  "x": 5,
  "y": 5
 },
 "color": "black",
 "isStructureLabel": true,
 "additionalInformation": {}
 "belongsTo": {
  "type": "structure",
  "id": 0,
  "atomLinks": [0, 1]
```

```
 }
},
...
]
```

**Table B.24:** JSON definition for annotations

| field | type | description |
|---|---|---|
| "id" | (Number\|mandatory) | unique identifier of the annotation within the scene |
| "label" | (String\|mandatory) | text to draw |
| "coordinates" | (Object\|mandatory) | x- and y-coordinates to place annotation at |
| "color" | (String\|optional) | valid CSS color of the drawn text (defaults to black or the color of splines when associated with a hydrophobic contact) |
| "isStructureLabel" | (Boolean\|optional) | true if the annotation should be hidden when the corresponding structure's current representation is "circle" |
| "additionalInformation" | (Object\|optional) | any additional information that is not required by the InteractionDrawer library directly and may be relevant to external sources during runtime |

| | | |
|---|---|---|
| "belongsTo" | (Object\|optional) | binds the annotation to an element of the scene, defined by its type (either a "structure" or a spline as type "structureSpline") and its identifier. This binding can also optionally be extended by "atomLinks": if one of the atoms referenced by its identifier in this JSON array is moved, the annotation follows the movement. If not set, the nearest atom is linked to the annotation |

### B.2.4 Technical Implementation

This subsection summarizes technical implementation details of components of the InteractionDrawer library. The InteractionDrawer library's functionality requires additional JavaScript libraries:

- D3.js JavaScript library for SVG manipulation [97]
- Fraction.js JavaScript library for drawing partial atomic charges as fractions [98]
- SmilesDrawer JavaScript library for adding new structures via SMILES strings [99, 100]
- Jasmine Ruby on Rails library (https://github.com/jasmine/jasmine-gem) for writing tests for the InteractionDrawer library and executing them in a browser [101]

### B.2.5 Code

This subsection describes the code of the InteractionDrawer library regarding its file and folder structure and class dependencies. The src folder contains the library's source code. Most files represent one specific JavaScript class named like the file. The test folder contains one corresponding file with tests for each file in the src folder. The library consists of 20 folders with 105 files and 106 classes. The directory tree and included files of the InteractionDrawer library and its prototype are shown in Figure B.1 and Figure B.2, respectively.

The InteractionDrawer class instance initializes the library's basic components (User-Interactionhandler, SvgDrawer, SceneData, SvgComponent) and dependent instances to expose its main functionalities to users, like the loading and file saving of diagrams. Furthermore, it subscribes several callbacks to enable the implementation of user-defined actions. For example, the selectionCallback callback provides users information about scene objects selected via the mouse pointer to postprocess that user interaction based on individual requirements.

The dependent UserInteractionhandler class instance binds event listeners to specific elements of the drawing area. Dependent instances of specialized event handlers process corresponding user interactions. For example, the ClickSelectionHandler processes the user-performed selection of scene objects via mouse click.

Based on the performed user interactions, the UserInteractionhandler class instance and the dependent instances of specialized event handlers update the data (SceneData) and the view (SvgComponent) via a SvgDrawer class instance and dependent instances of specialized drawers. For example, the AddHanlder's AtomDrawer class instance specifically stores and draws new atoms via the SceneData-dependent AtomsData class instance and SvgComponent-dependent AtomGroupsComponent class instance.

Furthermore, the UserInteractionhandler class instance and dependent instances of specialized event handlers always track the state of current user interactions, such as the drawing area's mouse hover via the InteractionState class instance. The above-described dependency of the primary classes of the InteractionDrawer library is illustrated in Figure B.3.

```
InteractionDrawer/src
├── InteractionDrawer.js
├── UserInteractionHandlers
│   ├── UserInteractionhandler.js
│   ├── AddHandler.js
│   ├── RemoveHandler.js
│   ├── HoverHandler.js
│   ├── Movement
│   │   ├── MirrorHandler.js
│   │   ├── RotationHandler.js
│   │   └── TranslationHandler.js
│   └── Selection
│   ├── ClickSelectionHandler.js
│   ├── LassoSelectionHandler.js
│   └── RectangleSelectionHandler.js
├── Drawers
│   ├── SvgDrawer.js
│   ├── AnnotationDrawer.js
│   ├── HydrophobicDrawer.js
│   ├── IntermolecularDrawer.js
│   ├── Structure
│   │   ├── StructureDrawer.js
│   │   ├── AtomDrawer.js
│   │   ├── EdgeDrawer.js
│   │   ├── RingDrawer.js
│   │   ├── StructureCircleDrawer.js
│   │   └── StructureRepresentationDrawer.js
│   ├── TextLabelDrawer.js
│   ├── HistoryDrawer.js
│   └── ViewerDrawer.js
├── Components
│   ├── SvgComponent.js
│   ├── TransformGroupsComponent.js
│   ├── AnnotationGroupsComponent.js
│   ├── HydrophobicGroupsComponent.js
│   ├── IntermolecularGroupsComponent.js
│   ├── Structure
│   │   ├── AtomGroupsComponent.js
│   │   ├── EdgeGroupsComponent.js
│   │   └── StructureCircleGroupsComponent.js
│   ├── Utils
│   │   ├── SvgUtils.js
│   │   ├── SelectorUtils.js
│   │   ├── GroupUtils.js
│   │   ├── TextUtils.js
│   │   ├── BaseUtils.js
│   │   ├── CircleUtils.js
│   │   └── LineUtils.js
```

```
│   └── Viewer
│       ├── AnnotationFormComponent.js
│       ├── BackgroundComponent.js
│       ├── DefsComponent.js
│       ├── EditAnnotationFormComponent.js
│       ├── EditAtomFormComponent.js
│       ├── EditEdgeFormComponent.js
│       ├── EditStructureFormComponent.js
│       ├── InteractionElementsGroupComponent.js
│       ├── IntermolecularLineComponent.js
│       ├── MirrorLineComponent.js
│       ├── SelectionLineComponent.js
│       └── StructureFormComponent.js
├── Data
│   ├── SceneData.js
│   ├── AnnotationsData.js
│   ├── HydrophobicData.js
│   ├── IntermolecularData.js
│   ├── Objects
│   │   ├── Structure.js
│   │   ├── Ring.js
│   │   ├── EdgeInterfaceBased.js
│   │   ├── TextLabelBased.js
│   │   ├── Spline.js
│   │   ├── Line.js
│   │   └── VerticalLine.js
│   └── Structure
│       ├── StructuresData.js
│       ├── AtomsData.js
│       ├── EdgesData.js
│       ├── RingsData.js
│       ├── AnnotationConnectionData.js
│       ├── HydrophobicConnectionData.js
│       ├── IntermolecularConnectionData.js
│       └── RepresentationsData.js
├── DataProcessing
│   ├── JsonBuilder.js
│   ├── JsonPreprocessor.js
│   ├── JsonValidator.js
│   ├── ChangeMapCreater.js
│   ├── ClosestObjectFinder.js
│   ├── CollisionFinder.js
│   ├── EdgeBuilder.js
│   ├── RemoveCollector.js
│   ├── StructureVisitor.js
│   └── TextBuilder.js
```

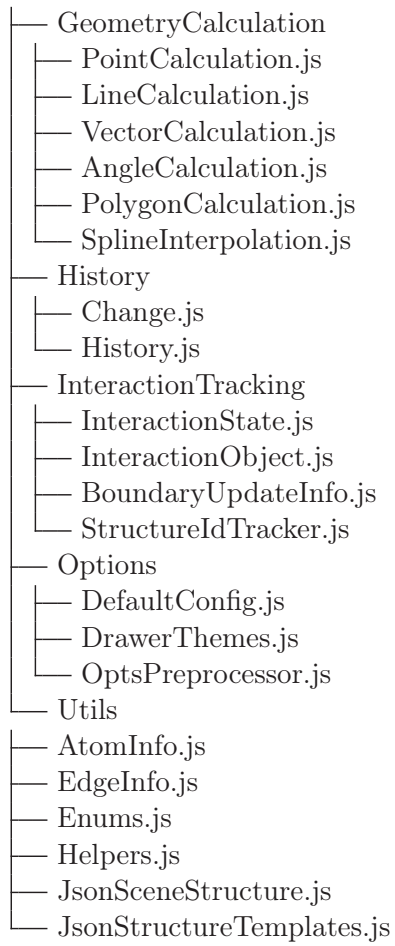



**Figure B.1:** Class/folder structure of the InteractionDrawer library

```
InteractionDrawer/src
├── Drawer.js
├── SvgDrawer.js
├── Structure.js
├── Ring.js
├── Atom.js
├── Edge.js
├── Spline.js
├── Simple_Objects.js
├── Geometry.js
├── History.js
├── Change.js
├── Interaction_Object.js
├── BoundaryUpdateInfo.js
├── DefaultConfig.js
├── Enums.js
├── Helpers.js
├── JSON_Base_Structures.js
├── StructureInfo.js
├── SMARTSFromStructure.js
├── SSSR.js
└── Scorer.js
```

**Figure B.2:** Folder structure of the InteractionDrawer library prototype

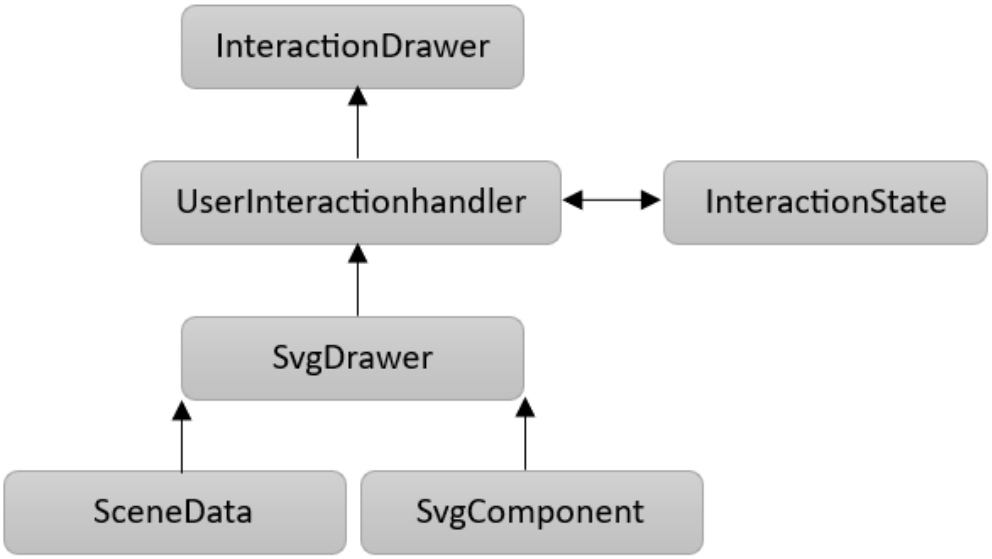



**Figure B.3:** Primary class dependency in the InteractionDrawer library

## B.3 Proteins*Plus* Web Server

### B.3.1 Usage

This subsection describes the Proteins*Plus* web server, its basic functionalities shared by all tools, and their usage. The Proteins*Plus* web server (https://proteins.plus) and the hosted computational tools are developed by the Computational Molecular Design Group (AMD) research group at the Center for Bioinformatics (ZBH). The tools of Proteins*Plus* offer a wide range of applications related to the analysis of proteins and ligands, including, for example, prediction and placement of hydrogen atoms by Protoss [33] and automated protein-ligand docking by JAMDA [31].

The Proteins*Plus* web server's landing page (Figure B.4) accepts several user-input types, including identifiers of structure entries from publicly available databases that are automatically downloaded as well as own uploaded structure files:

- four-letter identifier of a PDB database entry or Porteins*Plus* generated identifier of an uploaded PDB file (Figure B.4a)
- UniProt accession number of an AlphaFold database entry (Figure B.4b)
- PDB file (Figure B.4c)
- SDF file containing one or multiple ligands (Figure B.4d)

A keyword-based full-text search is available to query the entries of the AlphaFold and PDB databases to obtain user-input suggestions (Figure B.5). Each result list entry displays corresponding information like its title, organism, and release date. In addition, the results list can be further down-filtered based on these properties. While GeoMine works with ligand-bound and ligand-unbound input structures, PoseEdit requires ligand-bound structures as input to generate corresponding pose diagrams. All tools on Proteins*Plus*, including GeoMine and PoseEdit, can be directly used without a graphical user interface via Proteins*Plus*'s REST API [102]. The website's footer provides users with documentation for the website's and REST API's usage and further information about the Proteins*Plus* web server, including new developments.
The main page (Figure B.6) provides a 3D viewer on the left showing the input structure and a corresponding control panel located below, two scrollable lists in the middle, and a list of computational tools and corresponding descriptions on the right for analyzing the input structure.

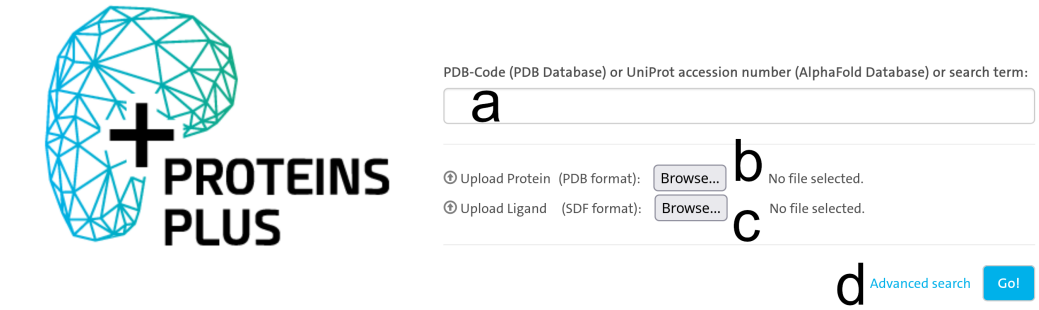



**Figure B.4:** Landing page and user-input options of Proteins*Plus*. **a** PDB identifier (PDB database) or UniProt accession number (AlphaFold database) input field. **b** structure file upload button (PDB format). **c** ligands file upload button (SDF format). **d** Linked keyword search functionality. Taken from [D1]

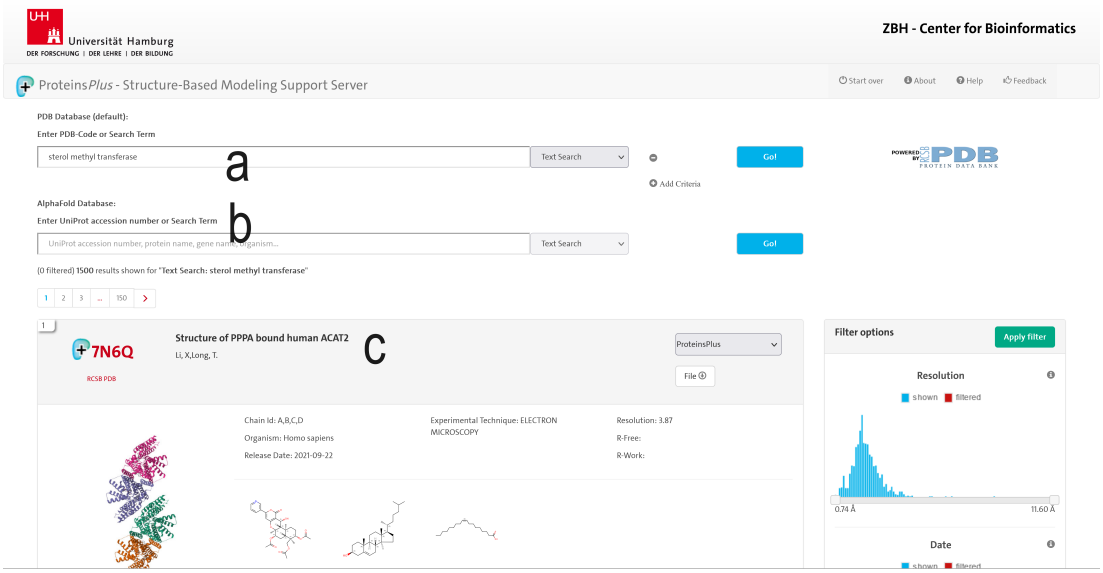



**Figure B.5:** Keyword search functionality of Proteins*Plus* demonstrated by searching the PDB with the multiword expression "sterol methyl transferase". **a** Keyword search input field for the Protein Data Bank. **b** Keyword search input field for the AlphaFold database. **c** Search results and corresponding information. Taken from [D5]

The central *Pockets* and *Ligands* lists display information about all ligands (simple ions, solvent molecules, small molecule inhibitors, etc.) of the input structure like their 2D structure diagrams and user-specified binding sites. A custom binding site can be built and loaded into the central list by users selecting residues in the 3D viewer-visualized input structure. Selecting a tool on the right, the tool-specific graphical user interface replaces the tools list. Depending on the tool, a tool's input and output are handled by the user via one or multiple of the main page's three principal user interface

components: the 3D viewer on the left visualizing the input structure, the two central *Ligands* list and *Pockets* list, and the tool-specific graphical user interface on the right. The website's header allows to change the size of the 3D viewer and to provide an email address, which is notified when calculation results of tools are available and provided with corresponding links to the results.

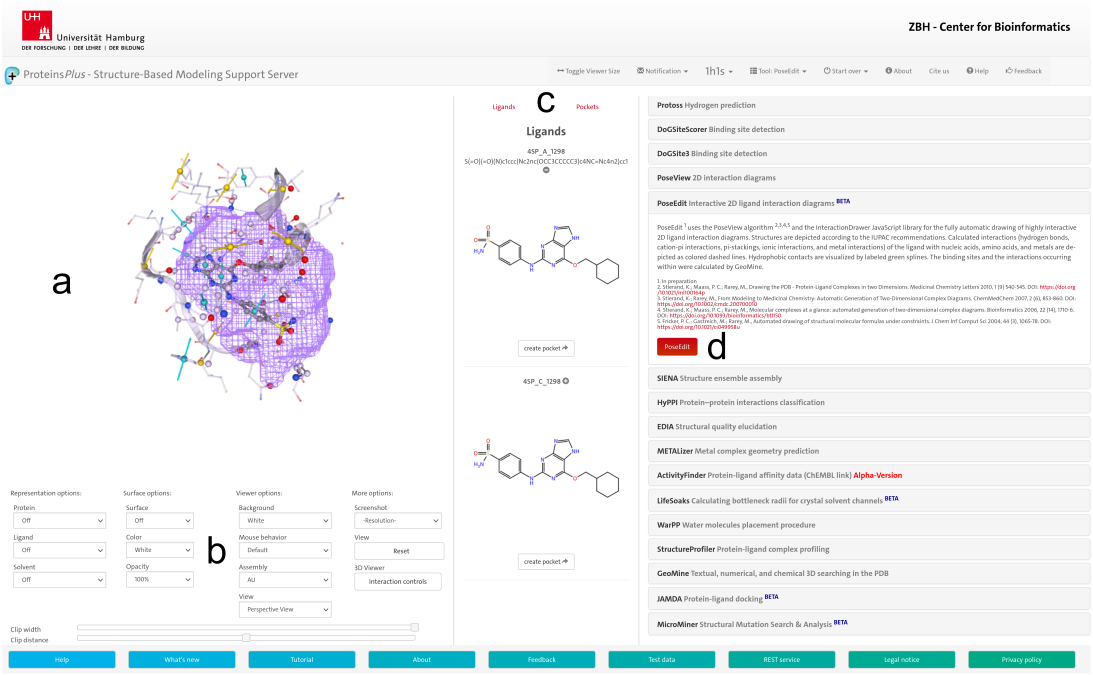



**Figure B.6:** Main page of Proteins*Plus*. **a** 3D viewer visualizing the 3D binding site of a cyclin-dependent kinase bound to an inhibitor [https://doi.org/10.1038/nsb842] (PDB code: 1H1S, Proteins*Plus* ligand identifier: 4SP_A_1298). **b** 3D viewer control panel. **c** Togglable *Pockets* and *Ligands* lists. **d** List of tools. Taken from [D1]

### B.3.2 Technical Implementation

This subsection describes implementation details of the Proteins*Plus* web server. The Proteins*Plus* web server uses the following primary backend and frontend technologies on which all hosted tools are based:

- Ruby on Rails web-app framework [103]
- MySQL database [104]
- DelayedJob Ruby on Rails library for the execution of background tasks [105]
- rspec-rails Ruby on Rails library for backend testing [106]

- rack-attack Ruby on Rails library to throttle and block abusive web server requests [107]
- HTML, Vanilla JavaScript, and the Bootstrap 3 library for general frontend design [108]
- NGL JavaScript library for the molecular 3D viewer [109–111]
- DataTables JavaScript library for creating interactive data tables [112]
- Chart JavaScript library to visualize data charts [113]
- Matomo as web analytics tool [114]
- The RCSB PDB RESTful Web Service interface to search and load PDB database entries [115]

## B.4 PoseEdit

### B.4.1 REST API Usage

This subsection describes the usage of PoseEdit's REST API. A rate limiting of 30 jobs a minute is applied to API endpoints addressing performance and security requirements. A custom PDB file can be uploaded with the following request employing the HTTP method POST. A successfully processed request returns the HTTP status code 200 and a Proteins*Plus* generated identifier for the uploaded PDB file that can be set for creating PoseEdit jobs like a PDB four-letter identifier.

```
curl -F pdb_file[pathvar]=@/path/myfile.pdb
-X POST https://proteins.plus/api/pdb_files_rest
-H "Accept: application/json"
```

A PoseEdit job is created via the HTTP method POST passing JSON data. If the request is a success, the response provides JSON data about the location and processing status of the results.

```
URL:

https://proteins.plus/api/poseview2_rest

Method:
```

```
POST

URL Params:

None

Data Params:

Required:

poseview2=[hash] - Contains the following parameters:

pdbCode=[string] - Set a four-letter identifier of the Protein Data Bank (PDB),
UniProt accession number, or a ProteinsPlus-generated identifier of a custom
PDB file obtained through its uploading.
ligand=[string] - Set the name of a ligand of the specified PDB structure.

Success Response:

Code: 200
Content: {
 status_code: 200,
 location: "",
 message: "Job already exists"
 }

OR

Code: 202
Content: {
 status_code: 202,
    message: "The job will be created in the specified location",
 location: ""
 }
```

```
OR

Code: 202
Content: {
 status_code: 202,
 message: "Job exists and is still in 'processing' state",
 location: ""
 }

Error Response:

Code: 400 BAD REQUEST
Content: {
 status_code: 400, error: "Bad Request",
 message: "Parameter values must be strings"
 }

OR

Code: 400 BAD REQUEST
Content: {
 status_code: 400, error: "Bad Request",
 message: "Invalid number of parameters or incorrect parameter name"
 }

OR

Code: 400 BAD REQUEST
Content: {
 status_code: 400, error: "Bad Request",
 message: "Invalid pdbCode"
 }

OR

Code: 400 BAD REQUEST
```

```
Content: {
 status_code: 400, error: "Bad Request",
 message: "Invalid ligand"
 }

OR

Code: 400 BAD REQUEST
Content: {
 status_code: 400, error: "Bad Request",
 message: "Job saving error"
 }

OR

Code: 400 BAD REQUEST
Content: {
 status_code: 400, error: "Bad Request",
 message: "Job loading error"
 }

OR

Code: 429 TOO MANY REQUESTS
Content: {
 status_code: 429,
 error: "Too Many Requests",
 message: "Throttle limit reached. Retry later."
 }

Sample Data:


{
 "poseview2": {
  "pdbCode":"1kzk",
```

```
  "ligand":"JE2_A_701"}
}

Sample Call (curl):


curl -d '{"poseview2": {"pdbCode":"1kzk","ligand":"JE2_A_701"}}'
-H "Accept: application/json" -H "Content-Type: application/json"
-X POST https://proteins.plus/api/poseview2_rest
```

The results location returns JSON data about a successfully processed PoseEdit job using the HTTP method GET.

```
URL

https://proteins.plus/api/poseview2_rest/:id

Method:

GET

URL Params:

Required:

id=[string]

Data Params:

None

Success Response:

Code: 200
Content: {
 status_code: 200,
 result_svg: ""
 }
```

```
OR

Code: 202
Content: {
 status_code: 202,
 message: "Job exists and is still in 'processing' state",
 location: ""
 }

Error Response:

Code: 400 BAD REQUEST
Content: {
 status_code: 400, error: "Bad Request",
 message: "Job loading error"
 }

OR

Code: 404 NOT FOUND
Content: {
 status_code: 404, error: "Not Found",
 message: "Invalid ID"
 }

OR

Code: 429 TOO MANY REQUESTS
Content: {
 status_code: 429,
 error: "Too Many Requests",
 message: "Throttle limit reached. Retry later."
 }

Sample Call (curl):
```

```
curl https://proteins.plus/api/poseview2_rest/ixenp5kLNHohrRbj56fbt4dd

Output:


result_svg - 2D diagram (SVG-file) generated by PoseView
result_json - Input JSON usable by the InteractionDrawer JavaScript library
pdbCode - Identifier of the input PDB structure
ligandName - Name of the input ligand
```

### B.4.2 Container Usage

This subsection describes the building and usage of the PoseEdit container. The following list contains the corresponding prerequisites for building and starting the container:

**Prerequisites:**

- the web server's Git repository. The server's folder includes a container folder with files needed for the build and start process.
- a container managing tool like Docker [116] or Podman [117]
- a web browser
- a running GeoMine database

**Build the container:**

- clone the Git repository
- move the .dockerignore and poseedit_webserver.dockerfile files from the server's container folder to the server's parent folder
- add to the server's bin folder some folders with the following names and corresponding tools: combinesfiles, extractligand, removeligand, preprocess, geomine, poseview2(PoseEdit), moleculejsonindexer
- go to the server's parent folder and run there the following commands to build the container and save it as a .tar file

```
docker build -f poseedit_webserver.dockerfile -t poseedit_webserver
docker save -o poseedit_webserver.tar poseedit_webserver
```

**Start the container:**

The PoseEdit container (poseedit_webserver.tar) is loaded with the following command:

```
docker load --input poseedit_webserver.tar
```

Move the poseedit folder, credentials.ini, server.env, and ./geomine.env files from the server's container folder to the server's parent folder and modify their content:

- credentials file (credentials.ini)
  - set the username and password of the GeoMine database
- environment file (geomine.env)
  - set the name and port (default: 5432) of the GeoMine database
  - set the host name/address (default: 127.0.0.1)
  - set the GeoMine license
  - set the number of web workers for simultaneously processing multiple searches (default: 3)
- environment file (server.env)
  - set the URL of the server (default: localhost:3333)
  - set the ssl protocol (default: http)

Subsequently, the web server is started with the command below:

```
docker run
--volume ./poseedit:/local/poseedit
--volume ./credentials.ini:/server/credentials.ini
--env-file ./geomine.env --env-file ./server.env
--name poseedit_webserver
-p 3333:3333
poseedit_webserver
```

The PoseEdit user interface can be accessed via a web browser on port 3333. The exposed default port of the web server (3333) can be forwarded to another port. For example, -p 4444:3333 forwards 3333 to port 4444. The container can be accessed by running the following command:

```
docker exec -it poseedit_webserver /bin/bash
```

The binary of PoseEdit is located in /server/bin/poseview2. The binary of GeoMine is located in /server/bin/geomine.

### B.4.3 Editor Features

This subsection describes all options provided by PoseEdit's 2D editor to directly modify the general graphical styles of the diagram. The list represented by Table B.25 is based on the InteractionDrawer library's configuration options, see B.2.2. Furthermore, this subsection lists all structures PoseEdit's 2D editor can add to the diagram, including amino acids, nucleic acid residues, metal ions, and water. Amino acids can be displayed with both their backbone and their side chain or only partially, depending on which of these structural parts are involved in intermolecular interactions. All structures are based on JSON templates provided by the InteractionDrawer library (see B.2.3).

**Table B.25:** Configuration options for graphical diagram styles of the PoseEdit 2D editor

| **option** | **effect** |
|---|---|
| **editor** | |
| Theme | apply a color theme (Default, Dark, Oldschool, Solarized, Solarized dark, Matrix, Cyberpunk, Gruvbox, Gruvbox dark) to the diagram to re-color all objects at once in a specific style |
| Default color | color of annotations that do not belong to a hydrophobic contact spline or with no preset color, of atoms with an unknown element (R or text labels of structures), and of intermolecular interactions and bonds with unknown type |
| Hover | hover highlighting color |
| Selection | selection highlighting color |

| | |
|---|---|
| Background | background color of the 2D scene |
| **text and lines** | |
| Text size | font size of diagram text like atom labels |
| Atom radius | minimal distance to atoms at which bonds are allowed to be visualized |
| Font family | font family of diagram text |
| Charge offset | y position of charge symbol |
| Charge size | font size of charge symbol |
| Hydrogen number offset | y position of the numbering of implicit hydrogens |
| Hydrogen number size | font size of the numbering of implicit hydrogens |
| Double bond space | space between double bond lines |
| Triple bond space | space between triple bond lines |
| Ring space | space bond lines of aromatic rings |
| Line width | width of lines |
| Gap length | gap length separating dashed line segments of intermolecular interactions and aromatic bonds |
| Dash length | length of dashed line segments of intermolecular interactions and aromatic bonds |
| **circle representation** | |
| Ligand radius | radius of ligand structure circles |
| Ligand color | color of ligand structure circles |
| Ligand border color | border color of ligand structure circles |
| Ligand opacity | color opacity of ligand structure circles |
| Ligand text color | color of the text label inside ligand structure circles |
| Non-ligand radius | radius of non-ligand structure circles |
| Non-ligand color | color of non-ligand structure circles |
| Non-ligand border color | border color of non-ligand structure circles |
| Non-ligand opacity | color opacity of non-ligand structure circles |

| | |
|---|---|
| Non-ligand text color | color of the text label inside non-ligand structure circles |
| Text label split | list of semicolon-separated characters at which text labels inside structure circles are split into new lines |

Furthermore, colors for frequently appearing atom elements (B, Br, C, Ca, Cl, Co, Cu, F, Fe, H, I, Mg, Mn, N, Ni, O, P, S, Zn) and supported intermolecular interaction types (hydrogen bond, ionic interaction, metal interaction, cation-pi interaction, pi-pi interaction, hydrophobic contact) can be set. The following structures can be added to the diagram:

- Water
- All - Backbone
- Arginine - Sidechain
- Arginine - Complete
- Asparagine - Sidechain
- Asparagine - Complete
- Aspartic acid - Sidechain
- Aspartic acid - Complete
- Cysteine - Sidechain
- Cysteine - Complete
- Glutamine – Sidechain
- Glutamine - Complete
- Glutamic acid – Sidechain
- Glutamic acid - Complete
- Histidine – Sidechain

- Histidine - Complete
- Lysine - Sidechain
- Lysine - Complete
- Phenylalanine - Sidechain
- Phenylalanine - Complete
- Serine - Sidechain
- Serine - Complete
- Threonine - Sidechain
- Threonine - Complete
- Tryptophan - Sidechain
- Tryptophan - Complete
- Tyrosine - Sidechain
- Tyrosine - Complete
- Adenosine
- Cytidine
- Guanosine
- Uridine
- Deoxyadenosine
- Deoxycytidine
- Deoxyguanosine
- Deoxythymidine
- Fe
- Ca
- Co

- Cu
- Mg
- Mn
- Ni
- Zn

## B.5 GeoMine

### B.5.1 Tool Usage

This subsection describes the command line interface functionality of GeoMine relevant to the new development achieved in this doctoral project. A description of all supported options can be obtained with the -h or –help options. The options -v or –verbosity set the level of information detail (Quiet, Error, Warning, Info, Steps) that GeoMine displays during runtime, for example, regarding the search algorithm's progress.

The first command executes GeoMine's search functionality for the web server, accepting a query in XML file format (see B.5.5) and returning search results data in JSON file format (see B.5.6):

```
./geomine -o database.sqlite -S -q input.xml -O outout.json
```

The next command generates PoseEdit diagram data for a specific PDB file and ligand, accepting a PDB file and ligand name as input and a MOL2 file to store the output (see B.1.2):

```
./geomine -o database.sqlite -M ligandname -I entry.pdb -O output.mol2
```

The next command generates binding site data for a specific PDB file accepting a PDB file as input and a JSON file to store the output (see B.5.6):

```
./geomine -o database.sqlite -Q -I entry.pdb -O outout.json
```

The last command stores the total number of contained PDB entries and binding sites of the database in a JSON string and writes it in the standard output:

```
./geomine -o database.sqlite -D
```

Adding the option -p, a PostgreSQL database can be set and searched instead of an SQLite database. A credentials file containing the username and password of the database can be provided via the option -K. Alternatively, -u and -k can be set to specify the username and password via environment variables. The PostgreSQL database's server IP address and port can be specified via the options -n and -P, respectively. Furthermore, the command-line interface can be used to create and update a database. See Table B.26 for a comprehensive description of all corresponding options.

**Table B.26:** Command line options for database creation

| **option** | **function** |
|---|---|
| -t [–threads] | number of threads employed for database creation |
| -d [–directory] | path to folder with PDB files or mmCIF files to save in the database |
| -l [–complexlist] | list of paths of PDB files to store in the database |
| -s [–chunkSizeDBCreation] | number of files that are collectively added to the database |
| -C [–complexPocketsDirName] | path to folder to store generated binding site data |
| -i [–input] | path of PDB file to add to the database |
| -r [–remove] | PDB identifier to remove from the database |
| -b [–bindingSiteType] | type of calculated binding site (radius-based or DoGSite3-based) |
| -m [–maxSASLigandRatio] | ligands with high solvent exposure are excluded (0.0 - 1.0) |
| -q [–query] | path of XML query file to search in the database. General information about the search results is displayed in the standard output |
| –statistics | path of text file to store statistics about search results |

### B.5.2 REST API Usage

This subsection describes the usage of GeoMine's REST API. A rate limiting of 30 jobs a minute is applied to API endpoints addressing performance and security requirements.

A GeoMine job is created via the HTTP method POST passing JSON data. If the request is a success, the response provides JSON data about the location and processing status of the results.

```
URL:

https://proteins.plus/api/geomine_rest

Method:

POST

URL Params:

None

Data Params:

Required:

geomine=[hash] - Contains the following parameters:

geomine_request=[string] - XML formatted query.
An XML query can be generated using the Geomine GUI.
3D query Coordinates can be optionally set for the overlay of results.
If not set, all results are superimposed onto the first result.

Success Response:

Code: 202
Content: {
 status_code: 202,
 message: "The job will be created in the specified location",
 location: ""
```

```
 }

Error Response:

Code: 400 BAD REQUEST
Content: {
 status_code: 400,
 error: "Bad Request",
 message: "Parameter values must be strings"
 }

OR

Code: 400 BAD REQUEST
Content: {
 status_code: 400,
 error: "Bad Request",
 message: "Invalid number of parameters or incorrect parameter name"
 }

OR

Code: 400 BAD REQUEST
Content: {
 status_code: 400,
 error: "Bad Request",
 message: "SMARTS pattern of ... is not correct! Please correct your pattern."
 }

OR

Code: 400 BAD REQUEST
Content: {
 status_code: 400,
 error: "Bad Request",
 message: "Query is to unspecific. Please refine the search."
```

```
 }

OR

Code: 400 BAD REQUEST
Content: {
 status_code: 400,
 error: "Bad Request",
 message: "Job saving error"
 }

OR

Code: 400 BAD REQUEST
Content: {
 status_code: 400,
 error: "Bad Request",
 message: "Job loading error"
 }

OR

Code: 429 TOO MANY REQUESTS
Content: {
 status_code: 429,
 error: "Too Many Requests",
 message: "Throttle limit reached. Retry later."
 }

 Sample Data:

 "<!DOCTYPE GeoMineFilterPresets>
 <GeoMineFilter
 xmlns:i=\"urn:naomi:InteractionDB\"
 xmlns:ip=\"urn:naomi:GeoMine\" xmlversion=\"6\"
 xmlns:propertydb=\"urn:naomi:PropertyDB\" name=\"\"
```

```
xmlns:m=\"urn:naomi:MoleculeDB\">
<propertydb:SubstringFilter_OR
rule=\"including\"
id=\"intPatterns.protein_pdb_id\"
subsettype=\"2\">
<substring_element substring=\"1KZK\"/>
</propertydb:SubstringFilter_OR>
<m:SMARTSFilter></m:SMARTSFilter>
<ip:ECFilter_Chain></ip:ECFilter_Chain>
<i:InteractionDBfilterchain>
<i:Pointfilter/>
<i:Interactionfilter/>
</i:InteractionDBfilterchain>
<ip:pointSMARTS_Chain/>
<ip:AngleFilter_Chain/>
</GeoMineFilter>"

Sample Call (curl):

curl -d
'{"geomine": {"geomine_request":
  "<!DOCTYPE GeoMineFilterPresets>
  <GeoMineFilter
  xmlns:i=\"urn:naomi:InteractionDB\"
  xmlns:ip=\"urn:naomi:GeoMine\" xmlversion=\"6\"
  xmlns:propertydb=\"urn:naomi:PropertyDB\" name=\"\"
  xmlns:m=\"urn:naomi:MoleculeDB\">
  <propertydb:SubstringFilter_OR rule=\"including\"
  id=\"intPatterns.protein_pdb_id\"
  subsettype=\"2\">
  <substring_element substring=\"1KZK\"/>
  </propertydb:SubstringFilter_OR>
  <m:SMARTSFilter></m:SMARTSFilter>
  <ip:ECFilter_Chain></ip:ECFilter_Chain>
  <i:InteractionDBfilterchain>
  <i:Pointfilter/>
```

```
   <i:Interactionfilter/>
   </i:InteractionDBfilterchain>
   <ip:pointSMARTS_Chain/>
   <ip:AngleFilter_Chain/>
   </GeoMineFilter>"}}'
-H "Accept: application/json"
-H "Content-Type: application/json"
-X POST https://proteins.plus/api/geomine_rest
```

The results location returns JSON data about a successfully processed GeoMine job using the HTTP method GET.

```
URL:

https://proteins.plus/api/geomine_rest/:id

Method:

GET

URL Params:

Required:

id=[string]

Data Params:

None

Success Response:

Code: 200
Content: {
 status_code: 200,
 result: ""
 }
```

```
OR

Code: 202
Content: {
 status_code: 202,
 message: "Job exists and is still in 'processing' state",
 location: ""
 }

Error Response:

Code: 400 BAD REQUEST
Content: {
 status_code: 400,
 error: "Bad Request",
 message: "Job loading error"
 }

OR

Code: 404 NOT FOUND
Content: {
 status_code: 404,
 error: "Not Found",
 message: "Invalid ID"
 }

OR

Code: 429 TOO MANY REQUESTS
Content: {
 status_code: 429,
 error: "Too Many Requests",
 message: "Throttle limit reached. Retry later."
 }
```

```
Sample Call (curl):


curl https://proteins.plus/api/geomine_rest/ixenp5kLNHohrRbj56fbt4dd

Output:


result -  Search statistics
          Result pockets (PDB format)
          Data for 3D viewer visualization (NGL) and table representation
          of search results (JSON format)
```

### B.5.3 Container Usage

This subsection describes the building and usage of the GeoMine container. The following list contains the corresponding prerequisites for building and starting the container:
**Prerequisites:**

- the web server's Git repository. The server's folder includes a container folder with files needed for the build and start process.
- a container managing tool like Docker [116] or Podman [117]
- a web browser
- a running GeoMine database

**Build the container:**

- clone the Git repository
- move the .dockerignore and geomine_webserver.dockerfile files from the server's container folder to the server's parent folder
- add to the server's bin folder some folders with the following names and corresponding tools: combinesfiles, extractligand, removeligand, preprocess, geomine, moleculejsonindexer
- go to the server's parent folder and run there the following commands to build the container and save it as a .tar file

```
docker build -f geomine_webserver.dockerfile -t geomine_webserver
docker save -o geomine_webserver.tar geomine_webserver
```

**Start the container:**

The GeoMine container (geomine_webserver.tar) is loaded with the following command:

```
docker load --input geomine_webserver.tar
```

Move the credentials.ini, server.env, and ./geomine.env files from the server's container folder to the server's parent folder and modify their content:

- credentials file (credentials.ini)
  - set the username and password of the GeoMine database
- environment file (geomine.env)
  - set the name and port (default: 5432) of the GeoMine database
  - set the host name/address (default: 127.0.0.1)
  - set the GeoMine license
  - set the number of web workers for simultaneously processing multiple searches (default: 3)
- environment file (server.env)
  - set the URL of the server (default: localhost:3333)
  - set the ssl protocol (default: http)

Subsequently, the web server is started with the command below:

```
docker run
--volume ./credentials.ini:/server/credentials.ini
--env-file ./geomine.env --env-file ./server.env
--name geomine_webserver
-p 3333:3333
geomine_webserver
```

The GeoMine user interface can be accessed via a web browser on port 3333. The exposed default port of the web server (3333) can be forwarded to another port. For example, -p 4444:3333 forwards 3333 to port 4444. The container can be accessed by running the following command:

```
docker exec -it geomine_webserver /bin/bash
```

The binary of GeoMine is located in /server/bin/geomine.

### B.5.4 Binding Site Calculation

This subsection provides calculation details of GeoMine's binding sites generation, indicating how they are stored in the database and visualized in the graphical user interface. The shape and structural composition of binding sites are predicted by DoGSite3 [32]. The volume of a valid binding site must be more than 100 $Å^3$. Furthermore, only the largest k pockets are included in the final set of binding sites. k equals two times the number of asymmetric unit chains in a PDB file. Valid binding site ligands are specified as molecular PDB file entries with more than 50% HETATM records and with more than five and less than 100 heavy atoms. A ligand is excluded from a DoGSite3-predicted binding site if less than 20% of its heavy atoms are contained. To obtain also binding sites for these excluded ligands, a binding site's shape and structural composition is alternatively based on a 6.5 Å radius of the ligand's heavy atoms.

Similarly, all other small molecules like water molecules and metal ions within the boundaries of a DoGSite3-predicted or radius-based binding site are included. Missing hydrogen atoms and the orientations of ambiguous amino acid side chains (asparagine, glutamine, and histidine) are predicted by Protoss [33]. Hydrogen atoms are necessary to determine the presence of hydrogen bonds. Various chemical atom properties and additional chemical features of interest, including intermolecular interactions, aromatic ring centers and the corresponding ring normals, secondary structure points (C$\alpha$ atoms of central or terminal amino acids of helices or $\beta$-sheet strands), and the directions of the corresponding secondary structure elements are subsequently determined for database storage and visualization. A central secondary structure point has two directions that are oriented towards the terminal secondary structure points of the same secondary structure element. For a terminal secondary structure point, one direction exists that is oriented towards the corresponding central secondary structure point. The supported intermolecular interaction types and the corresponding calculation criteria are based on the tool Pelikan [79] and are described in the following:

- hydrogen bonds: donor-acceptor distance between 2 Å - 3.8 Å, donor hydrogen-acceptor angle between -45° and 45°, donor-acceptor lone pair angle between -70° and 70°, hydrogen-lone pair distance between 0 Å and 2 Å
- cation-pi interactions: centroid-cation distance between 2 Å and 4 Å
- metal interactions: metal- coordinating atom distance between 1 Å and 3 Å
- pi-stacking interactions: centroid-centroid distance between 2.5 Å and 5 Å
- ionic interactions: anion-cation distance is within the sum of their van der Waals radii ±1 Å

Furthermore, based on the PoseView [34, 42, 43] criteria, hydrophobic contacts are specifically calculated for PoseEdit. The hydrophobic and surface-exposed residue atom-ligand atom distance must be below the sum of their van der Waals radii ±0.8 Å. Three or more ligand atoms must fulfill that criterion for one residue to form a corresponding hydrophobic contact

### B.5.5 XML

This subsection describes the XML representation of GeoMine queries. The XML of a query is defined by several XML elements and XML element-specific attributes and values enclosed by the <GeoMineFilter/> XML element. The additional enclosed XML elements include specifications for query points and the PDB filter:

- <i:InteractionDBfilterchain/>
- <ip:AngleFilter_Chain/>
- <ip:pointSMARTS_Chain/>
- <m:SMARTSFilter>
- <ip:ECFilter_Chain/>
- <ip:SimilarityFilter_Chain/>
- <ip:PocketHasLigandFilter_Chain/>
- <ip:filterSymmetricMatchesFilter/>
- <ip:RMSDfilter/>

- <ip:MaxMatchesPerPocketFilter/>
- <propertydb:PropertyFilter/>
- <propertydb:SubstringFilter_OR/>
- <propertydb:SubstringFilter_AND/>

See Table B.27 for all supported XML elements of query points and the corresponding XML attributes, valid XML attribute values, and chemical and spatial properties they represent.

All query point types represented by the XML elements <refligandfilter/>, <metalfilter/>, <aminoacidfilter/>, <nucleicacidfilter/>, and <waterfilter/> have the XML attribute id, which specifies a query point with a unique numerical identifier. A pair of query points can be combined by distances and intermolecular interactions using the enclosing <distancefilter/> and <interactionfilter> XML elements. These XML elements also have the XML attribute id for a unique numerical identifier, reference the respective numerical identifiers of the two associated query points, and specify the size of the distance range. The following illustrates an exemplary query distance between two query points representing ligand atoms in XML format.

```
<i:InteractionDBfilterchain>
 <distancefilter inter_point1="1" inter_point2="2" id="3"
inter_mindist="5" inter_maxdist="4">
  <firstpoint>
  <refligandfilter point_coord_x="15.48"
  point_coord_y="-4.73"
  point_coord_z="14.67"
  id="1" point_intertype="ANY"
  point_surface="0" point_element="6"
  point_fungroup="ANY"/>
  </firstpoint>
  <secondpoint>
  <refligandfilter "point_coord_x="18.57"
  point_coord_y="-5.56"
  point_coord_z="14.79"
  id="2" point_intertype="ANY"
  point_surface="0" point_element="7"
  point_fungroup="ANY/>
```

```
  </secondpoint>
 </distancefilter>
</i:InteractionDBfilterchain>
```

Similarly, a pair of distances and intermolecular interactions can be combined by an angle range with the enclosing XML element <AngleFilter/>. This XML element references the two corresponding numerical identifiers and specifies the size of the angle range. Ring normals of aromatic ring centers can also be included in angle ranges, setting their numerical identifier for the query angle's XML representation instead. An example of a query angle in XML format is given in the following.

```
<ip:AngleFilter_Chain>
 <AngleFilter firstinterid="3" secondinterid="4" id="9" min="30" max="90"/>
</ip:AngleFilter_Chain>
```

SMARTS expressions, which define the chemical environment of a protein or ligand query point, are defined in separate XML elements (<pointSMARTS/>) referencing the query point's numerical identifier and the SMARTS expression. The following provides an exemplary XML.

```
<ip:pointSMARTS_Chain>
 <pointSMARTS smarts="SMARTS" id="1"/>
</ip:pointSMARTS_Chain>
```

Table B.27 highlights the XML composition of textual, numerical, and chemical PDB filters. For all PDB filter except the ones with the XML elements <ip:RMSDfilter/>, <ip:filterSymmetricMatchesFilter/>, and <ip:MaxMatchesPerPocketFilter/>, the XML attribute rule can be added and set to "including" or "excluding" to specify if the PDB filter includes or excludes matching PDBs.

**Table B.27:** XML definition of query points

| **XML elements** | **XML attributes** | **XML attribute values** | **represented chemical and spatial properties** |
|---|---|---|---|
| If <refligandfilter/>, <metalfilter/>, <aminoacidfilter/>, <nucleicacidfilter/>, or <waterfilter/> | | | ligand atom, metal atom, protein atom, nucleic acid atom, water atom |
| | point_coord_z<br>point_coord_x<br>point_coord_y | all floating-point numbers | coordinates in 3D space |
| | point_element | 5 | boron |
| | | 35 | bromine |
| | | 20 | calcium |
| | | 6 | carbon |
| | | 17 | chlorine |
| | | 27 | cobalt |
| | | 29 | copper |
| | | 9 | fluorine |
| | | 53 | iodine |
| | | 26 | iron |
| | | 12 | magnesium |
| | | 25 | manganese |
| | | 28 | nickel |
| | | 7 | nitrogen |
| | | 8 | oxygen |
| | | 15 | phosphorus |

| | | | |
|---|---|---|---|
| | | 16 | sulfur |
| | | 30 | zinc |
| | point_intertype | ANY | any |
| | | ACCEPTOR | acceptor |
| | | ANION | anion |
| | | AROMATIC | aromatic ring center |
| | | CATION | cation |
| | | DONOR | donor |
| | | HYDROPHPBIC | hydrophobic |
| | | METAL | metal |
| if <aminoacidfilter/> | point_aminoacid | ALA | alanine |
| | | ARG | arginine |
| | | ASN | asparagine |
| | | ASP | aspartic acid |
| | | CSO | S-hydroxycysteine |
| | | CYS | cysteine |
| | | GLU | glutamic acid |
| | | GLN | glutamine |
| | | Gly | glycine |
| | | HIS | histidine |
| | | ILE | Isoleucine |
| | | LEU | leucine |
| | | LYS | lysine |
| | | MET | methionine |
| | | PHE | phenylalanine |
| | | PRO | proline |

| | | | |
|---|---|---|---|
| | | SER | serine |
| | | THR | threonine |
| | | TRP | tryptophan |
| | | TYR | tyrosine |
| | | VAL | valine |
| | | HYDROPHOBIC | hydrophobic |
| | | POLAR | polar |
| | | AROMATIC | aromatic |
| | | ACIDIC | acidic |
| | | BASIC | basic |
| | | NEUTRAL | neutral |
| | point_-aminoacid_class | HYDROPHOBIC | hydrophobic |
| | | POLAR | polar |
| | | AROMATIC | aromatic |
| | | ACIDIC | acidic |
| | | BASIC | basic |
| | | NEUTRAL | neutral |
| | point_secstruct | UNKNOWN | unknown |
| | | HELIX | $\alpha$-helix |
| | | SHEET | $\beta$-sheet |
| | | HELIX_END | C-$\alpha$ of terminal amino acid of a helix |
| | | HELIX_C_-TERMINUS | C-$\alpha$ of C-terminal amino acid of a helix |
| | | HELIX_N_-TERMINUS | C-$\alpha$ of N-terminal amino acid of a helix |

| | | | |
|---|---|---|---|
| | | HELIX_MID | C-$\alpha$ of central amino acid of a helix |
| | | STRAND_END | C-$\alpha$ of terminal amino acid of a $\beta$-sheet |
| | | STRAND_MID | C-$\alpha$ of central amino acid of a $\beta$-sheet |
| | | NO_STRUCTURE | no secondary structure |
| if <nucleicacidfilter/> | point_aminoacid | A | adenosine |
| | | C | cytidine |
| | | G | guanosine |
| | | I | inosine |
| | | N | nucleoside |
| | | U | uridine |
| | | DA | deoxyadenosine |
| | | DC | deoxycytidine |
| | | DG | deoxyguanosine |
| | | DN | deoxynucleoside |
| | | DT | deoxythymidine |
| if <aminoacidfilter/> or <nucleicacidfilter/> | point_backbone_sidechain | UNDEFINED | undefined |
| | | BACKBONE | backbone |
| | | SIDECHAIN | sidechain |

| | | | |
|---|---|---|---|
| | point_surface | all floating-point numbers $\geq 0$ | surface atom or not and its solvent exposure in $Å^2$ |
| if <refligandfilter/> | point_fungroup | Alcohol | alcohol |
| | | Aldehyde | aldehyde |
| | | Amide | amide |
| | | Amidine | amidine |
| | | Amine | amine |
| | | Azide | azide |
| | | Ester | ester |
| | | Ether | ether |
| | | Furane | furane |
| | | Guanidine | guanidine |
| | | Ketone | ketone |
| | | Nitrile | nitrile |
| | | Phenyl | phenyl |
| | | Pyridine | pyridine |
| | | Pyrrole | pyrrole |
| | | Thiophene | thiophene |
| <pointSMARTS/> (for aminoacidfilter or refligandfilter) | smarts | smarts expression | smarts expression, which defines the chemical environment of a protein or ligand point |

**Table B.28:** XML building block to define textual, numerical, and chemical filter on the PDB

| **XML example** | **filtered PDB property** |
|---|---|
| **ligand properties** | |
| <m:SMARTSFilter rule="including"><br><SMARTS>X</SMARTS><br></m:SMARTSFilter> | X = SMARTS expression |
| <ip:SimilarityFilter_Chain rule="including"><br><SimilarityFilter simPercentage="30" simMaxOrDiameter="10" simMin="1" simUSmiles="USMILES" variant="X"/><br></ip:SimilarityFilter_Chain> | minimal ligand similarity based on an unique SMILES and fingerprint. X = CSFP, tCSFP, or ECFPlike |
| <propertydb:PropertyFilter subsettype="0" id="intPatterns.X" min="0" max="10" rule="including"/> | minimal and maximal number of an specific atom element or functional group with X = Boron, Bromine, Carbon, Chlorine, Fluorine, Iodine, Nitrogen, Oxygen, Phosphorus, Sulfur, Alcohol, Aldehyde, Amide, Amidine, Amine, Azide, Ester, Ether, Furane, Guanidine, Ketone, Nitrile, Phenyl, Pyridine, Pyrrole, or Thiophene |

| | |
|---|---|
| <propertydb:PropertyFilter subsettype="3" id="intPatterns.X" min="0" max="10" rule="including"/> | minimal and maximal values of a specific binding site property with X = MW (molecular weight), Atoms (heavy atoms), RotB (rotatable bonds), Donors, Acceptors, TPSA (Topological Polar Surface Area), logP, Volume, Charge, Hetero (hetero atoms), AromAtoms (aromatic atoms), Halogens, Inorganic (inorganic atoms), LipinskiAcceptors, EZ (stereo bonds), Cyclomatic Number, CRTB (maximal continuous path of rotatable bonds), Rings, URFs (unique ring families), AroRings (aromatic rings), MaxRSsize (maximal ring system size), MaxRing (max ring size), Max Cyclomatic Number, Ringsystems, AroRingsystems (aromatic ring systems), RS (stereo centers), Volume protein properties |
| <ip:ECFilter_Chain rule="including"><br><ECFilter ecnumber="X"/><br></ip:ECFilter_Chain> | X = Enzyme Commission number |

| | |
|---|---|
| <propertydb:SubstringFilter_OR subset-type="2" id="intPatterns.protein_uniprot_-id" "rule="including"><br><substring_element substring="%X%"/><br><substring_element substring="%X%"/><br>...<br></propertydb:SubstringFilter_OR><br><propertydb:SubstringFilter_AND subset-type="2" id="intPatterns.protein_uniprot_-id" rule="excluding"><br><substring_element substring="%X%"/><br><substring_element substring="%X%"/><br>...<br></propertydb:SubstringFilter_ AND> | X = UniProt identifier; include results matching one of the UniProt identifiers or exclude all matching ones |
| **pocket properties** | |
| <ip:PocketHasLigandFilter_Chain rule="including"><br><PocketHasLigandFilter pocketHasLi-gand="true"/><br></ip:PocketHasLigandFilter_Chain> | one ligand or more are present in the binding site |
| <propertydb:PropertyFilter subset-type="3" id="intPatterns.MultiChain" rule="including"/> | a binding site consists of multiple chains |

| | |
|---|---|
| <propertydb:SubstringFilter_OR subset-type="3" id="intPatterns.ligandname" "rule="including"><br><substring_element substring="%X%"/><br><substring_element substring="%X%"/><br>...<br></propertydb:SubstringFilter_OR><br><propertydb:SubstringFilter_AND sub-settype="3" id="intPatterns.ligandname" rule="excluding"><br><substring_element substring="%X%"/><br><substring_element substring="%X%"/><br>...<br></propertydb:SubstringFilter_AND > | X = ligand name, e.g. JE2_A_701; include results matching one of the ligand names or exclude all matching ones |
| <propertydb:PropertyFilter subsettype="3 id="intPatterns.X " min="0" max="10" rule="including"/> | minimal and maximal number of specific amino acid property or amino acid type with X = Hydrophobicity, Metal, HeavyAtoms (protein), Volume, Surface, Surface-Volume-Ratio, Hydrophobicity, Depth, Donors, Acceptors, ala_Pocket, arg_Pocket, asn_Pocket, asp_Pocket, cys_Pocket, glu_Pocket, gln_Pocket, gly_Pocket, his_Pocket, cso_Pocket, ile_Pocket, leu_Pocket, lys_Pocket, met_Pocket, phe_Pocket, pro_Pocket, ser_Pocket, thr_Pocket, trp_Pocket, tyr_Pocket, val_Pocket |
| **PDB entry properties** | |

| | |
|---|---|
| <propertydb:SubstringFilter_OR subsettype="2" id="intPatterns.protein_pdb_id" rule="including"><br><substring_element substring="%X%"/><br><substring_element substring="%X%"/><br>...<br></propertydb:SubstringFilter_OR><br><propertydb:SubstringFilter_AND subsettype="2" id="intPatterns.protein_pdb_id" "rule="excluding"><br><substring_element substring="%X%"/><br><substring_element substring="%X%"/><br>...<br></propertydb:SubstringFilter_AND> | X = PDB identifier; include results matching one of the PDB identifiers or exclude all matching ones |
| <propertydb:PropertyFilter subsettype="2" rule="including" id="intPatterns.protein_-expType0" min="X" max="X"/> | experimental source with X = 1 (X-ray), 2 (neutron diffraction), 3 (fiber diffraction), 5 (NMR solution), 6 (NMR solid-state), 8 (electron microscopy), 9 (electron crystallography), 10 (solution scattering), or 11 (unknown) |

| | |
|---|---|
| <propertydb:SubstringFilter_OR subsettype="2" rule="including" id="intPatterns.protein_species"><br><substring_element substring="%X%"/><br><substring_element substring="%X%"/><br>...<br></propertydb:SubstringFilter_OR><br><propertydb:SubstringFilter_AND subsettype="2" rule="excluding" id="intPatterns.protein_species"><br><substring_element substring="%X%"/><br><substring_element substring="%X%"/><br>...<br></propertydb:SubstringFilter_AND> | X = organism name; include results matching one of the organism names or exclude all matching ones |
| <propertydb:PropertyFilter subsettype="2" rule="including" id="intPatterns.protein_-resolution" min="0.0" max="2"/> | range of experimental resolution in Å |
| <propertydb:SubstringFilter_OR subsettype="2" rule="including" id="intPatterns.protein_pdb_title"><br><substring_element substring="%X%"/><br><substring_element substring="%X%"/><br></propertydb:SubstringFilter_OR><br>...<br><propertydb:SubstringFilter_AND subsettype="2" rule="excluding" id="intPatterns.protein_pdb_title"><br><substring_element substring="%X%"/><br><substring_element substring="%X%"/><br>...<br></propertydb:SubstringFilter_AND> | X = PDB title keyword; include results matching one of the PDB title keywords or exclude all matching ones result properties |

| | |
|---|---|
| <ip:filterSymmetricMatchesFilter filterSymmetricMatches="true"/> | returns for multiple symmetrical results (i.e. consisting of the same detected points) detected in one binding site only one result |
| <ip:RMSDfilter min="0" max="2"/> | returns only results that fulfill the specified RMSD range |
| <ip:MaxMatchesPerPocketFilter max="2"/> | returns only the specified maximal number of results detected for one binding site |

### B.5.6 JSON

This subsection describes the JSON file content of search results based on a query file in XML format and of binding site data generated for a specific PDB file as a query template, see Table B.29-30 and Table B.31-32. Query template data provides 3D binding sites for 3D query selection and corresponding textual, numerical, and chemical binding site properties for specifying PDB filter.

#### B.5.6.1 Search Results Data


```
{
 "geometrical_search_mode": "true",
 "number_of_found_pdbs": 500,
 "number_of_found_pockets": 1000,
 "total_number_of_geometrical_results": 2000,
 "number_of_visualizable_results": 150,
 "number_of_point_pairs": 10,
 "result_table_content": [
 {
  "Class": "Hydrolase",
```

```
  "PDB": "1KZK",
  "PDB Title": "JE-2147-HIV PROTEASE COMPLEX",
  "Pocket": "JE2_A_701",
  "RMSD": "0.00",
  "Result ID": "0"
 }
 ],
 "statistics": "..."
 "ngl_visualization_data": {
  "1KZK_JE2_A_701" : {
   "pocketMetals": "...",
   "pocketRefLig": "...",
   "pocketMolecules": "...",
   "pocketResidues": "...",
   "pocketWaters": "...",
  },
  "1KZK_JE2_A_701_0" : {
   "pdbfile": "...",
   "relevantMetals": "...",
   "relevantResidues": "...",
   "relevantWaters": "...",
   "relevantRefLigAtoms": [...],
   "relevantMolAtoms": [...],
   "numberOfPocketInteractions": 30,
   "numberOfRelevantPocketInteractions": 0,
   "pocketInteractions": [...],
   "pocketRelevantInteractions": [...],
   "transformationDataMatrix4D": [...],
   "transformedFirstPointsOfPointPairsList": [...],
   "transformedSecondPointsOfPointPairsList": [...]
  }
 }
}
```

**Table B.29:** JSON definition of search results

| field | type | description |
|---|---|---|
| "geometrical_search_mode" | (Boolean) | is set to "true" if a 3D query is part of the GeoMine query and "false" if not |
| "number_of_found_pdbs" | (Number) | number of PDBs that were detected by the search |
| "total_number_of_geometrical_results" | (Number) | number of 3D query matches across all PDBs and binding sites |
| "number_of_visualizable_results" | (Number) | number of search results that can be visualized in 3D (maximum is 150) |
| "number_of_point_pairs" | (Number) | number of point-to-point constraints representing distances and intermolecular interactions in the 3D query |

| | | |
|---|---|---|
| "result_table_content" | (Array) | each JSON object in this JSON array represents one row in the results table that contains 3D visualizable matches (maximum 150) |
| "statistics" | (String) | results statistics that include not only the maximal 150 3D visualizable matches of the results table but the complete set of results |

**3D visualization data**:
Each JSON object in the JSON array of the "ngl_visualization_data" field describes one 3D visualizable result detected in one specific PDB entry and binding site. Each result entry provides the corresponding binding site in PDB format, additional data about the matching binding site, and individual 3D query matches of that binding site. Chemical structures, structural parts like functional groups and intermolecular interactions that were directly matched by the 3D query are called relevant in the JSON fields.

**Table B.30:** JSON definition of 3D visualization data of search results

| field | type | description |
|---|---|---|
| "pocketMetals" | (String) | NGL selection string for selecting all metal ions in the binding site |
| "pocketRefLig" | (String) | NGL selection string for selecting all ligands in the binding site |
| "pocketMolecules" | (String) | NGL selection string for selecting all small molecules that are not ligands or water in the binding site |
| "pocketResidues" | (String) | NGL selection string for selecting all residues in the binding site |
| "pocketWaters" | (String) | NGL selection string for selecting all water molecules in the binding site |
| "pdbfile" | (String) | PDB file content of binding site |
| "relevantMetals" | (String) | NGL selection string for all relevant metal ions in the resulting pocket |

| | | |
|---|---|---|
| "relevantResidues" | (String) | NGL selection string for all relevant residues in the binding site |
| "relevantWaters" | (String) | NGL selection string for all relevant water molecules in the binding site |
| "relevantRefLigAtoms" | (Array) | JSON array with serial identifiers of relevant atoms of the reference ligands |
| "relevantMolAtoms" | (Array) | JSON array with serial identifiers of relevant atoms of all other small molecules |
| "numberOfPocketInteractions" | (Number) | total number of intermolecular interactions the binding site |
| "numberOfRelevantPocketInteractions" | (Number) | number of relevant intermolecular interactions the binding site |

| | | |
|---|---|---|
| "pocketInteractions" | (Array) | JSON array with all intermolecular interactions of the binding site. Each JSON object stores the intermolecular interaction type, endpoint coordinates, and the interacting molecule and atom types |
| "pocketRelevantInteractions" | (Array) | JSON array with relevant intermolecular interactions of the binding site. Each JSON object stores the intermolecular interaction type, endpoint coordinates, and the interacting molecule and atom types |
| "transformationDataMatrix4D" | (Array) | JSON array with 3D points of a transformation matrix for the superimposition of the resulting binding sites onto the 3D query |

| | | |
|---|---|---|
| "transformedFirstPointsOfPointPairsList" | (Array) | JSON array with 3D coordinates of matching chemical features like atoms that are transformed onto the respective 3D query points (start points) |
| "transformedSecondPointsOfPointPairsList" | (Array) | JSON array with 3D coordinates of matching chemical features like atoms that are transformed onto the respective 3D query points (endpoints) |

#### B.5.6.2 Query Template Data


```
{
 "USmiles": {
  "4SP_C_1298": "S(=O)(=O)(N)c1ccc(Nc2nc(OCC3CCCCC3)c4N=CNc4n2)cc1"
 },
 "emptyPockets": {
  "Empty_Pocket_1": {...},
  ...
 },
 "pockets": {
  "4SP_C_1298": {
   "grid": "",
   "pocket": "",
   "numberOfPdbTemplatePocketInteractions": 26,
```

```
   "pdbTemplateAromaticRingCenters": [ ...],
   "pdbTemplatePocketInteractions": [...],
   "pocketDescriptor": {
    "Acceptors": "17",
    "Depth": "16.74",
    "Dogsite Simple Score": "0.43",
    "Donors": "13",
    "Hydrophobicity": "0.63",
    "Metal": "0",
    "Protein Heavy Atoms": "195",
    "Surface": "927.71",
    "Surface-Volume-Ratio": "1.15",
    "Volume": "806.40"
   },
   "secStruc": {...},
   "surfaceAtoms": [
   1,
   ...
   ]
  }
 },
 "surface": {
  "1": 28.35,
  ...
 }
}
```


**Table B.31:** JSON definition of query template data

| field | type | description |
|---|---|---|
| "USmiles" | (Object) | ligand names as JSON fields mapped to corresponding unique SMILES |

| | | |
|---|---|---|
| "surface" | (Object) | PDB serial atom identifiers of the atoms of all binding sites as JSON fields mapped to the corresponding solvent exposure values in $Å^3$ |

**Ligand-bound and ligand-unbound binding sites**:
Given as JSON objects of two JSON objects with the JSON fields "emptyPockets" and "pockets" respectively, each JSON object maps a binding site name to binding site data.

**Table B.32:** JSON definition of binding site data

| **field** | **type** | **description** |
|---|---|---|
| "grid" | (String) | 3D visualizable surface grid of the binding site in PDB format |
| "pocket" | (String) | 3D visualizable binding site in PDB format |
| "numberOfPdbTemplatePocketInteractions" | (Number) | total number of intermolecular interactions of the binding site |

| | | |
|---|---|---|
| "pdbTemplateAromaticRingCenters" | (Array) | JSON array of JSON objects that stores data about all aromatic ring centers of the binding site. Each JSON object contains an identifier, molecule type, and the coordinates of the ring center and ring normal endpoints |
| "pdbTemplatePocketInteractions" | (Array) | JSON array of JSON objects that stores data about all intermolecular interactions of the binding site. Each JSON object contains the intermolecular interaction type, endpoint coordinates, and the interacting molecule and atom types |

| | | |
|---|---|---|
| "secStruc" | (Object) | maps PDB sequence numbers of terminal and central residues of $\alpha$-helices and $\beta$-sheet strands included in the binding site to JSON objects that define the corresponding secondary structure point (C-$\alpha$ atoms). Each JSON object stores the atom coordinates, the endpoint coordinates of the secondary structure element's directions, residue type, secondary structure point type, and the PDB serial atom identifier |
| "pocketDescriptor" | (Object) | JSON object with fields representing binding site properties mapped to corresponding property values |

| | | |
|---|---|---|
| "surfaceAtoms" | (Array) | JSON array with PDB serial identifiers of all solvent-exposed atoms |

### B.5.7 Technical Implementation

This subsection summarizes details on the implementation of the tool GeoMine, including:

- the C++ programming language for the tool's implementation
- the Qt framework, which provides numerous C++ classes that facilitate the implementation of GeoMine [118]
- the in-house cheminformatics library NAOMI, which provides comprehensive functionalities to work with structural PDB file data and database APIs [119]
- the tool PELIKAN as code and feature basis of GeoMine [79]
- the Proteins*Plus* web server for representing the graphical user interface and handling GeoMine's functionality [36, D6, D7]
- the tool Protoss to set protonation states [33]
- the tool DoGSite3 to predict ligand-bound and unbound binding sites and collect additional corresponding data like depth and volume [32]
- PostgreSQL database to store the searchable data [120]

# Appendix C

# Journal Articles

## C.1 PoseEdit: enhanced ligand binding mode communication by interactive 2D diagrams

[D1] K. Diedrich, B. Krause, O. Berg, and M. Rarey. Journal of Computer-Aided Molecular Design 37 (2023), pp. 491–503. Available: https://doi.org/10.1007/s10822-023-00522-4. Open access article distributed under the terms of the Creative Commons CC BY license.

Journal of Computer-Aided Molecular Design (2023) 37:491–503
https://doi.org/10.1007/s10822-023-00522-4

# PoseEdit: enhanced ligand binding mode communication by interactive 2D diagrams

**Konrad Diedrich**[1] · **Bennet Krause**[1,2] · **Ole Berg**[1] · **Matthias Rarey**[1]

Received: 16 June 2023 / Accepted: 13 July 2023 / Published online: 29 July 2023


## Abstract

In this article, we present PoseEdit, a new, interactive frontend of the popular pose visualization tool PoseView. PoseEdit automatically produces high-quality 2D diagrams of intermolecular interactions in 3D binding sites calculated from ligands in complex with protein, DNA, and RNA. The PoseView diagrams have been improved in several aspects, most notably in their interactivity. Thanks to the easy-to-use 2D editor of PoseEdit, the diagrams are extensively editable and extendible by the user, can be merged with other diagrams, and even be created from scratch. A large variety of graphical objects in the diagram can be moved, rotated, selected and highlighted, mirrored, removed, or even newly added. Furthermore, PoseEdit enables a synchronized 2D-3D view of macromolecule-ligand complexes simplifying the analysis of structural features and interactions. The representation of individual diagram objects regarding their visualized chemical properties, like stereochemistry, and general graphical styles, like the color of interactions, can additionally be edited. The primary objective of PoseEdit is to support scientists with an enhanced way to communicate ligand binding mode information through graphical 2D representations optimized with the scientist's input in accordance with objective criteria and individual needs. PoseEdit is freely available on the Proteins*Plus* web server (https://proteins.plus).

**Keywords** Protein–ligand complexes · Molecular interactions · Mutable molecule visualization · 2D structure diagrams · Pose diagrams · PoseView · Protein Data Bank

## Abbreviations

| | |
|---|---|
| IUPAC | Union of Pure and Applied Chemistry |
| JSON | JavaScript Object Notation |
| TXT | Text Document |
| PDB | Protein Data Bank |
| REST API | Representational State Transfer Application Programming Interface |
| SMILES | Simplified Molecular Input Line Entry System |
| SVG | Scalable Vector Graphics |
| 3D | Three-dimensional |
| 2D | Two-dimensional |

## Introduction

In the broad field of life sciences, the analysis of ligand interactions in biomacromolecular binding sites is crucial. For instance, medicinal chemists are required to visually investigate the activity of their candidate compounds obtained by molecular docking or virtual screening during a drug design endeavor. Also, they might want to concisely present the activity of their final compounds to others in reports, presentations, or scientific publications. The use of graphical representations is a common medium for communicating such information to scientists. Despite the lack of geometrical details, two-dimensional (2D) depictions of macromolecule-ligand complexes and the corresponding interactions are widely used in scientific research and commonly preferably chosen over three-dimensional (3D) counterparts. The exploration of a binding site via a 3D viewer is usually more time-consuming, and the proper usage requires some practice. Furthermore, the amount of buried visual information in a 2D screenshot could prevent a concise overall picture of how and with what a ligand interacts. The dimensional simplification of a binding site towards a planar arrangement

✉ Matthias Rarey
matthias.rarey@uni-hamburg.de

[1] Universität Hamburg, ZBH—Center for Bioinformatics, 20146 Hamburg, Germany

[2] Present Address: Capgemini, 10785 Berlin, Germany

of its constituents limits the content of spatial information. Still, it brings the interactions of the ligand and the interaction partners into the scientist's focus. This type of visualization renders these critical aspects clearly visible and facilitates an instant overview that is not feasible in any 3D presentation.

To the best of our knowledge, there are only a few published and freely accessible tools that automatically generate 2D diagrams of ligand interactions from 3D input structures: LigPlot + [1, 2], LeView [3], and PoseView [4–6]. Furthermore, some commercial modeling and screening software packages like MOE [7] and the Python library ProLIF [8] contain related functionalities. All mentioned tools are desktop applications except PoseView, which is accessible via a web server in addition. It should also be noted that these tools are aged, given that they were released more than ten years ago. The low number, limited accessibility, and high age of the existing 2D ligand interaction visualizer indicate a potential for further development. Considering the high and continuously growing number of citations of MOE, LigPlot +, and PoseView in particular, it is evident that there is a high demand for these tools and, therefore, also a potential interest in further tool development. In our opinion, especially the tool's user interfaces would benefit from some improvements regarding design and functionality to better support key scientific tasks.

Various issues may become apparent while examining 2D macromolecule-ligand interaction diagrams, such as:

- Intersecting lines representing interactions
- Overlapping graphical objects like structural objects
- Interaction lines crossing other graphical objects like a text label
- Crowded diagrams due to numerous graphical objects located too close to each other
- Missing or unnecessary chemical information, such as an interaction with a specific residue or the protonation state of an atom
- Unattractive graphical styles in the diagram, like low aesthetic quality of the structural drawings, the font type of atom labels, etc.

Due to such deficiencies that might be objective, like in the case of graphical collisions, or more of subjective nature regarding, for instance, the chemical information content, the ligand's layout, the overall arrangement of residues or graphical styles, a ligand binding mode may not be satisfactorily represented for users. Therefore, they might consider a diagram inadequate for investigating or presenting the ligand's interactions.

The diagrams produced by the above-mentioned tools vary significantly regarding the chemical information content, graphical styles, and occurrence of objective aesthetic deficiencies. Figure 1 shows the diagrams of LeView, LigPlot +, PoseView, and MOE of the inhibitor 4-[[6-(cyclohexylmethoxy)-7H-purin-2-yl]amino]benzenesulfonamide in complex with a cyclin-dependent kinase (PDB code: 1H1S) [9] and Table 1 illustrates the diagrams generated by these tools in comparison.

Based on the differences of the diagrams, users can address the previously described issues to a limited extent by choosing the most suitable tool. Some examples are given in the following.

Users who want to display a ligand binding mode with a high level of chemical detail may consider PoseView or LigPlot + as an appropriate choice. Using PoseView, characteristics of the ligand binding mode, such as the interacting atoms, can be easily identified due to the drawing of structures in atomic detail. Moreover, structures are specified with further details like explicit polar hydrogen atoms and charge symbols. Only residues involved in hydrophobic contacts are represented by text labels that annotate splines placed around the ligand. Like PoseView, LigPlot + also draws complete structures except hydrophobically interacting residues in atomic detail but without showing further details like charges or polar hydrogen atoms. Residues with hydrophobic contacts to the ligand are drawn as labeled arcs with spikes extending in the ligand's direction.

If users are more interested in an overall picture of the binding site or in a collision-free layout, MOE and LeView should be used. MOE and LeView reduce the level of detail by showing only the ligand in atomic detail and interacting structures as circles or text labels. This approach reduces collisions and consequently permits the visualization of not only the ligand's interaction partners but also surrounding non-interacting structures. In MOE, all non-interacting residues, cofactors, and solvent molecules within a 4.5 Å cut-off radius of the ligand atoms are shown. In LeView, users can adjust the cut-off distance to include non-interacting residues and water molecules.

PoseView might be a convenient choice if the aesthetic quality of the chemical drawings is important to users. It is the only tool that strictly adheres to the Union of Pure and Applied Chemistry (IUPAC) [10] guidelines, which define a style of the depiction of chemical structures applied by the vast majority of scientists. The other tools tend to deviate from IUPAC ideals, resulting in issues like inconsistent bond lengths and angle sizes, or they draw structures as circles or text labels.

Users who want to visualize a wide variety of interaction types of a ligand binding mode may consider PoseView and MOE as the best options. Both tools include hydrophobic contacts, hydrogen bonds, pi-pi interactions, cation-pi interactions with protein and nucleic acid residues, and interactions with metals. LigPlot + considers fewer interaction types, including hydrophobic contacts and hydrogen bonds

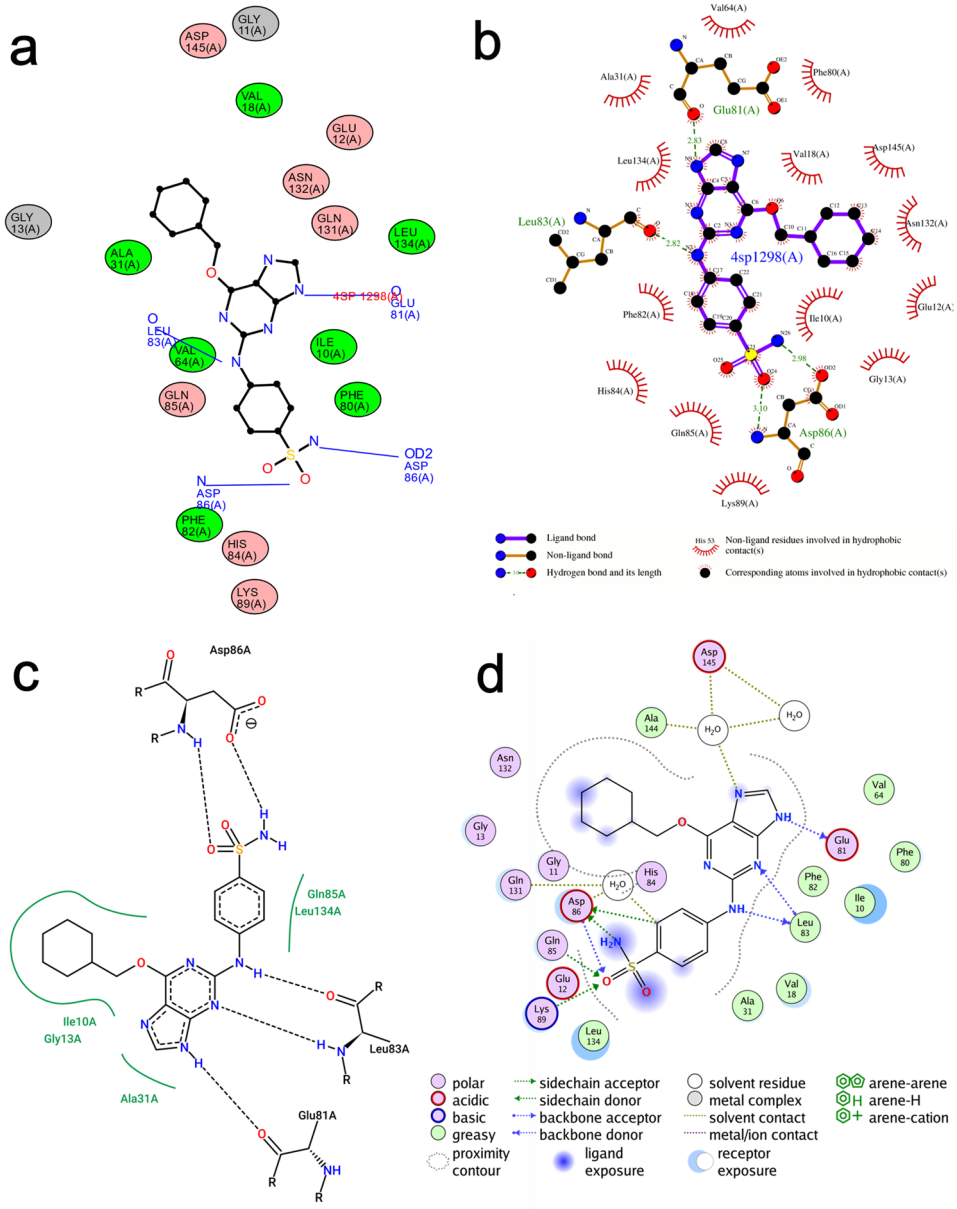



**Fig. 1** Diagrams of different tools of the inhibitor 4-[[6-(cyclohexylmethoxy)-7H-purin-2-yl]amino]benzenesulfonamide in complex with a cyclin-dependent kinase (PDB code: 1H1S). **a** LeView. **b** LigPlot+. **c** PoseView. **d** MOE

to protein and nucleic acid residues and metal interactions. LeView depicts hydrogen bonds only. In addition to direct hydrogen bonds, MOE, LigPlot+, and LeView can also display hydrogen bonds to residues that are mediated by water.

Despite the wide range of options offered by these tools, users may fail to find a viable workaround to the various issues previously described or be forced to an unsatisfactory compromise. For instance, users might want to reduce collision in a highly complex diagram by choosing MOE. Still, they want to keep certain favored aspects of other tools like the IUPAC-based depiction style of PoseView. In both cases, the users are forced to accept the diagram's graphical styles, chemical information content, and objective aesthetic deficiencies.

The manual modification of a diagram after its generation is an approach that could help users to satisfy objective aesthetic requirements and subjective preferences about what is displayed and how. For example, the appropriate rearrangements of the diagram content could resolve intersections, overlaps, and overcrowded scenes. While the diagrams of PoseView are static, this approach is feasible to a varying extent in LeView, LigPlot+, and MOE, which provide interactive diagrams through a 2D editor interface. Table 2 presents a comparison across the editing features offered by these three tools.

With the intent to create the most user-friendly and useful frontend possible for the manual post-processing of 2D ligand interaction diagrams, we compiled a list of features that specifically address the issues mentioned

**Table 1** Comparison of the diagrams created by LeView, LigPlot+, PoseView, and MOE

| | PoseView | LeView | LigPlot+ | MOE |
|---|---|---|---|---|
| Ligand representation | - Skeletal | - Skeletal | - Skeletal | - Skeletal |
| Interaction partner representation | - Skeletal | - Circle | - Skeletal | - Circle |
| Multiple ligands | – | – | + | – |
| IUPAC | + | – | – | – |
| Hydrogen bonds | + | +($H_2O$-mediated) | +($H_2O$-mediated) | +($H_2O$-mediated) |
| Hydrophobic contacts | + | Near residues | + | Near residues |
| pi–pi | + | – | – | + |
| Cation-pi | + | – | – | + |
| pi-H | – | – | – | + |
| Ionic | – | – | – | + |
| Metal | + | – | – | + |
| Covalent bonds | – | – | + | + |
| Charges | + | – | – | + |
| Explicit hydrogens | + | – | – | + |
| Bond order assignment | Automated | CIF-based | CIF-based | Automated |

**Table 2** Comparison of the 2D editor interfaces of PoseView, LeView, LigPlot+, and MOE

| | PoseView | LeView | LigPlot+ | MOE |
|---|---|---|---|---|
| Diagram modification features | | | | |
| Graphical styles (sizes, colors, etc.) | – | + | + | + |
| Interactive object types | – | – Structures | – Structures<br>– Atoms<br>– Text labels | – None |
| Object translation | – | + | + | – |
| Object rotation | – | – | + | – |
| Object removal | – | + | – | – |
| Mirror structure at bond | – | – | + | – |
| Merge multiple diagrams | – | – | + | – |
| Usability features | | | | |
| Editing history | – | – | +(undo of the last ten structural movements) | – |
| Diagram export | PDF, PNG, SVG | PNG, JPG, GIF, PDF, SVG, EPS, TXT | PS, DRW | PNG, JPG, EPS, PS, BMP, TIF, EMF+, SVG |
| Diagram import | – | – | DRW | – |
| 3D visualization | – | – | + | + |
| Interactions list | – | + | + | + |
| Diagram legend | – | – | + | + |
| Diagram rotation | – | + (45° intervals) | – | + |
| Diagram translation | – | – | + | – |
| Zoom in/out of diagram | – | – | + | – |
| Diagram reset | – | + | – | + |
| Diagram recentering | – | – | + | – |

above, as well as the respective limitations of the existing tools. Based on that list, we extended PoseView, resulting in a new graphical frontend PoseEdit, which we present in this paper. In addition, we also aimed to address some graphical and informational shortcomings in the PoseView diagrams and consequently modified those in this regard. In the following, we will primarily focus on the newly built 2D editor of PoseEdit and its features. We will then showcase the usage of the 2D editor and discuss

the benefits of its features for improving interactive 2D ligand interaction diagrams from a user's perspective.

## Methods

### Features

The PoseView diagrams have been enhanced regarding graphical style and chemical information content, but most importantly, their interactivity. The key features of PoseEdit and its improvements over PoseView and the other tools are summarized in the following and include:

- A maximized accessibility through its implementation as a web application, which is freely accessible as part of the Proteins*Plus* [11–13] web server's tool collection (https://proteins.plus)
- Interactive diagrams presented through a 2D editor with an intuitive interface design
- A large variety of interactive objects in the diagram, including all structures (the ligand, metal ions, protein and nucleic acid residues) and their atoms and bonds, hydrophobic contact splines and their spline control points, interactions, and text labels
- Extensive manual modification options through the translation, rotation, highlighting, hiding, mirroring, adding, removing, and editing of visualized chemical properties and graphical styles of interactive diagram objects
- The merging of multiple diagrams
- The export of the diagram and its legend in Scalable Vector Graphics (SVG) format
- Additional 2D editor features for a user-friendly overall diagram editing experience, such as the zooming, translation, rotation, and recentering of the diagram, the reset of the diagram to its initial unmodified state, the selection of multiple interactive diagram objects for editing them as a group, an editing history enabling undo/redo of all user changes and the export of diagrams in the JavaScript Object Notation (JSON) format that can be reimported for sharing and further editing
- An improved comprehension and exploration of the ligand binding mode through several interactive info sections of the editor and a simultaneously and synergistically inspectable 3D representation, which is synchronized with the 2D ligand interaction diagram
- An exportable JSON and Text Document (TXT) file that can be parsed for obtaining corresponding textual information
- An increased aesthetic quality of the PoseView diagram due to graphical style choices such as the depiction of bonds by using a color gradient, a minimal atom radius, within which bonds and interactions are not allowed to extend such that collisions are reduced, the drawing of interactions by colored lines with dashes of equal length and the visualization of amino acid side chains up to the Cα atom
- A more detailed description of the ligand binding mode by the depiction of covalent bonds of the ligand to residues and a new reparametrized interaction model based on the tools GeoMine [14–16] and Protoss [17, 18] that also annotates ionic interactions with residues and assigns pi-pi and cation-pi interactions to single aromatic rings rather than to entire aromatic ring systems

### Proteins*Plus* user interface and PoseEdit integration

The input for PoseEdit is provided on the Proteins*Plus* landing page (Fig. 2) through the specification of a Protein Data Bank (PDB) [19] identifier, UniProt accession number for accessing a structure in the AlphaFold [20] database (Fig. 2a), or by the upload of a structure file in the PDB format (Fig. 2b). Additionally, users can upload ligands in the Structural Data File (SDF) format that are docked into a binding site of the input structure (Fig. 2c). If the users do not have a structure of interest yet, they can obtain a list of potential input candidates by querying the PDB or AlphaFold databases with keywords via the linked keyword search

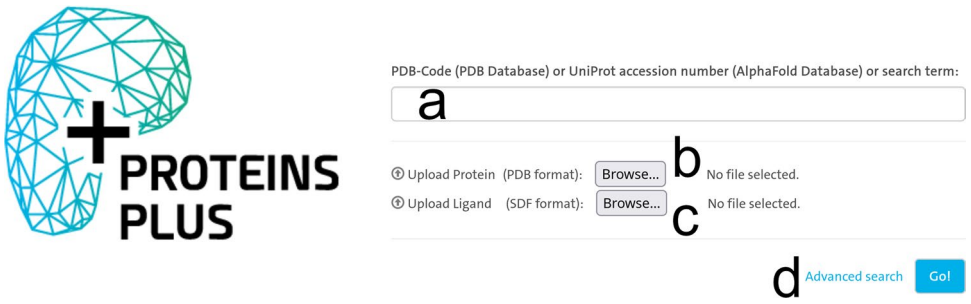



**Fig. 2** Input area of the landing page of the Proteins*Plus* web server. **a** Text field for the specification of an input structure via a Protein Data Bank identifier or UniProt accession number for the AlphaFold database. **b** Upload button for the upload of a structure file in PDB format. **c** Upload button for additional ligands in SDF format. **d** Link to the keyword search functionality

functionality (Fig. 2d). As an alternative to the Proteins*Plus* web site, PoseEdit can also be used in a more direct and automated way via the Representational State Transfer Application Programming Interface (REST API) of Proteins*Plus*, whose usage documentation can be found on the web page.

After Protein*Plus* has preprocessed the input, users are forwarded to the main page (Fig. 3), which is divided into three primary sections: the 3D visualization section (Fig. 3a) on the left shows the input structure in a 3D viewer, which can be set up via a control panel below (Fig. 3b). The users can switch between two scrollable lists in the central section (Fig. 3c). The names, Simplified Molecular Input Line Entry System (SMILES) strings, and 2D diagrams of all ions and small molecules of the input structure such as solvent molecules, cofactors, and inhibitors are included in the *Ligand list*. The *Pocket list* contains empty and ligand-bound binding sites that are calculated on-the-fly from the input structure [21] and which can be separately visualized in the 3D viewer, along with further highlighted details such as the intermolecular interactions. In the tools section (Fig. 3d) on the right, users can select PoseEdit from the tool list, specify an input ligand from the *Ligand list* and start the diagram calculation. After the calculation is finished, the 2D editor of PoseEdit with the 2D ligand interaction diagram appears in the tools section. Furthermore, a link for the later retrieval of the calculated and unmodified diagram is provided.

## The PoseEdit editor

The 2D editor of PoseEdit (Fig. 4) provides a top panel that consists of an info section with two toolbars below. The info section on the top (Fig. 4a) lists the names of all structures of the diagram. The two toolbars below contain buttons labeled with text and icons that indicate their functions. Users can select a diagram editing mode in the toolbar at the top (Fig. 4b). All modes are described in Table 3. A mode is activated by clicking its corresponding button, whose color then turns blue. Modes with an inverted triangle icon next to the text label of the corresponding button require further specification by the users for activation. When users click on such a mode button, a drop-down list appears allowing users to choose a mode-specific option. For example, users can define whether the Move mode affects single atoms, bonds, rings, or the complete structure. The activated mode can be applied by performing the

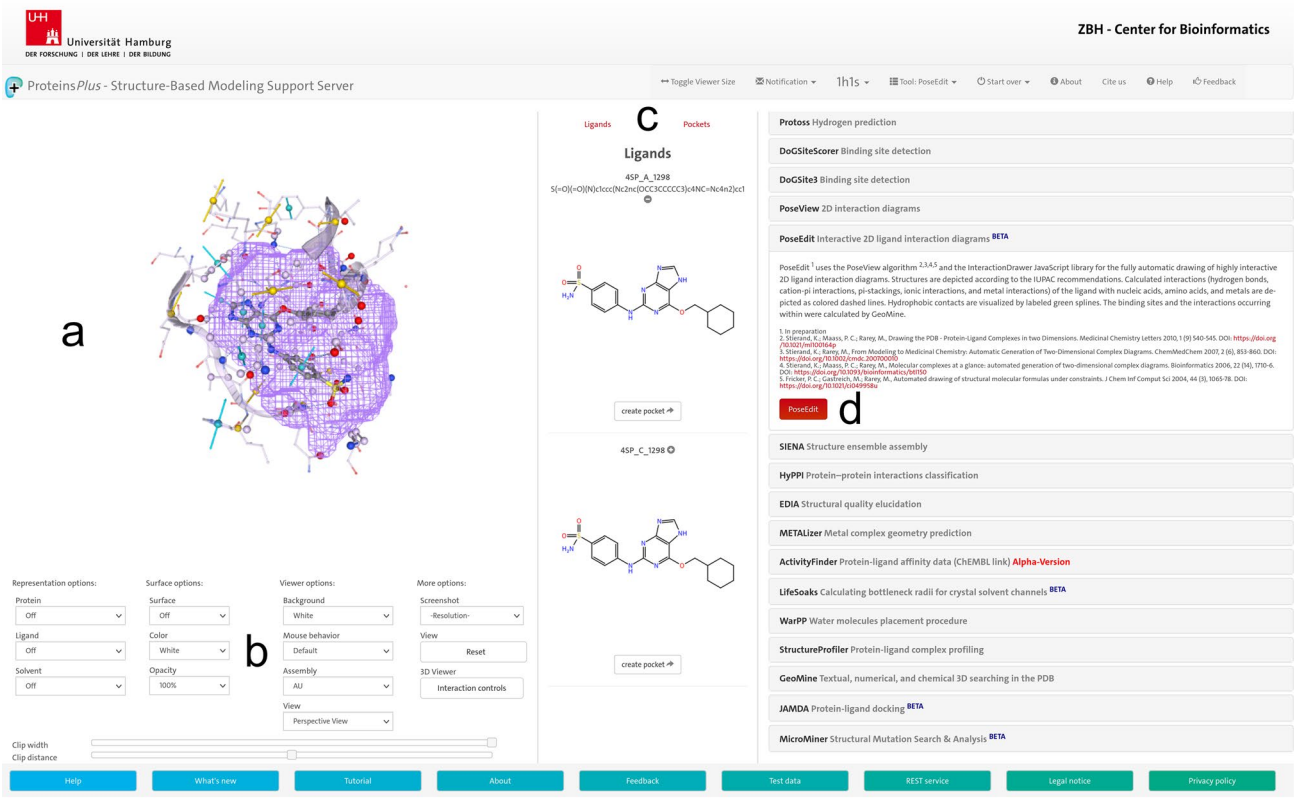



**Fig. 3** Main page of the Proteins*Plus* web server. **a** 3D viewer showing a 3D binding site of the inhibitor 4-[[6-(cyclohexylmethoxy)-7H-purin-2-yl]amino]benzenesulfonamide in complex with a cyclin-dependent kinase (PDB code: 1H1S). **b** 3D viewer control panel. **c** Togglable lists with ligands and on-the-fly calculated binding sites of the input structure. **d** list of tools, e.g., PoseEdit

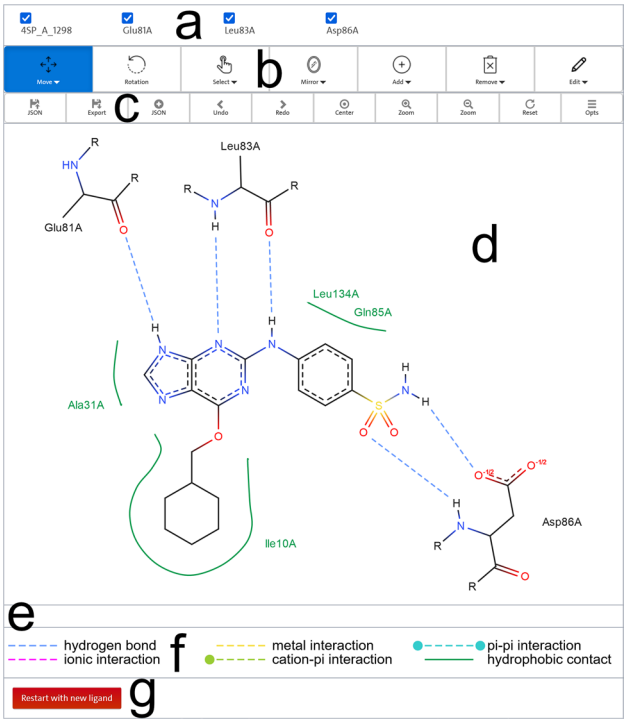



**Fig. 4** 2D editor of PoseEdit showing a diagram of the inhibitor 4-[[6-(cyclohexylmethoxy)-7H-purin-2-yl]amino]benzenesulfonamide with the internal Proteins*Plus* ID 4SP_A_1298 interacting via hydrogen bonds and hydrophobic contacts with a cyclin-dependent kinase (PDB code: 1H1S). **a** Info section that contains the names of all structures in the diagram. **b** Buttons for the activation of a diagram editing mode. **c** Buttons for the handling of diagram files and additional editor controls. **d** Drawing area displaying the 2D ligand interaction diagram. **e** Info section that shows information about atoms, bonds, and structures hovered over with the mouse pointer in the diagram or 3D viewer. **f** Legend that illustrates the supported interaction types. **g** Restart button

required left mouse click and click-and-drag operations in the drawing area. The second toolbar (Fig. 4c) below contains buttons for downloading and uploading a diagram in different file formats and buttons that execute actions that directly modify the drawing area, such as the reset of the diagram to its initial unmodified state. Below the top panel is the drawing area (Fig. 4d) that shows the calculated 2D ligand interaction diagram and two additional info sections. The first one (Fig. 4e) displays information about atoms, bonds, and structures that are hovered over with the mouse pointer in the diagram or the 3D viewer. The second section (Fig. 4f) contains a legend that illustrates the supported intermolecular interaction types and their corresponding colors. A new PoseEdit calculation can be performed with different ligands from the *Ligand list* by clicking the restart button below (Fig. 4g). By moving the mouse pointer over any control element of the 2D editor, such as the button of a diagram editing mode, a tooltip with corresponding usage information appears.

### Technical implementation details

The frontend was developed with HTML, Vanilla JavaScript, and the Bootstrap 3 library (https://getbootstrap.com). The 2D diagrams are implemented by Scalable Vector Graphics. SVGs are created and rendered interactive by the InteractionDrawer JavaScript library, which is based on D3.js (https://d3js.org). The InteractionDrawer library was newly developed for that purpose, and its code is available on GitHub (https://github.com/rareylab/InteractionDrawer). The SMILES parsing, required for adding new structures specified by SMILES, is achieved by integrating the SmilesDrawer [22] JavaScript library. The 3D viewer is implemented with the NGL library [23, 24]. The Ruby on Rails framework (https://rubyonrails.org) was used to develop the backend of the webserver.

## Application

Our showcase of PoseEdit's features is based on the structure of a lysine-specific histone demethylase 1A (LSD1, PDB code: 5LGT) in complex with the inhibitor 4-methyl-*N*-[2-[[4-(1-methylpiperidin-4-yl)oxyphenoxy]methyl]phenyl]thieno[3,2-b]pyrrole-5-carboxamide (Proteins*Plus* identifier: 6W3_A_902) and an flavin adenine dinucleotide (FAD) cofactor (Proteins*Plus* identifier: FAD_A_901) in the same binding site [25]. First, we will demonstrate how users can verify the chemical information content of a diagram. Next, we will show how users can optimize a diagram according to objective layout quality issues. Last, we will exemplify how users can further customize a diagram by editing its chemical information content and graphical restyling.

### Verification of the chemical information content

The affinity of a ligand does not depend on user's taste. What may depend on user's taste is the degree of focus to put on the various interactions contributing to ligand affinity. Interaction models are based on different studies and apply various different criteria to decide on the presence or absence of an interaction. The choice of the supported interaction types, their geometric parametrization, and the structure types as interaction partners may not always match the user's expectations. Depending on the individual thresholds, experienced users might come to different assessments on the presence of specific interactions. Therefore, they might be skeptical that the automatically generated diagram accurately represents the chemical information they would have picked themself, or they might already be aware of discrepancies. This section will show how to explore the inhibitor's environment beyond the pre-calculated diagram, providing the

**Table 3** Description of PoseEdit's diagram editing modes

| mode | options | function |
|---|---|---|
| Move | Structure freedom level | Move the scene, a structure, structure circle, hydrophobic contact spline and its control points, or annotation. When a structure or structure circle is moved, all linked hydrophobic contact splines and annotations are also moved. When *Structure freedom level* is set to *Atoms and bonds* or *Rings*, the mode affects not the complete structure but its atoms and bonds or rings, respectively |
| Rotation | – | Rotate the scene, a structure, structure circle, or hydrophobic contact spline around their midpoints. When a structure or structure circle is rotated, all linked hydrophobic contact splines and annotations are also rotated |
| Select | Click<br>Lasso<br>Rectangle | Select objects in the drawing area by mouse click or with a rectangular or lasso selection tool. Deselect an object by clicking on it again and deselect everything by clicking in the blank of the drawing area. Selected objects are highlighted, which is synchronized with the 3D viewer and visible in the downloadable SVG. Selected atoms, bonds, structures, and structure circles can be moved, rotated, and removed together |
| Mirror | Bond<br>Line | Mirror a structure at a bond or a structure or hydrophobic contact spline at a user-defined line that goes through its midpoint. When a structure is mirrored, all linked hydrophobic contact splines and annotations are also mirrored |
| Add | Annotation<br>Atom with covalent bond Atom–atom interaction Cation-pi interaction<br>Pi–pi interaction<br>Explicit H with covalent bond<br>Hydrophobic contact<br>Structure | Specify an object type and add a new object of this type to the diagram. For an atom, annotation, or structure, several properties can be specified via a form |
| Remove | Structure freedom level | Remove a structure, structure circle, hydrophobic contact spline or its control points, annotation, or interaction. When a structure or structure circle is removed, all linked interactions, hydrophobic contact splines, and annotations are also removed. When *Structure freedom level* is set to *Atoms and bonds* or *Rings*, the mode affects not the complete structure but its atoms and bonds or rings, respectively |
| Edit | Annotation<br>Atom<br>Bond<br>Structure | Specify an object type and edit the properties of a diagram object of this type via a form |

users with ideas to modify its chemical information content with PoseEdit. Figure 5 shows the PoseEdit diagram of the inhibitor 6W3_A_902 in complex with its target automatically generated by PoseView. The exported diagram in JSON format is available in the Supplementary Information (Online Resource 1).

Users can load and inspect the pose with the ligand-associated 3D binding site from the central *Pocket list* (Proteins-*Plus* pocket identifier: FAD_A_901_6W3_A_902) in the 3D viewer. The PoseEdit diagram of the inhibitor is an excerpt from this 3D binding site. While the diagram displays ligand interactions with protein and nucleic acid residues and metals, the 3D binding site also shows all non-interacting structural elements and additional interaction partners that are not included in a PoseEdit diagram by default, such as water molecules. Therefore, the 3D binding site is a suitable starting point for verifying the chemical information content of the diagram resulting in ideas of how to extend it. The 2D-3D synchronization feature supports the user's exploration of the ligand binding mode in both visualizations. Structural diagram objects in the 2D editor that are selected via the *Select mode* and consequently highlighted in dark green are automatically focused and highlighted in the 3D viewer as well. In addition, when users place the mouse pointer over any unselected or selected structural diagram object in the 2D editor or 3D viewer, it is highlighted with a light green color in both depictions. A *Select mode* option dictates how and what is selected. Users can select multiple atoms, bonds of structures, and structure circles via a rectangular and lasso selection tool with the *Rectangle* and *Lasso* option. With the *Click* option, users can pick single atoms, bonds, structure circles, as well as text labels.

Users can, for instance, select the atoms and bonds of all structures and the three text labels of the hydrophobic residues His564A, Phe538A, and Ala539A to highlight the corresponding structures in the 3D viewer. This feature enables an easier comprehension of already covered aspects of the 3D binding site and potential additional chemical information to be included. Interesting chemical information in the 3D binding site that is not depicted in the PoseEdit diagram

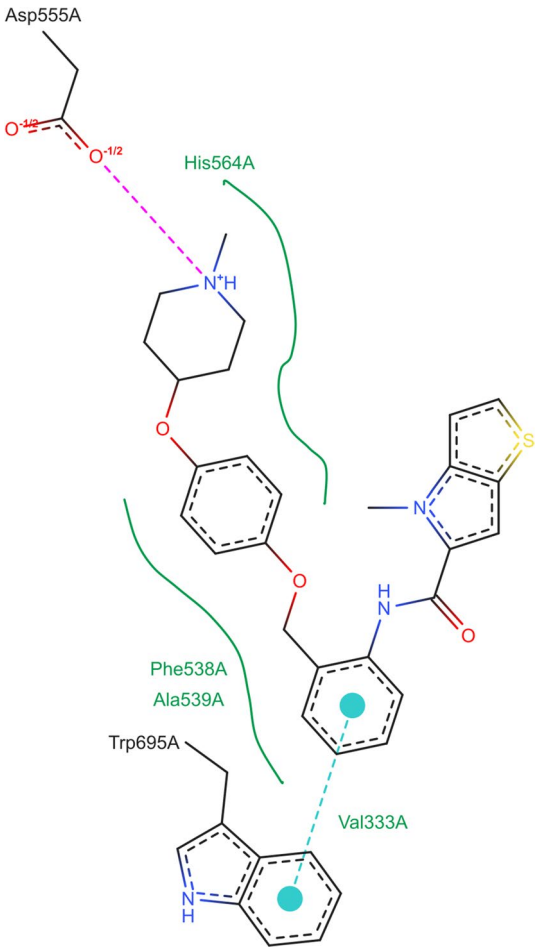



**Fig. 5** PoseEdit diagram of lysine-specific histone demethylase 1A in complex with an inhibitor 4-methyl-*N*-[2-[[4-(1-methylpiperidin-4-yl)oxyphenoxy]methyl]phenyl]thieno[3,2-b]pyrrole-5-carboxamide (PDB code: 5LGT; Proteins*Plus* identifier: 6W3_A_902). The following ligand interactions are depicted in the diagram: an ionic interaction with residue Asp555A, a pi–pi interaction with residue Trp695A, hydrophobic contacts with residues His564A, Phe538A, Ala539A, and Val333A

is, for instance, the ligand FAD_A_901, the FAD cofactor. This cofactor interacts not only with the protein binding site via ionic interactions and hydrogen bonds but also with the inhibitor via pi-stacking interactions. Interactions with cofactors other than metal like FAD are not included in a PoseEdit diagram by default but might be relevant. The following section will provide more insights into the ligand binding mode of the cofactor by inspecting its PoseEdit diagram.

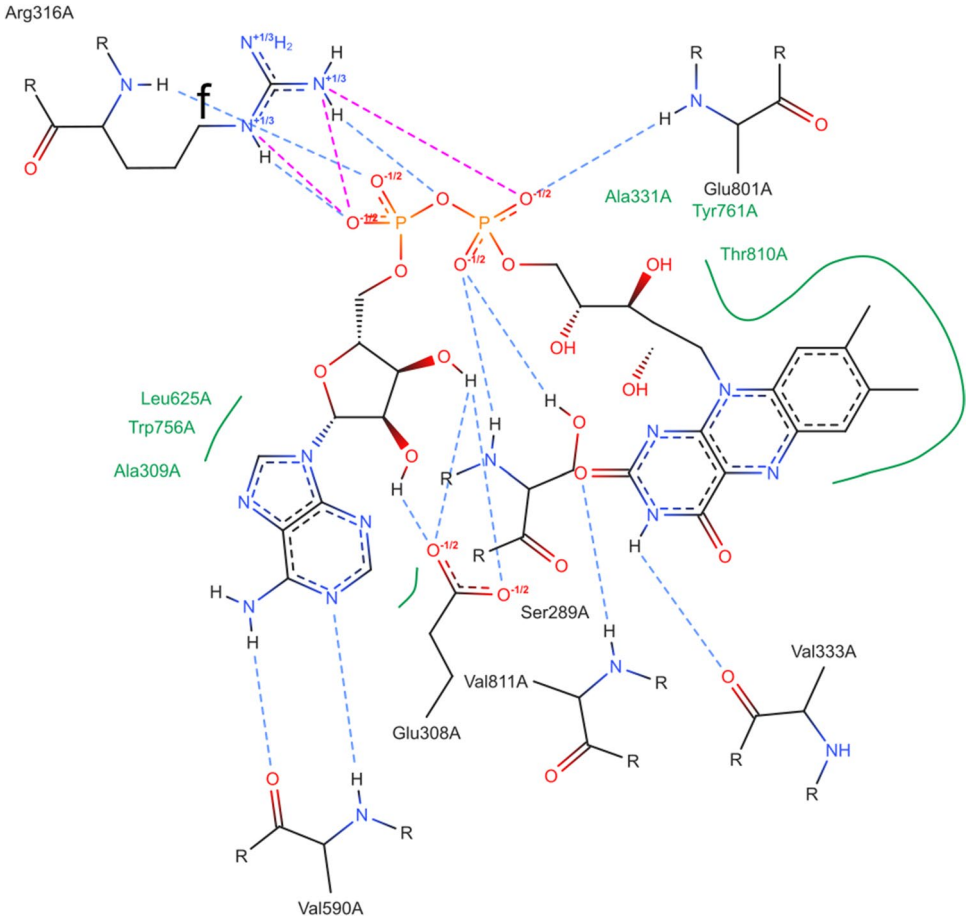
## Fixing of objective aesthetic deficiencies

Concerning the occurrence of overlaps and intersections of graphical objects, especially those diagrams that condense a large amount of chemical information closely arranged in 3D space may need to be manually revised. A highly complex interaction pattern makes it algorithmically more challenging to depict a diagram in 2D, which may result in a lower layout quality. The diagram of the FAD cofactor mentioned previously is an example of such an objectively suboptimal layout caused, for example, by the adjacent and non-planar diphosphate group undergoing numerous hydrogen bond and ionic interactions. Such functional groups often contribute to crowded diagrams. Figure 6 shows the cofactor's unmodified PoseEdit diagram. Figure 7 shows the diagram after manual optimization of the layout. The corresponding JSON files can be accessed in the Supplementary Information (Online Resource 2 and 3). A screen recording video that illustrates the following textually described diagram optimization procedure is given in Online Resource 4.

As the necessary modifications for optimizing a diagram may not be immediately visible, users might have to experiment with the editor's functionality. The editor's history, which is accessible via the *Undo* and *Redo* buttons enables, for instance, a trial-and-error approach. In addition, the possibility of hiding structures in the diagram via the editor's top structure list helps to focus on a specific aesthetic issue and, consequently, quickly find ways to solve it.

The first aesthetic problem to fix is the curved ligand structure. This representation could be more appealing. In addition, the bent ligand structure surrounds and squashes the residues Glu308A, Val811A, and Ser289A such that the two hydrogen bonds of Glu308A cross the structure of Ser289A. The ligand structure can be elongated with the *Mirror mode* and the *Bond* option. By left-clicking on a bond, users can cycle through all possible mirroring positions until the most appropriate one is found. In this case, the *Mirror mode* is applied once on both phosphate anhydride bonds. Thereby, the diphosphate group is unchanged, and the non-bridging oxygen atoms that interact with Arg316A are still on the same side of the ligand, which prevents the crossing of Arg316A intersections with the ligand structure. Using the *Rotation mode*, the ligand is then rotated counterclockwise into a horizontal position. Subsequently, intersection- and overlap-free positions can now be found for all residues except Arg316A with the *Move mode* and *Rotation mode*. A structure's mirroring, rotation, or movement also affects all associated interactions, hydrophobic contact splines, and text labels, simplifying such structural modifications. Further layout optimization can be achieved by reorienting the hydrogen atoms of the ligand towards the acceptor oxygen atoms of Glu308A by mirroring their bonds with the *Mirror mode* and by repositioning overlapping text labels of the hydrophobic contacts with the *Move mode*, which also creates more space for a better placement of Arg316A and Glu308A.

Next, we address the issue that the ligand's diphosphate group engages in several intersecting hydrophilic



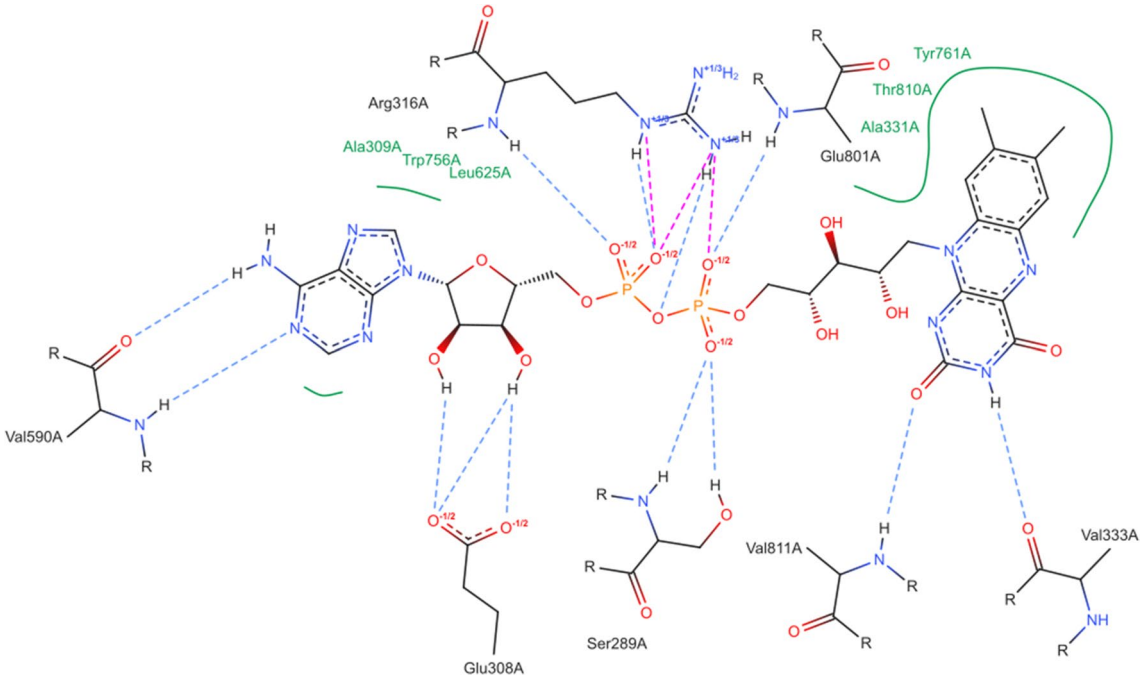




**Fig. 6** PoseEdit diagram of lysine-specific histone demethylase 1A in complex with a cofactor (PDB code: 5LGT; Proteins*Plus* identifier: FAD_A_901) with a suboptimal layout

**Fig. 7** PoseEdit diagram of lysine-specific histone demethylase 1A in complex with a cofactor (PDB code: 5LGT; Proteins*Plus* identifier: FAD_A_901) with a layout optimized by PoseEdit

interactions with Arg316A. No intersection-free position can be found for that residue by rotation and translation alone. By once mirroring the CB–CG bond of Arg316A with the *Mirror mode*, Arg316A can be moved and rotated such that its structure is not intersected anymore by the hydrogen bond that originates from its backbone. However, that hydrogen bond still intersects several interactions of the Arg316A side chain, which interact with a second non-bridging oxygen atom of the diphosphate group. Since the two non-bridging oxygen atoms are chemically equivalent in the diagram, users can avoid these intersections in two ways. Either users can remove the interactions of both non-bridging oxygen atoms with the *Remove mode* and add them with the *Add mode* to the equivalent one, or users can flip the positions of the two non-bridging oxygen atoms with the *Move mode*. The diagram is now free of overlaps and intersections and can be exported as a JSON file.

## Customization of the diagram

This section will exemplify how to obtain a personalized diagram in terms of information content and graphical styles based on the diagram of the inhibitor and the optimized one of the FAD cofactor. Figure 8 shows an example of an individually customized diagram. The exported JSON file of the diagram is deposited in the Supplementary Information (Online Resource 5).

Users might, for instance, prefer a single, comprehensive diagram that includes the inhibitor, the cofactor, the interactions between the ligands, and their interactions with the protein binding site rather than two distinct diagrams. The information of the two diagrams can be combined by two approaches. The first one is to individually add the structures, interactions, hydrophobic contact splines, and text labels displayed in one diagram to another one with the *Add mode*. Structures can be specified by SMILES strings or via a list containing a preselection of frequently appearing binding site structures, from which users can select, for example, the interacting residues. Ligands like the inhibitor or FAD cofactor are not in the list and must be added via the corresponding SMILES strings, which can be obtained from the central *Ligand list*. Subsequently, the *Add mode* can be used to draw all missing interactions, hydrophobic contact splines, and text labels. The second and more straightforward and efficient approach is the merging of the two diagrams with the JSON import feature of PoseEdit. Users can, for example, export the JSON file of the diagram of the inhibitor and import it into the optimized diagram of FAD via the button with the JSON text label and plus sign. The imported diagram is automatically placed next to the one of FAD. Multiple structures can be selected and subsequently moved and rotated, along with all linked interactions, hydrophobic contact splines, and text labels using the *Select mode*'s rectangle or lasso selection tool. Based on the 3D binding site information, users can select all structures of the diagram of the inhibitor and apply the *Move mode* and *Rotation mode* such that the two interacting aromatic ring

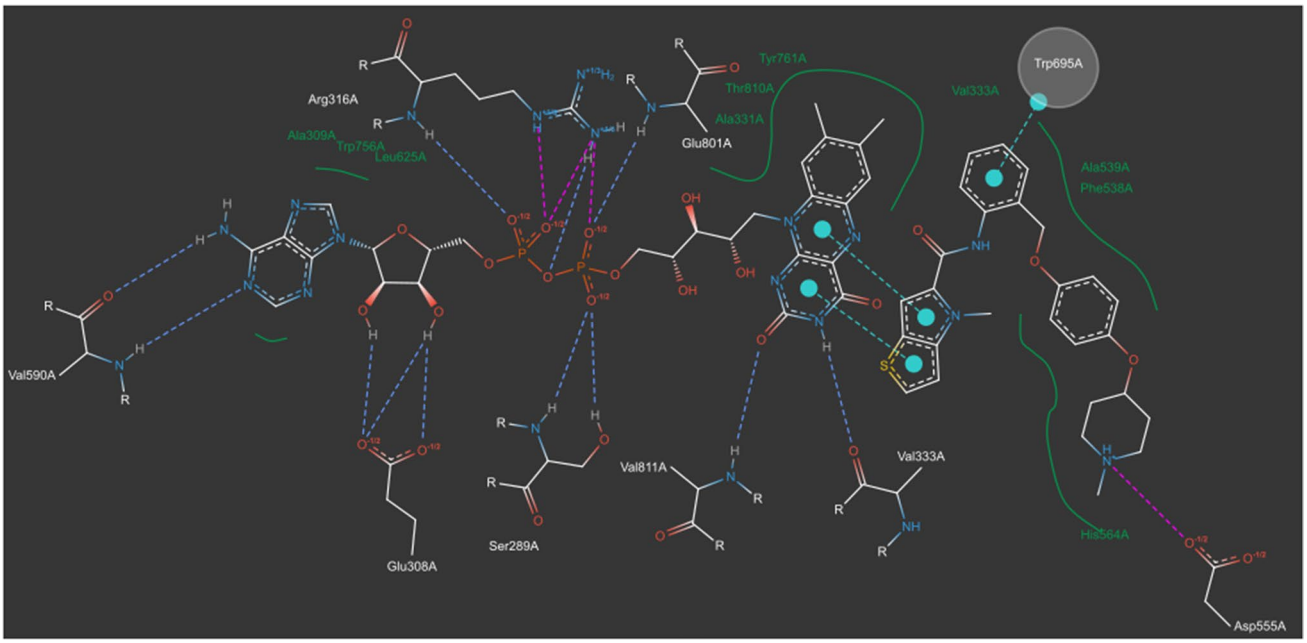



**Fig. 8** PoseEdit diagram obtained by the merging of the two diagrams of the LSD1 inhibitors with the ProteinsPlus identifier 6W3_A_902 and FAD and subsequential chemical editing and graphical restyling

systems of both ligands are adjacently placed. The missing pi-stacking interactions between the ring systems can then be drawn with the *Add mode*.

Based on the merged diagram, another subjective adjustment exemplified here regards the Nε nitrogen atom of the Arg 316A side chain. This atom is involved in a hydrogen bond as well as an ionic interaction with the same ligand atom. Users might want to keep only the stronger intermolecular force, the ionic interaction. With the *Remove mode*, users can remove the nitrogen atom's hydrogen bond and its explicitly drawn hydrogen atom. The nitrogen atom can then be annotated with one implicit hydrogen atom using the *Edit mode*.

Since the diagram is very complex, users might consider also reducing the diagram's complexity to avoid overloading the viewer with information or to focus on specific aspects like the atomic interactions with the protein residues. In this regard, the *Edit mode* can be useful, for example by changing the representation style of Trp695A, which is involved in a pi-stacking interaction, to the *Circle representation*.

Finally, users can modify all graphical styles in the diagram via a comprehensive configuration list to further individualize the diagram. The list is accessible via the *Opts* button and contains numerous styling options for atoms, bonds, interactions, structures, structure circles, the editor's control elements and the diagram background. Users can freely experiment with custom settings since the editor's editing history tracks all changes. PoseEdit also offers a list of several preset themes. To exemplify the various styling possibilities, the *Dark theme*, which might be an eye strain-reducing alternative for some users, was used to recolor the background and structures in the diagram. The customized and final diagram now contains the user-desired chemical information and graphical styles.

## Conclusion

In this work, we presented PoseEdit as a comprehensively extended and interactive version of the tool PoseView. This new development stands out from other published tools in several ways. User preferences and aesthetic ideals cannot always be fully satisfied by automatically generated diagrams. Users working with 2D ligand interaction diagrams, and in particular those who favor the ones calculated by PoseView will clearly benefit from the extended opportunities of PoseEdit, avoiding typical limitations of commonly used 2D interaction diagram generators. PoseEdit enables users to resolve all sorts of subjective and objective deficiencies that would otherwise impede the intended communication of the ligand binding mode or even prevent diagram usability. This key feature distinguishes PoseEdit from other tools that provide either static diagrams or diagrams with a much lower level of interactivity.

Furthermore, while previously existing tools are all desktop applications that are not all freely accessible, PoseEdit can be accessed without limitations on a web server. This makes PoseEdit's distinctive features usable for everyone and everywhere on various devices without any installation issues. The interactive diagrams are embedded in a 2D editor with an intuitive interface design making it easily accessible to all scientists. Several additional features, like the editing history or the diagram export/import functionality, render it a user-friendly alternative to other tools.

We hope that PoseEdit will increase the usage and quality of 2D ligand interaction diagrams by combining its algorithmically generated output with the valuable input of the scientist on a web-based platform. While the main application of PoseEdit is the generation of pose diagrams, the code base can be used for other purposes. For example, we are currently developing a two-dimensional editor for complex interaction patterns, which should substantially simplify the use of GeoMine [14, 15], our database of macromolecule-ligand complex structures.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s10822-023-00522-4.

**Acknowledgements** The authors thank the whole development team of the NAOMI library and Proteins*Plus* web server for forming the basis of this work.

**Author contributions** The concepts behind PoseEdit were developed by KD and MR. KD and BK implemented PoseEdit. The original draft of the manuscript was written by KD, the project and manuscript writing were supervised by MR. All authors read and approved the final manuscript.

**Funding** Open Access funding enabled and organized by Projekt DEAL. Development of ProteinsPlus was supported by de.NBI (in part); German Federal Ministry of Education and Research (BMBF) [031L0105 to K.S. and J.S.]; Development of GeoMine was supported by the German Federal Ministry of Education and Research as part of CompLS and de.NBI [031L0172, 031L0105 to K.D. and J.G.].

**Data availability** All data generated or analysed during this study are included in this published article and its supplementary information files.

**Software availability** The PoseEdit application is available at https://proteins.plus. The code of the InteractionDrawer library used for PoseEdit is available at https://github.com/rareylab/InteractionDrawer.

## Declarations

**Competing interests** The authors declare the following competing financial interest(s): ProteinsPlus and the NAOMI ChemBioSuite use some methods that are jointly owned and/or licensed to BioSolveIT GmbH, Germany, MR is a shareholder of BioSolveIT GmbH.

## References

1. Laskowski RA, Swindells MB (2011) LigPlot+: Multiple ligand-protein interaction diagrams for drug discovery. J Chem Inf Model 51:2778–2786. https://doi.org/10.1021/ci200227u
2. Wallace AC, Laskowski RA, Thornton JM (1995) LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. Protein Eng Des Sel 8:127–134. https://doi.org/10.1093/protein/8.2.127
3. Caboche S (2013) LeView: automatic and interactive generation of 2D diagrams for biomacromolecule/ligand interactions. J Cheminform. https://doi.org/10.1186/1758-2946-5-40
4. Stierand K, Rarey M (2010) Drawing the PDB—protein-ligand complexes in two dimensions. ACS Med Chem Lett 1:540–545. https://doi.org/10.1021/ml100164p
5. Stierand K, Rarey M (2007) From modeling to medicinal chemistry: automatic generation of two-dimensional complex diagrams. ChemMedChem 2:853–860. https://doi.org/10.1002/cmdc.200700010
6. Stierand K, Maass PC, Rarey M (2006) Molecular complexes at a glance: automated generation of two-dimensional complex diagrams. Bioinformatics 22:1710–1716. https://doi.org/10.1093/bioinformatics/btl150
7. Clark AM, Labute P (2007) 2D depiction of protein−ligand complexes. J Chem Inf Model 47:1933–1944. https://doi.org/10.1021/ci7001473
8. Bouysset C, Fiorucci S (2021) ProLIF: a library to encode molecular interactions as fingerprints. J Cheminform 13. https://doi.org/10.1186/s13321-021-00548-6
9. Davies TG, Bentley J, Arris CE, Boyle FT, Curtin NJ, Endicott JA, Gibson AE, Goldin BT, Griffin RJ, Hardcastle IR, Jewsbury P, Johnson LN, Mesguich V, Newell DR, Noble MEM, Tucker JA, Wang L, Whitfield HJ (2002) Structure-based design of a potent purine-based cyclin-dependent kinase inhibitor. Nat Struct Mol Biol 9:745–749. https://doi.org/10.1038/nsb842
10. McNaught AD, Wilkinson A (1997) IUPAC. Compendium of Chemical Terminology, 2nd ed. (the "Gold Book"), Blackwell Scientific Publications, Oxford
11. Schöning-Stierand K, Diedrich K, Ehrt C, Flachsenberg F, Graef J, Sieg J, Penner P, Poppinga M, Ungethüm A, Rarey M (2022) ProteinsPlus: a comprehensive collection of web-based molecular modeling tools. Nucleic Acids Res 50:611–615. https://doi.org/10.1093/nar/gkac305
12. Schöning-Stierand K, Diedrich K, Fährrolfes R, Flachsenberg F, Meyder A, Nittinger E, Steinegger R, Rarey M (2020) ProteinsPlus: interactive analysis of protein–ligand binding interfaces. Nucleic Acids Res 48:48–53. https://doi.org/10.1093/nar/gkaa235
13. Fährrolfes R, Bietz S, Flachsenberg F, Meyder A, Nittinger E, Otto T, Volkamer A, Rarey M (2017) ProteinsPlus: a web portal for structure analysis of macromolecules. Nucleic Acids Res 45:337–343. https://doi.org/10.1093/nar/gkx333
14. Graef J, Ehrt C, Diedrich K, Poppinga M, Ritter N, Rarey M (2022) Searching geometric patterns in protein binding sites and their application to data mining in protein kinase structures. J Med Chem 65:1384–1395. https://doi.org/10.1021/acs.jmedchem.1c01046
15. Diedrich K, Graef J, Schöning-Stierand K, Rarey M (2021) GeoMine: interactive pattern mining of protein-ligand interfaces in the Protein Data Bank. Bioinformatics 37:424–425. https://doi.org/10.1093/bioinformatics/btaa693
16. Inhester T, Bietz S, Hilbig M, Schmidt R, Rarey M (2017) Index-based searching of interaction patterns in large collections of protein-ligand interfaces. J Chem Inf Model 57:148–158. https://doi.org/10.1021/acs.jcim.6b00561
17. Bietz S, Urbaczek S, Schulz B, Rarey M (2014) Protoss: a holistic approach to predict tautomers and protonation states in protein-ligand complexes. J Cheminform. https://doi.org/10.1186/1758-2946-6-12
18. Lippert T, Rarey M (2009) Fast automated placement of polar hydrogen atoms in protein-ligand complexes. J Cheminform. https://doi.org/10.1186/1758-2946-1-13
19. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. Nucleic Acids Res 28:235–242. https://doi.org/10.1093/nar/28.1.235
20. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Hassabis D (2021) Highly accurate protein structure prediction with AlphaFold. Nature 596:583–589. https://doi.org/10.1038/s41586-021-03819-2
21. Graef J, Ehrt C, Rarey M (2023) Binding site detection remastered: enabling fast, robust, and reliable binding site detection and descriptor calculation with DoGSite3. J Chem Inf Model 63:3128–3137. https://doi.org/10.1021/acs.jcim.3c00336
22. Probst D, Reymond JL (2018) smilesdrawer: parsing and drawing SMILES-encoded molecular structures using client-side javascript. J Chem Inf Model 58:1–7. https://doi.org/10.1021/acs.jcim.7b00425
23. Rose AS, Bradley AR, Valasatava Y, Duarte JM, Prlić A, Rose PW (2018) NGL viewer: web-based molecular graphics for large complexes. Bioinformatics 34:3755–3758. https://doi.org/10.1093/bioinformatics/bty419
24. Rose AS, Hildebrand PW (2015) NGL Viewer: a web application for molecular visualization. Nucleic Acids Res 43:576–579. https://doi.org/10.1093/nar/gkv402
25. Vianello P, Sartori L, Amigoni F, Cappa A, Faga G, Fattori R, Legnaghi E, Ciossani G, Mattevi A, Meroni G, Moretti L, Cecatiello V, Pasqualato S, Romussi A, Thaler F, Trifiro P, Villa M, Botrugno OA, Dessanti P, Minucci S, Vultaggio S, Zagarri E, Varasi M, Mercurio C (2017) Thieno[3,2-b]pyrrole-5-carboxamides as New reversible inhibitors of histone lysine demethylase KDM1A/LSD1. Part 2: structure-based drug design and structure-activity relationship. J Med Chem 60:1693–1715. https://doi.org/10.1021/acs.jmedchem.6b01019

## C.2 GeoMine: interactive pattern mining of protein-ligand interfaces in the Protein Data Bank

[D2] K. Diedrich, J. Graef, K. Schöning-Stierand, and M. Rarey. Bioinformatics 37 (2020), pp. 424–425. Available: https://doi.org/10.1093/bioinformatics/btaa693. Reprinted with permission from [D2] and Oxford University Press.

*Bioinformatics*, 37(3), 2021, 424–425
doi: 10.1093/bioinformatics/btaa693
Advance Access Publication Date: 31 July 2020
Applications Note



Structural bioinformatics

# GeoMine: interactive pattern mining of protein–ligand interfaces in the Protein Data Bank

**Konrad Diedrich , Joel Graef , Katrin Schöning-Stierand and Matthias Rarey ***

Universität Hamburg, ZBH – Center for Bioinformatics, 20146 Hamburg, Germany

*To whom correspondence should be addressed.

Associate Editor: Arne Elofsson

Received on May 29, 2020; revised on July 12, 2020; editorial decision on July 22, 2020; accepted on July 24, 2020

## Abstract

**Summary:** The searching of user-defined 3D queries in molecular interfaces is a computationally challenging problem that is not satisfactorily solved so far. Most of the few existing tools focused on that purpose are desktop based and not openly available. Besides that, they show a lack of query versatility, search efficiency and user-friendliness. We address this issue with GeoMine, a publicly available web application that provides textual, numerical and geometrical search functionality for protein–ligand binding sites derived from structural data contained in the Protein Data Bank (PDB). The query generation is supported by a 3D representation of a start structure that provides interactively selectable elements like atoms, bonds and interactions. GeoMine gives full control over geometric variability in the query while performing a deterministic, precise search. Reasonably selective queries are processed on the entire set of protein–ligand complexes in the PDB within a few minutes. GeoMine offers an interactive and iterative search process of successive result analyses and query adaptations. From the numerous potential applications, we picked two from the field of side-effect analyze showcasing the usefulness of GeoMine.

**Availability and implementation:** GeoMine is part of the ProteinsPlus web application suite and freely available at https://proteins.plus.

**Contact:** rarey@zbh.uni-hamburg.de

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 Introduction

The understanding, manipulation and modulation of protein function require substantial structural knowledge of the protein binding sites. One of the main sources for structural data is the Protein Data Bank (PDB) (Berman, 2000). Despite its substantial growth, improvement of data quality and potential as a knowledge source, there is only a small number of tools featuring 3D geometric searching for protein–ligand interfaces based on user-defined queries (Angles *et al.*, 2020; Hendlich *et al.*, 2003; Korb *et al.*, 2016; Mobilio *et al.*, 2010; Weisel *et al.*, 2012). The first has been Relibase which was suspended in 2018. To our knowledge, Relibase was the only tool supporting atomic precision for both protein and ligand parts within the query. Besides GSP4PDB, the tools are desktop applications and not freely available. Reasonably short runtimes can be observed in the case of Prolix and CrossMiner due to their use of fingerprint techniques. The query versatility of all tools is however limited. In CrossMiner, a query consists of pharmacophore spheres which represent predefined features. Prolix and PRDB do not support an atom-level precision for protein parts of the query. GSP4PDB lacks atomic query precision. In Prolix and GSP4PDB, the user can design a query using a 2D sketcher. PRDB requires a query in SQL format. Only CrossMiner provides the possibility to construct queries in a 3D representation of a protein–ligand complex. Regarding the lack of existing solutions, we developed GeoMine, which is based on an enhanced version of PELIKAN (Inhester *et al.*, 2017).

GeoMine is publicly available via a web-interface and part of the ProteinsPlus (Fährrolfes *et al.*, 2017; Schöning-Stierand *et al.*, 2020) server (https://proteins.plus). A search of geometrical, textual and numerical queries can be easily performed on protein–ligand interfaces derived from complexes contained in the PDB in a reasonably short time. In the following, we will illustrate the features of GeoMine by different use cases. Detailed descriptions of the methods are available in the PELIKAN (Inhester *et al.*, 2017) publication.

## 2 Usage and output

According to the ProteinsPlus workflow, GeoMine is started with only a PDB structure as input. 3D query design is guided by a precalculated pocket with selectable features in the embedded NGL viewer (Rose *et al.*, 2018). The query generation process allows the user to select amongst others atoms and aromatic ring centers or to place such points at unoccupied positions. Point–point constraints, i.e. interatomic distance ranges or interactions, and angle constraints between any pair of connected point–point constraints or aromatic ring normals allow the definition of any geometric arrangement.

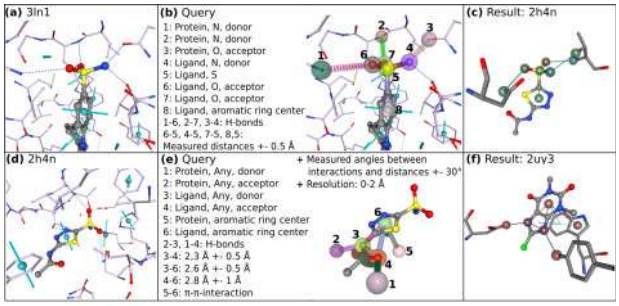



**Fig 1.** Query construction for the search of celecoxib (**a**, **b**) and acetazolamide (**d**, **e**) using GeoMine. (**c**, **f**) Example results from the searches of the queries defined in (b) and (e), respectively. The structures matching the query are highlighted in the results.

The resulting query is shown in the NGL viewer and simultaneously in tables for further modification by a variety of geometrical constraints, e.g. the range of a distance, and chemical properties. Main properties like the molecule type of an atom are automatically set while more discriminative ones, like the functional group of a ligand atom, can be explicitly defined by the user. Additionally, it is possible to query textual and numerical properties of the protein–ligand complex and its components, i.e. the depth of a pocket or the EC number. All search types can be used separately or together either on a PDB subselection or on the complete dataset. The first 150 matches are listed in a table and can be superimposed for visual analysis onto the 3D query in the viewer. An extensive statistics report, which allows a more sophisticated analysis of all results, as well as the pockets of the first 150 matches can be downloaded. The complete result set can be filtered continuing to search in it with the current query as a new starting point.

## 3 Applications

Since GeoMine is able to find structural similarities between binding sites of unrelated proteins, it is a valuable tool for off-target studies, e.g. with the aim of lead optimization, drug repurposing or explaining side effects. In the following, we will describe two different off-target searches showcasing the comprehensiveness of GeoMine results. Additional application examples are available in the PELIKAN (Inhester *et al.*, 2017) paper. The GeoMine database searches are performed using up to 30 cores of a 2× Intel Xeon Gold 6248 processor (20 cores/2.5 GHz), 200 GB of main memory and a Dell 1.6TB NVMe HHHL AIC PM1725b solid state drive with an xfs file system.

The identification of protein–ligand complexes with similar interaction patterns like a given query complex can generate ideas about potential off-target proteins. For our first example application, we choose the COX-2 selective inhibitor celecoxib (PDB code: 3LN1;Fig. 1a). In the precalculated pocket, the unsubstituted arylsulfonamide moiety of celecoxib interacts with the protein environment via 4 hydrogen bonds (Fig. 1a). A query describing partially this interacting moiety (Fig. 1b) took 45 s and resulted in 43 matches (see statistics report in Supplementary Material S1). A variety of different protein classes emerged by this search, for example carbonic anhydrase (CA II) complexed with the inhibitor acetazolamide (PDB code: 2H4N; Fig. 1c). According to studies, CA II is an off-target for unsubstituted sulfonamides like celecoxib (Weber *et al.*, 2004) implicating the enzyme in celecoxib side effects. Validation results for this query are 3LN1, 5JW1 (celecoxib cocrystallized with COX-2) and 1OQ5 (celecoxib cocrystallized with CA II). 1OQ5 was found removing one hydrogen bond from the query.

For the illustration of the second off-target search, we choose the previously found CA II complexed with acetazolamide (PDB code: 2H4N; Fig. 1d). Dependent on the arrangement of available functional groups in a protein pocket, ligands may form different interaction patterns. To find similar geometric arrangements, we constructed a query from the complex 2H4N with parts of the ligand and hypothetical alternative interactions. It includes the ligands' thiadiazole ring center and the donor and acceptor of its acetamide fragment (Fig. 1e). Potential interaction directions are defined by angles. Geometrical flexibility is achieved by relatively large tolerance values for angles and distances. Keeping the atom elements unspecified by describing only their interaction and molecule types ensures chemical fuzziness. Low-quality results are prevented by a numerical filter. The search took 52 s and resulted in 57 matches (see statistics report in Supplementary Material S2), for instance, chitinase that binds the inhibitor theophylline (PDB code: 2UY3; Fig. 1f). According to a study, chitinase is an off-target for acetazolamide, which results as a promising lead for antifungal drug development (Schüttelkopf *et al.*, 2010). A validation result can be found in PDB file 2UY4 (acetazolamide cocrystallized with chitinase).

## 4 Conclusions

GeoMine addresses the computational challenge of efficient geometrical data mining of protein–ligand binding sites. Reasonable queries can be answered by GeoMine within seconds up to a few minutes. All structures that match the query are found and presented in a comprehensive manner. The search infrastructure of GeoMine is easy to use and publicly available as part of the ProteinsPlus web service.

## Funding

This work was supported by the German Federal Ministry of Education and Research as part of the German Network for Bioinformatics Infrastructure – de.NBI [031L0172, 031L0105].

*Conflict of Interest:* ProteinsPlus and in the NAOMI ChemBio Suite use some methods that are jointly owned and/or licensed to BioSolveIT GmbH, Germany, M.R. is a shareholder of BioSolveIT GmbH.

## References

Angles,R. *et al.* (2020) GSP4PDB: a web tool to visualize, search and explore protein–ligand structural patterns. *BMC Bioinformatics*, **21**, 85.
Berman,H.M. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
Fährrolfes,R. *et al.* (2017) ProteinsPlus: a web portal for structure analysis of macromolecules. *Nucleic Acids Res.*, **45**, 337–343.
Hendlich,M. *et al.* (2003) Relibase: design and Development of a database for comprehensive analysis of protein–ligand interactions. *J. Mol. Biol.*, **326**, 607–620.
Inhester,T. *et al.* (2017) Index-based searching of interaction patterns in large collections of protein–ligand interfaces. *J. Chem. Inf. Model.*, **57**, 148–158.
Korb,O. *et al.* (2016) Interactive and versatile navigation of structural databases. *J. Med. Chem.*, **59**, 4257–4266.
Mobilio,D. *et al.* (2010) A protein relational database and protein family knowledge bases to facilitate structure-based design analyses. *Chem. Biol. Drug Des.*, **76**, 142–153.
Rose,A.S. *et al.* (2018) NGL viewer: web-based molecular graphics for large complexes. *Bioinformatics*, **34**, 3755–3758.
Schöning-Stierand,K. *et al.* (2020) ProteinsPlus: interactive analysis of protein–ligand binding interfaces. *Nucleic Acids Res.*, 48, 48-53.
Schüttelkopf,A.W. *et al.* (2010) Acetazolamide-based fungal chitinase inhibitors. *Bioorg. Med. Chem.*, **18**, 8334–8340.
Weber,A. *et al.* (2004) Unexpected nanomolar inhibition of carbonic anhydrase by COX-2-selective celecoxib: new pharmacological opportunities due to related binding site recognition. *J. Med. Chem.*, **47**, 550–557.
Weisel,M. *et al.* (2012) PROLIX: rapid mining of protein–ligand interactions in large crystal structure databases. *J. Chem. Inf. Model.*, **52**, 1450–1461.

# Supplementary Information

# GeoMine: Interactive Pattern Mining of Protein-Ligand Interfaces in the Protein Data Bank

## Supplementary material 1: First search example (query defined in 3LN1)

Occuring PDB codes:

| | | | | |
|---|---|---|---|---|
| 1NMQ | 3KK6 | 3S1H | 4JNK | 6DM1 |
| 1OIT | **3LN1** | 3W1F | 4WEK | 6DPT |
| 1PL0 | 3R9D | 4E47 | 5HNA | 6HMK |
| 2ABJ | 3R9H | 4F93 | 5IEY | 6HML |
| 2BTS | 3R9O | 4GCJ | **5JW1** | 6HMM |
| **2H4N** | 3RAK | 4HWO | 5LHB | 6HMN |
| 2IW8 | 3RD9 | 4HWP | 5VQK | 6O9G |
| 2UZB | 3RMF | 4HWR | 6DLZ | |
| 2UZD | 3RPY | 4HWS | 6DM0 | |

Occuring Ligands:

14 occurences: Clc1c(S(=O)(=O)N)cc2S(=O)(=O)N[C@H](Nc2c1)[C@@H]3[C@H]4C[C@@H](C3)CC4

5 occurences: S(=O)(=O)(N)c1ccc(N2N=C(C=C2c3ccc(cc3)C)C(F)(F)F)cc1

2 occurences: S(=O)(=O)(NC(=O)[C@@H](N)[C@H](O)C)c1cc(c2cc3ncnc(N)c3cc2)ccc1

2 occurences: Clc1cc2c(OC(C(=O)NNS(=O)(=O)c3c(cccc3)C(F)(F)F)=C2)cc1

2 occurences: S(=O)(=O)(NC(=O)[C@@H](N)[C@H](O)C)c1cc(c2cc3c(NN=C3)cc2)ccc1

2 occurences: S(=O)(=O)(NC(=O)[C@@H](N)[C@H](O)C)c1cc(c2cc3nc([nH+]c(N)c3cc2)C)ccc1

2 occurences: Clc1nc(N)c2c(n1)cc(c3cc(S(=O)(=O)NC(=O)[C@@H](N)[C@H](O)C)ccc3)cc2

1 occurences: S(=O)(=O)(N)c1ccc(NC=2SC(C(=O)c3ccc(OC)cc3)=C(N2)N)cc1

1 occurences: S(=O)(=O)(N)c1c(ccc(c1)C2=NNc3c2cc(OCC)c(c3)C=4C=NN(C4)C)C

1 occurences: S(=O)(=O)(NC1(CC1)C)c2cc3c(N(C(=O)N(C3=O)C)Cc4ccccc4)cc2

1 occurences: S(=O)(=O)(N)c1ccc(NC=2SC(C(=O)N)=C(N2)N)cc1

1 occurences: S(=O)(=O)(N[C@@H](C(=O)N1CCCC1)Cc2cc(C#N)ccc2)c3cc4c(cc3)C[NH2+]CC4

1 occurences: S(=O)(=O)(N)c1ccc(NC=2SC(C(=O)c3cc4c(cc3)cccc4)=C(N2)N)cc1

1 occurences: S(=O)(=O)(N)c1ccc(NC=2SC(C(=O)c3c([N+]([O-])=O)cccc3)=C(N2)N)cc1

1 occurences: Clc1c(S(=O)(=O)N)cc(c(OCc2cc(OCCCCCCCCCCC)cc(O)c2)c1)C([O-])=O

1 occurences: S(=O)(=O)(N)c1ccc(Nc2nc(N[C@@H]([C@H](O)C)C)ccn2)cc1

1 occurences: S(=O)(=O)(NC1(CC1)C)c2cc3c(N(C(=O)N(C3=O)CC=4SC(=NC4)C)CC5CC5)cc2

1 occurences: S(=O)(=O)(N)c1ccc(cc1)C=2OC(=CC2)CC=3SC(=NC3O)N

1 occurences: S(=O)(=O)(NC)c1ccc(cc1)CNC(=O)c2c3c(N(N=C3)c4ccc(F)cc4)c(OC)nc2

1 occurences: S(=O)(=O)(N)c1cc(N2N=C(C=C2O)C)ccc1

1 occurences: Clc1c(O)c(S(=O)(=O)Nc2c(SC3=NN=CN3C)cccc2)ccc1

1 occurences: S(=O)(=O)(N)c1ccc(Nc2nc(OCC3CCCCC3)c4N=CNc4n2)cc1

1 occurences: S(=O)(=O)(N)c1ccc(NC=2SC(C(=O)Nc3c(F)cccc3F)=C(N2)N)cc1

1 occurences: S(=O)(=O)(NC)c1ccc(cc1)C=2OC(/C=C/3\SC(=NC3=O)N)=CC2

1 occurences: S(=O)(=O)(NC(C)(C)C)c1cc2c(cc1)C(=O)c3c(cccc3)C2=O

1 occurences: S(=O)(=O)(NC1(CC1)C)c2cc3c(N(C(=O)N(C3=O)CC=4C=NN(C4)C)CC=5SC(=NC5C)C)cc2

1 occurences: S(=O)(=O)(N)c1ccc(NC=2SC(C(=O)c3ccccc3)=C(N2)N)cc1

1 occurences: S(=O)(=O)(N[C@H](C(=O)C)CC([O-])=O)c1cc(ccc1)CSSCCNC(=O)C=2SC(S(=O)(=O)C)=CC2

1 occurences: Clc1c(Cl)ccc(c1)C=2N=C(S[C@@H](C(=O)Nc3ccc(S(=O)(=O)N)cc3)C)NC(=O)C2C#N

1 occurences: [?R?]C(=O)[C@@H](NC(=O)C(=NOC(C([O-])=O)(C)C)C=1N=C(SC1)N)[C@@H](NC(=O)NS(=O)(=O)N2N=C(N(C2=O)CCCS(=O)(=O)C)C=3[N-]C=C(O)C(=O)C3)CC

1 occurences: S(=O)(=O)([NH-])C=1SC(=NN1)NC(=O)C

1 occurences: S(=O)(=O)(N)c1ccc(N)cc1

1 occurences: S(=O)(=O)(NC1(CC1)C)c2cc3c(N(C(=O)N(C3=O)CC=4SC(=NC4)C)C)cc2

1 occurences: S(=O)(=O)(N)c1ccc(Nc2nc(ccn2)C=3N4C(=NC3)C=CC=C4)cc1

1 occurences: S(=O)(=O)(N)c1ccc(NC=2SC(C(=O)Nc3cc(F)ccc3)=C(N2)N)cc1

1 occurences: S(=O)(=O)(Nc1cc2c(N=C(NC2=O)N)cc1)c3ccc(cc3)C(=O)N

1 occurences: S(=O)(=O)(N)c1ccc(NC=2SC(=CN2)C(C)C)cc1

1 occurences: S(=O)(=O)(N)c1ccc(NC=2SC(C(=O)Nc3cc(F)cc(F)c3)=C(N2)N)cc1

Statistics for Points in the Result:

Statistics for Point No.14-Donor

No. of detected points 135

Element Distribution:

Nitrogen: 135 (100.00%)

Molecule Distribution:

Protein: 135 (100.00%)

Aminoacid Distribution for Protein Points:

Alanine: 2 (1.48%)

Arginine: 27 (20.00%)

Asparagine: 4 (2.96%)

Aspartic acid: 15 (11.11%)

Glutamine: 15 (11.11%)

Glycine: 1 (0.74%)

Histidine: 2 (1.48%)

Isoleucine: 6 (4.44%)

Leucine: 1 (0.74%)

Lysine: 37 (27.41%)

Phenylalanine: 5 (3.70%)

Serine: 16 (11.85%)

Threonine: 3 (2.22%)

Tyrosine: 1 (0.74%)

Secondary Structure Distribution for Protein Points:

No Sec Structure: 39 (28.89%)

Helix: 46 (34.07%)

Sheet: 50 (37.04%)

Interaction Point Type Distribution:

Donor: 135 (100.00%)

Statistics for Point No.12-Acceptor

No. of detected points 65

Element Distribution:

Oxygen: 65 (100.00%)

Molecule Distribution:

Protein: 65 (100.00%)

Aminoacid Distribution for Protein Points:

Alanine: 1 (1.54%)

Asparagine: 1 (1.54%)

Aspartic acid: 14 (21.54%)

Glutamine: 15 (23.08%)

Glutamic acid: 6 (9.23%)

Glycine: 2 (3.08%)

Histidine: 1 (1.54%)

Isoleucine: 1 (1.54%)

Leucine: 4 (6.15%)

Serine: 17 (26.15%)

Threonine: 3 (4.62%)

Secondary Structure Distribution for Protein Points:

No Sec Structure: 30 (46.15%)

Helix: 12 (18.46%)

Sheet: 23 (35.38%)

Interaction Point Type Distribution:

Acceptor: 65 (100.00%)

Statistics for Point No.1-Aromatic

No. of detected points 60

Element Distribution:

Any Element: 60 (100.00%)

Molecule Distribution:

Reference Ligand: 60 (100.00%)

Functional Group Distribution for Ligand/RefLigand Points:

No functional group: 60 (100.00%)

Interaction Point Type Distribution:

Aromatic: 60 (100.00%)

Statistics for Point No.2-Any

No. of detected points 60

Element Distribution:

Sulfur: 60 (100.00%)

Molecule Distribution:

Reference Ligand: 60 (100.00%)

Functional Group Distribution for Ligand/RefLigand Points:

No functional group: 60 (100.00%)

Interaction Point Type Distribution:

van der waals Contact: 60 (100.00%)

Statistics for Point No.4-Acceptor

No. of detected points 120

Element Distribution:

Oxygen: 120 (100.00%)

Molecule Distribution:

Reference Ligand: 120 (100.00%)

Functional Group Distribution for Ligand/RefLigand Points:

No functional group: 120 (100.00%)

Interaction Point Type Distribution:

Acceptor: 120 (100.00%)

Statistics for Point No.6-Donor

No. of detected points 60

Element Distribution:

Nitrogen: 60 (100.00%)

Molecule Distribution:

Reference Ligand: 60 (100.00%)

Functional Group Distribution for Ligand/RefLigand Points:

No functional group: 60 (100.00%)

Interaction Point Type Distribution:

Donor: 60 (100.00%)

Statistics for Point No.8-Acceptor

No. of detected points 120

Element Distribution:

Oxygen: 120 (100.00%)

Molecule Distribution:

Reference Ligand: 120 (100.00%)

Functional Group Distribution for Ligand/RefLigand Points:

No functional group: 120 (100.00%)

Interaction Point Type Distribution:

Acceptor: 120 (100.00%)

Statistics for Point No.10-Donor

No. of detected points 135

Element Distribution:

Nitrogen: 135 (100.00%)

Molecule Distribution:

Protein: 135 (100.00%)

Aminoacid Distribution for Protein Points:

Alanine: 2 (1.48%)

Arginine: 27 (20.00%)

Asparagine: 4 (2.96%)

Aspartic acid: 15 (11.11%)

Glutamine: 15 (11.11%)

Glycine: 1 (0.74%)

Histidine: 2 (1.48%)

Isoleucine: 6 (4.44%)

Leucine: 1 (0.74%)

Lysine: 37 (27.41%)

Phenylalanine: 5 (3.70%)

Serine: 16 (11.85%)

Threonine: 3 (2.22%)

Tyrosine: 1 (0.74%)

Secondary Structure Distribution for Protein Points:

No Sec Structure: 39 (28.89%)

Helix: 46 (34.07%)

Sheet: 50 (37.04%)

Interaction Point Type Distribution:

Donor: 135 (100.00%)

Statistics for Interactions in the Result

Statistics for Interaction No.15

No. of detected interactions 135

Distance Distribution:

| | | |
|---|---|---|
| 3.01 | 3.43 | 2.68 |
| 2.77 | 3.30 | 3.10 |
| 3.14 | 3.47 | 3.34 |
| 3.00 | 2.46 | 3.39 |
| 3.25 | 3.07 | 3.11 |
| 3.32 | 3.37 | 3.06 |
| 3.62 | 2.88 | 3.62 |
| 2.43 | 3.26 | 3.04 |
| 3.00 | 3.28 | 3.50 |
| 2.92 | 3.37 | 3.23 |
| 2.92 | 3.24 | 2.86 |
| 3.22 | 3.12 | 2.83 |
| 3.14 | 3.24 | 2.45 |
| 3.53 | 3.12 | 3.07 |
| 3.54 | 2.99 | 3.13 |
| 3.57 | 3.21 | 3.35 |
| 3.06 | 3.36 | 3.09 |

| | | |
|---|---|---|
| 3.18 | 3.62 | 3.29 |
| 3.24 | 3.41 | 3.04 |
| 3.13 | 3.50 | 3.27 |
| 2.83 | 2.95 | 3.07 |
| 2.93 | 3.14 | 3.65 |
| 3.35 | 3.11 | 3.16 |
| 3.55 | 3.34 | 2.92 |
| 3.13 | 3.08 | 3.15 |
| 3.55 | 3.57 | 2.88 |
| 3.17 | 2.46 | 3.27 |
| 3.32 | 3.05 | 2.83 |
| 2.94 | 2.98 | 3.11 |
| 3.25 | 3.18 | 3.07 |
| 3.00 | 3.09 | 3.19 |
| 2.99 | 3.14 | 3.21 |
| 2.46 | 3.34 | 2.43 |
| 3.05 | 2.98 | 2.98 |
| 3.67 | 3.13 | 3.09 |
| 3.15 | 3.14 | 3.42 |
| 3.73 | 2.94 | 2.54 |
| 2.81 | 3.28 | 2.77 |
| 3.18 | 3.32 | 2.99 |
| 3.10 | 2.88 | 3.13 |
| 3.13 | 3.60 | 3.25 |
| 3.08 | 3.14 | 3.32 |
| 3.35 | 3.56 | 3.15 |
| 2.74 | 2.97 | 3.22 |
| 3.34 | 2.82 | 3.14 |

InteractionType Distribution:

H-bond: 135 (100.00%)

Statistics for Interaction No.5

No. of detected interactions 120

Distance Distribution:

| | | |
|---|---|---|
| 1.53 | 1.43 | 1.51 |
| 1.54 | 1.43 | 1.47 |
| 1.60 | 1.45 | 1.43 |
| 1.63 | 1.46 | 1.44 |
| 1.45 | 1.56 | 1.51 |
| 1.46 | 1.57 | 1.50 |
| 1.50 | 1.42 | 1.42 |
| 1.47 | 1.42 | 1.45 |
| 1.44 | 1.43 | 1.47 |
| 1.46 | 1.43 | 1.47 |
| 1.50 | 1.43 | 1.44 |
| 1.47 | 1.43 | 1.45 |
| 1.44 | 1.51 | 1.42 |
| 1.43 | 1.47 | 1.42 |
| 1.43 | 1.45 | 1.50 |
| 1.46 | 1.44 | 1.47 |
| 1.45 | 1.42 | 1.46 |
| 1.46 | 1.41 | 1.48 |
| 1.51 | 1.43 | 1.43 |
| 1.47 | 1.43 | 1.44 |
| 1.48 | 1.43 | 1.51 |
| 1.48 | 1.44 | 1.47 |
| 1.42 | 1.44 | 1.44 |
| 1.43 | 1.44 | 1.43 |
| 1.50 | 1.51 | 1.51 |
| 1.47 | 1.47 | 1.47 |
| 1.50 | 1.52 | 1.50 |
| 1.47 | 1.48 | 1.47 |
| 1.44 | 1.51 | 1.45 |
| 1.45 | 1.52 | 1.41 |

| | | |
|---|---|---|
| 1.60 | 1.60 | 1.47 |
| 1.62 | 1.61 | 1.48 |
| 1.47 | 1.46 | 1.57 |
| 1.49 | 1.52 | 1.59 |
| 1.43 | 1.43 | 1.59 |
| 1.43 | 1.43 | 1.61 |
| 1.42 | 1.46 | 1.38 |
| 1.40 | 1.43 | 1.36 |
| 1.43 | 1.50 | 1.50 |
| 1.44 | 1.47 | 1.47 |

InteractionType Distribution:

No interaction: 120 (100.00%)

Statistics for Interaction No.3

No. of detected interactions 60

Distance Distribution:

| | | |
|---|---|---|
| 3.14 | 3.20 | 3.18 |
| 3.05 | 3.04 | 3.05 |
| 3.14 | 3.15 | 3.19 |
| 3.20 | 2.88 | 3.10 |
| 3.03 | 3.15 | 3.16 |
| 3.22 | 3.21 | 3.21 |
| 3.01 | 2.98 | 3.15 |
| 3.19 | 3.05 | 3.01 |
| 3.14 | 3.15 | 3.21 |
| 3.21 | 3.17 | 3.03 |
| 3.11 | 3.14 | 3.21 |
| 3.08 | 3.21 | 3.21 |
| 3.22 | 3.14 | 3.03 |
| 3.22 | 3.11 | 3.27 |
| 3.18 | 3.21 | 3.16 |
| 3.15 | 3.04 | 3.13 |

| | | |
|---|---|---|
| 3.01 | 3.14 | 3.17 |
| 3.04 | 3.02 | 3.05 |
| 3.03 | 3.20 | 3.04 |
| 3.13 | 3.07 | 3.22 |

InteractionType Distribution:

No interaction: 60 (100.00%)

Statistics for Interaction No.7

No. of detected interactions 60

Distance Distribution:

| | | |
|---|---|---|
| 1.56 | 1.62 | 1.75 |
| 1.56 | 1.75 | 1.61 |
| 1.72 | 1.61 | 1.75 |
| 1.75 | 1.61 | 1.75 |
| 1.61 | 1.62 | 1.58 |
| 1.75 | 1.64 | 1.67 |
| 1.61 | 1.60 | 1.70 |
| 1.68 | 1.75 | 1.61 |
| 1.71 | 1.67 | 1.59 |
| 1.75 | 1.77 | 1.61 |
| 1.71 | 1.75 | 1.56 |
| 1.59 | 1.61 | 1.61 |
| 1.75 | 1.63 | 1.63 |
| 1.75 | 1.60 | 1.58 |
| 1.66 | 1.69 | 1.75 |
| 1.64 | 1.73 | 1.53 |
| 1.69 | 1.67 | 1.71 |
| 1.58 | 1.75 | 1.57 |
| 1.66 | 1.72 | 1.59 |
| 1.61 | 1.60 | 1.75 |

InteractionType Distribution:

No interaction: 60 (100.00%)

Statistics for Interaction No.9

No. of detected interactions 120

Distance Distribution:

| | | |
|---|---|---|
| 1.54 | 1.45 | 1.48 |
| 1.53 | 1.44 | 1.52 |
| 1.63 | 1.43 | 1.52 |
| 1.60 | 1.43 | 1.51 |
| 1.46 | 1.46 | 1.47 |
| 1.45 | 1.45 | 1.51 |
| 1.47 | 1.57 | 1.44 |
| 1.50 | 1.56 | 1.43 |
| 1.46 | 1.42 | 1.50 |
| 1.44 | 1.42 | 1.51 |
| 1.47 | 1.43 | 1.45 |
| 1.50 | 1.43 | 1.42 |
| 1.43 | 1.43 | 1.47 |
| 1.44 | 1.43 | 1.47 |
| 1.46 | 1.47 | 1.45 |
| 1.43 | 1.51 | 1.44 |
| 1.46 | 1.44 | 1.42 |
| 1.45 | 1.45 | 1.42 |
| 1.47 | 1.41 | 1.47 |
| 1.51 | 1.42 | 1.50 |
| 1.48 | 1.43 | 1.48 |
| 1.48 | 1.43 | 1.46 |
| 1.43 | 1.44 | 1.44 |
| 1.42 | 1.43 | 1.43 |
| 1.47 | 1.44 | 1.47 |
| 1.50 | 1.44 | 1.51 |
| 1.47 | 1.47 | 1.43 |
| 1.50 | 1.51 | 1.44 |

| | | |
|---|---|---|
| 1.47 | 1.40 | 1.47 |
| 1.51 | 1.42 | 1.50 |
| 1.47 | 1.44 | 1.48 |
| 1.50 | 1.43 | 1.47 |
| 1.41 | 1.61 | 1.59 |
| 1.45 | 1.60 | 1.57 |
| 1.62 | 1.52 | 1.61 |
| 1.60 | 1.46 | 1.59 |
| 1.49 | 1.43 | 1.36 |
| 1.47 | 1.43 | 1.38 |
| 1.43 | 1.43 | 1.47 |
| 1.43 | 1.46 | 1.50 |

InteractionType Distribution:

No interaction: 120 (100.00%)

Statistics for Interaction No.11

No. of detected interactions 135

Distance Distribution:

| | | |
|---|---|---|
| 2.77 | 3.57 | 3.24 |
| 3.14 | 3.53 | 3.12 |
| 3.01 | 3.43 | 3.24 |
| 3.25 | 3.06 | 3.21 |
| 3.00 | 3.47 | 2.99 |
| 3.62 | 3.30 | 2.68 |
| 3.32 | 3.07 | 3.36 |
| 3.00 | 2.46 | 3.34 |
| 2.43 | 2.88 | 3.10 |
| 2.92 | 3.37 | 3.11 |
| 2.92 | 3.28 | 3.39 |
| 3.14 | 3.37 | 3.62 |
| 3.22 | 3.26 | 3.04 |
| 3.54 | 3.12 | 3.06 |

| | | |
|---|---|---|
| 3.23 | 3.08 | 3.14 |
| 3.50 | 3.18 | 2.82 |
| 2.83 | 3.10 | 2.97 |
| 2.86 | 2.74 | 3.27 |
| 3.07 | 3.35 | 3.07 |
| 2.45 | 3.62 | 3.29 |
| 3.35 | 3.34 | 3.04 |
| 3.13 | 3.50 | 3.16 |
| 3.24 | 3.41 | 3.65 |
| 3.13 | 3.14 | 3.15 |
| 3.09 | 2.95 | 2.92 |
| 3.18 | 3.34 | 3.27 |
| 2.93 | 3.11 | 2.88 |
| 2.83 | 3.57 | 3.11 |
| 3.55 | 3.08 | 2.83 |
| 3.35 | 3.05 | 3.19 |
| 3.55 | 2.46 | 3.21 |
| 3.13 | 3.18 | 3.07 |
| 3.32 | 3.09 | 2.98 |
| 3.17 | 2.98 | 2.43 |
| 3.25 | 3.34 | 3.42 |
| 2.94 | 3.14 | 3.09 |
| 2.99 | 3.13 | 2.77 |
| 3.00 | 2.98 | 2.54 |
| 3.05 | 3.28 | 3.13 |
| 2.46 | 3.32 | 2.99 |
| 3.15 | 3.14 | 3.32 |
| 3.67 | 2.94 | 3.15 |
| 2.81 | 3.60 | 3.25 |
| 3.73 | 2.88 | 3.14 |
| 3.13 | 3.56 | 3.22 |

InteractionType Distribution:

H-bond: 135 (100.00%)

Statistics for Interaction No.13

No. of detected interactions 65

Distance Distribution:

| | | |
|---|---|---|
| 2.99 | 3.32 | 2.83 |
| 3.12 | 2.85 | 2.87 |
| 3.08 | 2.87 | 2.96 |
| 2.82 | 2.92 | 3.15 |
| 2.79 | 2.96 | 3.23 |
| 3.08 | 2.75 | 2.91 |
| 3.22 | 3.42 | 2.81 |
| 3.66 | 2.86 | 3.15 |
| 3.18 | 3.42 | 2.74 |
| 3.11 | 3.16 | 2.95 |
| 2.76 | 2.68 | 2.87 |
| 2.87 | 2.70 | 3.10 |
| 2.83 | 2.87 | 3.03 |
| 3.05 | 2.79 | 2.73 |
| 3.21 | 3.07 | 2.95 |
| 3.21 | 3.03 | 2.80 |
| 3.16 | 2.63 | 2.75 |
| 2.64 | 3.08 | 3.62 |
| 2.92 | 2.71 | 3.01 |
| 2.98 | 2.86 | 2.92 |
| 2.54 | 3.23 | 3.22 |
| 2.96 | 2.73 | |

InteractionType Distribution:

H-bond: 65 (100.00%)

## Supplementary material 2: Second search example (query defined in 2H4N)

Occuring PDB codes:

| | | | |
|---|---|---|---|
| 1M3Q | 4GLX | 4UCS | 6GLU |
| **2UY3** | 4HRQ | 4Z6O | 6PI5 |
| **2UY4** | 4HUO | 4Z6R | 6PII |
| 2XO1 | 4HUP | 4Z6U | |
| 2XTK | 4HV3 | 5ANW | |
| 2YKJ | 4JV3 | 5D7I | |
| 2YR6 | 4LVZ | 5H9I | |
| 3BAA | 4M0Z | 5HH9 | |
| 3FO4 | 4M13 | 5M51 | |
| 3FZT | 4M14 | 5M53 | |
| 3GOT | 4M15 | 5NGS | |
| 3PX8 | 4M5N | 5OSZ | |
| 3PX9 | 4MX1 | 5QIS | |
| 3QRJ | 4MX5 | 5QJK | |
| 4BS0 | 4PJ5 | 6EQ2 | |
| 4EFE | 4PJE | 6EQ7 | |
| 4ESI | 4PJH | 6GLG | |
| 4FEO | 4PUM | 6GLH | |

Occuring Ligands:

7 occurences: O=C1[NH+]=C(NC=2N=CNC12)N

6 occurences: [?R?]=Cc1nc2c(nc1)NC(=NC2=O)NC(=O)C

4 occurences: [nH+]1c2NC(=Nc2ccc1)N

4 occurences: n1c(N2CCC2)c3N=CNc3nc1

3 occurences: S(=O)(=O)(N)C=1SC(=NN1)NC(=O)C

2 occurences: n1c(nc(N)c2N=CNc12)N

2 occurences: [O-]C(=O)c1c(N)cccc1

2 occurences: O=C(NC=1C(=NN(c2c3c(ccc(OC(C)C)c3)ccc2)C1)C(=O)N)N

2 occurences: S=C1N=C2NC=NC2=C(N1)N

2 occurences: [O-][N+](=O)c1cc2NN=Nc2cc1

1 occurences: O=C(NC=1C(=NN(c2c3c(ccc(OC)c3)ccc2)C1)C(=O)N)N

1 occurences: O=C(N[C@H]1c2c(c3c1cccc3)c(ccc2)C4=Nc5c(N4)cncc5)c6cc(ncc6)N

1 occurences: Brc1c(nc2nc(N)c(cc2c1)C(=O)N)C(F)(F)F

1 occurences: O=C(N(C)C)c1cc(Nc2nc(OCC3CCCCC3)c4N=CNc4n2)ccc1

1 occurences: O=C1NC(=NC=2N=C(NC21)N)N

1 occurences: [O-
]C(=O)[C@@H](NC(=O)[C@@H](NC(=O)c1nc2NC(=NC(=O)c2nc1)N)CO)CC=3c4c(NC3)cccc4

1 occurences: [O-][N+](=O)c1cc([O-])c(O)cc1

1 occurences: [?R?]=Cc1nc2c([nH+]c1)NC(=NC2=O)NC(=O)C

1 occurences: O=C1[N-]C=2NC=NC2C(=O)N1

1 occurences: O=C1NC(=NC=2NN=NC12)N

1 occurences: [O-
]C(=O)c1c(O)c(NC(=O)CC[C@@]2(C(=O)C=C[C@@]34[C@H]2C[C@@H](C(=C)C3)CC4)C)c(O)cc1

1 occurences: O=C(N)c1c(nc2nc(nc(NCCC=3NC=NC3)c2c1)C(C)(C)C)N

1 occurences: O=C1c2c(nc(O[C@@H]3[C@@H](CCCC3)C)nc2N)N(C=C1)CC#N

1 occurences: [O-]C(=O)NCCNC(=O)c1nc2NC(=NC(=O)c2nc1)N

1 occurences: Fc1nc2NC=Nc2c(n1)N

1 occurences: S([O-])(=O)(=O)c1cc([O-])c(O)cc1

1 occurences: Fc1c(c2c3N=C(Nc3[nH+]cc2)N)ccnc1

1 occurences: O=C(N)c1cc(Nc2nc(OCC3CCCCC3)c4N=CNc4n2)ccc1

1 occurences: S(c1nc([nH+]c2NC=Nc12)N)CCc3ccccc3

1 occurences: O=C(N(C)C)Cc1ccc(c2nc([nH+]c3c2cccc3)N)cc1

1 occurences: O=C(NC1=NNC(=C1)COc2cnccc2)NC=3N(N=C(C3)C(C)(C)C)c4ccc(cc4)C

1 occurences: [O-]C(=O)c1nc2NC(=NC(=O)c2nc1)N

1 occurences: O=C(N1N=C(N=C1N)C=2OC=CC2)N

1 occurences: O=C1N=C(Nc2nc(cnc21)C(=O)NCC=3NN=NC3)N

1 occurences: ClC1=NC=2N(C(=O)N(C(=O)C2N1)C)C

1 occurences: [?R?]CC(=O)NC1=NOC(=C1)C

1 occurences: O=C1N=C(Nc2nc(cnc21)C(=O)NCCNC(O)=Nc3ccccc3)N

1 occurences: [O-]c1c(O)ccc(c1)CC([O-])=O

1 occurences: O=C(NC=1C(=NN(c2c3c(ccc(OCCC)c3)ccc2)C1)C(=O)N)N

1 occurences: O=C1N=C(Nc2c1cc3NC(=[NH+]c3c2)NC)N

1 occurences: P([O-])([O-])(=O)OCc1c(c(O)c([nH+]c1)C)C=O

1 occurences: [O-]C(=O)[C@@H](NC(=O)CNC(=O)c1nc2NC(=NC(=O)c2nc1)N)Cc3ccccc3

1 occurences: n1c2NC=Nc2c(nc1)NC

1 occurences: Fc1c(NC(=O)NC=2N(N=C(C2)C(C)(C)C)c3cc4c(nccc4)cc3)ccc(Oc5cc(ncc5)C(=O)NC)c1

1 occurences: [O-]C(=O)[C@@H](NC(=O)[C@@H](NC(=O)CNC(=O)c1nc2NC(=NC(=O)c2nc1)N)Cc3ccccc3)Cc4ccccc4

1 occurences: Clc1c(c2c(cccc2)CC)ccc(c1)C[NH2+]CCC3=Nc4c(N3)cccc4

1 occurences: O=C1N=C(Nc2nc(cnc21)C(=O)NCC=3OC=CC3)N

1 occurences: Clc1nc(nc2NC=Nc12)N

Statistics for Points in the Result

Statistics for Point No.26-Acceptor

No. of detected points 76

Element Distribution:

Oxygen: 76 (100.00%)

Molecule Distribution:

Protein: 76 (100.00%)

Aminoacid Distribution for Protein Points:

Asparagine: 3 (3.95%)

Aspartic acid: 17 (22.37%)

Glutamic acid: 17 (22.37%)

Glycine: 11 (14.47%)

Leucine: 2 (2.63%)

Serine: 9 (11.84%)

Tyrosine: 7 (9.21%)

Valine: 5 (6.58%)

COULD NOT FIND NAME: 5 (6.58%)

Secondary Structure Distribution for Protein Points:

No Sec Structure: 33 (43.42%)

Helix: 24 (31.58%)

Sheet: 19 (25.00%)

Interaction Point Type Distribution:

Acceptor: 76 (100.00%)

Statistics for Point No.27-Aromatic

No. of detected points 123

Element Distribution:

Any Element: 123 (100.00%)

Molecule Distribution:

Protein: 123 (100.00%)

Aminoacid Distribution for Protein Points:

Histidine: 6 (4.88%)

Phenylalanine: 20 (16.26%)

Tryptophan: 56 (45.53%)

Tyrosine: 36 (29.27%)

COULD NOT FIND NAME: 4 (3.25%)

COULD NOT FIND NAME: 1 (0.81%)

Secondary Structure Distribution for Protein Points:

No Sec Structure: 44 (35.77%)

Helix: 35 (28.46%)

Sheet: 44 (35.77%)

Interaction Point Type Distribution:

Aromatic: 123 (100.00%)

Statistics for Point No.23-Donor

No. of detected points 72

Element Distribution:

Nitrogen: 68 (94.44%)

Oxygen: 4 (5.56%)

Molecule Distribution:

Reference Ligand: 72 (100.00%)

Functional Group Distribution for Ligand/RefLigand Points:

Amide: 5 (6.94%)

Amine: 6 (8.33%)

Alcohol: 4 (5.56%)

No functional group: 57 (79.17%)

Interaction Point Type Distribution:

Donor: 72 (100.00%)

Statistics for Point No.22-Acceptor

No. of detected points 72

Element Distribution:

Nitrogen: 51 (70.83%)

Oxygen: 21 (29.17%)

Molecule Distribution:

Reference Ligand: 72 (100.00%)

Functional Group Distribution for Ligand/RefLigand Points:

Aldehyde: 1 (1.39%)

Ketone: 1 (1.39%)

Amide: 8 (11.11%)

Ester: 2 (2.78%)

No functional group: 60 (83.33%)

Interaction Point Type Distribution:

Acceptor: 72 (100.00%)

Statistics for Point No.25-Donor

No. of detected points 74

Element Distribution:

Nitrogen: 60 (81.08%)

Oxygen: 14 (18.92%)

Molecule Distribution:

Protein: 74 (100.00%)

Aminoacid Distribution for Protein Points:

Arginine: 14 (18.92%)

Asparagine: 16 (21.62%)

Aspartic acid: 7 (9.46%)

Cysteine: 1 (1.35%)

Glutamine: 2 (2.70%)

Histidine: 1 (1.35%)

Isoleucine: 2 (2.70%)

Leucine: 4 (5.41%)

Lysine: 4 (5.41%)

Threonine: 1 (1.35%)

Tyrosine: 8 (10.81%)

Valine: 8 (10.81%)

COULD NOT FIND NAME: 5 (6.76%)

Any aminoacid: 1 (1.35%)

Secondary Structure Distribution for Protein Points:

No Sec Structure: 35 (47.30%)

Helix: 5 (6.76%)

Sheet: 34 (45.95%)

Interaction Point Type Distribution:

Donor: 74 (100.00%)

Statistics for Point No.21-Aromatic

No. of detected points 72

Element Distribution:

Any Element: 72 (100.00%)

Molecule Distribution:

Reference Ligand: 72 (100.00%)

Functional Group Distribution for Ligand/RefLigand Points:

No functional group: 72 (100.00%)

Interaction Point Type Distribution:

Aromatic: 72 (100.00%)

Statistics for Interactions in the Result

Statistics for Interaction No.31

No. of detected interactions 76

Distance Distribution:

| | | | | |
|---|---|---|---|---|
| 2.71 | 2.98 | 2.69 | 3.18 | 3.13 |
| 2.84 | 3.10 | 2.72 | 2.66 | 2.68 |
| 2.95 | 2.89 | 2.63 | 2.57 | 2.70 |
| 2.68 | 3.30 | 2.47 | 2.81 | 3.08 |
| 2.72 | 3.04 | 2.46 | 3.40 | 2.67 |
| 2.71 | 2.47 | 2.63 | 2.47 | 2.67 |
| 2.97 | 2.58 | 2.64 | 3.40 | 3.19 |
| 2.98 | 2.68 | 2.72 | 2.58 | 2.94 |
| 2.79 | 2.76 | 3.25 | 2.98 | 2.52 |
| 2.83 | 2.78 | 2.70 | 2.94 | 2.92 |
| 2.81 | 3.11 | 2.49 | 2.92 | 2.70 |
| 2.42 | 3.25 | 2.72 | 2.63 | 2.56 |
| 2.49 | 2.71 | 2.53 | 3.20 | |
| 2.69 | 3.04 | 3.09 | 3.12 | |
| 3.23 | 2.75 | 2.80 | 3.24 | |
| 2.97 | 3.06 | 3.12 | 3.28 | |

InteractionType Distribution:

H-bond: 76 (100.00%)

Statistics for Interaction No.30

No. of detected interactions 123

Distance Distribution:

| | | | | |
|---|---|---|---|---|
| 3.83 | 3.54 | 3.70 | 3.56 | 3.63 |
| 3.51 | 3.49 | 4.12 | 3.77 | 3.51 |
| 3.79 | 3.54 | 4.56 | 3.48 | 3.53 |
| 3.72 | 4.89 | 3.84 | 4.45 | 3.64 |

| | | | | |
|---|---|---|---|---|
| 4.01 | 3.66 | 3.64 | 3.45 | 3.75 |
| 3.51 | 3.47 | 3.50 | 4.57 | 3.88 |
| 3.58 | 3.85 | 3.66 | 3.44 | 4.13 |
| 4.03 | 3.76 | 3.41 | 4.31 | 3.93 |
| 3.41 | 3.64 | 3.89 | 3.57 | 4.05 |
| 4.54 | 3.84 | 4.19 | 4.95 | 4.84 |
| 3.53 | 4.85 | 4.19 | 4.58 | 3.97 |
| 3.81 | 3.97 | 4.76 | 3.46 | 4.20 |
| 3.51 | 4.25 | 3.58 | 3.56 | 3.37 |
| 3.79 | 4.30 | 4.01 | 4.22 | 4.34 |
| 4.48 | 3.55 | 3.38 | 4.12 | 3.76 |
| 4.28 | 3.96 | 3.61 | 3.56 | 3.86 |
| 4.85 | 3.54 | 4.14 | 4.96 | 4.50 |
| 3.51 | 3.82 | 3.90 | 3.46 | 4.95 |
| 3.63 | 4.12 | 4.58 | 3.53 | 3.45 |
| 3.73 | 4.25 | 3.57 | 3.60 | 3.55 |
| 3.83 | 3.99 | 4.87 | 3.61 | 3.65 |
| 3.57 | 3.68 | 4.10 | 4.22 | 3.90 |
| 4.05 | 3.68 | 3.97 | 4.10 | 4.15 |
| 3.45 | 4.08 | 3.58 | 3.79 | |
| 3.60 | 3.50 | 4.11 | 3.58 | |

InteractionType Distribution:

pi-pi: 123 (100.00%)

Statistics for Interaction No.28

No. of detected interactions 74

Distance Distribution:

| | | | | |
|---|---|---|---|---|
| 3.36 | 2.86 | 3.13 | 3.40 | 2.80 |
| 3.22 | 3.00 | 2.80 | 2.20 | 3.21 |
| 3.47 | 3.61 | 2.73 | 2.86 | 2.57 |
| 2.61 | 3.10 | 3.00 | 3.50 | 3.22 |
| 3.06 | 2.80 | 3.49 | 3.07 | 2.91 |

| | | | | |
|---|---|---|---|---|
| 2.78 | 2.72 | 3.01 | 3.02 | 2.89 |
| 2.86 | 2.80 | 2.86 | 2.88 | 3.02 |
| 2.95 | 2.68 | 2.84 | 3.46 | 2.84 |
| 2.75 | 2.84 | 2.80 | 2.94 | 3.04 |
| 2.97 | 3.07 | 2.61 | 3.24 | 2.65 |
| 3.50 | 2.86 | 2.60 | 2.92 | 2.46 |
| 2.97 | 2.68 | 2.97 | 2.73 | 2.97 |
| 2.89 | 2.88 | 3.54 | 2.94 | 2.83 |
| 2.81 | 3.09 | 2.42 | 2.99 | 3.14 |
| 2.72 | 3.10 | 2.83 | 3.04 | |

InteractionType Distribution:

H-bond: 74 (100.00%)

Statistics for Interaction No.34

No. of detected interactions 72

Distance Distribution:

| | | | | |
|---|---|---|---|---|
| 2.17 | 2.25 | 2.40 | 2.17 | 2.18 |
| 2.28 | 2.67 | 2.16 | 2.17 | 2.40 |
| 2.40 | 2.37 | 2.18 | 2.24 | 2.39 |
| 2.29 | 2.39 | 2.23 | 2.71 | 2.65 |
| 2.34 | 2.24 | 2.23 | 2.20 | 2.37 |
| 2.26 | 2.24 | 2.24 | 2.38 | 2.21 |
| 2.27 | 2.28 | 2.21 | 2.79 | 2.33 |
| 2.36 | 2.17 | 2.36 | 2.73 | 1.96 |
| 2.26 | 2.25 | 2.33 | 2.38 | 2.65 |
| 2.27 | 2.17 | 2.20 | 2.23 | 2.22 |
| 2.24 | 2.32 | 2.25 | 2.38 | 2.75 |
| 2.23 | 2.19 | 2.40 | 2.68 | 2.30 |
| 2.19 | 2.75 | 2.18 | 2.24 | |
| 2.34 | 2.37 | 2.78 | 2.31 | |
| 2.16 | 2.39 | 2.19 | 2.19 | |

InteractionType Distribution:

No interaction: 72 (100.00%)

Statistics for Interaction No.32

No. of detected interactions 72

Distance Distribution:

| | | | | |
|---|---|---|---|---|
| 2.73 | 3.42 | 3.68 | 2.69 | 2.74 |
| 2.68 | 3.77 | 2.73 | 2.68 | 3.69 |
| 3.63 | 3.58 | 2.74 | 2.65 | 3.69 |
| 3.63 | 3.37 | 2.65 | 2.77 | 3.78 |
| 3.61 | 2.67 | 2.63 | 3.79 | 3.64 |
| 3.72 | 2.73 | 2.64 | 3.59 | 2.78 |
| 2.78 | 3.70 | 2.79 | 2.73 | 3.69 |
| 3.58 | 2.72 | 3.64 | 3.70 | 2.72 |
| 3.63 | 3.78 | 3.72 | 3.62 | 2.73 |
| 2.71 | 2.68 | 2.73 | 2.77 | 2.67 |
| 2.67 | 3.75 | 2.66 | 3.65 | 3.34 |
| 2.67 | 2.76 | 3.69 | 3.80 | 3.57 |
| 2.74 | 3.72 | 2.73 | 2.68 | |
| 3.61 | 3.65 | 3.55 | 3.73 | |
| 2.74 | 3.64 | 2.75 | 2.74 | |

InteractionType Distribution:

No interaction: 72 (100.00%)

Statistics for Interaction No.33

No. of detected interactions 72

Distance Distribution:

| | | | | |
|---|---|---|---|---|
| 2.70 | 2.69 | 2.68 | 2.52 | 2.68 |
| 2.75 | 2.72 | 2.72 | 2.73 | 2.59 |
| 2.70 | 2.49 | 2.70 | 2.79 | 2.73 |
| 2.59 | 2.71 | 2.57 | 2.55 | 2.78 |
| 2.69 | 2.74 | 2.81 | 2.73 | 2.79 |
| 2.55 | 2.72 | 2.71 | 2.82 | 2.74 |

| | | | | |
|---|---|---|---|---|
| 2.70 | 2.78 | 2.74 | 2.75 | 2.59 |
| 2.72 | 2.73 | 2.76 | 2.59 | 2.69 |
| 2.74 | 2.78 | 2.73 | 2.73 | 2.77 |
| 2.71 | 2.75 | 2.79 | 2.71 | 2.63 |
| 2.72 | 2.74 | 2.80 | 2.72 | 2.47 |
| 2.74 | 2.76 | 2.71 | 2.72 | 2.76 |
| 2.80 | 2.68 | 2.74 | 2.77 | |
| 2.69 | 2.67 | 2.70 | 2.69 | |
| 2.59 | 2.71 | 2.74 | 2.79 | |

InteractionType Distribution:

No interaction: 72 (100.00%)

Statistics for Anglefilters in the Result

Statistics for Anglefilter No.36

No. of detected angles 123

Angle Distribution:

| | | | | |
|---|---|---|---|---|
| 104.50 | 106.61 | 99.57 | 88.86 | 93.72 |
| 79.24 | 107.10 | 99.89 | 111.81 | 101.17 |
| 110.23 | 82.57 | 78.17 | 70.21 | 81.38 |
| 73.36 | 72.37 | 108.73 | 99.95 | 103.14 |
| 101.84 | 101.89 | 81.70 | 77.71 | 84.94 |
| 108.75 | 110.05 | 65.24 | 101.86 | 100.82 |
| 101.81 | 99.69 | 65.97 | 76.83 | 81.51 |
| 78.97 | 60.18 | 105.52 | 107.43 | 99.54 |
| 77.80 | 89.76 | 101.50 | 76.42 | 58.43 |
| 98.54 | 98.33 | 72.59 | 75.93 | 87.20 |
| 113.83 | 59.36 | 103.33 | 67.36 | 99.49 |
| 95.74 | 88.37 | 99.27 | 82.72 | 85.21 |
| 77.19 | 77.36 | 60.26 | 70.52 | 98.69 |

| | | | | |
|---|---|---|---|---|
| 83.33 | 102.65 | 86.15 | 74.93 | 68.26 |
| 78.72 | 57.98 | 102.49 | 89.17 | 78.16 |
| 100.80 | 73.99 | 78.98 | 75.84 | 77.89 |
| 69.54 | 78.40 | 100.77 | 106.21 | 85.06 |
| 84.88 | 95.78 | 70.15 | 79.43 | 110.55 |
| 97.76 | 58.62 | 102.14 | 101.17 | 99.23 |
| 60.06 | 87.18 | 75.30 | 79.34 | 91.63 |
| 89.39 | 85.94 | 92.05 | 99.14 | 104.78 |
| 78.76 | 100.39 | 107.98 | 65.43 | 77.31 |
| 62.11 | 110.77 | 75.49 | 80.99 | 100.22 |
| 89.53 | 101.72 | 105.28 | 99.23 | |
| 72.02 | 73.96 | 67.22 | 81.92 | |

Statistics for Anglefilter No.35

No. of detected angles 123

Angle Distribution:

| | | | | |
|---|---|---|---|---|
| 117.39 | 98.43 | 108.22 | 80.95 | 68.27 |
| 79.07 | 104.25 | 76.97 | 104.29 | 104.50 |
| 110.25 | 78.17 | 107.32 | 94.95 | 76.47 |
| 91.76 | 68.00 | 79.28 | 67.39 | 65.61 |
| 95.97 | 97.36 | 66.89 | 118.82 | 95.09 |
| 104.81 | 78.06 | 96.63 | 69.10 | 87.40 |
| 97.59 | 66.10 | 105.63 | 77.23 | 64.42 |
| 66.21 | 95.71 | 88.73 | 82.09 | 94.47 |
| 63.92 | 108.65 | 95.17 | 77.00 | 103.07 |
| 117.68 | 93.53 | 65.60 | 67.39 | 93.03 |
| 81.21 | 115.49 | 114.52 | 96.32 | 94.94 |
| 114.14 | 77.41 | 79.50 | 76.28 | 103.36 |
| 78.29 | 109.22 | 109.90 | 94.58 | 104.89 |
| 110.35 | 64.21 | 106.89 | 105.12 | 76.09 |
| 98.63 | 67.95 | 104.67 | 79.82 | 66.39 |
| 100.99 | 95.04 | 67.21 | 65.63 | 94.88 |
| 93.02 | 98.82 | 65.72 | 117.01 | 104.68 |

| | | | | |
|---|---|---|---|---|
| 105.75 | 80.73 | 102.70 | 93.67 | 114.48 |
| 88.92 | 95.38 | 109.20 | 65.64 | 101.87 |
| 94.95 | 66.08 | 75.86 | 116.79 | 93.22 |
| 62.88 | 93.80 | 107.67 | 79.60 | 105.37 |
| 107.96 | 106.92 | 65.49 | 66.11 | 66.97 |
| 102.39 | 91.22 | 116.70 | 77.56 | 114.56 |
| 85.23 | 97.84 | 64.97 | 68.66 | |
| 117.28 | 67.39 | 114.95 | 105.76 | |

Statistics for Anglefilter No.38

No. of detected angles 76

Angle Distribution:

| | | | | |
|---|---|---|---|---|
| 146.99 | 141.77 | 135.12 | 120.68 | 154.09 |
| 156.12 | 159.45 | 140.48 | 150.93 | 134.61 |
| 128.92 | 125.57 | 151.08 | 149.03 | 132.66 |
| 149.09 | 119.99 | 144.89 | 122.89 | 156.70 |
| 137.93 | 144.05 | 146.45 | 122.51 | 136.04 |
| 125.68 | 143.56 | 157.71 | 150.65 | 169.12 |
| 134.43 | 157.89 | 166.94 | 151.27 | 115.71 |
| 153.19 | 140.81 | 136.78 | 138.60 | 138.88 |
| 125.41 | 147.55 | 116.63 | 147.89 | 149.70 |
| 153.54 | 142.60 | 170.22 | 132.40 | 154.53 |
| 158.54 | 119.84 | 144.34 | 128.76 | 155.27 |
| 143.35 | 116.31 | 158.13 | 153.05 | 139.08 |
| 146.99 | 146.20 | 165.47 | 117.71 | |
| 166.35 | 154.89 | 150.65 | 117.14 | |
| 130.79 | 147.91 | 168.05 | 132.97 | |
| 151.18 | 130.42 | 119.15 | 154.79 | |

Statistics for Anglefilter No.37

No. of detected angles 74

Angle Distribution:

| | | | | |
|---|---|---|---|---|
| 122.83 | 159.66 | 152.14 | 126.33 | 134.97 |
| 126.16 | 147.00 | 138.12 | 166.22 | 122.03 |

| | |
|---|---|
| 126.90 | 170.12 |
| 171.49 | 130.23 |
| 174.21 | 171.05 |
| 171.20 | 124.57 |
| 153.63 | 144.53 |
| 132.22 | 148.33 |
| 169.85 | 158.46 |
| 151.86 | 168.37 |
| 163.32 | 118.16 |
| 118.43 | 139.94 |
| 174.71 | 158.63 |
| 128.80 | 121.26 |
| 156.62 | 132.29 |
| 118.40 | 149.46 |
| 157.08 | 162.80 |
| 154.84 | 162.80 |
| 133.22 | 152.10 |
| 157.72 | 131.00 |
| 143.08 | 132.69 |
| 146.78 | 156.75 |
| 155.93 | 122.49 |
| 150.20 | 149.14 |
| 156.01 | 149.33 |
| 149.66 | 155.92 |
| 172.73 | 151.82 |
| 169.45 | 152.44 |
| 120.08 | 136.57 |
| 126.80 | 171.57 |
| 151.77 | 121.72 |
| 149.20 | 137.61 |
| 135.46 | 137.47 |
| 146.47 | 147.26 |

## C.3 Searching Geometric Patterns in Protein Binding Sites and Their Application to Data Mining in Protein Kinase Structures

[D3] J. Graef, C. Ehrt, K. Diedrich, M. Poppinga, N. Ritter, and M. Rarey. Journal of Medicinal Chemistry 65 (2022), pp. 1384–1395. Available: https://doi.org/10.1021/acs.jmedchem.1c01046. Reprinted with permission from [D3]. 

Journal of Medicinal Chemistry

pubs.acs.org/jmc

Article

# Searching Geometric Patterns in Protein Binding Sites and Their Application to Data Mining in Protein Kinase Structures

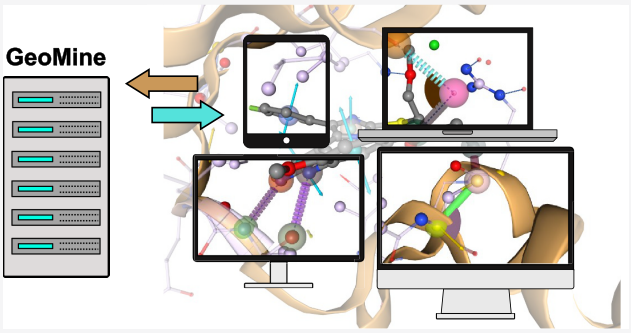
Joel Graef, Christiane Ehrt, Konrad Diedrich, Martin Poppinga, Norbert Ritter, and Matthias Rarey*

Cite This: J. Med. Chem. 2022, 65, 1384−1395

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

**ABSTRACT:** The ever-growing number of protein-ligand complex structures can give fundamental insights into protein functions and protein-ligand interactions, especially in the field of protein kinase research. The number of tools to mine this data for individually defined structural motifs is restricted due to the challenging task of developing efficient index structures for 3D data in relational databases. Herein we present GeoMine, a database system with web front-end mining of more than 900 000 binding sites. It enables database searches for geometric (interaction) patterns in protein-ligand interfaces by, for example, textual, numerical, substructure, similarity, and 3D searches. GeoMine processes reasonably selective user-defined queries within minutes. We demonstrate its usability for advancing protein kinase research with a special emphasis on unusual interactions, their use in designing selective kinase inhibitors, and the analysis of reactive cysteine residues that are amenable to covalent kinase inhibitors. GeoMine is freely available as part of our modeling support server at https://proteins.plus.






## INTRODUCTION

The analysis of protein-ligand interactions is a key element for understanding the structure-to-function relationships and selectivity profiles of protein kinase inhibitors. The identification and optimization of small-molecule binders as a central task in early drug discovery relies on the detailed knowledge of molecular recognition. Therefore, an analysis and comparison of spatial arrangements in protein-ligand interfaces is of the utmost relevance for life science research. Searching for spatial atomic arrangements is highly useful for numerous applications. Novel ligands can be designed by employing similar binding sites, allowing the transfer of interacting functional groups from one ligand to another. Similar geometric patterns can give significant insights into a protein's function and selectivity. Analyzing the environment of functional groups helps to obtain a better understanding of interaction geometries, to name just the most intuitive use cases that come to mind.

Due to the rising number of protein structures and known protein kinase-ligand complexes, efficient three-dimensional (3D) search algorithms are required but are challenging to develop. Even with a focus on a single protein class, proteins share structural similarities such that data analysis tools should be able to cope with the entire Protein Data Bank (PDB).[1] The search method of choice must enable a search for spatial atomic arrangements in a reasonable period of time. In addition, the search must be flexible so that complex queries can be composed for multiple structural features as well as classical textual, numerical, and substructural elements. The method must be chemistry-aware to allow detailed atomic interaction modeling. Finally, a web application is desirable to avoid complex installation processes and offer an easy-to-use interface.

Up to now, several tools have been developed to search for geometric spatial arrangements, including CSD-CrossMiner,[2] PRDB,[3] PROLIX,[4] Relibase and Relibase+,[5] PDBeMotif and MSDmotif,[6] PELIKAN,[7] and GSP4PDB.[8] Recently, a new motif search was proposed by Bittrich et al. called strucmotif-search, which uses an efficient index for inter-amino acid distances.[9] The differences between all these tools can be roughly divided into the following five categories: (1) query variability, in other words, which types of textual, numerical, chemical, and geometric features can be used to create a search query. This also refers to the precision of the query, i.e., whether fuzziness can be introduced; (2) structure database, i.e., which data collections are available for search; (3) data processing and storage, i.e., how the data is extracted from raw PDB[1] files and what database technologies are used; (4) search capabilities and algorithms, i.e., how the query is evaluated and the precision of the search, and (5) result presentation, i.e., how results are reported and visualized.

Special Issue: New Horizons in Drug Discovery - Understanding and Advancing Kinase Inhibitors

Received: June 11, 2021
Published: September 7, 2021





https://doi.org/10.1021/acs.jmedchem.1c01046
J. Med. Chem. 2022, 65, 1384−1395

Table 1 gives an overview of the existing tools related to their query variability and searchable structures. More detailed information can be found in Tables S3 and S4. Overall, the two most prominent disadvantages are that some tools are not publicly available and others are even discontinued and retired, e.g., Relibase. In addition, some of the tools are desktop applications that have to be installed, and the databases have to be updated in a semiautomated fashion. None of the tools include predicted sites, solvent exposure, and secondary structure information. An overview of the capabilities of the mentioned tools is given in Table 2.

In the following, we will introduce our new approach GeoMine, a flexible, geometrically reliable, and efficient method to search for spatial arrangements in protein-ligand interfaces and predicted binding sites and apply them to several use cases related to protein kinase research. Based on our earlier desktop application PELIKAN, GeoMine combines a flexible relational database with efficient search algorithms and a new easy-to-use web-based front-end for query generation and result browsing. This enables structural investigations on protein binding sites on-the-fly and offers a user-friendly environment for efficiently searching the binding site space. Several experiments demonstrate the performance and reliability of the search engine.

## RESULTS AND DISCUSSION

GeoMine offers a huge variety of possibilities to mine protein-ligand complexes and binding sites. A full description of GeoMine's comprehensive search capabilities can be found in the Experimental Section. The following applications of the tools toward protein kinase inhibitor design highlight the impact of these features for the structural investigations of protein kinase binding sites and the subsequent runtime analyses, showing that such analyses can be performed within seconds to minutes and offering a new way to work with large amounts of structural data in an interactive and easy-to-use manner. All queries described below are available in a machine-readable format in the Supporting Information.

**Exploiting Unusual Interactions for Selective Inhibitor Design.** During protein kinase inhibitor design, the question might arise as to whether previously unexplored unusual protein-ligand interactions play a major role in protein kinase selectivity. Based on a comprehensive study of such interactions,[10] we applied GeoMine to scan available protein-ligand complex structures for such interactions. We investigated the occurrence of interactions involving halogen atoms attached to aromatic rings with aromatic ring systems in protein side chains. To this end, we searched for appropriate geometric arrangements in all ligand-occupied pockets as stored in the PDB. As a template, we chose the structure with the PDB-ID 3q3k.[11] We defined three points: the aromatic ring center of the Tyr228 side chain, the location of the chlorine atom, and the aromatic ring center of the aromatic ring the halogen atom is attached to. All points were connected by distances with a tolerance of 2 Å for the distance between the halogen atom and the aromatic center of a protein residue's side chain (distance $d_1$) and tolerances of 1 Å for the remaining two distances. To model the relative orientation of the aromatic ring systems, we define two angles $\alpha$ and $\beta$ with tolerances of 40° and 15°, respectively. Additionally, we reduce this search to pockets occupied by ligands with chlorine atoms. This search retrieved 695 3D matches in 629 pockets of 491 PDB structures (calculation time of 1.24 min). The results of the analysis are depicted in Figure 1 (top). A visual inspection of the results showed expected

Table 1. Overview of Existing Tools for 3D Searching in Structural Data Related to Query Variability and Precision[a]

| | query variability | structure database |
|---|---|---|
| **Relibase/Relibase+** (availability: not available) | atom-level precision, constraint ranges, nongeometric attributes | PDB, no information about adding own structure files, BS = 7 Å of the ligand atoms |
| **PDBeMotif/MSDmotif** (availability: public, web server) | atom-level precision only for ligands | PDB, based on user-defined selection of structure files, BS = 16 Å of the ligand atoms |
| **PRDB** (availability: not available) | distances between amino acids only between $\alpha$ carbons, nongeometric attributes | PDB, no information about adding own structure files, BS = 8 Å of the ligand atoms |
| **PROLIX** (availability: not available) | atom-level precision only for ligands, constraint range | PDB, Roche in-house X-ray structure database, no information about adding own structure files, BS = 4.5 Å of the ligand atoms |
| **CSD-CrossMiner** (availability: commercial, standalone) | only predefined feature types, pharmacophore query with user-defined tolerance spheres, nongeometric attributes, no angles | CSD, PDB (ligand-based binding sites), based on user-defined selection of structure files. BS = 6 Å of the ligand atoms |
| **PELIKAN** (availability: academic use, standalone) | atom-level precision, constraint ranges, nongeometric attributes | precompiled databases such as scPDB 2013 and PDB (November 2016) available on the ZBH web site, structure files, BS = 6.5 Å of the ligand atoms |
| **GSP4PDB** (availability: public, web server) | ligands as three-letter codes or "any"; protein as one of the 20 natural amino acids or "undefined", "any", "polar", etc.; constraint ranges, gaps, and next in sequence; nongeometric attributes | PDB, BS = 7 Å of the ligand atoms |
| **strucmotif-search** (availability: public, web server (research collaboration for structural bioinformatics, RCSB), standalone) | amino acid defined by the distance between $\alpha$-carbons as the side chain and $\beta$-carbons as backbone representatives, nongeometric attributes using RCSB functions | PDB, structure files in standalone, distances between amino acid pairs 15 Å at maximum |

[a]BS, binding site.



https://doi.org/10.1021/acs.jmedchem.1c01046
J. Med. Chem. 2022, 65, 1384−1395

Table 2. Overview of GeoMines Capabilities in Contrast to Other Existing Tools for 3D Searching[a]

| tool or GeoMine capabilities | user-defined angles | interaction detection | protonation | solvent exposure | atom-based queries | SSEs | ligand features | non-geometric filters | predicted sites | GUI |
|---|---|---|---|---|---|---|---|---|---|---|
| Relibase/Relibase+ | x | | x (ligand) | | x | | x | x | | x |
| PDBeMotif/ MSDmotif | | x | x | | x[b] | | x | | | x |
| PRDB | x | | | | x | | x | x | | x (ADOpt) |
| PROLIX | | x | | | | | | | | x |
| CSD-CrossMiner | | | | | x | | | x | | x |
| PELIKAN | x | x | x | | x | | x | x | | x |
| GSP4PDB(2) | | | | | | | | x | | x |
| strucmotif-search | | | | | | | x (via PDB) | x (via PDB) | | x |

[a]SSEs, secondary structure elements; GUI, graphical user interface. [b]Restricted to Cα and the end of side chains.

interaction geometries between aromatic side chains and chlorine atoms attached to aromatic (hetero)cycles. We found that 51% of the ligands interact with tyrosine side chains and 26% interact with phenylalanine side chains, and we also found interactions with tryptophan (17%) and histidine side chains (5%). An analysis of the distance and angle ranges revealed broadly spread distributions of the distance $d_1$ and the angle $\alpha$. Based on these findings, we can refine our query ($d_1 = 4 \pm 1$ Å, $\alpha = 25° \pm 25°$) and apply this query to search in the published structures assembled in the third version of the Kinase-Ligand Interaction Fingerprints and Structures (KLIFS) database[12] (https://klifs.net/[13]) to investigate the role of this interaction in protein kinases. The GeoMine search with the adjusted query resulted in 45 3D matches in 37 pockets of 34 PDB structures (0.42 min). Many of the identified interactions are located in binding sites in complex with well-known selective kinase inhibitors (Figure 1, mid), e.g., CDK2 in complex with the chemical probe 5,6-dichlorobenzimidazone-1-β-D-ribofuranoside (PDB-ID 3my5[14]), but we could also identify small fragments that seem to undergo halogen-aromatic interactions, e.g., DYRK1A in complex with a chlorobenzothiazole fragment (PDB-ID 5a4q[15]). Concerning the type of residues undergoing these interactions, we find that approximately 51% involve phenylalanine and 27% involve tyrosine side chains as the most important interaction partners. Many of the hits indicate the usability of this interaction type for selective inhibition of protein kinases.

As a consequence, the question arises whether there already exist known inhibitors of a protein kinase of interest with an aromatic ring in the proximity of an aromatic side chain. Such an aromatic ring might be exploited as a selectivity anchor by adding chlorine atoms. To this end, the query had to be slightly adjusted, and instead of the chlorine atom we chose the attached carbon atom for the query and adjusted the distances according to the bond length of an aromatic carbon-chlorine bond of approximately 1.7 Å (see Figure 1, mid). We further restricted the character of the ligand carbon atom to an aromatic carbon atom with exactly two connections, excluding implicit hydrogen atoms (SMARTS pattern "[c;D2]"). This ensured that the aromatic carbon atoms of the retrieved ligands are not connected to halogen atoms and are instead connected to a hydrogen atom. A search with this query in all ligand-bound protein kinase structures retrieved 218 pockets of 178 PDB structures (1.23 min), highlighting that there are numerous kinase inhibitors with known binding modes that could be used as potential starting points to improve the compound selectivity.

For this showcase study, we picked cyclin-dependent-like kinase 5 (CDK5) as an example for a pharmacologically relevant protein kinase target[16] and searched with the pattern against all available complex structures of this target. This search retrieved a structure of CDK5 in complex with an ATP analogue (PDB-ID 3o0g) with a $K_i$ value of 600 nM, which is likely not highly selective[17] but which might become more selective upon addition of a chlorine atom that interacts with Phe80 (Figure 1, bottom).

To support this hypothesis, we constructed a query for this site of CDK5 with the binding site properties, which are crucial for the interaction with this inhibitor (Figure 1, bottom). This time, we did not restrict our search to ligand-occupied pockets but searched for the pocket feature arrangement in all binding sites (predicted and ligand-based) of human protein kinases to identify as many potential off-targets as possible. This search retrieved 4801 3D matches in 3287 pockets of 2322 PDB structures (23.02 min, 227 distinct protein kinases have similar geometric arrangements of potentially interacting residues), highlighting the expected missing selectivity. However, the addition of the aromatic center of Phe80 as an additional interaction point for a potential halogen aromatic interaction reduced the number of hits by a factor of 8 (616 3D matches in 494 pockets of 392 PDB structures, 9.58 min, 33 distinct protein kinases are still similar based on the GeoMine query), pinpointing residue Phe80 as a potential selectivity anchor. Although there are still several similar protein kinase binding sites, they belong to similar kinase families, and additional ideas for further selectivity anchors can be derived from a visual inspection of the corresponding superimpositions. This example not only highlights the applicability of GeoMine for selectivity profiling but also the importance of enabling user-specific query design and database profiling in a large-scale manner.

Taken together, this showcase study exemplifies the capabilities of the GeoMine tool as an idea generator for protein kinase drug design. In the first step, typical and unusual interaction patterns can be explored and investigated. Subsequently, the queries can be adapted to find potential starting points for selective inhibitor design through the establishment of such interactions based on already known protein kinase complex structures. Moreover, GeoMine might also assist in improving our understanding of potential reasons for inhibitor selectivity, as discussed in the next example.

**Searching for Selectivity Anchors in Protein Kinases.** GeoMine can also assist in the search for potential off-targets. This is not restricted to potential off-target structures with bound ligands, as empty predicted pockets are also included in the database. As an example, we picked two structures of epidermal growth factor receptor (EGFR) in complex with two well-characterized inhibitors. For the first structure of EGFR in


https://doi.org/10.1021/acs.jmedchem.1c01046
J. Med. Chem. 2022, 65, 1384−1395

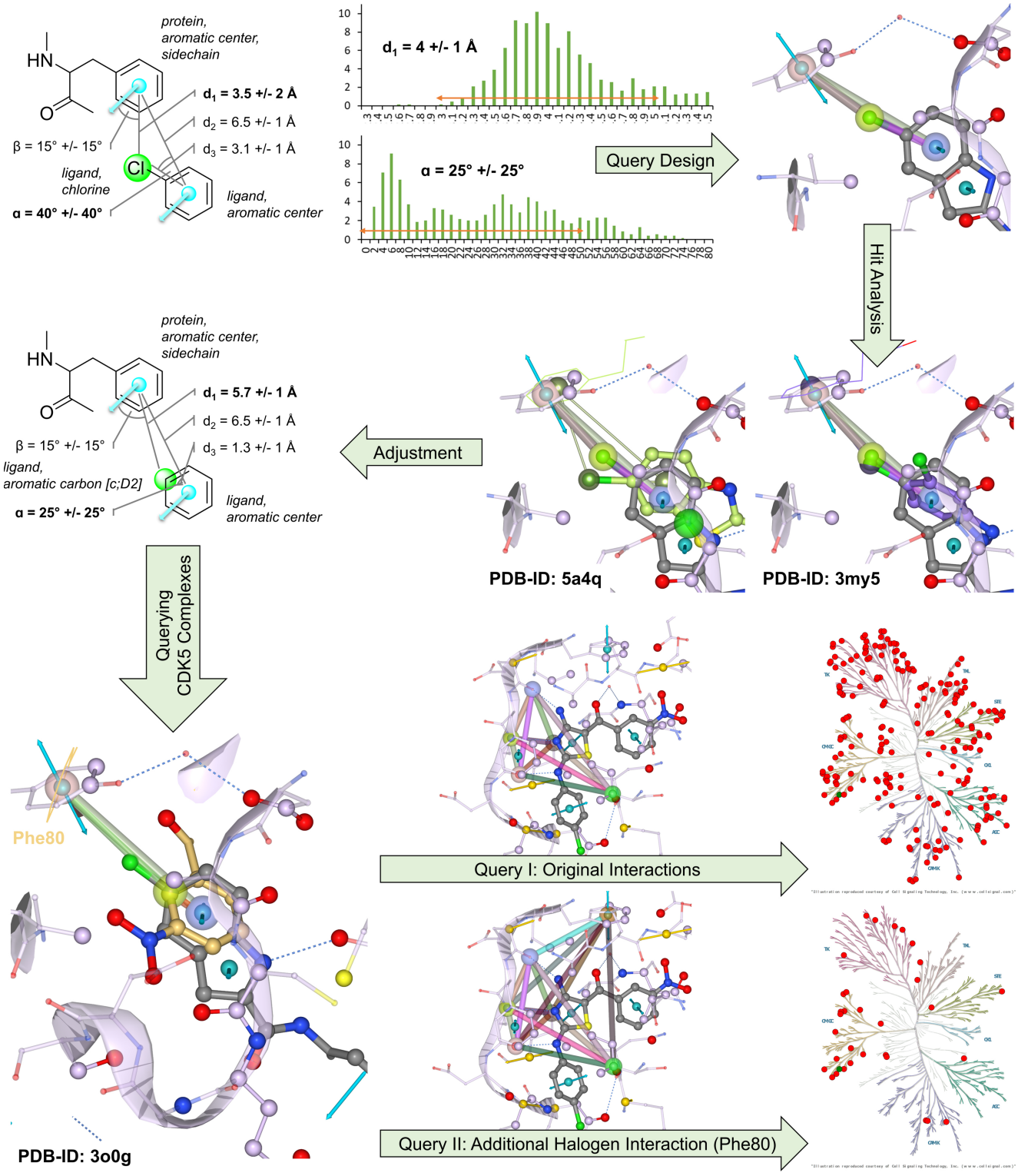



**Figure 1.** GeoMine as an idea generator for advancing protein kinase inhibitor design. In the presented workflow, the user will first search for a certain protein-ligand interaction pattern. In this case, we investigated the uncommon interaction between chlorine atoms and aromatic ring systems (top). Based on the derived geometric data, the user can adjust the query accordingly and search for this interaction type in a predefined set of ligand-bound structures (mid). After visual inspection, the user can subsequently try to find potential starting points for exploiting this interaction type to advance known inhibitors with respect to selectivity. In our showcase study, we picked the structure of CDK5 in complex with an ATP analogue (bottom). The interacting residues of the original compound are spread across the whole kinome. However, the inclusion of an interaction with Phe80 of CDK5 might lead to an improved selectivity profile, as highlighted by the GeoMine hits of the corresponding queries using KinMap.[18]

complex with the inhibitor gefitinib (PDB-ID 4wkq), we generated a query consisting of the protein's relevant interacting residues, i.e., a backbone nitrogen atom as hydrogen bond donor in the hinge region (Met793), a hydrophobic side chain carbon



https://doi.org/10.1021/acs.jmedchem.1c01046
J. Med. Chem. 2022, 65, 1384−1395

atom in the N-lobe (Ala743), a hydrophobic side chain carbon atom in the C-lobe (Leu844), and a hydrophobic side chain carbon atom of Lys745. This ligand-independent search query was used to screen all known protein kinases as stored in the KLIFS database but was restricted to structures from the organism *Homo sapiens*. The search resulted in 10 918 3D matches in 4451 pockets of 3049 PDB structures, pinpointing unselective inhibition (1.45 min). Subsequently, this initial query was extended by one side chain of residue Thr790 as a potential selectivity anchor as known from the structure of EGFR in complex with lapatinib (PDB-ID 1xkk[19]). Intriguingly, this search retrieved only 32 3D matches in 22 pockets of 20 PDB structures (1.16 min), highlighting the importance of Thr790 as a potential selectivity-introducing residue. This is in line with the finding that mutation T790 M leads to a significant increase in the $IC_{50}$ values for the highly selective inhibitor lapatinib ($IC_{50}$(EGFR) = 4.9 nM, $IC_{50}$(T790M) = 850 nM, and $IC_{50}$(T790M/L858R) = 8500 nM) .[20] A comparison of the hits with additional kinase profiling data[20] further underlines the validity of the result (Table 3).

**Table 3. Selectivity Profiling Results for Two Different EGFR GeoMine Queries**[a]

| | gefitinib | | lapatinib | |
|---|---|---|---|---|
| protein kinase | GeoMine result based on the query for 4wkq | $IC_{50}$ (nM)[20] | GeoMine result based on the query for 1xkk[19] | $IC_{50}$ (nM)[20] |
| EGFR | hit | 0.51 | hit | 4.9 |
| ERBB2 (ErbB2) | hit | 3100 | hit | 9.8 |
| ERBB4 (ErbB4) | hit | 7.6 | hit | 24 |
| LCK | hit | 390 | hit | n.d. |
| LYN | hit | 350 | hit | n.d. |
| DDR1 | hit | 37 | not found | 4400 |
| DDR2 | hit | 570 | not found | n.d. |
| EPHA5 (EphA5) | hit | 740 | not found | n.d. |
| EPHA7 (EphA7) | hit | 990 | not found | n.d. |
| EPHA8 (EphA8) | hit | 730 | not found | n.d. |
| EPHB2 (EphB2) | hit | 890 | not found | n.d. |
| EPHB4 (EphB4) | hit | 420 | not found | n.d. |
| FLT3 | hit | 730 | not found | n.d. |
| MKNK1 (MNK1) | hit | 130 | not found | n.d. |
| MKNK2 (MNK2) | hit | 150 | not found | n.d. |
| PDGFRA (PDGFRa) | hit | 600 | not found | n.d. |
| PTK6 (BRK) | hit | 860 | not found | 1100 |
| SLK | hit | 1300 | not found | n.d. |
| STK10 (LOK) | hit | 430 | not found | n.d. |
| TNK2 (ACK) | hit | 1100 | not found | n.d. |

[a]The queries were based on the structure in complex with gefitinib (PDB-ID 4wkq; interacting residues Met793, Ala743, Leu844, and Lys745) and lapatinib (PDB-ID 1xkk; interacting residues Met793, Ala743, Leu844, Lys745, Thr790).

**Searching for Reactive Cysteines in Protein Kinases.** GeoMine enables the inclusion of secondary structure features in the search query. A very prominent example of how to use this feature is the search for compounds in the vicinity of reactive cysteine residues for the structure-based design of covalent inhibitors. Cysteine residues at the amino terminus of $\alpha$-helices are frequently characterized by a high nucleophilicity.[21] This fact is often exploited for the design of covalent protein kinase inhibitors.[22] In our last example, we demonstrate how the inclusion of secondary structure elements and solvent exposure enables the search for protein kinases with reactive cysteines in the neighborhood of known inhibitors. Based on the crystal structure of the protein kinase EGFR with reactive cysteine Cys797,[23] we constructed a query to search for known kinase inhibitors in the proximity of an amino-terminal solvent-exposed cysteine (PDB-ID 3poz[24]). This query consists of the solvent-exposed cysteine sulfur atom, the cysteine's backbone nitrogen atom, the amino-terminal helix end, and any ligand atom at well-defined distance intervals from these points (see Figure 2, top). Additionally, we used the angle between the helix vector and the segment between the helix terminus and the backbone nitrogen atom to further restrict the search. We used this query to search in a PDB subselection that contained all PDB-IDs of protein kinases in the KLIFS database.[12] The query resulted in 75 3D matches in 54 pockets of 36 PDB structures and took 54 s. The results not only reveal promising inhibitors that can be extended by suitable reactive groups, so-called covalent warheads,[25] to address the reactive cysteine in EGFR, but also hint at other protein kinases, e.g., janus kinase 3 (JAK3), that also harbor a reactive cysteine in this position and might be potential off-targets for covalent inhibitors to address this cysteine residue (Figure 2, bottom).

**Comparison to Other Methods.** None of the three exemplary analyses could be performed by any other tool listed in Table 2, hampering a comparison for the selected examples. The search for unusual protein-ligand interactions by other tools is prevented by the missing functionality to restrict the geometric search by angles. PRDB is the only database that enables the definition of angles in the search query. However, this database is no longer accessible. In our second example, we define a query with interaction features, e.g., a hydrophobic residue atom. PROLIX is the only other tool that enables the use of such features. However, the definition of queries is restricted to distinct residues such that matching is only possible between identical residues. The main limitation of similar methods for the analyses of reactive amino-terminal cysteine residues, as shown in our last example, is the lack of ability to include secondary structure information in the search queries. GeoMine combines the unique strengths of the individual tools in Table 2, thereby creating a versatile user-friendly method for multiple purposes.

**Query Computing Time.** The performance of GeoMine was tested with a standard PELIKAN benchmark[7] on all protein-ligand complexes in the scPDB[28,29] (2017 version). The database was constructed with only those PDB files that contained at least one reference ligand (16 561 PDB files), and unoccupied binding sites were excluded for comparability to PELIKAN. All queries were designed using the protein-ligand complex with the PDB-ID 1j7u[30] so that every query resulted in at least one match. There are the following three kinds of geometric queries: (1) four points that are linearly arranged, (2) six points in a star shape, and (3) four points in a tetrahedron shape (see Figure 3). The point-point constraints are used with a


https://doi.org/10.1021/acs.jmedchem.1c01046
J. Med. Chem. 2022, 65, 1384−1395

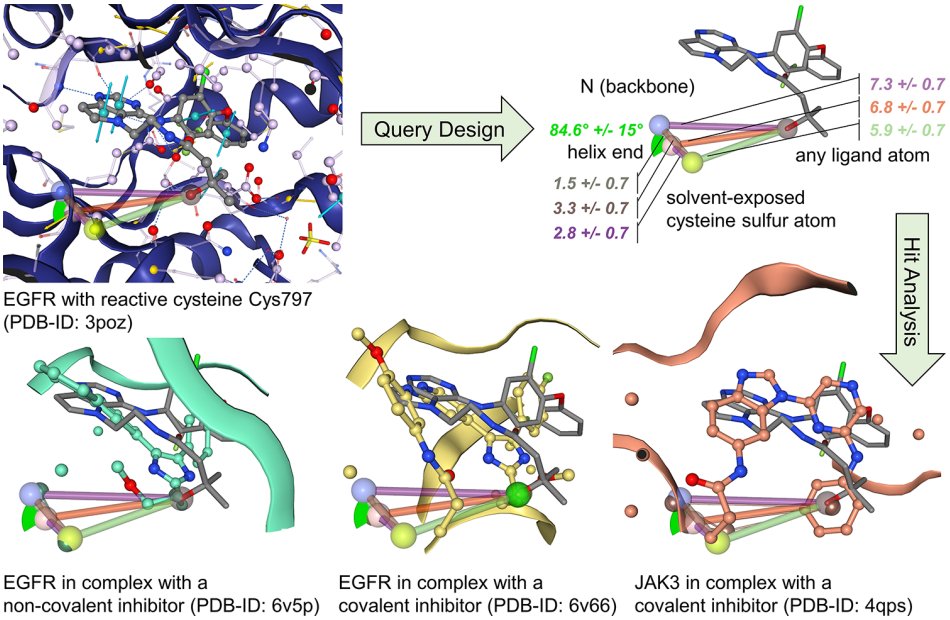



**Figure 2.** Profiling for protein kinases with reactive cysteines at the amino terminus of helices. (Top) The query was generated based on the structure of EGFR with a small-molecule inhibitor (PDB-ID 3poz[24]). The query consists of four points, six distances, and one angle characteristic for reactive cysteine in the proximity of ligands. (Bottom) Selected hits are presented here. (Left) Structures of EGFR in complex with a noncovalent and covalent inhibitor (PDB-IDs 6v5p and 6v66,[26] respectively). (Right) Structure of JAK3 harboring a reactive cysteine at the same location in complex with a covalent inhibitor as potential off-target (PDB-ID 4qps[27]).

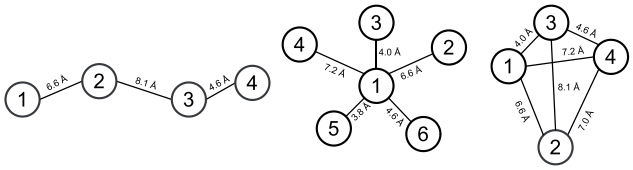


**Figure 3.** Geometric layout of the test queries for the runtime analysis, which are linear, star-shaped, or tetrahedral. Numbered points represent the search points with their IDs. Black lines describe the distance constraints. Queries were created using the PDB structure 1j7u as template structure.

tolerance of 0.5 Å. All three query types are used with different element types to assess the influence of specific attributes. The "standard" queries consist of oxygen and nitrogen atoms and, in case of the star-shaped queries, a carbon atom; the "metal" queries are those where one of the query points is changed to a magnesium ion; the "metal, water" queries are those with a magnesium ion and a water query point, and "metal, water, phosphorus" queries were those where another point is changed to a phosphorus atom. More detailed definitions of all test queries can be found in the Supporting Information (see Paragraph S2).

GeoMine and PELIKAN use different database technologies. While SQLite writes and reads data directly from disk, PostgreSQL has to establish a database connection first. In general, this network overhead is larger than that when reading from the solid-state drive. GeoMine was up to 10× faster than PELIKAN (see Figure 4) in our runtime tests. This is due to reimplemented key functionalities in the C++ code, adapted structure query language (SQL) queries, such as using JOINs on smaller parts of tables and more bundled database transactions, and making more use of the hardware by exploiting the advantages of the PostgreSQL database system.

In particular, queries with a linear topology are much faster, which is due to the lower number of distance constraints. In general, the distance constraint checks are the slowest in PELIKAN and GeoMine. Although the time required for this step was reduced with our query optimizations, it remains the most time-consuming one. The queries named "standard" are overall the slowest because they consist of carbon and nitrogen atoms, which are the most common elements in the database. Therefore, there are numerous possible points where all distance and angle constraints have to be checked. Overall, most of the query runtimes were significantly improved, and no query took longer using GeoMine. Using the solvent exposure of protein points reduces the runtimes even further.

**Upscaling.** In PELIKAN, databases had to be created by the user. When PELIKAN was developed, the PDB contained about 80 000 structures. Currently, there are more than twice as many. The focus of the application was on searching subsets of the PDB, e.g., the scPDB. With GeoMine, the entire PDB is searchable, and the database will be kept up-to-date in the future. As the PDB is continuously growing, GeoMine was designed to handle this upscaling. Since the number of available pockets and search points in the database increased by a factor of 3.5, retrieving a potential result point (PRP) list from the indexing structure results in substantially increased runtimes. This is not surprising, since the number of indexed bins defined by typical atom arrangements is constant. Therefore, the PELIKAN indexing structure is no longer used in GeoMine. To enable as many users as possible to search the database quickly, searches are performed on a server using up to 30 cores of a 2x Intel Xeon Gold 6248 processor (20 cores/2.5 GHz), 200 GB of main memory, and a Dell 1.6TB NVMe HHHL AIC PM1725b solid-state drive with an XFS file system. The parallelization of the queries is managed by PostgreSQL. This ensures that the hardware is used in the best possible way at each point in time and with a varying numbers of users. In addition, the use of


https://doi.org/10.1021/acs.jmedchem.1c01046
J. Med. Chem. 2022, 65, 1384−1395

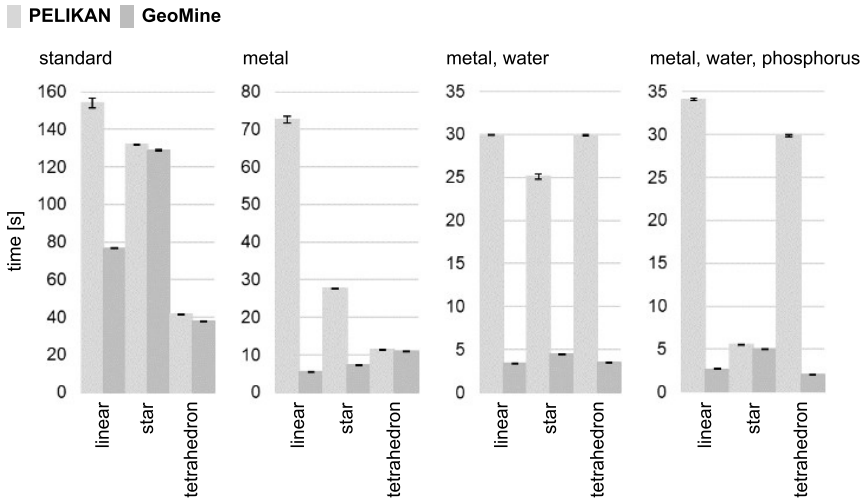



**Figure 4.** Average query runtimes of test queries. Each bar displays the mean value of five independent experiments. For each geometric query type, there is a "standard" query consisting of oxygen, nitrogen, and carbon atoms, a "metal" query where one of the query points is changed to a magnesium ion, a "metal, water" query with a magnesium ion and a water query point, and a "metal, water, phosphorus" query where a third point is also changed to a phosphorus atom. Calculations were performed on a PC equipped with an Intel i5-9500 (3.0 GHz) processor, 16 GB of main memory (6 GB usable by the PostgreSQL RDBMS), and a Toshiba BG4 PCIe solid-state drive (512 GB, model nvme) with a btrfs file system. The platform runs with a standard configuration of an openSUSE LEAP 15.0 with either PELIKAN (left, light gray) or GeoMine (right, dark gray).

PostgreSQL ensures that the method will keep working in the future even if there is an exponential growth of the number of structures in the PDB. Currently, the database is about 154 GB in size with the biggest table which stores all points of the pockets using about 72 GB of this data, while PostgreSQL reports a maximum table size of 32 TB.[31]

## CONCLUSIONS

Structure-based data mining has a huge potential in modern rational drug design, offering a large variety of data analytics. However, a database-driven method for efficient access to geometrical features in protein-ligand binding sites is required. For this task, we developed GeoMine and implemented it for practical validation. To our knowledge, it is the first method of its kind that enables searches in the entire PDB, including empty binding sites. Even purely ligand-based searches and template-free queries are possible. Our approach allows highly flexible definitions of search points that are not limited to predefined motifs. Point-point constraints and angle constraints allow the definition of geometrically precise and vague parts within a query. Textual and numerical properties of ligands, pockets, proteins, and complexes can be used to define a query in more detail and restrict the runtime.

GeoMine offers a comprehensive web interface. Searches with sufficiently specific queries are answered within seconds to minutes. The results can be displayed as superimpositions in an NGL viewer,[32,33] and statistics of the matched points, distances, interactions and angles can be downloaded. Through the extensive possibilities of query generation and GeoMine's public availability as part of the Proteins*Plus* web service, our method is highly useful for numerous applications.

Particularly for protein kinase inhibitor design with its wealth of structural data, GeoMine enables a rational and data-driven molecular design approach. Similarities in binding sites can be used to design novel ligands. Selectivity patterns in protein kinases can be analyzed and investigated based on known binding sites even in the absence of ligands. Environments of functional groups can be analyzed to gain a better understanding of interaction geometries. Geometric data mining enables the exploitation of specific interactions to selectively address protein kinase binding sites. Such target assessments and selectivity analyses are facilitated through flexible and time-efficient searches with intuitively generated template-based and template-free queries.

In the future, we plan to extend GeoMine by integrating protein-protein interfaces, 2D query design, and an automated query generation to extract commonalities and differences between given protein structure collections to formulate queries. The major benefits of GeoMine include the possibilities to design tailor-made queries, rendering it a versatile tool for multiple challenges in protein kinase inhibitor design, the inclusion of textual and numerical filters, its applicability to ligand-bound and predicted binding sites, and its short almost interactive computing time. These are unique features that enable a new way to deal with large amounts of structural data in drug design. Therefore, we hope that the tool will assist in numerous applications scenarios and will provide a unique means to explore and annotate protein kinases.

## EXPERIMENTAL SECTION

To make GeoMine accessible for many life scientists independent of their previous experience in software usage, the system is based on a database server with a web-based graphical user interface. As described in detail below, the database contains precalculated data on protein-ligand complexes and supports efficient access based on a highly flexible query engine. Briefly speaking, GeoMine is based on the PostgreSQL database management system due to its large SQL and full ACID (atomicity, consistency, isolation, and durability) compliance, good multiuser management, extensibility, and multitude of features such as the support for different security authentications. A back-end software written in C++ on top of the NAOMI library[34,35] preprocesses the PDB. The process is fully automated, including the calculation of binding sites[36] and the handling of protonation states and tautomerism.[37] Using the graphical user interface of the web service, a query is generated in extensible markup language (XML) and sent to the back-end server, which initiates the database search with an iterative approach. For computational chemists interested in automation, a representational


https://doi.org/10.1021/acs.jmedchem.1c01046
J. Med. Chem. 2022, 65, 1384−1395

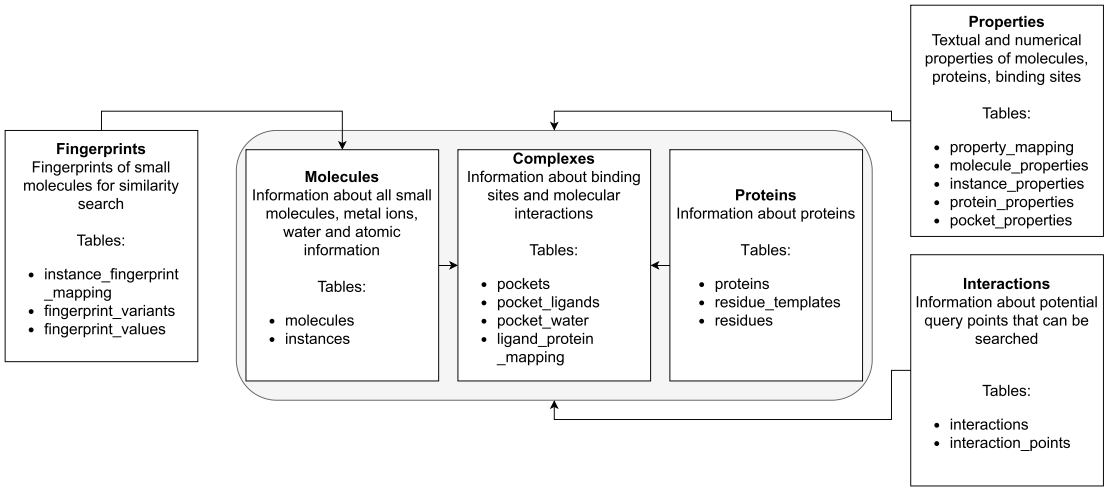



**Figure 5.** Illustration of the database schema. Tables are grouped as either molecule, protein, complex, property, fingerprint, or interaction. Arrows depict dependencies between groups.

state transfer application programming interface (REST API) is available to directly submit a query to the server.

**Data Preprocessing and Knowledge Extraction.** The process of building the GeoMine database consists of four subsequent phases. In the first phase, PDB files are read and converted into objects that represent the complex, its small molecules, water molecules, and metal ions. Among others, the PDB-ID, compound names, source organism, experimental method, and resolution are extracted from the header section. The second phase is dominated by data preparation. First, missing hydrogen atoms in the complex are identified and their coordinates are optimized using Protoss.[37] Protoss analyzes rotatable hydrogen atoms of terminal groups (e.g., hydroxy and amino groups), tautomers and protonation states of all chemical moieties (including ligand molecules), and flips of ambiguous residue side chain orientations (Asn, Gln, and His) and evaluates alternative orientations of water molecules. In addition, alternative conformations that might be annotated in the original protein structure are removed as they could hinder the analysis of molecular interactions. Second, all chains with at least 50% HETATM entries in the PDB file and more than 5 and less than 100 heavy atoms are converted to ligands. Third, the pocket detection algorithm DoGSite is applied.[36] DoGSite is a grid-based approach where each grid point is labeled depending on its spatial overlap with any protein atom. With a difference-of-Gaussian filter, small sphere-like cavities are identified, which are subsequently clustered to potential subpockets. Lastly, adjacent subpockets are merged into pockets. Fourth, all ligands that have at least six heavy atoms are associated with the detected pockets. This is done by finding the pocket that contains at least 20% of all the small molecule atoms. If there is no precalculated pocket for the small molecule, we calculate a new one using all heavy atoms within a 6.5 Å radius of any of the molecule atoms. The chosen radius represents a reasonable trade-off between specificity and runtime. We decided to use all atoms instead of the ligand's geometric center and the radius thereof to more accurately capture the binding site shape. The pockets that are not associated with a ligand are now filtered based on two criteria. First, the pocket volume has to exceed 100 Å$^3$. Pocket volumes are calculated using the DoGSite pocket grids. The threshold is motivated by the fact that at least three water molecules should be accommodated by the pocket. Second, the largest $k$ pockets are selected, where $k$ is limited to two times the number of protein chains in the asymmetric unit.

Subsequently, several pocket characteristics,[36] secondary structure assignments from the PDB annotation or those calculated by an in-house version of DSSP[38] if the helix and sheet assignments are missing in the PDB file, and all protein-ligand interactions are calculated. Interactions include hydrogen bonds based on the predicted protonation or tautomerism patterns, aromatic interactions between rings, ionic interactions, metal interactions, and $\pi$-$\pi$ interactions. The interactions are calculated according to Inhester et al.[7] A hydrogen bond is assigned if the distance between the corresponding donor and acceptor atoms is between 2 and 3.8 Å, the hydrogen-donor-acceptor angle is between −45° and 45°, the donor-acceptor-lone pair angle is between −70° and 70°, and the distance between the hydrogen atom and the lone pair is between 0 and 2 Å. $\pi-\pi$ interactions are assigned if the centroids of the interacting ring systems are between 2.5 and 5 Å apart. For $\pi$-cation interactions, the distance between the ring center and the cation has to be between 2 and 4 Å. Metal interactions are identified if the metal ion is between 1 and 3 Å away from any coordinating atom. Ionic hydrogen bonds are identified if the distance between the interaction partners corresponds to the sum of their corresponding van der Waals radii ±1 Å. Further, the solvent-accessible surface area[39,40] of each pocket atom is calculated as described for the scoring function HYDE[41−43] so that buried atoms can be differentiated from solvent-exposed atoms in queries.

**Database Content and Design.** In GeoMine, there are multiple tables stored that can be divided into four groups (see Figure 5). This schema was already used in GeoMine's predecessor PELIKAN and was only slightly altered because some features that only exist in SQLite and not in PostgreSQL were used, e.g., the possibility of storing multiple data types in a single column. The first group stores general information about all small molecules, metal ions, water, proteins, and binding sites. Additionally, the molecular interactions and atomic information are stored there. With this information, the protein-ligand complex can be searched and also reconstructed when visualizing results. The second group stores fingerprints of all small molecules to enable similarity searching. In the third group, the textual and numerical properties of the ligands, proteins, and pockets such as the ligand names and resolution or the pocket volume are stored. Lastly, the fourth group contains all information and data about the potential query points, i.e., heavy atoms, ring centers, secondary structure points, and interactions. Each query point is stored with a unique ID and, among others, its coordinates, its chemical element, a foreign key, i.e., a key that links to the complexes table with the pocket properties, its solvent accessibility, its originating molecule (protein or ligand), and vectors in case of ring normals and secondary structure element end- or midpoints. Interactions are defined by the interaction type and the two query point IDs of the interacting atoms. Using the B-tree database indexes available in PostgreSQL enables fast access to this data and geometric queries.

**GeoMine Query Types.** GeoMine enables the search for properties of interest and geometric patterns in different ways. All query types can be combined.

**Textual and Numerical Searches.** The most basic search of GeoMine is uses textual and numerical properties. This is helpful for a preselection or restriction of the search or the selection of a query template structure. A variety of filters for ligands, proteins, pockets, and


https://doi.org/10.1021/acs.jmedchem.1c01046
J. Med. Chem. 2022, 65, 1384−1395

protein-ligand complexes are available. For ligands, an element count, i.e., the minimum and maximum number of specific chemical elements, a functional group count, and molecular properties such as the molecular weight or the number of hydrogen bond donors and acceptors can be defined. Proteins can be filtered by their UniProt ID,[44] EC number, or source organisms. A minimum and maximum count of specific amino acids the search space to relevant protein can be set to filter the pockets. Additionally, ligand names and pocket properties such as hydrophobicity, volume, or depth can be specified. Finally, the number of complexes can be reduced by title entries, resolution, organism, EC number, and experimental source. Handling and querying these data is straightforward in a relational database.

**Chemical Substructure Searches.** Substructure searches can be carried out using simplified molecular input line entry system (SMILES) arbitrary target specification (SMARTS) strings.[45] Large SMARTS patterns (patterns describing at least five atoms) and those containing rare elements are highly discriminative and can therefore be used to reduce the number of pockets in the search space early on. SMARTS patterns that do not fulfill these criteria will be evaluated at the end of the search algorithm because they will likely result in numerous matches, causing an increased runtime.

**Ligand Similarity Searches.** All ligands of the PDB structures are stored with their extended connectivity fingerprint (ECFP)-like[46] Morgan[47] fingerprints and CSFP (connected subgraph fingerprint) and tCSFP (topological connected subgraph fingerprint)[48] fingerprints in the database. The latter are new fingerprints that, in contrast to other methods, capture all connected subgraphs as structural features of a molecule. This property gives these fingerprints a complete feature space and a high adaptive potential. Especially in standard similarity-driven virtual screening settings, the CSFP has substantial advantages.[48] All fingerprints enable the user to define a certain similarity level to a given query ligand specified in SMILES,[49] which further helps to reduce the search space to relevant protein-ligand complexes. Furthermore, they can be used standalone to efficiently search small molecules in GeoMine.

**Geometry-Based Searches.** The unique key feature of GeoMine is its capability for precise geometric searching in binding sites. All atoms, helix and strand mid- and endpoints, and ring centers are possible query points in GeoMine. These points can either be selected using a template structure (loaded PDB file) or specified template-free. Queries based on template structures can be enriched by artificial search points, enabling the generation of hypothetical queries. Furthermore, there are many properties associated with query points that can either be automatically determined or manually adjusted. These properties are the chemical element, the interaction type, the molecule type (protein, ligand, metal, or water), the secondary structure class (helix, sheet, or none), whether they are in the protein backbone or a side chain, and the amino acid or amino acid type, e.g., polar. In the case of the secondary structure annotation, the user can also define a C$\alpha$ atom as an end- or midpoint of a helix or strand. Also, functional groups of ligands and the environment can be specified using SMARTS strings. GeoMine allows connecting atoms in SMARTS expressions with query points such that substructures with certain geometric orientations can be searched. Note that all property specifications are optional. All points that are at most 15 Å apart from each other can be connected by distance and interaction constraints. This value was chosen so that residue atoms on opposite sides of the pocket can be selected while the runtime of the search algorithm remains feasible. The distance constraints enable defined minimum and maximum ranges, while the interaction constraints are used to indicate specific interactions such as hydrogen bonds and ionic, metal, $\pi$-cation, and $\pi$-$\pi$ interactions. In addition, angles can be defined between two distances, interactions, or ring normals with minimum and maximum values to allow different constraint levels.

**Overview of the Search Process.** From the perspective of a user, it is important to get a rough idea of the search process so that the impact of query types and parameters on runtime and results can be estimated. An iterative search algorithm was developed to reduce the potential result set as early as possible while executing the fast search steps before the time-consuming ones. This algorithm consists of four distinct steps, which are displayed in Figure 6. Note that the algorithm is exact, i.e., complexes and binding sites are retrieved if and only if they

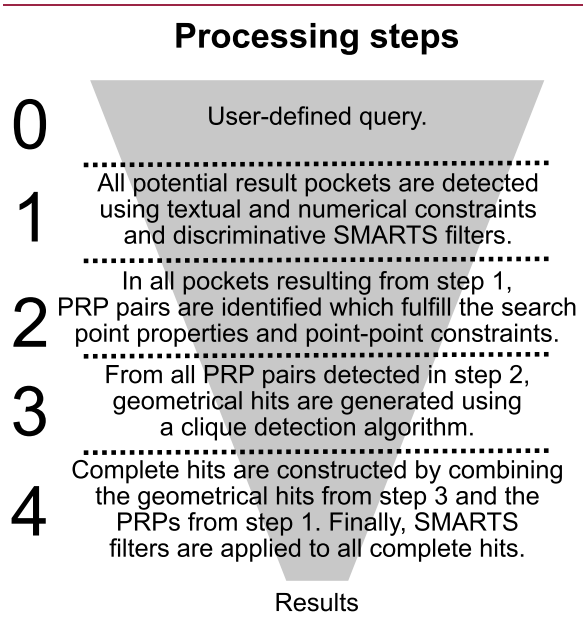



**Figure 6.** Overview of the search process in GeoMine.

fulfill the specified user query. The algorithm differs from the one used in GeoMine's predecessor PELIKAN[7] in several aspects. In the latter, a tailor-made indexing structure was used to identify potential hits (named potential result points, PRPs) quickly. In GeoMine, accessing and checking matching properties of search points in the database are considerably improved with respect to runtime. Details about these improvements are given in the Query Computing Time section.

*Step 1.* All query features regarding textual and numerical properties, ligand similarity searches, and specific SMARTS-based substructure searches, which include more than four atoms or contain at least one non-carbon, non-nitrogen or non-oxygen atom, are performed. This step results in a list of potentially matching binding sites and is passed on to the next step.

*Step 2.* Point-point or distance and interaction constraints are processed sequentially. To reduce the runtime of this step, the processing order of these constraints is optimized beforehand. Here, the order is ascending with respect to the number of expected results, which is estimated by the product of the database count of different elements and interaction types for each search point. In each point-point constraint processing step, all PRPs that match all properties of the search points and are in the list of potential pockets are detected in the database. The interaction type is used if the point-point constraint is an interaction constraint. Set distance ranges are evaluated by calculating the Euclidean distance between the two PRPs. If no matching PRPs are found in a specific pocket for a point-point constraint, this pocket is removed from the list of potential pockets. The list of all potential PRPs is updated in each processing step. By doing this, the size of these lists of possible pockets and PRPs steadily decreases.

This step results in a list containing all possible PRPs that match all properties and point-point constraints for each search point. At this step of the algorithm, the possible PRPs are not yet assigned to pockets.

*Step 3.* The list of all detected PRPs from the previous step is used to construct a product graph for each pocket[50] (see Figure 7). Herein, each combination of a point-point constraint with a matching PRP pair generates a node. These nodes are connected if the PRP assignments do not contradict each other and the angle constraints are fulfilled. To find all matches for the whole query, the maximal cliques are calculated on the product graph.


https://doi.org/10.1021/acs.jmedchem.1c01046
J. Med. Chem. 2022, 65, 1384−1395

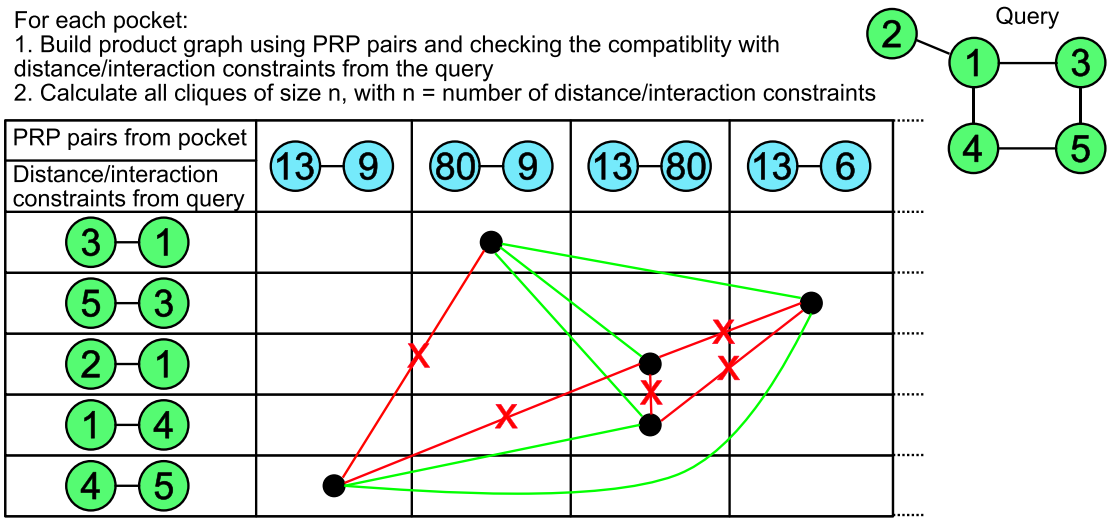



**Figure 7.** Step 3 of GeoMine's search process. For each pocket, a product graph is built using all compatible results (black dots) of pocket PRP pairs (blue circles) to the distance and interaction constraints (green circles) of the user-defined query. Numbers in green and the blue circles represent PRP and search point IDs, respectively. Cliques are calculated for all compatible results. Green lines indicate that two results fulfill the angle constraints, and red lines that they do not. Adapted in part with permission from Inhester et al.[7] Copyright 2017, American Chemical Society.

*Step 4.* SMARTS patterns that were not processed in the first step are evaluated on the resulting PRPs. This includes, for example, SMARTS patterns that describe parts of an atom environment or chemical relations between search points. All these patterns are checked for each of the matching atoms of step 3.

**Database Statistics.** As mentioned before, GeoMine stores information about small molecules in the database. Conceptually, the topology and 3D coordinates of each inserted small molecule are separated. The two properties are stored in different tables so that the different topologies define unique ligands while the 3D coordinates define instances in exactly one binding site. Based on the RCSB PDB (https://www.rcsb.org/),[51] which contains 166 296 entries, there are 44 786 unique ligand topologies in the database. This number is a result of the ligand definition and the adjustment of HETATM annotations in the PDB files mentioned above. Water is the most abundant small molecule. The most frequently occurring small molecules are inorganic ions such as sulfate, zinc, and magnesium ions. The most commonly occurring organic small molecules comprise, for example, heme, *N*-acetylglucosamine, and glycerol. Of the 914 408 available pockets, 412 924 contain a ligand. This means that about 55% of all available pockets contain no reference ligand and should be considered as hypothetical binding sites. On average, the pockets have a volume of 270 Å$^3$ and comprise 200 heavy atoms with a solvent-accessible surface of about 325 Å$^2$. There are about 17 solvent-accessible hydrogen bond acceptors and 12 hydrogen bond donors on average, and the ratio of hydrophobic pocket residues to all pocket residues is 0.376.

## SOFTWARE AND DATA AVAILABILITY

All data used were generated from the Protein Data Bank, which is freely available at PDB, https://www.rcsb.org/. GeoMine is available as a free web service that can be accessed using the link https://proteins.plus. All queries developed throughout this study are available in the Supporting Information.

## ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jmedchem.1c01046.

List of all properties of search points and their value ranges, list of all textual and numerical filters that can be used in GeoMine and their value ranges, exact query definitions of the application examples, and results for the application example queries downloaded with the GeoMine GUI (PDF)

List of all PDB-IDs in the GeoMine database as of May 28, 2021 and all query files in a machine-readable JavaScript object notation (JSON) format that can be imported in GeoMine to reproduce the queries and their results downloaded from GeoMine (ZIP)

## AUTHOR INFORMATION


### Corresponding Author

**Matthias Rarey** − *ZBH Center for Bioinformatics, Universität Hamburg, 20146 Hamburg, Germany;* orcid.org/0000-0002-9553-6531; Email: matthias.rarey@uni-hamburg.de

### Authors

**Joel Graef** − *ZBH Center for Bioinformatics, Universität Hamburg, 20146 Hamburg, Germany;* orcid.org/0000-0001-8327-4936

**Christiane Ehrt** − *ZBH Center for Bioinformatics, Universität Hamburg, 20146 Hamburg, Germany;* orcid.org/0000-0003-1428-0042

**Konrad Diedrich** − *ZBH Center for Bioinformatics, Universität Hamburg, 20146 Hamburg, Germany;* orcid.org/0000-0001-8171-0888

**Martin Poppinga** − *ZBH Center for Bioinformatics, Universität Hamburg, 20146 Hamburg, Germany*

**Norbert Ritter** − *ZBH Center for Bioinformatics, Universität Hamburg, 20146 Hamburg, Germany*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jmedchem.1c01046


### Author Contributions

J.G. implemented the C++ PostgreSQL interface of GeoMine, developed faster SQL queries and core functionalities, and integrated the ligand similarity search, empty binding sites, secondary structure points, and surface properties to the search capabilities. K.D. implemented the graphical user interface and the functions that provide the results to the GUI. C.E. designed and performed all exemplary protein kinase use cases and



https://doi.org/10.1021/acs.jmedchem.1c01046
J. Med. Chem. 2022, 65, 1384−1395

analyzed the results. M.P. and N.R. participated in database design and query optimization. M.R. participated in the development of concepts and supervised the project. J.G., C.E., and M.R. wrote the manuscript.

### Notes

The authors declare the following competing financial interest(s): Proteins*Plus* and the NAOMI ChemBioSuite use some methods that are jointly owned and/or licensed to BioSolveIT GmbH, Germany, M.R. is a shareholder of BioSolveIT GmbH.

## ■ ACKNOWLEDGMENTS

The authors thank the whole development team of the NAOMI library for forming the basis of this work as well as the developers of PostgreSQL for making their database system available. This work was supported by the German Federal Ministry of Education and Research as part of CompLS and de.NBI (031L0172 and 031L0105). C.E. is funded by Data Science in Hamburg, Helmholtz Graduate School for the Structure of Matter (Grant HIDDS-0002).

## ■ ABBREVIATIONS USED

ACID, atomicity, consistency, isolation, and durability;; BS, binding site; CSFP, connected subgraph fingerprint; ECFP, extended connectivity fingerprint; GUI, graphical user interface; JSON, JavaScript object notation; KLIFS, Kinase-Ligand Interaction Fingerprints and Structures; PRP, potential result point; RCSB, research collaboration for structural bioinformatics; REST API, representational state transfer application programming interface; SMARTS, SMILES arbitrary target specification; SMILES, simplified molecular input line entry system; SQL, structure query language; SSE, secondary structure element; tCSFP, topological connected subgraph fingerprint; XML, extensible markup language

## ■ REFERENCES

(1) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.

(2) Korb, O.; Kuhn, B.; Hert, J.; Taylor, N.; Cole, J.; Groom, C.; Stahl, M. Interactive and Versatile Navigation of Structural Databases. *J. Med. Chem.* **2016**, *59*, 4257−4266.

(3) Mobilio, D.; Walker, G.; Brooijmans, N.; Nilakantan, R.; Denny, R. A.; DeJoannis, J.; Feyfant, E.; Kowticwar, R. K.; Mankala, J.; Palli, S.; Punyamantula, S.; Tatipally, M.; John, R. K.; Humblet, C. A Protein Relational Database and Protein Family Knowledge Bases to Facilitate Structure-Based Design Analyses. *Chem. Biol. Drug Des.* **2010**, *76*, 142−153.

(4) Weisel, M.; Bitter, H.-M.; Diederich, F.; So, W. V.; Kondru, R. PROLIX: Rapid Mining of Protein−Ligand Interactions in Large Crystal Structure Databases. *J. Chem. Inf. Model.* **2012**, *52*, 1450−1461.

(5) Hendlich, M.; Bergner, A.; Günther, J.; Klebe, G. Relibase: Design and Development of a Database for Comprehensive Analysis of Protein−Ligand Interactions. *J. Mol. Biol.* **2003**, *326*, 607−620.

(6) Golovin, A.; Henrick, K. MSDmotif: Exploring Protein Sites and Motifs. *BMC Bioinf.* **2008**, *9*, 312.

(7) Inhester, T.; Bietz, S.; Hilbig, M.; Schmidt, R.; Rarey, M. Index-Based Searching of Interaction Patterns in Large Collections of Protein−Ligand Interfaces. *J. Chem. Inf. Model.* **2017**, *57*, 148−158.

(8) Angles, R.; Arenas-Salinas, M.; García, R.; Reyes-Suarez, J. A.; Pohl, E. GSP4PDB: A Web Tool to Visualize, Search and Explore Protein-Ligand Structural Patterns. *BMC Bioinf.* **2020**, *21*, 85.

(9) Bittrich, S.; Burley, S. K.; Rose, A. S. Real-Time Structural Motif Searching in Proteins Using an Inverted Index Strategy. *PLoS Comput. Biol.* **2020**, *16*, No. e1008502.

(10) Kuhn, B.; Gilberg, E.; Taylor, R.; Cole, J.; Korb, O. How Significant Are Unusual Protein−Ligand Interactions? Insights from Database Mining. *J. Med. Chem.* **2019**, *62*, 10441−10455.

(11) Yoshikawa, K.; Yoshino, T.; Yokomizo, Y.; Uoto, K.; Naito, H.; Kawakami, K.; Mochizuki, A.; Nagata, T.; Suzuki, M.; Kanno, H.; Takemura, M.; Ohta, T. Design, Synthesis and SAR of Novel Ethylenediamine and Phenylenediamine Derivatives as Factor Xa Inhibitors. *Bioorg. Med. Chem. Lett.* **2011**, *21*, 2133−2140.

(12) Kanev, G. K.; de Graaf, C.; Westerman, B. A.; de Esch, I. J. P.; Kooistra, A. J. KLIFS: An Overhaul After the First 5 Years of Supporting Kinase Research. *Nucleic Acids Res.* **2021**, *49*, D562−D569.

(13) KLIFS. https://klifs.net/ (accessed 2021-03-24).

(14) Baumli, S.; Endicott, J. A.; Johnson, L. N. Halogen Bonds Form the Basis for Selective P-TEFb Inhibition by DRB. *Chem. Biol.* **2010**, *17*, 931−936.

(15) Rothweiler, U.; Stensen, W.; Brandsdal, B. O.; Isaksson, J.; Leeson, F. A.; Engh, R. A.; Svendsen, J. S. M. Probing the ATP-Binding Pocket of Protein Kinase DYRK1A with Benzothiazole Fragment Molecules. *J. Med. Chem.* **2016**, *59*, 9814−9824.

(16) Do, P. A.; Lee, C. H. The Role of CDK5 in Tumours and Tumour Microenvironments. *Cancers* **2021**, *13*, 101.

(17) Ahn, J. S.; Radhakrishnan, M. L.; Mapelli, M.; Choi, S.; Tidor, B.; Cuny, G. D.; Musacchio, A.; Yeh, L.-A.; Kosik, K. S. Defining Cdk5 Ligand Chemical Space with Small Molecule Inhibitors of Tau Phosphorylation. *Chem. Biol.* **2005**, *12*, 811−823.

(18) Eid, S.; Turk, S.; Volkamer, A.; Rippmann, F.; Fulle, S. KinMap: A Web-Based Tool for Interactive Navigation Through Human Kinome Data. *BMC Bioinf.* **2017**, *18*, 16.

(19) Wood, E. R.; Truesdale, A. T.; McDonald, O. B.; Yuan, D.; Hassell, A.; Dickerson, S. H.; Ellis, B.; Pennisi, C.; Horne, E.; Lackey, K.; Alligood, K. J.; Rusnak, D. W.; Gilmer, T. M.; Shewchuk, L. A Unique Structure for Epidermal Growth Factor Receptor Bound to GW572016 (Lapatinib). *Cancer Res.* **2004**, *64*, 6652−6659.

(20) Kitagawa, D.; Yokota, K.; Gouda, M.; Narumi, Y.; Ohmoto, H.; Nishiwaki, E.; Akita, K.; Kirii, Y. Activity-Based Kinase Profiling of Approved Tyrosine Kinase Inhibitors. *Genes to Cells* **2013**, *18*, 110−122.

(21) Kortemme, T.; Creighton, T. E. Ionisation of Cysteine Residues at the Termini of Model α-Helical Peptides. Relevance to Unusual Thiol pKa Values in Proteins of the Thioredoxin Family. *J. Mol. Biol.* **1995**, *253*, 799−812.

(22) Liu, Q.; Sabnis, Y.; Zhao, Z.; Zhang, T.; Buhrlage, S.; Jones, L.; Gray, N. Developing Irreversible Inhibitors of the Protein Kinase Cysteinome. *Chem. Biol.* **2013**, *20*, 146−159.

(23) do Amaral, D. N.; Lategahn, J.; Fokoue, H. H.; da Silva, E. M. B.; Sant'Anna, C. M. R.; Rauh, D.; Barreiro, E. J.; Laufer, S.; Lima, L. M. A Novel Scaffold for EGFR Inhibition: Introducing N-(3-(3-Phenylureido)Quinoxalin-6-yl) Acrylamide Derivatives. *Sci. Rep.* **2019**, *9*, 14.

(24) Aertgeerts, K.; Skene, R.; Yano, J.; Sang, B.-C.; Zou, H.; Snell, G.; Jennings, A.; Iwamoto, K.; Habuka, N.; Hirokawa, A.; Ishikawa, T.; Tanaka, T.; Miki, H.; Ohta, Y.; Sogabe, S. Structural Analysis of the Mechanism of Inhibition and Allosteric Activation of the Kinase Domain of HER2 Protein. *J. Biol. Chem.* **2011**, *286*, 18756−18765.

(25) Petri, L.; Egyed, A.; Bajusz, D.; Imre, T.; Hetényi, A.; Martinek, T.; Ábrányi Balogh, P.; Keseru, G. An Electrophilic Warhead Library for Mapping the Reactivity and Accessibility of Tractable Cysteines in Protein Kinases. *Eur. J. Med. Chem.* **2020**, *207*, 112836.

(26) Heppner, D. E.; Günther, M.; Wittlinger, F.; Laufer, S. A.; Eck, M. J. Structural Basis for EGFR Mutant Inhibition by Trisubstituted Imidazole Inhibitors. *J. Med. Chem.* **2020**, *63*, 4293−4305.

(27) Goedken, E. R.; Argiriadi, M. A.; Banach, D. L.; Fiamengo, B. A.; Foley, S. E.; Frank, K. E.; George, J. S.; Harris, C. M.; Hobson, A. D.; Ihle, D. C.; Marcotte, D.; Merta, P. J.; Michalak, M. E.; Murdock, S. E.; Tomlinson, M. J.; Voss, J. W. Tricyclic Covalent Inhibitors Selectively Target Jak3 through an Active Site Thiol. *J. Biol. Chem.* **2015**, *290*, 4573−4589.

(28) Desaphy, J.; Bret, G.; Rognan, D.; Kellenberger, E. sc-PDB: A 3D-database of Ligandable Binding Sites—10 years on. *Nucleic Acids Res.* **2015**, *43*, D399—D404.
(29) sc-PDB. http://bioinfo-pharma.u-strasbg.fr/scPDB/ (accessed 2020-07-30).
(30) Burk, D. L.; Hon, W. C.; Leung, A. K.-W.; Berghuis, A. M. Structural Analyses of Nucleotide Binding to an Aminoglycoside Phosphotransferase. *Biochemistry* **2001**, *40*, 8756—8764.
(31) The PostgreSQL Global Development Group. Appendix K. PostgreSQL Limits. *PostgreSQL*. https://www.postgresql.org/docs/12/limits.html (accessed on 2021-03-19).
(32) Rose, A. S.; Hildebrand, P. W. NGL Viewer: A Web Application for Molecular Visualization. *Nucleic Acids Res.* **2015**, *43*, W576—W579.
(33) Rose, A. S.; Bradley, A. R.; Valasatava, Y.; Duarte, J. M.; Prlić, A.; Rose, P. W. NGL viewer: Web-Based Molecular Graphics for Large Complexes. *Bioinformatics* **2018**, *34*, 3755—3758.
(34) Urbaczek, S.; Kolodzik, A.; Fischer, J. R.; Lippert, T.; Heuser, S.; Groth, I.; Schulz-Gasch, T.; Rarey, M. NAOMI: On the Almost Trivial Task of Reading Molecules from Different File formats. *J. Chem. Inf. Model.* **2011**, *51*, 3199—3207.
(35) Bietz, S.; Inhester, T.; Lauck, F.; Sommer, K.; von Behren, M. M.; Fährrolfes, R.; Flachsenberg, F.; Meyder, A.; Nittinger, E.; Otto, T.; Hilbig, M.; Schomburg, K. T.; Volkamer, A.; Rarey, M. From Cheminformatics to Structure-Based Design: Web Services and Desktop Applications Based on the NAOMI Library. *J. Biotechnol.* **2017**, *261*, 207—214.
(36) Volkamer, A.; Kuhn, D.; Grombacher, T.; Rippmann, F.; Rarey, M. Combining Global and Local Measures for Structure-Based Druggability Predictions. *J. Chem. Inf. Model.* **2012**, *52*, 360—372.
(37) Bietz, S.; Urbaczek, S.; Schulz, B.; Rarey, M. Protoss: A Holistic Approach to Predict Tautomers and Protonation States in Protein-Ligand Complexes. *J. Cheminf.* **2014**, *6*, 12.
(38) Kabsch, W.; Sander, C. Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers* **1983**, *22*, 2577—2637.
(39) Lee, B.; Richards, F. The Interpretation of Protein Structures: Estimation of Static Accessibility. *J. Mol. Biol.* **1971**, *55*, 379—400.
(40) Richards, F. M. Areas, Volumes, Packing, and Protein Structure. *Annu. Rev. Biophys. Bioeng.* **1977**, *6*, 151—176.
(41) Reulecke, I.; Lange, G.; Albrecht, J.; Klein, R.; Rarey, M. Towards an Integrated Description of Hydrogen Bonding and Dehydration: Decreasing False Positives in Virtual Screening with the HYDE Scoring Function. *ChemMedChem* **2008**, *3*, 885—897.
(42) Schneider, N.; Hindle, S.; Lange, G.; Klein, R.; Albrecht, J.; Briem, H.; Beyer, K.; Claußen, H.; Gastreich, M.; Lemmen, C.; Rarey, M. Substantial Improvements in Large-Scale Redocking and Screening Using the Novel HYDE Scoring Function. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 701—723.
(43) Schneider, N.; Lange, G.; Hindle, S.; Klein, R.; Rarey, M. A Consistent Description of HYdrogen Bond and DEhydration Energies in Protein-Ligand Complexes: Methods Behind the HYDE Scoring Function. *J. Comput.-Aided Mol. Des.* **2013**, *27*, 15—29.
(44) The UniProt Consortium. UniProt: A Worldwide Hub of Protein Knowledge. *Nucleic Acids Res.* **2019**, *47*, D506—D515.
(45) Daylight Chemical Information Systems, Inc. 4. SMARTS - A Language for Describing Molecular Patterns. *Daylight Theory Manual*. http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html (accessed on 2020-04-21).
(46) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742—754.
(47) 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.
(48) Bellmann, L.; Penner, P.; Rarey, M. Connected Subgraph Fingerprints: Representing Molecules Using Exhaustive Subgraph Enumeration. *J. Chem. Inf. Model.* **2019**, *59*, 4625—4635.
(49) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Model.* **1988**, *28*, 31—36.
(50) Bron, C.; Kerbosch, J. Algorithm 457: Finding All Cliques of an Undirected Graph. *Commun. ACM* **1973**, *16*, 575—577.
(51) RCSB PDB. https://www.rcsb.org/ (accessed 2020-12-16).


https://doi.org/10.1021/acs.jmedchem.1c01046
*J. Med. Chem.* 2022, 65, 1384—1395

# Supporting Information

# Searching Geometric Patterns in Protein Binding Sites and its Application to Data Mining in Protein Kinase Structures

Author Names: Joel Graef, Christiane Ehrt, Konrad Diedrich, Martin Poppinga, Norbert Ritter, Matthias Rarey*

Author Address: Universität Hamburg, Center for Bioinformatics, Bundesstrasse 43, 20146 Hamburg, Germany

E-mail: matthias.rarey@uni-hamburg.de

**Table of Contents**



**Table S1**. Properties which can be assigned to a search point in the geometrical query and their possible values.

| **Property** | **Possible choices** |
|---|---|
| Original Molecule | Ligand, Metal, Protein, Water |
| Element | Alpha Carbon, Boron, Bromine, Calcium, Carbon, Chlorine, Cobalt, Copper, Fluorine, Iodine, Iron, Magnesium, Manganese, Nickel, Nitrogen, Oxygen, Phosphorus, Sulfur, Zinc |

| | | |
|---|---|---|
| Interaction type | | Acceptor, Anion, AromaticRingCenter, Cation, Donor, Hydrophobic, Metal |
| If Original Molecule = Protein | Amino acid | Ala, Arg, Asn, Asp, Cso, Cys, Glu, Gln, Gly, His, Ile, Leu, Lys, Met, Phe, Pro, Ser, Thr, Trp, Tyr, Val, Hydrophobic, Polar, Aromatic, Acidic, Basic, Neutral |
| | Location in amino acid | Backbone, Sidechain |
| | Secondary structure | Helix, Sheet, Helix End, Helix Mid, Sheet End, Sheet Mid, no secondary structure |
| If Original Molecule = Ligand | Functional group | Alcohol, Aldehyde, Amide, Amidine, Amine, Azide, Ester, Ether, Furane, Guanidine, Ketone, Nitrile, Phenyl, Pyridine, Pyrrole, Thiophene |
| If Original Molecule = Ligand or Protein | Atom description | SMARTS |
| | Minimal surface | All floating-point numbers $\geq 0$ Å$^2$ |

**Table S2**. Textual and numerical properties which can be added to a query and their possible values.

| **Category** | **Property** | **Possible choices** |
|---|---|---|
| Ligand filter | Element | Boron, Bromine, Carbon, Chlorine, Fluorine, Iodine, Nitrogen, Oxygen, Phosphorus, Sulfur, and a count (min, max) |
| | Functional group | Alcohol, Aldehyde, Amide, Amidine, Amine, Azide, Ester, Ether, Furane, Guanidine, Ketone, Nitrile, Phenyl, Pyridine, Pyrrole, Thiophene, and a count (min, max) |
| | Molecule property | Acceptors, aromatic atoms, aromatic rings, aromatic ringsystems, charge, cyclomatic number, donors, halogens, heavy atoms, hetero atoms, inorganic atoms, Lipinski acceptors, logP, molecular weight (MW), Max continuous path of rotatable bonds, max cyclomatic number, max ring size, max ringsystem size, rings, ringsystems, rotatable bonds, stereo bonds (E/Z), stereo centers (R/S), topological polar surface area, unique ring families (URFs), volume, and a count (min, max) |

| | | |
|---|---|---|
| | Similarity | CSFP, tCSFP, ECFPlike, a similarity percentage, the pocket name to which other ligands should be similar to and a min and max for CSFP and tCSFP or radius for ECFPlike Morgan fingerprint |
| | SMARTS | SMARTS string |
| Protein filter | Uniprot ID | Free text |
| | Organism | Free text |
| | EC number | All four numbers can be set individually |
| Pocket filter | Ligand name | Free text |
| | Amino acids | Ala, Arg, Asn, Asp, Cso, Cys, Glu, Gln, Gly, His, Ile, Leu, Lys, Met, Phe, Pro, Ser, Thr, Trp, Tyr, Val, and a count (min, max) |
| | Property | Acceptors, depth, donors, heavy atoms, hydrophobicity, metal, DoGSite simple score, surface, surface-volume-ratio, volume, and a value range (min, max) |
| | Has ligand | No parameter. If filter is set only pockets containing a ligand are searched. |
| P-L-Complex-filter | PDB title entry | Free text |
| | Resolution | Value range (min, max) |
| | Experimental source | Unknown, electron crystallography, electron microscopy, fiber diffraction, NMR solid state, NMR solution, neutron diffraction, solution scattering, X-ray |

**Paragraph S1**. Additional filters

In addition to the filters above there exists a PDB subselection option. The user can provide a list of PDB IDs to limit the search to specific PDBs.

**Table S3.** Overview of existing tools for 3D searching in structural data related to data processing and storage.

| | Dataprocessing and storage |
|---|---|
| Relibase/Relibase+ | - C++-based self-written database system with objects stored in multiple archives in a B-tree data structure |

| | |
|---|---|
| | - binding site definition: 7 Å of the ligand atoms |
| PDBeMotif/MSDmotif | - PostgreSQL<br>- binding site definition: 16 Å of the ligand atoms |
| PRDB | - information about database type not provided in the publication<br>- binding site definition: 8 Å of the ligand atoms |
| PROLIX | - information about database type not provided in the publication<br>- binding site definition: 4.5 Å of the ligand atoms |
| CSD-CrossMiner | - SQLite<br>- HET groups with less than 5 or more than 100 atoms are removed<br>- binding site definition: 6 Å of the ligand atoms |
| PELIKAN | - SQLite<br>- binding site definition: 6.5 Å of the ligand atoms |
| GSP4PDB | - PostgreSQL<br>- HET entries as ligands<br>- binding site definition: 7.0 Å of the ligand atoms |
| strucmotif-search | - inverted index implemented by a file system-based approach - no database<br>- motifs of at least 3 and at most 10 amino acids<br>- distances between animo acid pairs of 15 Å at maximum<br>- all ligands are removed |

**Table S4.** Overview of existing tools for 3D searching in structural data related to search capabilities and algorithms, and result presentation.

| | Search capabilities and algorithms | Result presentation |
|---|---|---|
| Relibase/Relibase+ | - incremental<br>- start with fingerprints | - superimposition based on sequence<br>- list of results and statistics |
| PDBeMotif/MSDmotif | - no information in publication other than the search being based on Cα coordinates or end of sidechain coordinates | - superimpose similar proteins (only available for sequence-based searches)<br>- list of results and statistics |
| PRDB | - conversion from search query to SQL query or directly as database SQL | - list of results |

| | | |
|---|---|---|
| | query without using the interface | |
| PROLIX | - incremental<br>- start with ligand fingerprints | - list of results and statistics |
| CSD-CrossMiner | - incremental search which starts with fingerprints | - superimpositions based on the geometric query<br>- list of results |
| PELIKAN | - incremental<br>- start with points from index structure using environment as triangles<br>- exact (including symmetric matches) | - superimpose results based on the geometric query<br>- list of results and statistics |
| GSP4PDB | - graph-based structural pattern query is transformed into an SQL query | - 2D and 3D alignment<br>- list of results |
| strucmotif-search | - motifs are split into residue pairs, similar occurrences are retrieved for geometric descriptors via an inverted index lookup, checking candidates for resemblance to query motif | - Alignment of motifs to query as well as complete structure that contains query or motif according to publication[1]<br>- list of results with a score based on geometric properties of residues in the query and the matched structure pair |

**Table S5**. Queries and results of application examples in paper

Files for all queries in the paper are given in the supporting information. JSON files are queries which were exported in GeoMine and can be used with with the Import function (Import button next to the search button in GeoMine on ProteinsPlus. Results of the queries downloaded from GeoMine are in the correspondingly named ZIP files. Below are the names and a short description of the queries.

| **Filename** | **Description** |
|---|---|
| Exploiting_Unusual_Interactions _pdb_search | Application example: Exploiting Unusual Interactions for Selective Inhibitor Design<br>Query: Search for appropriate geometric arrangements of halogen atoms attached to aromatic rings with aromatic ring systems in protein sidechains in all ligand-occupied pockets as stored in the PDB. |
| Exploiting_Unusual_Interactions _KLIFS | Application example: Exploiting Unusual Interactions for Selective Inhibitor Design |

[1] Bittrich, S.; Burley, S. K.; Rose, A. S. Real-time structural motif searching in proteins using an inverted index strategy. bioRxiv 2020

| | |
|---|---|
| | Query: Investigation of interaction geometries between aromatic sidechains and chlorine atoms attached to aromatic (hetero-)cycles in the published structures assembled in the third version of the Kinase–Ligand Interaction Fingerprints and Structures (KLIFS) database (https://klifs.net/, accessed on 03/24/2021). |
| Exploiting_Unusual_Interactions Adapated_query_KLIFS | Application example: Exploiting Unusual Interactions for Selective Inhibitor Design<br>Query: Adaption of the KLIFS query in all ligand-bound protein kinase structures highlighting that there are numerous kinase inhibitors with known binding modes that could be used as potential starting points to improve compound selectivity. |
| Exploiting_Unusual_Interactions _CDK5 | Application example: Exploiting Unusual Interactions for Selective Inhibitor Design<br>Query: There are numerous kinase inhibitors with known binding modes that could be used as potential starting points to improve compound selectivity issues. Cyclin-dependent-like kinase 5 (CDK5) is used as an example for a pharmacologically relevant protein kinase target. This query continues the search in all PDB-IDs that result from the 'Adapated_query_KLIFS'. |
| Exploiting_Unusual_Interactions _3o0g_selectivity_permissive | Application example: Exploiting Unusual Interactions for Selective Inhibitor Design<br>Query: Investigation if CDK5 in complex with an ATP analogue (PDB-ID 3o0g) might become more selective upon addition of a chlorine atom which interacts with Phe80. No restriction to ligand-occupied pockets but all predicted and ligand-based to identify as many potential off-targets as possible. |
| Exploiting_Unusual_Interactions _3o0g_selectivity_restrictive | Application example: Exploiting Unusual Interactions for Selective Inhibitor Design<br>Query: Addition of the aromatic center of Phe80 as additional interaction point for a potential halogen aromatic interaction. |
| Searching_for_Selectivity_Anchors _in_Protein_Kinases_4wkq | Application example: Searching for Selectivity Anchors in Protein Kinases<br>Query: Search for potential off-targets in empty predicted binding sites by using EGFR in complex with inhibitor gefitinib (PDB-ID 4wkq) as template and screening the protein kinases stored in the KLIFS database. |

| | |
|---|---|
| Searching_for_Selectivity_Anchors _in_Protein_Kinases_1xkk | Application example: Searching for Selectivity Anchors in Protein Kinases<br>Query: Extension of the application examples query in 4wkq by one sidechain of residue Thr790 as potential selectivity anchor as known from the structure of EGFR in complex with lapatinib (PDB-ID 1xkk). |
| Searching_for_Reactive_Cysteins _in_Protein_Kinases_3poz | Application example: Searching for Reactive Cysteines in Protein Kinases<br>Query: Demonstration of how the inclusion of secondary structure elements and solvent exposure enables the search for protein kinases with reactive cysteines in the neighborhood of known inhibitors based on the crystal structure of the protein kinase EGFR (PDB-ID 3poz). |

**Paragraph S2**. Query definitions of runtime analysis and comparison between PELIKAN and GeoMine

*Linear - standard*
Search point 1: Original Molecule = Reference ligand, Element = Oxygen, Interaction type = Undefined, Functional group = any, Amino acid = Any, Atom name = Any, Sec. structure = Any, Amino acid location = Any
Search point 2: Original Molecule = Protein, Element = Nitrogen, Interaction type = Undefined, Functional group = any, Amino acid = Any, Atom name = Any, Sec. structure = Any, Amino acid location = Any
Search point 3: Original Molecule = Reference ligand, Element = Nitrogen, Interaction type = Undefined, Functional group = any, Amino acid = Any, Atom name = Any, Sec. structure = Any, Amino acid location = Any
Search point 4: Original Molecule = Protein, Element = Oxygen, Interaction type = Undefined, Functional group = any, Amino acid = Any, Atom name = Any, Sec. structure = Any, Amino acid location = Any
Distance constraint between search points 1-2: min = 6.1 Å, max = 7.1 Å
Distance constraint between search points 2-3: min = 7.6 Å, max = 8.6 Å
Distance constraint between search points 3-4: min = 4.1 Å, max = 5.1 Å

*Linear – metal*
Search point 1: Original Molecule = Reference ligand, Element = Oxygen, Interaction type = Undefined, Functional group = any, Amino acid = Any, Atom name = Any, Sec. structure = Any, Amino acid location = Any
Search point 2: Original Molecule = Protein, Element = Nitrogen, Interaction type = Undefined, Functional group = any, Amino acid = Any, Atom name = Any, Sec. structure = Any, Amino acid location = Any
Search point 3: Original Molecule = Metal, Element = Magnesium, Interaction type = Metal, Functional group = any, Amino acid = Any, Atom name = Any, Sec. structure = Any, Amino acid location = Any

Search point 4: Original Molecule = Protein, Element = Oxygen, Interaction type = Undefined, Functional group = any, Amino acid = Any, Atom name = Any, Sec. structure = Any, Amino acid location = Any
Distance constraint between search points 1-2: min = 6.1 Å, max = 7.1 Å
Distance constraint between search points 2-3: min = 9 Å, max = 10 Å
Distance constraint between search points 3-4: min = 4.1 Å, max = 5.1 Å

*Linear – metal, water*
Search point 1: Original Molecule = Reference ligand, Element = Oxygen, Interaction type = Undefined, Functional group = any, Amino acid = Any, Atom name = Any, Sec. structure = Any, Amino acid location = Any
Search point 2: Original Molecule = Protein, Element = Nitrogen, Interaction type = Undefined, Functional group = any, Amino acid = Any, Atom name = Any, Sec. structure = Any, Amino acid location = Any
Search point 3: Original Molecule = Metal, Element = Magnesium, Interaction type = Metal, Functional group = any, Amino acid = Any, Atom name = Any, Sec. structure = Any, Amino acid location = Any
Search point 4: Original Molecule = Water, Element = Undefined, Interaction type = Undefined, Functional group = any, Amino acid = Any, Atom name = Any, Sec. structure = Any, Amino acid location = Any
Distance constraint between search points 1-2: min = 6.1 Å, max = 7.1 Å
Distance constraint between search points 2-3: min = 9 Å, max = 10 Å
Distance constraint between search points 3-4: min = 6 Å, max = 7 Å

*Linear – metal, water, phosphorus*
Search point 1: Original Molecule = Reference ligand, Element = Phosphorus, Interaction type = Undefined, Functional group = any, Amino acid = Any, Atom name = Any, Sec. structure = Any, Amino acid location = Any
Search point 2: Original Molecule = Protein, Element = Nitrogen, Interaction type = Undefined, Functional group = any, Amino acid = Any, Atom name = Any, Sec. structure = Any, Amino acid location = Any
Search point 3: Original Molecule = Metal, Element = Magnesium, Interaction type = Metal, Functional group = any, Amino acid = Any, Atom name = Any, Sec. structure = Any, Amino acid location = Any
Search point 4: Original Molecule = Water, Element = Undefined, Interaction type = Undefined, Functional group = any, Amino acid = Any, Atom name = Any, Sec. structure = Any, Amino acid location = Any
Distance constraint between search points 1-2: min = 6.7 Å, max = 7.7 Å
Distance constraint between search points 2-3: min = 9 Å, max = 10 Å
Distance constraint between search points 3-4: min = 6 Å, max = 7 Å

*Star - standard*
Search point 1: Original Molecule = Reference ligand, Element = Oxygen, Interaction type = Undefined, Functional group = any, Amino acid = Any, Atom name = Any, Sec. structure = Any, Amino acid location = Any
Search point 2: Original Molecule = Protein, Element = Nitrogen, Interaction type = Undefined, Functional group = any, Amino acid = Any, Atom name = Any, Sec. structure = Any, Amino acid location = Any
Search point 3: Original Molecule = Reference ligand, Element = Nitrogen, Interaction type = Undefined, Functional group = any, Amino acid = Any, Atom name = Any, Sec. structure = Any, Amino acid location = Any

Search point 4: Original Molecule = Protein, Element = Oxygen, Interaction type = Undefined, Functional group = any, Amino acid = Any, Atom name = Any, Sec. structure = Any, Amino acid location = Any
Search point 5: Original Molecule = Reference ligand, Element = Nitrogen, Interaction type = Undefined, Functional group = any, Amino acid = Any, Atom name = Any, Sec. structure = Any, Amino acid location = Any
Search point 6: Original Molecule = Protein, Element = Carbon, Interaction type = Undefined, Functional group = any, Amino acid = Any, Atom name = Any, Sec. structure = Any, Amino acid location = Any
Distance constraint between search points 1-2: min = 6.1 Å, max = 7.1 Å
Distance constraint between search points 1-3: min = 3.5 Å, max = 4.5 Å
Distance constraint between search points 1-4: min = 6.7 Å, max = 7.7 Å
Distance constraint between search points 1-5: min = 3.3 Å, max = 4.3 Å
Distance constraint between search points 1-6: min = 4.1 Å, max = 5.1 Å

*Star – metal*
Search point 1: Original Molecule = Reference ligand, Element = Oxygen, Interaction type = Undefined, Functional group = any, Amino acid = Any, Atom name = Any, Sec. structure = Any, Amino acid location = Any
Search point 2: Original Molecule = Protein, Element = Nitrogen, Interaction type = Undefined, Functional group = any, Amino acid = Any, Atom name = Any, Sec. structure = Any, Amino acid location = Any
Search point 3: Original Molecule = Metal, Element = Magnesium, Interaction type = Metal, Functional group = any, Amino acid = Any, Atom name = Any, Sec. structure = Any, Amino acid location = Any
Search point 4: Original Molecule = Protein, Element = Oxygen, Interaction type = Undefined, Functional group = any, Amino acid = Any, Atom name = Any, Sec. structure = Any, Amino acid location = Any
Search point 5: Original Molecule = Reference ligand, Element = Nitrogen, Interaction type = Undefined, Functional group = any, Amino acid = Any, Atom name = Any, Sec. structure = Any, Amino acid location = Any
Search point 6: Original Molecule = Protein, Element = Carbon, Interaction type = Undefined, Functional group = any, Amino acid = Any, Atom name = Any, Sec. structure = Any, Amino acid location = Any
Distance constraint between search points 1-2: min = 6.1 Å, max = 7.1 Å
Distance constraint between search points 1-3: min = 4 Å, max = 5 Å
Distance constraint between search points 1-4: min = 6.7 Å, max = 7.7 Å
Distance constraint between search points 1-5: min = 3.3 Å, max = 4.3 Å
Distance constraint between search points 1-6: min = 4.1 Å, max = 5.1 Å

*Star – metal, water*
Search point 1: Original Molecule = Reference ligand, Element = Oxygen, Interaction type = Undefined, Functional group = any, Amino acid = Any, Atom name = Any, Sec. structure = Any, Amino acid location = Any
Search point 2: Original Molecule = Protein, Element = Nitrogen, Interaction type = Undefined, Functional group = any, Amino acid = Any, Atom name = Any, Sec. structure = Any, Amino acid location = Any
Search point 3: Original Molecule = Metal, Element = Magnesium, Interaction type = Metal, Functional group = any, Amino acid = Any, Atom name = Any, Sec. structure = Any, Amino acid location = Any

Search point 4: Original Molecule = Water, Element = Undefined, Interaction type = Undefined, Functional group = any, Amino acid = Any, Atom name = Any, Sec. structure = Any, Amino acid location = Any
Search point 5: Original Molecule = Reference ligand, Element = Nitrogen, Interaction type = Undefined, Functional group = any, Amino acid = Any, Atom name = Any, Sec. structure = Any, Amino acid location = Any
Search point 6: Original Molecule = Protein, Element = Carbon, Interaction type = Undefined, Functional group = any, Amino acid = Any, Atom name = Any, Sec. structure = Any, Amino acid location = Any
Distance constraint between search points 1-2: min = 6.1 Å, max = 7.1 Å
Distance constraint between search points 1-3: min = 4 Å, max = 5 Å
Distance constraint between search points 1-4: min = 5.3 Å, max = 6.3 Å
Distance constraint between search points 1-5: min = 3.3 Å, max = 4.3 Å
Distance constraint between search points 1-6: min = 4.1 Å, max = 5.1 Å

*Star – metal, water, phosphorus*
Search point 1: Original Molecule = Reference ligand, Element = Phosphorus, Interaction type = Undefined, Functional group = any, Amino acid = Any, Atom name = Any, Sec. structure = Any, Amino acid location = Any
Search point 2: Original Molecule = Protein, Element = Nitrogen, Interaction type = Undefined, Functional group = any, Amino acid = Any, Atom name = Any, Sec. structure = Any, Amino acid location = Any
Search point 3: Original Molecule = Metal, Element = Magnesium, Interaction type = Metal, Functional group = any, Amino acid = Any, Atom name = Any, Sec. structure = Any, Amino acid location = Any
Search point 4: Original Molecule = Water, Element = Undefined, Interaction type = Undefined, Functional group = any, Amino acid = Any, Atom name = Any, Sec. structure = Any, Amino acid location = Any
Search point 5: Original Molecule = Reference ligand, Element = Nitrogen, Interaction type = Undefined, Functional group = any, Amino acid = Any, Atom name = Any, Sec. structure = Any, Amino acid location = Any
Search point 6: Original Molecule = Protein, Element = Carbon, Interaction type = Undefined, Functional group = any, Amino acid = Any, Atom name = Any, Sec. structure = Any, Amino acid location = Any
Distance constraint between search points 1-2: min = 6.7 Å, max = 7.7 Å
Distance constraint between search points 1-3: min = 3 Å, max = 4 Å
Distance constraint between search points 1-4: min = 5.2 Å, max = 6.2 Å
Distance constraint between search points 1-5: min = 4.5 Å, max = 5.5 Å
Distance constraint between search points 1-6: min = 4.2 Å, max = 5.2 Å

*Tetrahedron – standard*
Search point 1: Original Molecule = Reference ligand, Element = Oxygen, Interaction type = Undefined, Functional group = any, Amino acid = Any, Atom name = Any, Sec. structure = Any, Amino acid location = Any
Search point 2: Original Molecule = Protein, Element = Nitrogen, Interaction type = Undefined, Functional group = any, Amino acid = Any, Atom name = Any, Sec. structure = Any, Amino acid location = Any
Search point 3: Original Molecule = Reference ligand, Element = Nitrogen, Interaction type = Undefined, Functional group = any, Amino acid = Any, Atom name = Any, Sec. structure = Any, Amino acid location = Any

Search point 4: Original Molecule = Protein, Element = Oxygen, Interaction type = Undefined, Functional group = any, Amino acid = Any, Atom name = Any, Sec. structure = Any, Amino acid location = Any
Distance constraint between search points 1-2: min = 6.1 Å, max = 7.1 Å
Distance constraint between search points 1-3: min = 3.5 Å, max = 4.5 Å
Distance constraint between search points 1-4: min = 6.7 Å, max = 7.7 Å
Distance constraint between search points 2-3: min = 7.6 Å, max = 8.6 Å
Distance constraint between search points 2-4: min = 6.5 Å, max = 7.5 Å
Distance constraint between search points 3-4: min = 4.1 Å, max = 5.1 Å

*Tetrahedron – metal*
Search point 1: Original Molecule = Reference ligand, Element = Oxygen, Interaction type = Undefined, Functional group = any, Amino acid = Any, Atom name = Any, Sec. structure = Any, Amino acid location = Any
Search point 2: Original Molecule = Protein, Element = Nitrogen, Interaction type = Undefined, Functional group = any, Amino acid = Any, Atom name = Any, Sec. structure = Any, Amino acid location = Any
Search point 3: Original Molecule = Metal, Element = Magnesium, Interaction type = Metal, Functional group = any, Amino acid = Any, Atom name = Any, Sec. structure = Any, Amino acid location = Any
Search point 4: Original Molecule = Protein, Element = Oxygen, Interaction type = Undefined, Functional group = any, Amino acid = Any, Atom name = Any, Sec. structure = Any, Amino acid location = Any
Distance constraint between search points 1-2: min = 6.1 Å, max = 7.1 Å
Distance constraint between search points 1-3: min = 5.7 Å, max = 6.7 Å
Distance constraint between search points 1-4: min = 6.7 Å, max = 7.7 Å
Distance constraint between search points 2-4: min = 6.5 Å, max = 7.5 Å
Distance constraint between search points 2-3: min = 9 Å, max = 10 Å
Distance constraint between search points 4-3: min = 4.2 Å, max = 5.2 Å

*Tetrahedron – metal, water*
Search point 1: Original Molecule = Reference ligand, Element = Oxygen, Interaction type = Undefined, Functional group = any, Amino acid = Any, Atom name = Any, Sec. structure = Any, Amino acid location = Any
Search point 2: Original Molecule = Protein, Element = Nitrogen, Interaction type = Undefined, Functional group = any, Amino acid = Any, Atom name = Any, Sec. structure = Any, Amino acid location = Any
Search point 3: Original Molecule = Metal, Element = Magnesium, Interaction type = Metal, Functional group = any, Amino acid = Any, Atom name = Any, Sec. structure = Any, Amino acid location = Any
Search point 4: Original Molecule = Water, Element = Undefined, Interaction type = Undefined, Functional group = any, Amino acid = Any, Atom name = Any, Sec. structure = Any, Amino acid location = Any
Distance constraint between search points 1-2: min = 6.1 Å, max = 7.1 Å
Distance constraint between search points 1-3: min = 5.7 Å, max = 6.7 Å
Distance constraint between search points 1-4: min = 5.3 Å, max = 6.3 Å
Distance constraint between search points 2-3: min = 9 Å, max = 10 Å
Distance constraint between search points 2-4: min = 4.2 Å, max = 5.2 Å
Distance constraint between search points 3-4: min = 6 Å, max = 7 Å

*Tetrahedron – metal, water, phosphorus*
Search point 1: Original Molecule = Reference ligand, Element = Phosphorus, Interaction type = Undefined, Functional group = any, Amino acid = Any, Atom name = Any, Sec. structure = Any, Amino acid location = Any
Search point 2: Original Molecule = Protein, Element = Nitrogen, Interaction type = Undefined, Functional group = any, Amino acid = Any, Atom name = Any, Sec. structure = Any, Amino acid location = Any
Search point 3: Original Molecule = Metal, Element = Magnesium, Interaction type = Metal, Functional group = any, Amino acid = Any, Atom name = Any, Sec. structure = Any, Amino acid location = Any
Search point 4: Original Molecule = Water, Element = Undefined, Interaction type = Undefined, Functional group = any, Amino acid = Any, Atom name = Any, Sec. structure = Any, Amino acid location = Any
Distance constraint between search points 2-3: min = 9 Å, max = 10 Å
Distance constraint between search points 2-4: min = 4.2 Å, max = 5.2 Å
Distance constraint between search points 1-2: min = 6.7 Å, max = 7.7 Å
Distance constraint between search points 3-4: min = 6 Å, max = 7 Å
Distance constraint between search points 1-3: min = 4.2 Å, max = 5.2 Å
Distance constraint between search points 1-4: min = 5.2 Å, max = 6.2 Å

## C.4 Database and Workflow Optimizations for Spatial-Geometric Queries in GeoMine

[D4] M. Poppinga, J. Graef, K. Diedrich, M. Rarey, and N. Ritter. Proceedings of the LWDA 2023 Workshops: FGDB, FGBIA, FGKDML, FGWM, and FGIR (LWDA 2023). Maarburg, Deutschland: CEUR-WS.org. Available: https://ceur-ws.org/Vol-3630/LWDA2023-paper8.pdf. Reprinted with permission from [D4] and CEUR Workshop Proceedings.

# Database and Workflow Optimizations for Spatial-Geometric Queries in GeoMine

Martin Poppinga[1,2], Joel Graef[2], Konrad Diedrich[2], Matthias Rarey[2] and Norbert Ritter[1]

[1]Universität Hamburg, Fachbereich Informatik, 22527 Hamburg, Germany
[2]Universität Hamburg, ZBH – Center for Bioinformatics, 20146 Hamburg, Germany

Abstract

Addressing computational problems in science often involves customized algorithmic approaches, which can lead to overlooking well-established solutions in data management and storage. When scientific datasets grow, these customized approaches may struggle to query data efficiently. Effective data management is essential for ensuring accurate and fast analysis of scientific data. Describing changes in the *GeoMine* software, this paper highlights the potential for improvements in data-driven science.

*GeoMine* enables spatial-geometric searches in three-dimensional molecular space, facilitating tasks such as pharmaceutical drug discovery by finding similar geometric patterns in protein-ligand complexes. The original *GeoMine* application utilized a relational database solely for fundamental data storage and combined it with a tailored algorithmic pattern-matching strategy, leaving room for improvements. This work presents a technical overview of database and workflow optimizations in *GeoMine* to handle the increasing data size. Our improvements focus on moving the main computational tasks from the application level to the database system and optimizing the database utilization. A new query design, better utilization of indexes, and optimizations in textual queries led to a 15x speedup in our experiments, reducing the mean runtime of queries to under 8 seconds.

The presented improvements are essential for *GeoMine* to be offered as a service-oriented web application. The success of these improvements highlights the significance of database optimization in science, demonstrating the potential and necessity of proper data management.

**Keywords**

Database optimization, query optimization, data management, databases for bioinformatics

## 1. Introduction

Mining huge datasets is a central task in research. Analyzing molecular interactions between proteins and small organic molecules is essential for understanding disease treatments and advancing medical research. This includes searching for spatial similarities and geometric arrangements, which can provide vital insights into the functional aspects of proteins. Results can be used for further research, for example, in pharmaceutical drug discovery or biotechnology [1]. With the growth of accessible datasets, searching for patterns in this data becomes

*LWDA'23: Lernen, Wissen, Daten, Analysen. October 09–11, 2023, Marburg, Germany*
martin.poppinga@uni-hamburg.de (M. Poppinga); graef@zbh.uni-hamburg.de (J. Graef); diedrich@zbh.uni-hamburg.de (K. Diedrich); rarey@zbh.uni-hamburg.de (M. Rarey); norbert.ritter@uni-hamburg.de (N. Ritter)
0000-0001-8529-8376 (M. Poppinga); 0000-0001-8327-4936 (J. Graef); 0000-0001-8171-0888 (K. Diedrich); 0000-0002-9553-6531 (M. Rarey); 0000-0002-1502-1395 (N. Ritter)

increasingly challenging [2, 3]. Besides the continuous growth of available experimental data, machine-learning-based structure predictions add millions of new structural models [4].

*GeoMine* [3] is an application enabling a visual-guided geometric pattern search of molecular data in three-dimensional space. It is embedded in the *proteins.plus*[1] server [5], a collection of different web-based tools for various tasks in protein-based research. The server is a free service based on publicly available datasets handling over half a million page requests per year. The back end of GeoMine was derived in prior work from the *PELIKAN* application developed in the same group [6], which was utilizing a custom algorithmic approach for query processing. With the Protein Data Bank (PDB) [7] as a fast-growing dataset underlying GeoMine and the shift from a desktop application to a server-based approach, GeoMine required an overhaul of the original query workflow to maintain the ability to provide results in a fast manner.

With this work, we investigate the potential of adopting a database-driven architecture, focusing on the database as the main part of query execution and reducing application-side processing. We were able to reduce the mean runtime in our experiments from about 2 minutes per query to less than 8 seconds, utilizing changes in the workflow and database optimizations. As we present in this work, a substantial performance enhancement has been achieved by shifting to a more database-centric method.

The paper is organized as follows: Section 2 provides an overview of the field of work, the data structure, and the query design; Section 3 details the improvements made to the query workflow and database optimizations; Section 4 presents and discusses the experimental results; Section 5 concludes the paper and outlines future work.

## 2. Background and Related Work

### 2.1. Data Management and Storage

Data management in scientific research involves the systematic collection, organization, storage, and sharing of data to facilitate its reusability and ensure the reproducibility of research findings. In the context of our work, which focuses on querying structured data sets, the storage aspect is particularly important. In the scientific domain, many existing applications are designed for single-user usage, often locally storing data in various formats or utilizing object stores with limited retrieval possibilities [8, 9]. For structured data, *Relational Database Management Systems* (RDBMS) are the most commonly used systems, providing robust and efficient solutions. Commonly, embedded systems are used, such as SQLite [10] for applications with smaller or medium-sized data sizes or DuckDB [9] for analytical workloads. For *Online Transaction Processing* (OLTP) workloads which require fast query performance and regular updates, server-based RDBMS are a popular choice. Large analytical queries are often served by designated *Online Analytical Processing* (OLAP) systems such as data warehouses, which are often proprietary solutions. For handling large-scale semi-structured datasets, NoSQL systems are frequently used, with columnar and graph databases being popular for analytical queries. The choice of data management and storage solutions is crucial to ensure efficient processing, reduced resource consumption, and accurate and fast analysis of scientific data.

[1] https://proteins.plus

**PostgreSQL** GeoMine utilizes PostgreSQL [11], a robust and widely accessible open-source database management system. As multiple users can access a web-based application such as GeoMine at the same time, the ability of a client-server-based database system to handle multiple queries efficiently in parallel is required. PostgreSQL's widespread adoption [12] enables cloud-agnostic hosting on every major platform since most cloud platforms offer PostgreSQL solutions or other PostgreSQL-compatible scalable databases. Additionally, setting up on-premise or local instances is straightforward. PostgreSQL is suited for OLTP and also OLAP workloads [13]. The required workloads here can be depicted in the area of OLAP, given the potential complexity of the designed queries. However, given the use case of an interactive search mask for a web service, fast responses are a requirement. PostgreSQL's efficient query planning and extensibility for additional approaches (e.g., PostGIS [14] for spatial data or Citus [15] for distributed and columnar storage) make it a suitable foundation for GeoMine's use case.

## 2.2. Protein-Ligand Interactions and Binding Pockets

Protein-ligand interactions are of particular interest in biomolecular and pharmaceutical research. Ligands are small molecules that can interact and bind to the generally much larger proteins. Protein complexes can contain multiple pockets of varying sizes, partly containing ligands. Drug molecules used as pharmaceuticals are generally designed to target specific proteins. Researchers can gain valuable insights by investigating specific three-dimensional structures and searching for potential candidates to bind with these proteins.

**Protein Data Bank** The PDB [7, 2], established in 1971, is a comprehensive repository of 3D structural data of proteins and nucleic acids. The structural information is primarily obtained through experimental methods, predominantly X-ray crystallography, from research facilities worldwide [2]. As a freely available resource, the PDB has become vital for research in various fields by providing atomic-scale structural insights for drug design and understanding biological processes, containing more than 200,000 structures as of April 2023. Further, with the advantage of *Computed Structure Models*, which are protein structure predictions, for example, by *AlphaFold2* [4], additional datasets with about 1,000,000 structures are available now [2].

## 2.3. GeoMine

Discovering similar structures across distinct complexes or finding molecules that bind to a specific pocket of interest is a major task in medical research. GeoMine is able to construct comprehensive databases derived from the PDB and supports exploring these databases with a web-based search interface. [3]

The preprocessing and database creation procedures employ components of the NAOMI library [16]. For example, pockets are classified in a complex preprocessing pipeline when constructing the database [3]. Central components are the DoGSite algorithm [17], which identifies empty binding pockets within protein structures, and the calculation of interactions [18].

The central part of the search and unique key feature is the ability to specify geometric properties, for instance, distances and angles between any points, such as atoms. Further, point properties can be specified, such as an atom's chemical element and interactions between

points. This way, precise structural motifs (structural patterns) in protein-ligand complexes can be searched. While GeoMine's predecessor *PELIKAN* was a single-user application based on an integrated *SQLite* [10] database, the *GeoMine* back end is aimed at a server-focused architecture. In the initial development of GeoMine [3], the query execution capabilities of PELIKAN were extended for new functionality but were not changed in structure to adapt to the new architecture.

**Database Design** For our experiments in Section 4, we used a PostgreSQL15 database created with the PDB dataset from October 2022. For querying the dataset, the database can be considered read-only. The database requires approximately 165GB of disk space.

For the geometric search, we focus on two tables. The first table, the *point table*, comprises all atoms and other definable points, such as the center of aromatic rings. It contains 340,716,693 searchable entries. These points are distributed across 1,382,853 distinct pockets, which serve as containers for groups of points. The largest pocket identified in our dataset contains 20,306 points, while the smallest pocket only holds 9 points. Each entry in the point table has a unique identifier, references the containing pocket, and contains various other fields with properties per point. Some properties, such as the accessible surface area of an atom, are floating point numbers. Other attributes, such as the chemical element, contain only a few distinct values, represented as integers or short strings.

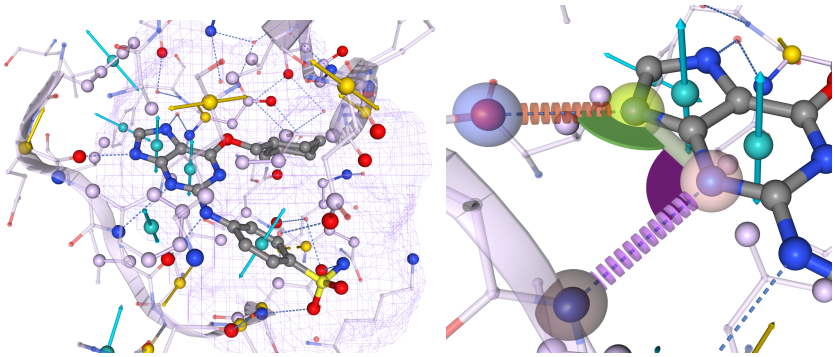
The second table, the *interaction table*, stores pre-calculated interactions [18]. These interactions represent noteworthy connections between two points, for example, hydrogen bonds. 13,018,225 point pairs are stored here.

**Query Creation** When creating a query, users can specify multiple constraints. The most fundamental categories encompass *Textual and Numerical Searches*, wherein metadata filters at the protein structure or pocket level can be defined. Users can directly pre-select several structures or create various filters, such as the minimum number of particular chemical elements or a certain molecular weight range for the ligand. It also enables filtering using patterns that describe a local environment using the chemical substructure language SMARTS strings [19].

The central search element and origin of GeoMine's name are geometry-based searches. To build the query, users may interactively select points in the web front end [21] (see Figure 1), utilizing an arbitrary PDB file as a template structure or define them without a template.

Users may select an arbitrary number of points, which can be filtered based on different properties. Moreover, the specification of distance ranges between two points and angles between specified distances is possible. Further, interactions between points, as stored in the interaction table, can be added to the query. Together they resemble an atomic substructure, which will be searched for. Each pocket can be examined individually as the interactions between one ligand and an individual pocket in a protein are of interest.

**Query Execution** The initial approach for query execution was first described for the predecessor tool PELIKAN by Inhester et al. [6]. The most significant enhancement for the runtime in developing the original GeoMine approach — utilizing a PostgreSQL database instead of SQLite — did not change the workflow of the searching process. The approach remained mostly



(a) View of a pocket (violet mesh) in structure *1H1S*

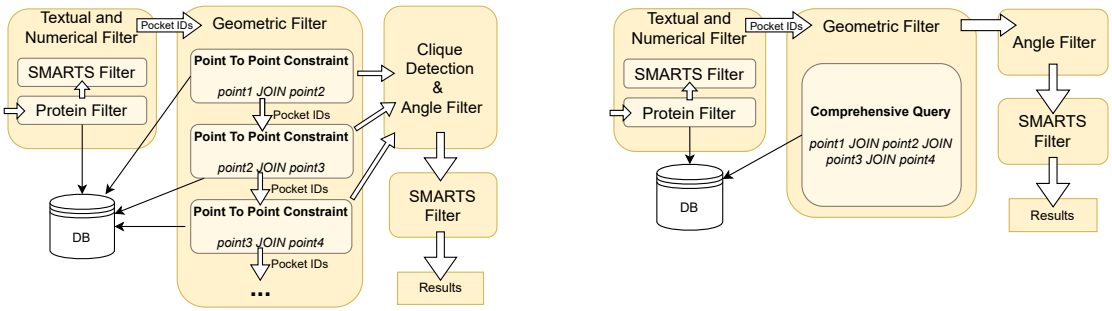
(b) Specifing points, distances, and angles for the query

**Figure 1:** GeoMine's three-dimensional view of a binding pocket based on the *NGL viewer* [20]. Users can interactively select atoms and other points and specify distances and interactions between them to generate the query. Here, a pocket around a ligand (bold bonds) is shown, together with the surrounding atoms of the protein.

algorithmic focused, with all major computational steps performed within the application (see Figure 2a), as the original PELIKAN software was designed to be a standalone desktop application. In the original approach of GeoMine [3], four major steps were performed strictly sequentially for each query to filter the potential results:

1. Textual and Numerical Constraints - A filter eliminates all proteins and pockets that do not meet specified properties or do not correspond to a given restrictive SMARTS filter. This step yields a list of all matching proteins and their pockets.
2. Obtaining all point pairs - For each point pair in the query, all possible results are returned, and distances, as well as interaction constraints, are checked.
3. Clique detection - An algorithm reconstructs the coherent component graph for all obtained point pairs and checks all defined angle constraints.
4. Less restrictive SMARTS filters for points were applied to the now-generated results.

Steps two and three of the query processing presented particular challenges. All point and point pair constraints were queried individually in the database. Since a single constraint for a point pair is often not very specific, it leads to big intermediate results. Only by chaining several constraints the number of points is sufficiently reduced. The need to cross-verify each point with all matching points in its pocket demanded significant computational resources, especially if the filter for the points were unspecific. The list of potential pockets needed to be recreated for each pair, as only pockets which contained results in prior pair subqueries remained in the search space. This caused the search to be strictly sequential and required the serialization and deserialization of long pocket-ID lists for the SQL WHERE clauses. As the application and database system are separate processes or running on separate servers, the required repeated transfer of these lists also affected the performance. Because some point-to-point constraints were specific (less frequent in the dataset) and others were unspecific (frequent in the dataset), a hand-crafted scoring function was utilized to estimate the best ordering of queries, starting




(a) Original workflow

(b) Improved workflow

**Figure 2:** The original and the improved processing workflow of GeoMine (Simplified) for a given search

with the most specific queries to reduce the search space early [6]. Although this improved the join order in many cases, it had the disadvantage of preventing the database system from executing classical optimizations, such as parallelism and join order optimization.

Further, an additional algorithm was required since the results from the preceding steps consisted only of point pairs. The *Bron-Kerbosch algorithm* [22], a graph-based backtracking algorithm for clique detection, was used. This algorithm recursively verified whether all discovered point pairs constituted a complete graph and checked for angle constraints. This demanded substantial computational effort, taking several hours on large potential result sets.

## 3. Optimizations

This research aims to achieve optimal performance and ease of setup across various environments. Alongside the contributions of this work, the application has transitioned to a containerized setup for cloud environments. The optimizations presented in this work are essential for facilitating the deployment of a scalable application. In this section, we will distinguish between the *original approach* in GeoMine [3] and the *improved approach* we present in this work. The yielded results for each query remained identical.

### 3.1. Optimizing SQL Queries

The most significant change from the original approach was the redesign of the SQL query generation. Sequential processing of each constraint within a query led to severely limited query-level parallelism and long processing times as described in Section 2.3. Therefore, all SQL queries are now designed to make use of PostgreSQL's internal planning and optimization. In contrast to the original approach, where each point-to-point constraint was queried separately, a single comprehensive query containing all attributes and constraints for geometrical patterns is now constructed, see Figure 2b. This reduces overhead by eliminating the need to repeatedly serialize extensive lists of pocket IDs or create temporary tables. To achieve this, the point table joins itself as often as points were specified in the query, usually 5-15 times. As a match occurs inside a single pocket, we only need to join points within the same pocket. With information about the distribution of properties like the chemical element, the RDBMS can estimate which

part of the query restricts the search space the most and improve the join order. The original approach required running the checks on all points within all remaining pockets, not being able to skip points that were not matched in earlier subqueries. Intermediate results now remain within the database system and do not require serialization for application transfer. Additionally, merging all constraints (points, distances, and interactions) into one query eliminates the need for clique detection, as the output of the RDBMS is a connected and valid result.

Among all the geometric properties, only the angle checking between point pairs remains a separate step in the application, as this increases the complexity of the query without showing the benefits of an early reduced search space in our tests. Textual and numerical filters remain in a separate query to allow prior filtering, as SMARTS patterns require in-application processing. Allowing the RDBMS to determine the join order and the parallel execution resulted in a significant speedup of benchmark queries. The results are detailed in Section 4.

### 3.2. Enhanced Utilization of PostgreSQL Indexes

In the original approach, a single extensive index structure was created, covering 15 out of 17 table columns. Although PostgreSQL allows for the construction of multi-column indexes with a large number of attributes, these structures are only effective in certain situations due to their size and depending on the used attributes. However, using multiple single-column indexes and allowing PostgreSQL to combine them as recommended in the documentation [23] did not achieve the desired performance improvement.

Only the combination of several attributes could substantially reduce the number of yielded points. The best-found solution for our workload was a balanced compromise between index size and utilization, including only the most frequently used columns in a multi-column index. We identified two separate cases for index usage. Firstly, the earliest scheduled subquery focused solely on the attributes, disregarding their pocket, in cases without textual and numerical filters. Secondly, an index for subsequent subqueries was needed to filter for pocket IDs required for the join. In almost all instances, the optimizer determined to filter for the pocket ID in the second subquery. In some instances, a parallel index scan was performed. Filtering by the pocket ID reduced the search space best in these cases since the most restrictive subquery had already been executed as the first scheduled subquery. Therefore, we introduced a second index with the pocket identifier positioned first in the index. For both structures, we utilized PostgreSQL's default *B-Tree index* as other index structures seemed not beneficial in our tests. As pockets usually contain only a few hundred points, spatial indexes, like r-trees provided by PostGIS [14], did not provide the desired benefits. Filtering points and calculating all distances performed better in our tests than spatial operations due to the overhead of utilizing a spatial column. Index creation only needed a few minutes, but additional indexes for specific queries would no longer fit into the filesystem read cache and reduce performance.

### 3.3. Improving Text Search

The initial step of the workflow involves filtering structures based on textual and numerical attributes. These filters target various properties, the most important being the PDB identifiers used to select a pre-defined or user-defined subset of protein structures. A short alphanumeric

**Table 1**
Experiment overview. Showing enabled improvments between baseline ***ex01*** and all improvments ***ex06***

| Improvement | **ex01** | ex02 | ex03 | ex04 | ex05 | **ex06** | ex07 | ex08 | ex09 | ex10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Index Improvement | | x | | | | x | x | x | x | |
| No Wildcards | | | x | | | x | x | x | | x |
| New Query Design | | | | x | | x | x | | x | x |
| No ILIKE | | | | | x | x | | x | x | x |

code identifies each structure.

Previously, an SQL *ILIKE* (case insensitive match) statement with a wildcard match at the beginning and end of the string was executed to check for the desired properties. For the PDB codes, we could make two changes. We could discard the wildcards in the query unless explicitly desired, which enables the utilization of a search index. And as the codes are not case-sensitive, we can replace the *ILIKE* with a *LIKE*, allowing for a case-sensitive search and resulting in a substantial speedup, as demonstrated in Section 4.

# 4. Evaluation and Discussion

## 4.1. Methods

To evaluate the impact of each modification suggested for GeoMine, several experiments were derived from the original GeoMine approach *ex01* (see Table 1). Experiments *ex02* to *ex05* each contain only one of the improvements, *ex06* contains all improvements, while experiments *ex07* to *ex10* contain all except one. This way, we show which change impacts the performance most, as different improvements benefit from each other.

For evaluating the performance across different workloads, we used a set of nine queries already used in previous work [3], designed to highlight available features, show examples for common applications and estimate the runtime of different patterns common in GeoMine practical applications. They emitted between 2 and 7117 results.

We used a PostgreSQL15 database system. All data was stored on an SSD. Unless otherwise specified, a dedicated server with 400GiB RAM and 80 Cores was used (PostgreSQL 128GB *sharedbuffers*, 16 *parallel workers*). Podman [24] was used to deploy the system. Each experiment was repeated five times. The GeoMine application was executed on the same node as the PostgreSQL database. We configured PostgreSQL to utilize less memory than available, as GeoMine required a high amount of working memory for some workloads. Additionally, we conducted tests on commodity systems by employing two setups (*small*/*medium*) using virtual servers. Both setups stored data on SSDs and were equipped with 12 cores and 24GB RAM, resp. 18 cores and 48GB RAM.

## 4.2. Results

Figure 3 shows the mean runtime of the nine test queries for each experiment as depicted in Table 1. Each change led to better performance, with the highest performance gain occurring

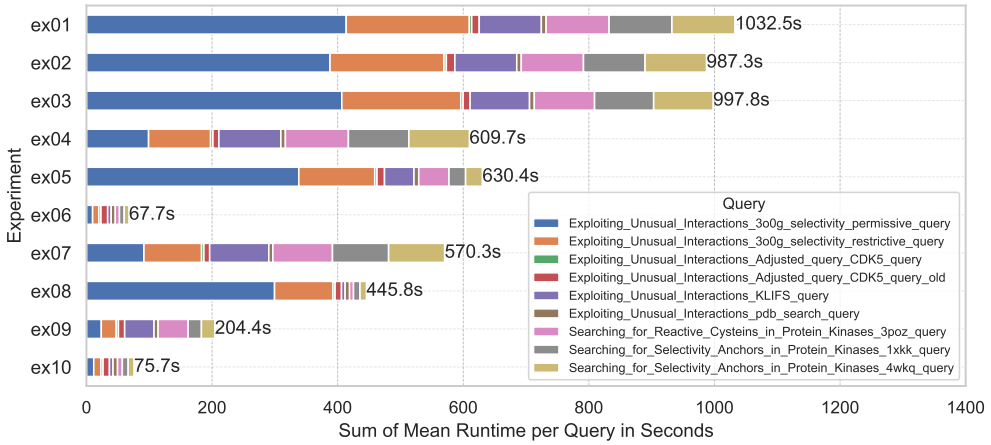


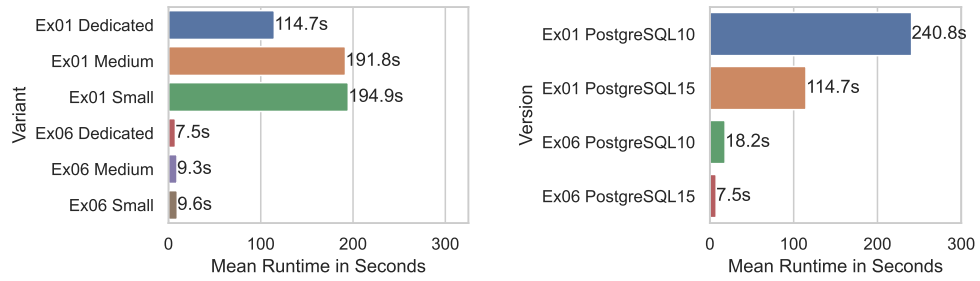

**Figure 3:** The sum of mean runtimes in seconds for each experiment as described in Section 4.1. Each color represents one distinct query

when all changes were applied together. The required time for performing all nine queries decreased from 1033sec of the original approach (*ex01*) to 68sec with all improvements (*ex06*).

The new query design (*ex04*) had the most substantial impact on performance, particularly visible in the long-running queries. Also, the transition from the ILIKE to the LIKE statement notably reduced runtime. The performance gain is most noticeable on the medium-running queries containing a long list of PDB IDs for a preselection. The experiments 02 and 03, the *new index* and *no wildcards* in the PDB ID selection showed only a small improvement. However, experiment 09, which contains all changes except the wildcard improvement, shows that it has an impact on the overall runtime, presumably benefiting from the switch to the LIKE statement. The changes in index structures showed less impact than expected, demonstrating that PostgreSQL can handle indexes with an inflated number of columns. However, the performance was drastically worse if no index was used or index structures did not combine multiple attributes. For instance, combining one index per attribute led to an increase of the sum of the mean runtimes from 68sec (*ex06*) to 134sec.

**Unspecific Queries** Some of the used queries include a protein filter to reduce the number of searched pockets. When removing these filters and searching the whole dataset, the original approach reached its set limits (needing more than 100GB RAM or 1h time) on some of these and other queries with less restrictive geometric filters. With the improved approach, some queries with extensive intermediate results could now be computed for the first time, often within minutes.

**Alternative Setups** As large database instances are not always accessible, for example, due to cost constraints in cloud environments, we also conducted our experiment on two smaller virtual servers. As shown in Figure 4a, the performance gains were also visible on these smaller server instances. These tests were performed on shared hardware, so they can only show a general trend rather than precise comparative data. However, they demonstrate the feasibility




(a) Virtual servers and dedicated hardware (b) PostgreSQL in Version 10 and 15

**Figure 4:** Mean runtimes in alternative configurations of experiment 01 and 06

of processing on shared virtual servers. Additionally, we observed a substantial speedup while transitioning from PostgreSQL10 to PostgreSQL15 as displayed in Figure 4b. Combined with our improvements, we achieved a speedup factor of 32.

## 5. Conclusion and Future Work

GeoMine is a unique application for geometric searches in large collections of protein-ligand complexes with high relevance for life-science research. We showed that it was possible to achieve a large speedup on our query processing by moving major parts of the processing from a custom-written logic inside the software to a PostgreSQL database system. Additionally, different approaches in database optimization contributed to further performance gain. Overall, these achievements are critical for the practical use of the system handling the growing dataset. Some queries could be executed for the first time on our setup due to these changes. In this work, we focused on optimizations of the database and query design. We demonstrated the substantial benefits of database optimizations in scientific applications, achieving a fifteen-fold speedup in GeoMine. Coupled with a halving of the runtime through the use of a newer PostgreSQL version, we managed to reduce the average runtime from minutes to seconds.

Looking ahead, we plan to explore additional database paradigms, such as distributed or column-based systems, and establish schema changes for further optimizations. The caching of intermediate results, as well as determining the join order by extended statistics or by utilizing machine learning, may potentially provide additional benefits. This way, we aim to achieve even better performance for searching scientific data with a service-oriented web service.

## Acknowledgments

This work was supported by the German Federal Ministry of Education and Research as part of CompLS and de.NBI (031L0172 and 031L0105).

## References


[1] T. Inhester, M. Rarey, Protein–ligand interaction databases: advanced tools to mine activity data and interactions on a structural level, WIREs Computational Molecular Science 4 (2014) 562–575. doi:10.1002/wcms.1192.

[2] S. K. Burley, C. Bhikadiya, C. Bi, S. Bittrich, H. Chao, L. Chen, e. a. Craig, RCSB Protein Data Bank (RCSB.org): delivery of experimentally-determined PDB structures alongside one million computed structure models of proteins from artificial intelligence/machine learning, Nucleic Acids Research 51 (2022).

[3] J. Graef, C. Ehrt, K. Diedrich, M. Poppinga, N. Ritter, M. Rarey, Searching Geometric Patterns in Protein Binding Sites and Their Application to Data Mining in Protein Kinase Structures, Journal of Medicinal Chemistry 65 (2022) 1384–1395. doi:10.1021/acs.jmedchem.1c01046.

[4] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, et al., Highly accurate protein structure prediction with alphafold, Nature 596 (2021) 583–589. doi:10.1038/s41586-021-03819-2.

[5] K. Schöning-Stierand, K. Diedrich, R. Fährrolfes, F. Flachsenberg, A. Meyder, E. Nittinger, R. Steinegger, M. Rarey, Proteins plus: interactive analysis of protein–ligand binding interfaces, Nucleic acids research 48 (2020) W48–W53. doi:10.1093/nar/gkaa235.

[6] T. Inhester, Mining of Interaction Geometries in Collections of Protein Structures, Ph.D. thesis, Universität Hamburg, 2017.

[7] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne, The Protein Data Bank, Nucleic Acids Research 28 (2000) 235–242. doi:10.1093/nar/28.1.235.

[8] C. Tenopir, N. M. Rice, S. Allard, L. Baird, J. Borycz, L. Christian, B. Grant, R. Olendorf, R. J. Sandusky, Data sharing, management, use, and reuse: Practices and perceptions of scientists worldwide, PloS one 15 (2020) e0229003.

[9] M. Raasveldt, H. Mühleisen, Data management for data science-towards embedded analytics., in: CIDR, 2020.

[10] R. D. Hipp, SQLite, 2020. URL: https://www.sqlite.org/.

[11] The PostgreSQL Global Development Group, Postgresql: The world's most advanced open source relational database, 2023. URL: https://www.postgresql.org.

[12] solid IT gmbh, Db-engines ranking, 2023. URL: https://db-engines.com/en/ranking.

[13] A. Conrad, Database of the year: Postgres, IEEE Software 38 (2021) 130–132. doi:10.1109/MS.2021.3089730.

[14] The PostGIS Development Group, Postgis, 2023. URL: https://postgis.net.

[15] U. Cubukcu, O. Erdogan, S. Pathak, S. Sannakkayala, M. Slot, Citus: Distributed postgresql for data-intensive applications, in: Proceedings of the 2021 International Conference on Management of Data, SIGMOD '21, 2021, p. 2490–2502. doi:10.1145/3448016.3457551.

[16] S. Urbaczek, A. Kolodzik, J. R. Fischer, T. Lippert, S. Heuser, I. Groth, T. Schulz-Gasch, M. Rarey, Naomi: On the almost trivial task of reading molecules from different file formats, Journal of Chemical Information and Modeling 51 (2011) 3199–3207. doi:10.1021/ci200324e.

[17] J. Graef, C. Ehrt, M. Rarey, Binding site detection remastered: Enabling fast, robust,

and reliable binding site detection and descriptor calculation with dogsite3, Journal of Chemical Information and Modeling 63 (2023) 3128–3137. doi:10.1021/acs.jcim.3c00336, pMID: 37130052.
[18] T. Inhester, S. Bietz, M. Hilbig, R. Schmidt, M. Rarey, Index-based searching of interaction patterns in large collections of protein–ligand interfaces, Journal of Chemical Information and Modeling 57 (2017) 148–158.
[19] I. Daylight Chemical Information Systems, Smarts-a language for describing molecular patterns, 2007.
[20] A. S. Rose, A. R. Bradley, Y. Valasatava, J. M. Duarte, A. Prlić, P. W. Rose, NGL viewer: web-based molecular graphics for large complexes, Bioinformatics 34 (2018) 3755–3758. doi:10.1093/bioinformatics/bty419.
[21] K. Diedrich, J. Graef, K. Schöning-Stierand, M. Rarey, GeoMine: interactive pattern mining of protein–ligand interfaces in the Protein Data Bank, Bioinformatics 37 (2020) 424–425. doi:10.1093/bioinformatics/btaa693.
[22] C. Bron, J. Kerbosch, Algorithm 457: finding all cliques of an undirected graph, Communications of the ACM 16 (1973) 575–577.
[23] PostgreSQL 15, Documentation, 2023. URL: https://www.postgresql.org/docs/15/.
[24] Containers, podman, 2023. URL: https://podman.io/.

## C.5 User-centric design of a 3D search interface for protein-ligand complexes

[D5] K. Diedrich, C. Ehrt, J. Graef, M. Poppinga, N. Ritter, and M. Rarey. Journal of Computer-Aided Molecular Design 38 (2024). Available: https://doi.org/10.1007/s10822-024-00563-3. 

Journal of Computer-Aided Molecular Design (2024) 38:23
https://doi.org/10.1007/s10822-024-00563-3

# User-centric design of a 3D search interface for protein-ligand complexes

Konrad Diedrich[1] · Christiane Ehrt[1] · Joel Graef[1] · Martin Poppinga[2] · Norbert Ritter[2] · Matthias Rarey[1]

Received: 3 April 2024 / Accepted: 17 May 2024


**Abstract**
In this work, we present the frontend of GeoMine and showcase its application, focusing on the new features of its latest version. GeoMine is a search engine for ligand-bound and predicted empty binding sites in the Protein Data Bank. In addition to its basic text-based search functionalities, GeoMine offers a geometric query type for searching binding sites with a specific relative spatial arrangement of chemical features such as heavy atoms and intermolecular interactions. In contrast to a text search that requires simple and easy-to-formulate user input, a 3D input is more complex, and its specification can be challenging for users. GeoMine's new version aims to address this issue from the graphical user interface perspective by introducing an additional visualization concept and a new query template type. In its latest version, GeoMine extends its query-building capabilities primarily through input formulation in 2D. The 2D editor is fully synchronized with GeoMine's 3D editor and provides the same functionality. It enables template-free query generation and template-based query selection directly in 2D pose diagrams. In addition, the query generation with the 3D editor now supports predicted empty binding sites for AlphaFold structures as query templates. GeoMine is freely accessible on the Proteins*Plus* web server (https://proteins.plus).

**Keywords** Protein-ligand complexes · Molecular interactions · 2D query editor · GeoMine · 3D search engine · PoseEdit · Protein Data Bank · AlphaFold

## Introduction

A large number of experimentally determined three-dimensional (3D) structures of biological macromolecules are publicly available thanks to the substantial growth of the Protein Data Bank (PDB) [1] and are easily accessible through its web service. This wealth of data is a fundamental scientific resource for understanding macromolecule-ligand interactions and their functional impact. However, to fully exploit this data resource, search engines have to go beyond basic querying on a textual level and enable direct searching of the most central part of the data, the 3D structures themselves. The capability to retrieve all structures with a similar relative spatial arrangement of chemical features like atoms, functional groups, or intermolecular interactions from the PDB can support numerous applications in life science research. For example, searching a query that covers a ligand binding mode within a binding site may result in potential off-target binding sites with similarly interacting ligands, thereby explaining side effects, mining for interaction geometries [2], searching for residue motifs [3], and assisting drug repurposing [4].

In addition to the web service of the PDB itself [5], several tools have been developed that enable specific types of spatial queries for the PDB: CSD-CrossMiner [6], PRDB [7], PROLIX [8], Relibase and Relibase+ [9], PDBeMotif [10], PELIKAN [11], GSP4PDB [12], GeoMine [13, 14], and nAPOLI [15]. In addition, some commercial and unpublished software applications such as Proasis4 [16] and 3decision [17] offer similar search capabilities. Of the published tools, PRDB, PROLIX, Relibase, Relibase+, PDBeMotif, and nAPOLI are no longer available. CSD-CrossMiner and PELIKAN are desktop applications, while GSP4PDB and GeoMine are accessible on the web. GSP4PDB and GeoMine

✉ Matthias Rarey
matthias.rarey@uni-hamburg.de

[1] Universität Hamburg, ZBH - Center for Bioinformatics, Albert-Einstein-Ring 8-10, 22761 Hamburg, Germany

[2] Universität Hamburg, Department of Informatics, Vogt-Kölln-Straße 30, 22527 Hamburg, Germany

Published online: 30 May 2024

are freely available. CrossMiner and PELIKAN require a commercial or academic license, respectively. The tools differ significantly regarding the supported query content and what regions of the structures in the PDB are searchable. For example, while the PDB web service allows searching the relative spatial arrangement of α- and β-carbon atoms of specific residues in complete protein structures, PELIKAN permits a query that describes a relative spatial arrangement of arbitrary user-specified heavy atoms and intermolecular interactions to screen ligand-bound binding sites defined by a radius of 6.5 Å of the ligand's heavy atoms. For a comprehensive overview of the query differences and technical aspects of the different tools, like the underlying data storage approach, see [13].

Due to the multidimensional nature of the data and the varying complexity of the supported 3D query, spatial searches are highly challenging, not only from the developer's point of view but also from the user's perspective. In contrast to the simple text-based user input of keywords, scalar values, sequences, or even substructures, specifying relative spatial arrangements of chemical features is a complex task. In general, drawing with a graphical editor substantially simplifies query generation, in contrast to defining the query purely textually.

A two-dimensional (2D) or 3D editor provides a more intuitive interface for placing and specifying chemical features and their geometric constraints. Additionally, both editor types already give life scientists a familiar environment for visualizing chemical structures. Generating queries with such an editor can be further simplified by visualizing a structure of interest as a template, in which the user can select the arrangement of chemical features to search for. Nevertheless, an additional textual specification of the 3D query as a manual post-processing step is useful for adapting its chemical and spatial precision to individual needs.

The PDB web service, CSD-CrossMiner, GeoMine, and PELIKAN provide a 3D editor. A template structure for query design can be used in all tools. While a query can be designed anywhere in a loaded 3D representation of a PDB entry using the PDB web service, the query options in CSD-CrossMiner, GeoMine, and PELIKAN are limited to corresponding 3D-visualized binding sites as structural templates. Query generation from scratch is possible with CSD-CrossMiner and GeoMine via the 3D editor and PELIKAN via a textual and tabular representation. PROLIX enables a purely textual approach. All other tools offer a 2D editor for template-free query generation.

Considering a 2D and 3D editor in comparison, both visualization concepts have advantages and disadvantages for generating spatial queries. A 3D environment is a natural choice because it provides precise spatial information. However, the drawback of a 3D editor is that its usage requires practice and time, especially when using a template for query selection. Due to the high amount of visualized structural information, the chemical features of interest might be visually buried inside the structure and must, therefore, be focused on by users by zooming, translating, and rotating the scene extensively. Therefore, query generation can still be challenging, even though a 3D visualization provides all required information.

In contrast, a 2D environment provides only distorted spatial information due to the dimensionality reduction. Furthermore, converting a 3D template structure into a planar representation prevents the visualization of the entire structure due to consequential structural overlaps causing suboptimal 2D layout quality. Therefore, 2D visualization requires a reduction in the amount of visualized structural information. Even though a 2D visualization provides less information than a 3D visualization, it visualizes and highlights only the most relevant chemical information a user might want to search for. Furthermore, 2D visualization offers chemical structure representation as structure diagrams that are very familiar to scientists. A 2D visualization permits an instant overview of the most relevant selectable chemical features, simplifying query generation.

In this article, we will introduce the latest version of GeoMine. First, we will provide a user-focused overview of GeoMine, including its new features: the 2D query editor and the 3D template type based on the artificial intelligence-predicted AlphaFold structures [18] that are retrieved from the corresponding database at https://alphafold.ebi.ac.uk. We will then present the 2D editor in more detail and showcase the application of the new features of the latest tool version, which exploits all the above-mentioned query generation approaches to design a graphical user interface with the highest usability possible for spatial searching within known and predicted binding sites.

## Methods

### Features overview

The key features of the most recent release of GeoMine are summarized in the list below. Subsequently, some of these points are illustrated in detail, including the integration of the new features of GeoMine, the 2D query editor, and the AlphaFold-based 3D template type:

- A graphical user interface that is freely accessible via the Proteins*Plus* [19–21] web server (https://proteins.plus).
- A fast and precise search functionality that enables 3D querying of ligand-bound and predicted empty binding sites of protein or nucleic acid structures in the entire

PDB. In the new GeoMine version, binding sites are predicted by DoGSite3 [22]. The binding sites are post-processed by Protoss [23, 24] to calculate the presence and coordinates of polar hydrogen atoms.

- On-the-fly loading of ligand-bound and predicted empty binding sites as query templates created from a PDB structure, an AlphaFold structure, or an uploaded custom structure file in PDB format.
- An interactive and user-friendly query generation process in a 2D and 3D editor that allows synchronized query selection in a ligand-bound or predicted empty 3D template binding site and 2D ligand interaction diagram, respectively, as well as its generation from scratch.
- A large number of selectable chemical features that include all buried and solvent-exposed heavy atoms of all ligands (e.g., solvent molecules, cofactors, small molecules), simple ions like metal ions, and biomolecular residues (amino acid and nucleic acid residues) in a binding site, as well as visualized aromatic ring centers, secondary structure elements, and hydrogen bond, pi-stacking, cation-pi, metal, and ionic interactions. GeoMine allows combining all of these chemical features into a single complex 3D query.
- The placement of hypothetical chemical features in 2D and 3D space for template-free query generation.
- A comprehensive specification of the spatial relationships between chemical features through geometric constraints, which include orientations, distance ranges, and angle ranges.
- A simple verification and arbitrarily precise specification of the query due to its additional representation in tables, which show various properties of the chemical features and geometric constraints that can be adjusted in detail or kept more generic. For example, users can specify whether a selected atom of the polar residue serine matches only serine residues or all polar residues or residues of any type and class.
- The automatic loading of the primary or even all properties of a chemical feature into the query table by chemical feature selection.
- A clear visual correspondence of the query visualized in the 2D and 3D editors and the tables achieved by synchronized mouse-over highlighting and individual coloring for the chemical features and geometric constraints of the query.
- A user-specified ulterior restriction of the search in the PDB by an optional list of PDB identifiers and by the inclusion or exclusion of results based on 53 additional textual and numerical filter criteria, such as the source organism, the protein class, or the root-mean-square deviation (RMSD) between the match and query points.
- An iterative search process of query editing and subsequent searching in already-detected results enabled by a refinement functionality and results history.
- The download and upload of a GeoMine query in JavaScript Object Notation (JSON) file format for sharing, archiving, reusing, and later editing.
- A comprehensive presentation and comparative analysis of the resulting binding sites by a table with information about the 150 best results and the visualization of these in the 3D editor together with the 3D template binding site as superimpositions of the template and matching binding sites with various 3D visualization options. The ranking and superimpositions of results are based on the RMSD between the chemical point features of the query (atoms, aromatic ring centers, secondary structure elements) and the corresponding ones of the matches.
- The download of the table content in JSON or comma-separated values (CSV) format, of the superposed binding sites of the 150 best results in PDB format, and of a file that contains the statistics for all matches.

## New features integration

The user can specify a template structure for query generation on the Proteins*Plus* landing page (Fig. 1) in several ways. Besides the specification of a Protein Data Bank structure by its 4-letter identifier (Fig. 1a) or a custom structure by a file in PDB format (Fig. 1b), the user can now directly access predicted structures in the AlphaFold database through their UniProt accession numbers. Additional ligands can be uploaded in Structural Data File (SDF) format (Fig. 1c) for the specified template structure. The linked advanced search functionality (Fig. 1d) allows the user to query the Protein Data Bank (Fig. 2a) and AlphaFold database (Fig. 2b) by keywords to search for potential input structures (Fig. 2c).

After input confirmation on the Proteins*Plus* landing page (Fig. 1e), the user is forwarded to the Proteins*Plus* main page (Fig. 3), which consists of three scrollable sections. The user can select GeoMine from the tool list in the right section to access the tool-specific graphical user interface components, including the new 2D query interface. The central section provides two scrollable lists: the *Pockets* and *Ligands* lists (Fig. 3b). The *Ligands* list contains information about all ions and small molecules of the input structure. The *Pockets* list provides information about on-the-fly calculated ligand-bound and DoGSite3-predicted empty binding sites. Ligand-bound binding sites are predicted with a so-called “ligand-bias” option, i.e., the solvent grids are biased by the buried fragments of the ligand to enforce these parts to be included in the predicted sites (ligand-biased

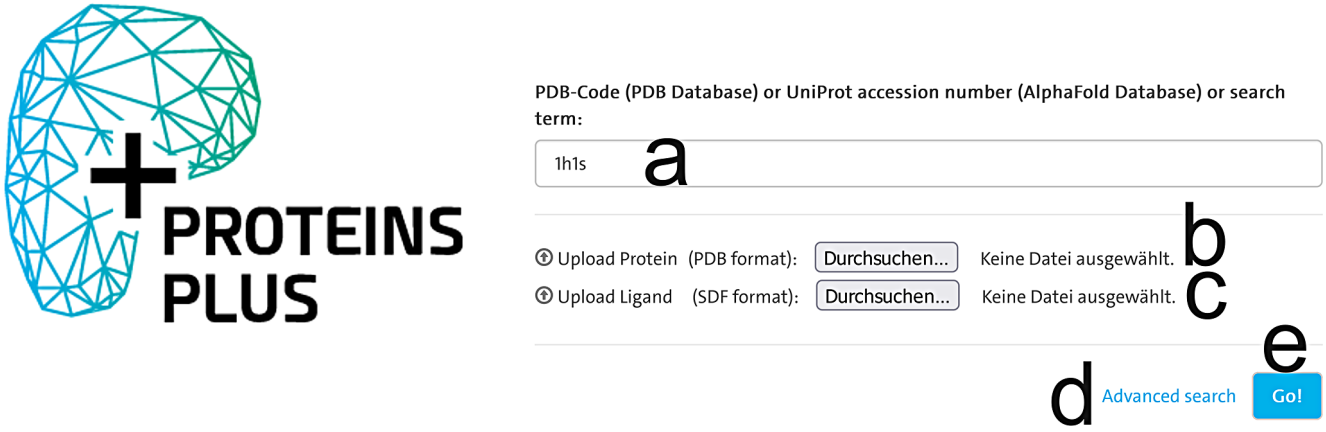



**Fig. 1** Excerpt of the landing page of Proteins*Plus*. **a** Text field for the specification of a Protein Data Bank or AlphaFold structure as input. **b** Upload button for a PDB file with a custom input structure. **c** Upload button for an SDF file with additional ligands. **d** Link to the advanced search functionality. **e** Button for the confirmation of the input structure

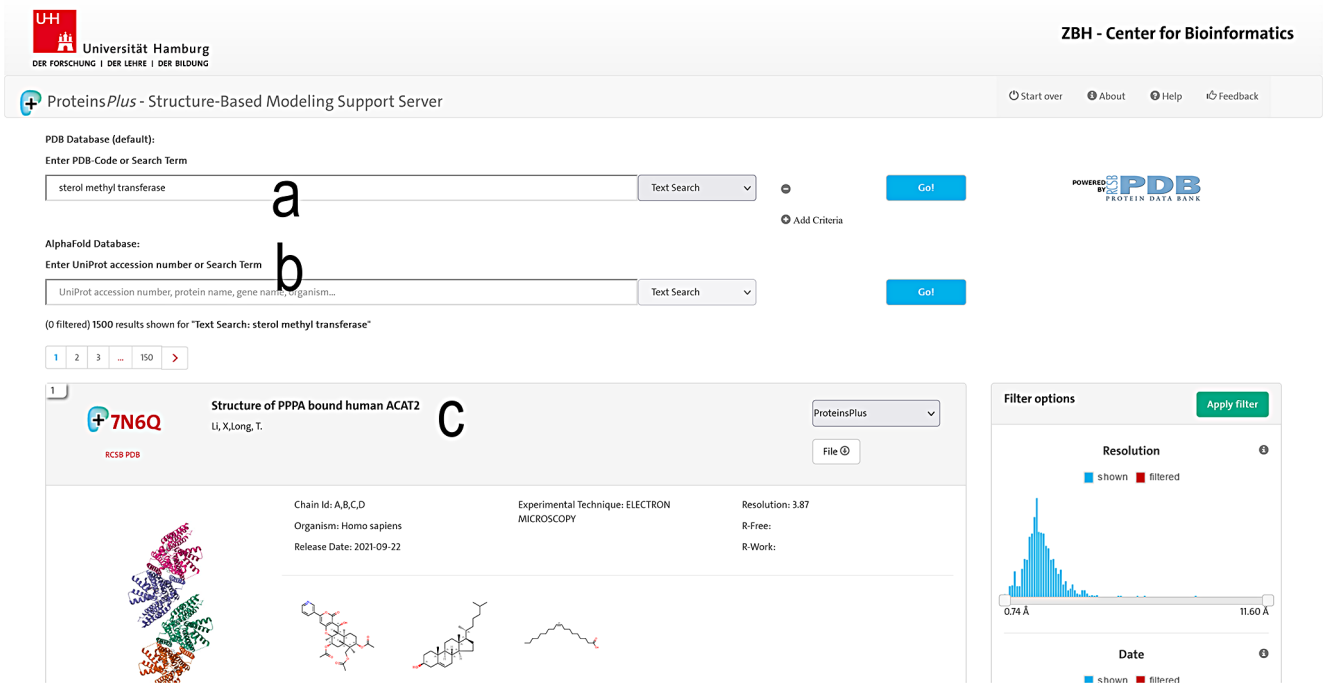



**Fig. 2** Advanced search functionality of Proteins*Plus*. **a** Text field for the keyword-based querying of the Protein Data Bank. **b** Text field for the keyword-based querying of the AlphaFold database. **c** List of search results

predicted sites, see [22] for details). A ligand might not be contained in any DoGSite3-predicted binding site, i.e., less than 20% of its heavy atoms lie in the pocket. In this case, a ligand radius-based binding site is created instead, including the ligand and all residues, other small molecules, and simple ions within a radius of 6.5 Å of the ligand's heavy atoms. In the case of AlphaFold-based input, only predicted empty binding sites are available, as those structures do not contain ligands. A 2D ligand interaction diagram created with PoseEdit [25] and PoseView [26–28] as a template for query selection can be loaded for a user-specified ligand from the *Ligands* list into the 2D editor on the right (Fig. 3c). The corresponding ligand-bound 3D binding site from the *Pockets* list is then automatically visualized in the 3D editor on the left (Fig. 3a).

Figure 4 provides a detailed view of the 2D ligand interaction diagram content, the 2D editor functionality, and the supported components of the query. A 2D ligand interaction diagram (Fig. 4b) shows an excerpt of the corresponding ligand-bound 3D binding site. The selectable chemical feature types are the same as in the 3D binding sites, but the visualized content is restricted to a

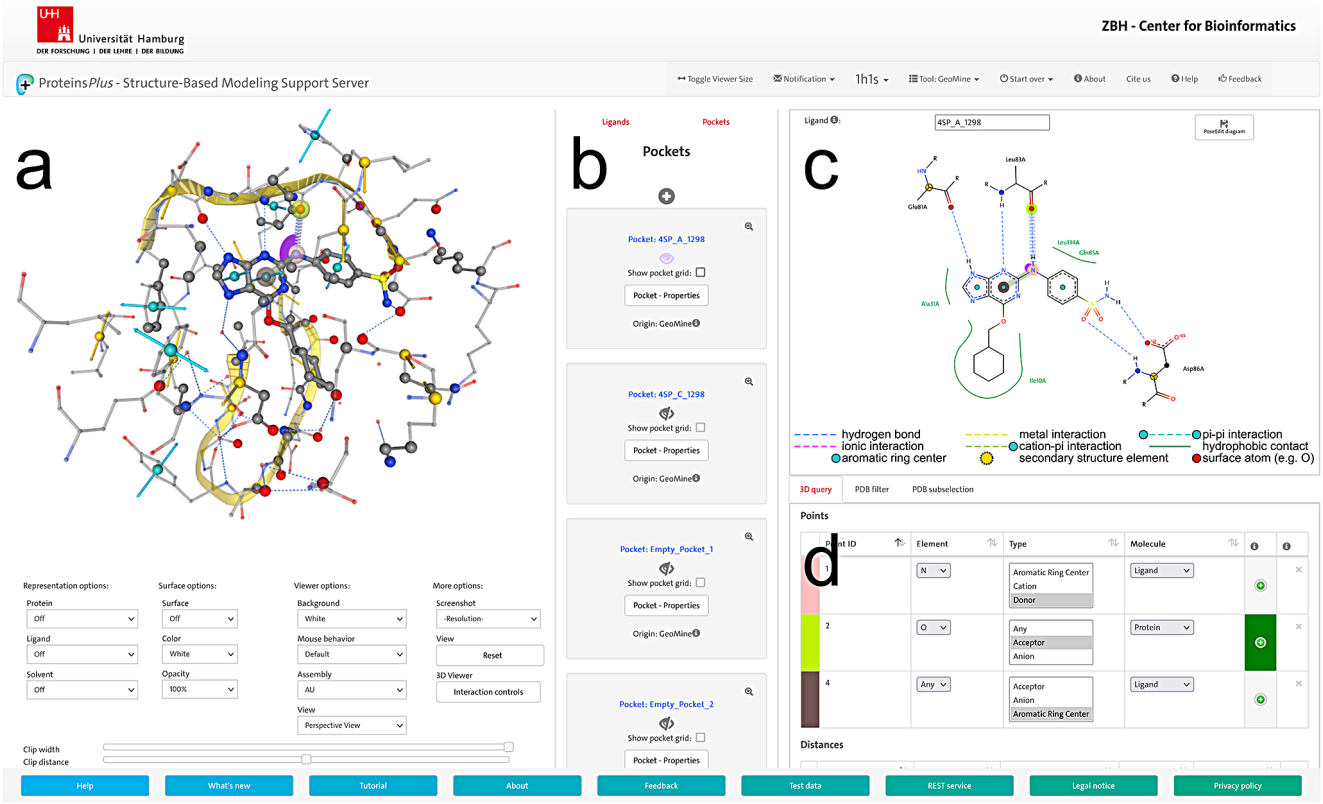



**Fig. 3** Main page of Proteins*Plus* showing the main components of the graphical user interface of GeoMine and a query. **a** 3D viewer showing a ligand-bound 3D binding site and the query. **b** List of ligand-bound and predicted binding sites of the input structure for visualization in the 3D viewer. A toggleable list of small molecules and ions of the input structure for visualizing the corresponding 2D interaction diagrams in the 2D viewer can be shown by clicking on "Ligands". **c** 2D viewer showing a 2D ligand interaction diagram and the query. **d** Scrollable section for the query representation by tables, with separate tables for query points, distances, angles, and interactions. In the figure, the focus is on the points table

specific ligand and directly interacting metals and macromolecular residues. Hydrophobic contacts with residues are not visualized in atomic detail but are indicated by green splines labeled by the corresponding residue identifiers. It is not possible to generate 2D diagrams for predicted empty binding sites. The substantial quantity of solvent-exposed residues in such a binding site cannot be effectively limited, as it is difficult to automatically specify which residues might be more important than others for query selection. A 2D diagram that displays all binding site residues is overly crowded and does not provide any chemical reference point to the user on what to select, rendering the query formulation in 2D space an ineffective alternative. In contrast, a query selection in a predicted empty 3D binding site is more feasible since residues are distinguishable on a spatial level. For example, a user might want to select specific solvent-exposed atoms of nearby residues surrounding a distinct subsection of the binding site. However, for ligand-bound 3D binding sites, it is possible to highlight the ligand, its interaction partners, and the intermolecular interactions in a 2D diagram. This focus increases the clarity of 2D diagrams while providing chemical information useful for query selection even without spatial information.

In addition to the input specification via the *Ligands* list, users can upload a diagram file in JSON format (Fig. 4d). This upload functionality is particularly useful when users want to improve the automatically generated 2D layout for query selection. With the 2D diagram editing tool PoseEdit, which is also accessible on Proteins*Plus*, the user can load and visualize the same 2D diagram to manually rearrange its content for resolving graphical issues like overlapping residues or intersecting intermolecular interactions. The optimized 2D diagram can be downloaded from PoseEdit as a JSON file and can then be uploaded into the 2D editor of GeoMine.

The 2D editor has the same query-building functionality as the 3D editor. Furthermore, the 2D editor is synchronized with the 3D editor and the query tables regarding query generation, visualization, mouse-over highlighting, and coloring. This synchronization allows the simultaneous usage of all query input types in a complementary manner. The query

consists of chemical features and geometric constraints that can be added without a template or selected in a template via several user modes (Fig. 4a). A legend below the 2D drawing area (Fig. 4c) explains the precalculated chemical features.

In the *Point* mode, the user can select so-called points, i.e., heavy atoms, aromatic ring centers, and secondary structure elements, represented by α-carbon atoms of central or terminal protein residues in helices and strands. Solvent-exposed heavy atoms are highlighted by big colored spheres. Like in the 3D editor, hypothetical points can be placed and moved in 2D space. They are automatically placed in the center of the ligand-bound 3D binding site that corresponds to the 2D ligand interaction diagram. The relative position of a hypothetical point can be adjusted via the 3D editor and by distance ranges. Intermolecular interactions are visualized by colored dashed lines and can be selected in the *Interaction* mode. It is also possible to specify a hypothetical intermolecular interaction between two points in that mode. Any two points can be connected by a distance range in the *Distance* mode. Lastly, angle ranges can be placed between connected distance pairs and interactions in the *Angle* mode.

The corresponding tables in the scrollable section below the 2D editor list defined points, distances, interactions, and angles (Fig. 3d). The tables allow further verification and modification of their properties, for example, the residue an atom belongs to or the tolerance value of a distance range. The user can specify that all properties of a selected chemical feature are automatically recognized and set in its corresponding table entry after selection by enabling the checkbox next to the list of modes. Otherwise, only its main properties are set automatically, i.e., the element for atoms and the molecule type for atoms, aromatic ring centers, and secondary structure elements. For a screen recording video

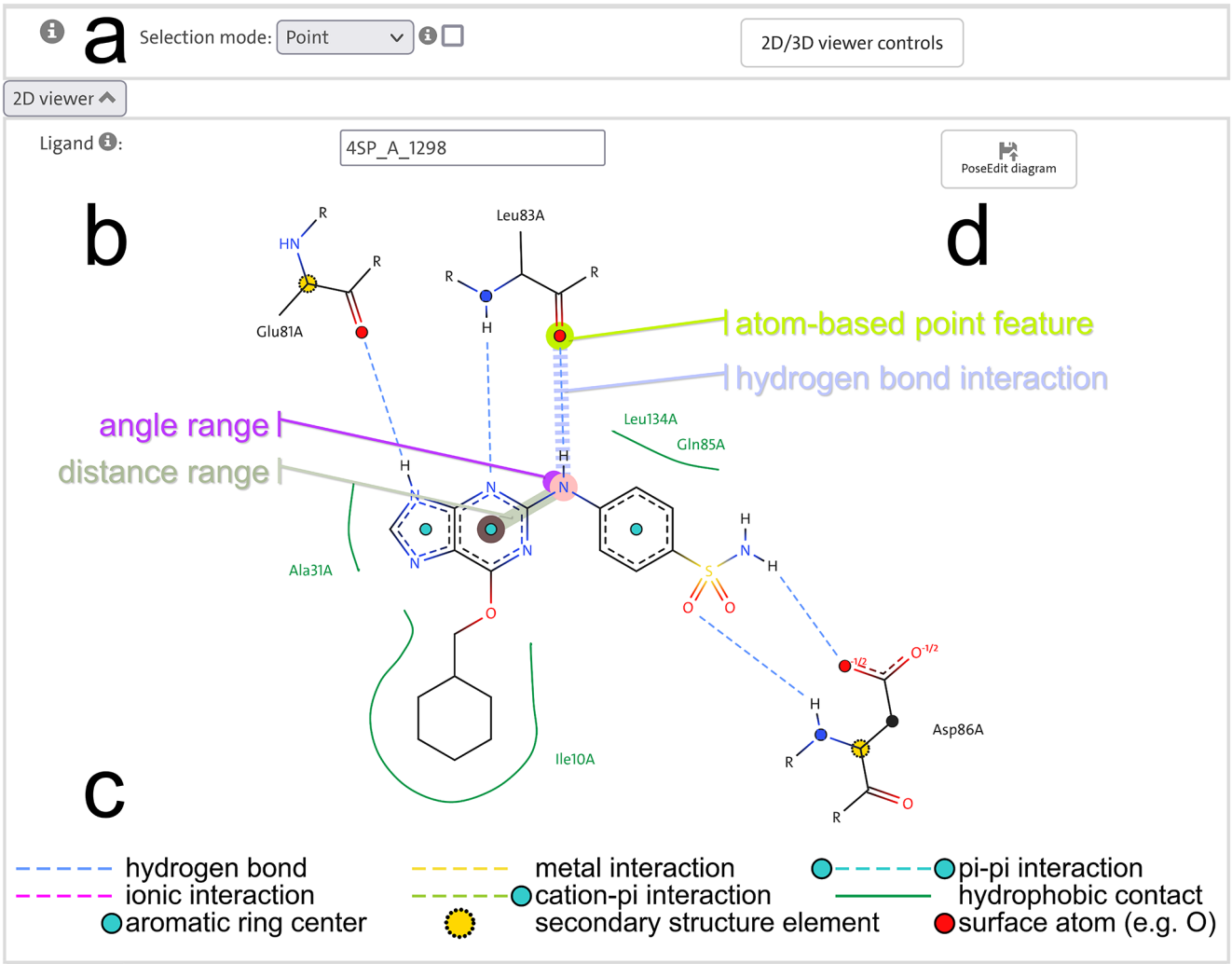



**Fig. 4** Excerpt of the Proteins*Plus* main page, showing the 2D editor of GeoMine with a 2D ligand interaction diagram and all possible query components. **a** List of modes for query generation in 2D and 3D. **b** Drawing area displaying a 2D diagram of the inhibitor with the internal Proteins*Plus* ID 4SP_A_1298 interacting with a cyclin-dependent kinase (PDB code: 1H1S) [29]. **c** Legend illustrating chemical features. **d** Button for uploading diagram files

demonstrating how to apply the user modes, see Online Resource 1. A 2D diagram of the inhibitor with the internal Proteins*Plus* ID 4SP_A_1298 interacting with a cyclin-dependent kinase (PDB code: 1H1S) is shown in the video to exemplify query generation with the 2D editor.

### Technical implementation details

The graphical user interface is primarily implemented with HTML, Vanilla JavaScript, and the Bootstrap 3 library (https://getbootstrap.com). Several JavaScript libraries were used to integrate specific frontend components. The 3D viewer uses the NGL library [30, 31] (https://nglviewer.org). The query tables employ the DataTables library (https://datatables.net). The 2D editor is based on the InteractionDrawer JavaScript library (https://github.com/rarey-lab/InteractionDrawer), which draws interactive 2D ligand interaction diagrams in Scalable Vector Graphics (SVG) format. The web server's backend is implemented using the Ruby on Rails framework (https://rubyonrails.org) and a MySQL database (https://www.mysql.com).

GeoMine's searches are performed on a server using a PostgreSQL (https://www.postgresql.org) database, 200 GB of main memory, up to 30 cores of a 2x Intel Xeon Gold 6248 processor (2.5 GHz), and a Dell 1.6 TB NVMe HHHL AIC PM1725b solid-state drive with an XFS file system.

## Application

### Binding site function prediction and off-target analyses for methyltransferases in *Leishmania*

In our case study, we want to illustrate how GeoMine can be used to analyze AlphaFold models and assist in suggesting ligands and their binding modes for a predicted protein structure of interest. The resulting complexes can subsequently be used to assess the uniqueness of the 3D arrangement of ligand-interacting binding site atoms using 2D query design. Here, we want to focus on neglected tropical diseases threatening millions worldwide [32]. Their treatment is restricted to a few medications that often harbor severe side effects [33]. Causative agents for these diseases are, among others, parasites of the genus *Leishmania*. The search for potential therapeutic agents became the focus of academic infection research, which identified several pharmaceutically promising targets [34]. Understanding their structure and function is crucial for future early-phase drug discovery and development.

The protein of interest in this case study is an enzyme called sterol 24-C methyltransferase (SMT) in *Leishmania* species. The enzyme uses S-adenosyl methionine (SAM) as a cosubstrate and catalyzes the C-C bond formation between a methyl group and the C24 of zymosterol to form ergosterol [35], the major sterol component of these parasites. Several substrate-based inhibitors of the enzyme from *Leishmania amazonensis* are known [36] and a recent computational study aimed to design novel inhibitors [37]. The authors focused their analyses on the zymosterol binding site of the protein to find novel inhibitors. In contrast, we wondered whether the SAM-binding site might provide a suitable starting point for structure-based design. Due to the unavailability of experimental structures, we used the AlphaFold model of the enzyme from *L. donovani* (UniProt Accession Q6RW42).

Upon loading the structure on Proteins*Plus* by entering its UniProt Accession, we can see the ligand-free structure of the protein. In the Pockets tab, we see two pockets predicted by DoGSite3 for the structure. The first is very large, with a volume of 587 Å³ (P1), while the second is much smaller and mainly occupied by charged residues (P2). We conclude that the first pocket might be the active site responsible for SAM and zymosterol binding. DoGSite3 detects three subpockets in this binding site: a large one with many aromatic atoms and a hydrophobicity ratio of 0.76, which is flanked by residues with low pLDDT scores (P1_1), and two smaller ones with lower predicted hydrophobicity and high pLDDT scores (P1_2 and P1_3, Fig. 5). Therefore, we hypothesized that the smaller subpockets might be the site binding to SAM and rely on these subpockets with an overall higher predicted accuracy in terms of pLDDT.

We performed a molecular docking of SAM with JAMDA [38, 39] into these combined subpockets. However, we obtained highly diverse potential poses partially extending to the P1_1 subpocket. Due to structural uncertainties of the structural model representing a considerable challenge for molecular docking [40], the best-scored pose might not correspond to the native binding mode. To find the most probable of the predicted binding poses, we built a GeoMine model based on flanking solvent-exposed binding site residues (Fig. 5) and a point indicating the position of the ligand and screened for similar binding sites in complex with SAM, its enzymatic product S-adenosyl homocysteine (SAH), or their analog sinefungin (SFG). The corresponding query file in JSON format is available in the Supplementary Information (Online Resource 2). The search finished in 21 s. Intriguingly, we found only one protein ligand-complex with SAH that did not clash considerably with the query protein residues: the SAM-binding pocket of ribosomal RNA large subunit methyltransferase K/L from *Escherichia coli* (strain K12, PDB code 3v97). The JAMDA pose on rank 6 is similar to the one in the RNA methyltransferase aligned with GeoMine and might provide a reliable binding hypothesis.

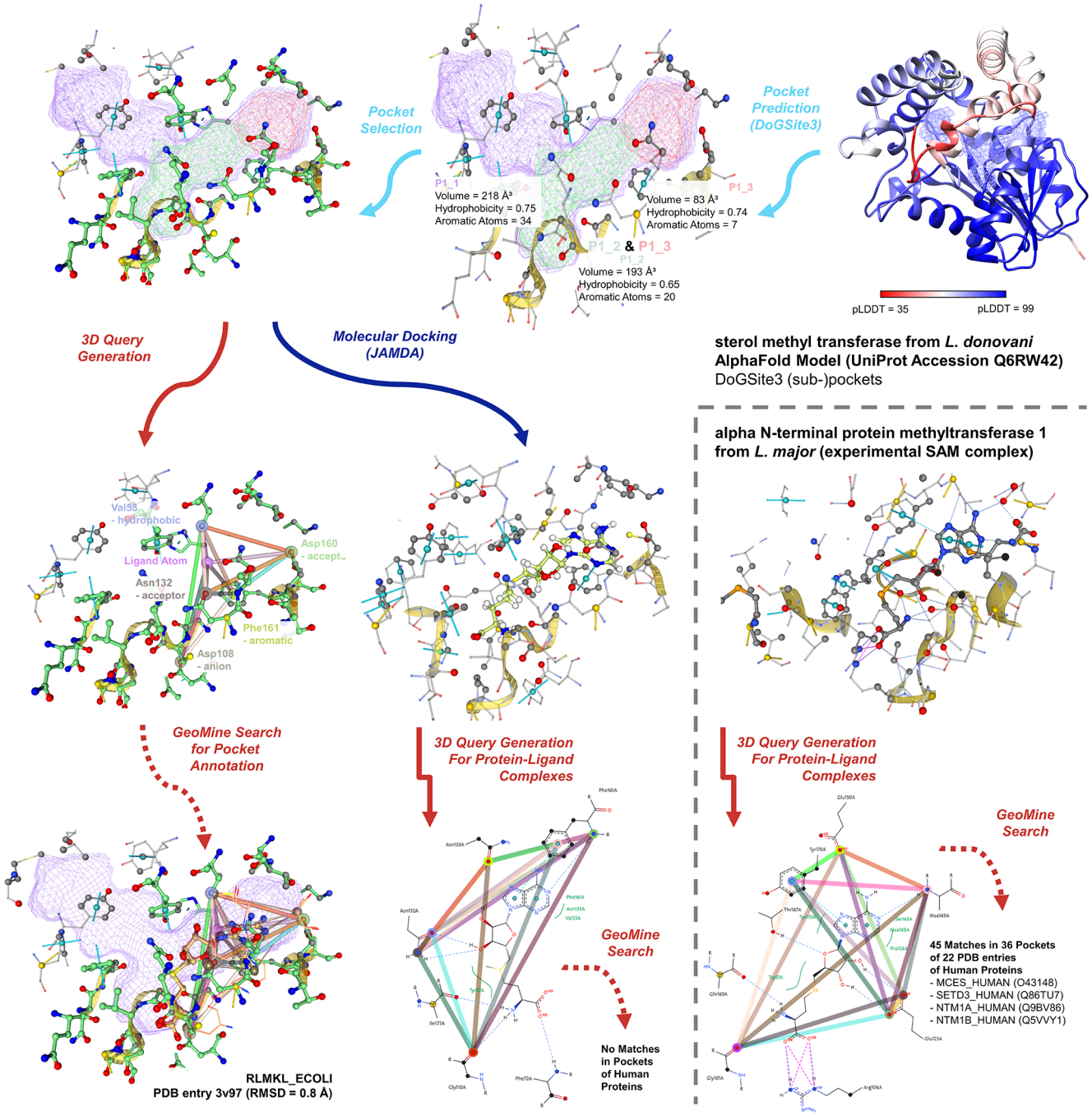



**Fig. 5** Ligand annotation and off-target prediction for binding sites of *Leishmania donovani* sterol C-24 methyltransferase. The explored workflow involves binding site prediction by DoGSite3 (light blue), molecular docking with JAMDA (dark blue), and the search for related binding sites of SAM, SAH, or SFG in the complete PDB for binding pose comparison and a subset of human structures for searching potential off-targets using GeoMine (predicted site 3D query generation and complex 2D query generation, red). The query for the initial GeoMine search with a predicted site was based on manually chosen solvent-exposed atoms. The pharmacophoric properties of the selected atoms were used as query points (hydrogen bond acceptors and donors, aromatic centers, and hydrophobic atoms). For the prediction of related sites in human protein structures, i.e., pockets with similar interaction patterns to the cosubstrate SAM, with the 2D editor, the point features of oxygen and nitrogen atoms were set as solvent-accessible hydrogen bond donors and acceptors, respectively, if they are involved in hydrogen bonds with the ligand. Aromatic centers were modeled if they undergo pi-pi interactions with the ligand. Independent of the query type, all modeled points connected by distances below 14 Å were annotated by distance restraints with tolerances of 1 Å. Note that only point-point distance restraints up to 15 Å can be defined in the frontend

One well-known issue of targeting SAM-binding sites is the comparatively high risk of off-target effects and corresponding toxicity when addressing similar conserved interaction patterns in related enzymes [41]. Although we find highly specialized classes of SAM-binding enzymes in nature [42], a close examination of the interaction pattern similarities might help to identify selectivity-mediating site properties and prevent the design of non-selective inhibitors. Therefore, we further explored the unique features of the binding site. We saved the JAMDA pose on rank 6 and uploaded it as complex to Proteins*Plus*. The corresponding PDB file is available in the Supplementary Information (Online Resource 3). Next, we used the 2D query feature of GeoMine to model residue atoms potentially interacting with SAM-related compounds. As the binding site is highly buried and the number of interactions is high, it is more convenient to use the 2D representation of the interacting partners in this case. We modeled the pharmacophoric properties of all interacting atoms except for the residues interacting with the carboxylic group of the methionyl moiety and backbone atom of Ile177, as those atoms are far apart from the adenosyl moiety. The resulting query was used to screen for related binding sites of human protein structures in the PDB. The corresponding query file in JSON format is available in the Supplementary Information (Online Resource 4). The search took 19 s. Intriguingly, we could not identify similarities in the SAM binding mode predicted for SMT to the one observed for any human enzymes of known structure in complex with SAM, SAH, or SFG, indicating a unique interaction pattern in this protein.

To compare this finding to the results of similar approaches with other SAM-binding enzymes from *L. donovani*, we used another SAM binding site of the enzyme alpha N-terminal protein methyltransferase 1 (UniProt accession number A0A3S7X350). A SIENA [43] search in the PDB revealed a highly related SAM-bound structure of the enzyme of *L. major* (PDB entry 1xtp by the Structural Genomics of Pathogenic Protozoa Consortium). The tool searches for closely related binding sites of other proteins based on perfect k-mer sequence matches in an indexed database of the PDB. As the residues of both active sites overlap nearly perfectly and there are no mutations or gaps in a 5 Å environment, we used a similar GeoMine search strategy to find potentially related sites for this target. As described previously for SMT, we modeled all interacting residue atoms and their distances, omitting the atoms interacting with the carboxylic group of the methionyl moiety and the backbone oxygen atom of Gln165. We omitted the oxygen atom of Thr167 as the ether might represent a comparably weak acceptor. The corresponding query file in JSON format is available in the Supplementary Information (Online Resource 5). The search was performed in 31 s. In contrast to our findings for SMT, we identify several binding sites in human enzymes that are structurally highly related (Fig. 6). The low RMSD values of the matched points indicate a high validity of the hits in terms of matching interacting atoms. A visual inspection of the matches highlights that mainly human N-terminal Xaa-Pro-Lys N-methyltransferase 1, N-terminal Xaa-Pro-Lys N-methyltransferase 2, and mRNA cap guanine-N7 methyltransferase should be considered potential off-targets of compounds addressing similar interacting residues of the SAM site. The match with actin-histidine N-methyltransferase does not lead to a convincing ligand alignment, indicating that the site of this enzyme is different regarding the atoms interacting with SAM. This result suggests that selectively inhibiting this binding site might be more challenging than addressing the one for SMT with an SAM-competitive inhibitor.

In summary, this study illustrates how GeoMine can support the analysis of protein structures concerning ligand binding in just one of the numerous imaginable workflows. Using DoGSite3, putative sites, e.g., from predicted protein structures, can be used as starting points. The fully integrated 2D and 3D query design options paired with the efficient database search capabilities of GeoMine enable on-the-fly structural investigations exploiting data from hundreds of thousands of protein structures. The new functionalities provide easy access to binding site function prediction and automated searches for potential off-targets.

## Conclusion

In this article, we present features and exemplary applications of the new version of GeoMine, a search engine for 3D searching in ligand-bound and predicted empty protein binding sites. Exploiting the full capabilities of such a search engine is a considerable challenge from the user's perspective due to the complexity of 3D molecular arrangements on the atomistic level being part of the query. In related tools, the 3D query formulation is based on either 2D, 3D, or text input. Each of these input types has advantages and disadvantages.

The new version of GeoMine seamlessly integrates all three input types to maximize the usability of the complex 3D query-building process. The newly implemented 2D editor enables a simplified template-free query generation and template-based query selection for ligand-bound binding sites. The 2D templates make optimal use of the editor's limited 2D space by highlighting only those chemical aspects of the binding site that are most relevant to the ligand's interaction with a macromolecule and, therefore, particularly interesting to search for. The 2D editor is instantaneously synchronized with the 3D editor and the textual

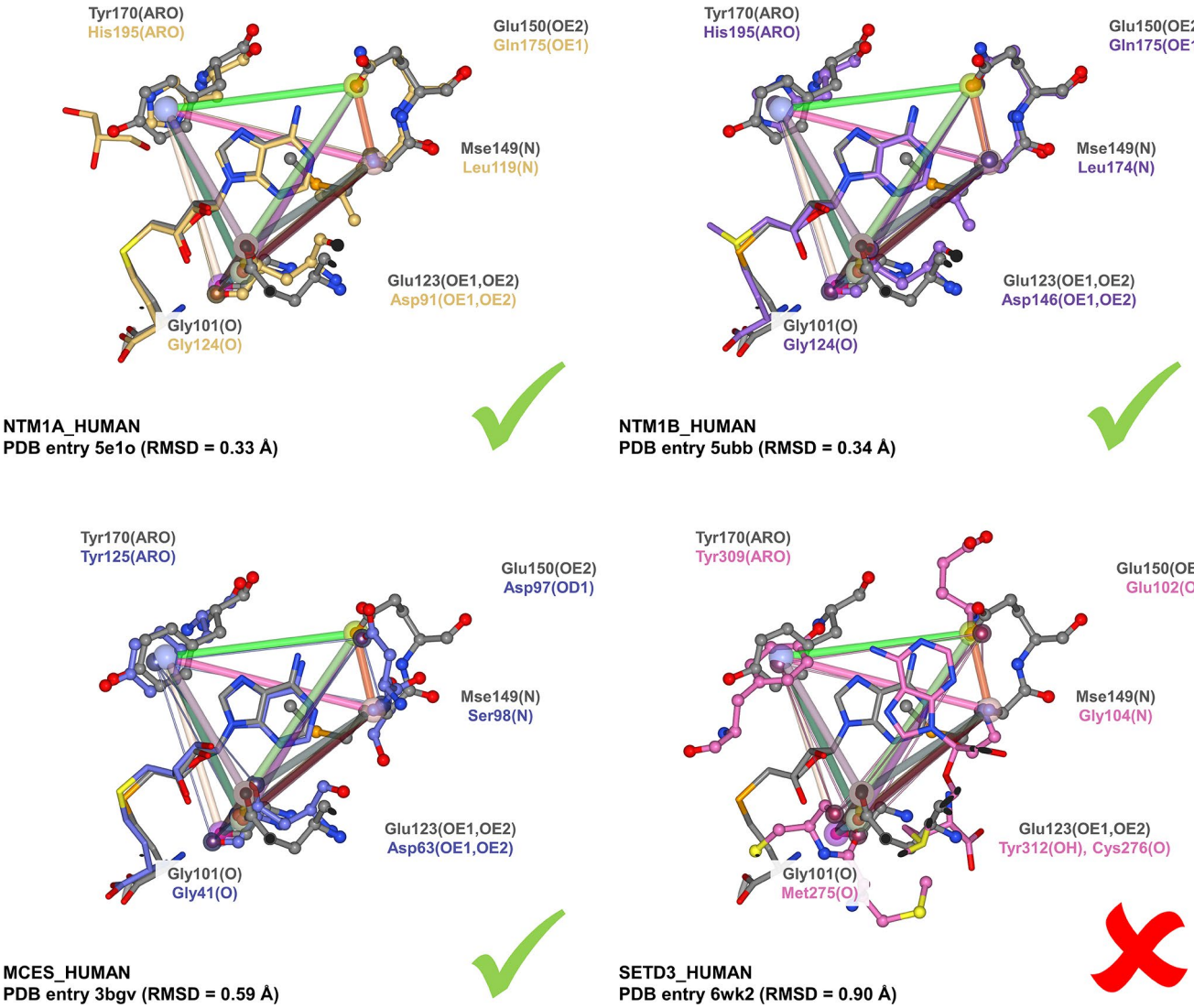



**Fig. 6** Off-Target Prediction for the SAM-binding site of alpha N-terminal protein methyltransferase 1 from *L. major*. The presented aligned matches are based on the query in Fig. 5 for the binding site of PDB entry 1xtp. The aligned matches of N-terminal Xaa-Pro-Lys N-methyltransferase 1 (PDB entry 5e1o), N-terminal Xaa-Pro-Lys N-methyltransferase 2 (PDB entry 5ubb), and mRNA cap guanine-N7 methyltransferase (PDB entry 3bgv) show striking correspondences in the spatial arrangement of interacting atoms and residue types. They should be considered potential off-targets. In contrast, the match with actin-histidine N-methyltransferase does not lead to a convincing alignment. Additionally, the ligand clashes with binding site residues of the query, indicating a different interaction pattern with SAM

query representation in tables, enabling a synergistic query generation process complemented by all three input types. A seamless integration of the PoseEdit features into GeoMine might further improve the usability of the 2D interface. Finally, predicted empty binding sites of artificial intelligence-based protein structure models can now be used as 3D templates in the 3D editor, giving the user a new starting point to tailor queries of interest to elucidate potential ligands.

The search engine's extended graphical user interface will support life scientists in effortlessly generating structural 3D queries on the PDB for the functional analysis of macromolecule-ligand interfaces.

## Abbreviations

| | |
|---|---|
| 3D | Three-dimensional |
| PDB | Protein Data Bank |
| 2D | Two-dimensional |
| RMSD | Root-mean-square deviation |
| JSON | JavaScript Object Notation |
| CSV | Comma-separated values |
| SDF | Structural Data File |
| SVG | Scalable Vector Graphics |
| SMT | Sterol 24-C methyltransferase |
| SAM | S-adenosyl methionine |
| SAH | S-adenosyl homocysteine |
| SFG | Sinefungin |

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s10822-024-00563-3.

**Acknowledgements** The authors thank the whole development team of the NAOMI library and Proteins*Plus* web server for forming the basis of this work.

**Author contributions** The concepts behind GeoMine were developed by KD, JG, MP and MR, the concept of the GeoMine web-based interface was developed and implemented by KD. The case studies were designed by CE. The original draft of the manuscript was written by KD and CE, the project and manuscript writing were supervised by MR, database design by JG and MP, supervised by NR and MR. All authors read and approved the final manuscript.

**Funding** Open Access funding enabled and organized by Projekt DEAL. Development of Proteins*Plus* was supported by de.NBI (in part); German Federal Ministry of Education and Research (BMBF) [031L0105]; Development of GeoMine was supported by the German Federal Ministry of Education and Research as part of CompLS and de.NBI [031L0172, 031L0105 to KD and JG]; CE acknowledges financial support from grant HIDSS-0002 DASHH (Data Science in Hamburg - Helmholtz Graduate School for the Structure of Matter).

**Data availability** All data generated or analyzed during this study are included in this published article and its supplementary information files.

## Declarations

**Competing interests** The authors declare the following competing financial interest(s): Proteins*Plus* and the NAOMI ChemBioSuite use some methods that are jointly owned and/or licensed to BioSolveIT GmbH, Germany, MR is a shareholder of BioSolveIT GmbH.

**Software availability** The GeoMine application is available at https://proteins.plus. The code of the InteractionDrawer library used for the drawing of interactive 2D diagrams is available at https://github.com/rareylab/InteractionDrawer.



## References

1. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. Nucleic Acids Res 28:235–242. https://doi.org/10.1093/nar/28.1.235
2. Kuhn B, Gilberg E, Taylor R, Cole J, Korb O (2019) How significant are unusual protein–ligand interactions? Insights from Database Mining. J Med Chem 62:10441–10455. https://doi.org/10.1021/acs.jmedchem.9b01545
3. Meng EC, Polacco BJ, Babbitt PC (2009) 3D motifs. Rigden. D.J. (eds) From protein structure to function with Bioinformatics. Springer, Dordrecht, pp 187–216. https://doi.org/10.1007/978-1-4020-9058-5_8
4. Ehrt C, Brinkjost T, Koch O (2016) Impact of binding site comparisons on Medicinal Chemistry and Rational Molecular Design. J Med Chem 59:4121–4151. https://doi.org/10.1021/acs.jmedchem.6b00078
5. Bittrich S, Burley SK, Rose AS (2020) Real-time structural motif searching in proteins using an Inverted Index Strategy. PLoS Comput Biol 16:e1008502. https://doi.org/10.1371/journal.pcbi.1008502
6. Korb O, Kuhn B, Hert J, Taylor N, Cole J, Groom C, Stahl M (2016) Interactive and versatile Navigation of Structural databases. J Med Chem 59:4257–4266. https://doi.org/10.1021/acs.jmedchem.5b01756
7. Mobilio D, Walker G, Brooijmans N, Nilakantan R, Denny RA, DeJoannis J, Feyfant E, Kowticwar RK, Mankala J, Palli S, Punyamantula S, Tatipally M, John RK, Humblet C (2010) Protein relational database and protein family knowledge bases to facilitate structure-based design analyses. Chem Biol Drug Des 76:142–153. https://doi.org/10.1111/j.1747-0285.2010.00994.x
8. Weisel M, Bitter HM, Diederich F, So WV, Kondru R (2012) PROLIX: Rapid Mining of protein–ligand interactions in large crystal structure databases. J Chem Inf Model 52:1450–1461. https://doi.org/10.1021/ci300034x
9. Hendlich M, Bergner A, Günther J, Klebe G (2003) Relibase: design and development of a database for Comprehensive Analysis of protein–ligand interactions. J Mol Biol 326:607–620. https://doi.org/10.1016/S0022-2836(02)01408-0
10. Golovin A, Henrick K (2008) MSDmotif: exploring protein sites and motifs. BMC Bioinf 9:312. https://doi.org/10.1186/1471-2105-9-312
11. Inhester T, Bietz S, Hilbig M, Schmidt R, Rarey M (2017) Index-based Searching of Interaction patterns in large collections of protein-ligand interfaces. J Chem Inf Model 57:148–158. https://doi.org/10.1021/acs.jcim.6b00561
12. Angles R, Arenas-Salinas M, García R, Reyes-Suarez JA, Pohl E (2020) GSP4PDB: a web Tool to visualize, search and explore protein-ligand structural patterns. BMC Bioinf 21:85. https://doi.org/10.1186/s12859-020-3352-x
13. Graef J, Ehrt C, Diedrich K, Poppinga M, Ritter N, Rarey M (2022) Searching geometric patterns in protein binding sites and their application to Data Mining in protein kinase structures. J Med Chem 65:1384–1395. https://doi.org/10.1021/acs.jmedchem.1c01046
14. Diedrich K, Graef J, Schöning-Stierand K, Rarey M (2021) GeoMine: interactive pattern mining of protein-ligand interfaces in the Protein Data Bank. Bioinformatics 37:424–425. https://doi.org/10.1093/bioinformatics/btaa693
15. Fassio AV, Santos LH, Silveira SA, Ferreira RS, de Melo-Minardi RC (2020) nAPOLI: a graph-based strategy to detect and visualize conserved protein-ligand interactions in large-scale. TCBB 17:1317–1328. https://doi.org/10.1109/TCBB.2019.2892099
16. Desert Scientific Software (DesertSci) Proasis4. https://desertsci.com. Accessed 7 May 2024
17. Discngine 3decision. https://3decision.discngine.com. Accessed 7 May 2024
18. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen

S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D (2021) Highly accurate protein structure prediction with AlphaFold. Nature 596:583–589. https://doi.org/10.1038/s41586-021-03819-2
19. Schöning-Stierand K, Diedrich K, Ehrt C, Flachsenberg F, Graef J, Sieg J, Penner P, Poppinga M, Ungethüm A, Rarey M (2022) ProteinsPlus: a comprehensive collection of web-based molecular modeling tools. Nucleic Acids Res 50:611–615. https://doi.org/10.1093/nar/gkac305
20. Schöning-Stierand K, Diedrich K, Fährrolfes R, Flachsenberg F, Meyder A, Nittinger E, Steinegger R, Rarey M (2020) ProteinsPlus: interactive analysis of protein–ligand binding interfaces. Nucleic Acids Res 48:48–53. https://doi.org/10.1093/nar/gkaa235
21. Fährrolfes R, Bietz S, Flachsenberg F, Meyder A, Nittinger E, Otto T, Volkamer A, Rarey M (2017) ProteinsPlus: a web portal for structure analysis of macromolecules. Nucleic Acids Res 45:337–343. https://doi.org/10.1093/nar/gkx333
22. Graef J, Ehrt C, Rarey M (2023) Binding site detection remastered: enabling fast, robust, and Reliable binding site detection and descriptor calculation with DoGSite3. J Chem Inf Model 63:3128–3137. https://doi.org/10.1021/acs.jcim.3c00336
23. Bietz S, Urbaczek S, Schulz B, Rarey M (2014) Protoss: a holistic approach to predict tautomers and protonation states in protein-ligand complexes. J Cheminform 6:12. https://doi.org/10.1186/1758-2946-6-12
24. Lippert T, Rarey M (2009) Fast automated placement of polar hydrogen atoms in protein-ligand complexes. J Cheminform 1:13. https://doi.org/10.1186/1758-2946-1-13
25. Diedrich K, Krause B, Berg O, Rarey M (2023) PoseEdit: enhanced ligand binding mode communication by interactive 2D diagrams. J Comput Aided Mol Des 37:491–503. https://doi.org/10.1007/s10822-023-00522-4
26. Stierand K, Rarey M (2010) Drawing the PDB - protein-ligand complexes in two dimensions. ACS Med Chem Lett 1:540–545. https://doi.org/10.1021/ml100164p
27. Stierand K, Rarey M (2007) From modeling to Medicinal Chemistry: Automatic Generation of two-Dimensional Complex diagrams. ChemMedChem 2:853–860. https://doi.org/10.1002/cmdc.200700010
28. Stierand K, Maass PC, Rarey M (2006) Molecular complexes at a glance: automated generation of two-dimensional complex diagrams. Bioinformatics 22:1710–1716. https://doi.org/10.1093/bioinformatics/btl150
29. Davies TG, Bentley J, Arris CE, Boyle FT, Curtin NJ, Endicott JA, Gibson AE, Goldin BT, Griffin RJ, Hardcastle IR, Jewsbury P, Johnson LN, Mesguich V, Newell DR, Noble MEM, Tucker JA, Wang L, Whitfield HJ (2002) Structure-based design of a potent purine-based cyclin-dependent kinase inhibitor. Nat Struct Mol Biol 9:745–749. https://doi.org/10.1038/nsb842
30. Rose AS, Bradley AR, Valasatava Y, Duarte JM, Prlić A, Rose PW (2018) NGL viewer: web-based molecular graphics for large complexes. Bioinformatics 34:3755–3758. https://doi.org/10.1093/bioinformatics/bty419
31. Rose AS, Hildebrand PW (2015) NGL viewer: a web application for molecular visualization. Nucleic Acids Res 43:576–579. https://doi.org/10.1093/nar/gkv402
32. Mitra AK, Mawson AR (2017) Neglected tropical diseases: epidemiology and global burden. Trop Med Infect Dis 2:36. https://doi.org/10.3390/tropicalmed2030036
33. Editorial (2023) Leishmania: an urgent need for new treatments. EBioMedicine 87:104440. https://doi.org/10.1016/j.ebiom.2023.104440
34. Chawla B, Madhubala R (2010) Drug targets in Leishmania. J Parasit Dis 34:1–13. https://doi.org/10.1007/s12639-010-0006-3
35. Nes WD (2000) Sterol methyl transferase: enzymology and inhibition. Mol Cell Biol Lipids 1529:63–88. https://doi.org/10.1016/S1388-1981(00)00138-4
36. Orenes Lorente S, Rodrigues JCF, Jiménez Jiménez C, Joyce-Menekse M, Rodrigues C, Croft SL, Yardley V, de Luca-Fradley K, Ruiz-Pérez LM, Urbina J, de Souza W, González Pacanowska D, Gilbert IH (2004) Novel azasterols as potential agents for treatment of Leishmaniasis and Trypanosomiasis. Antimicrob Agents Chemother 48:2937–2950. https://doi.org/10.1128/aac.48.8.2937-2950.2004
37. Sakyi PO, Broni E, Amewu RK, Miller WA, Wilson MD, Kwofie SK (2023) Targeting Leishmania Donovani sterol methyltransferase for leads using pharmacophore modeling and computational molecular mechanics studies. Inf Med Unlocked 37:101162. https://doi.org/10.1016/j.imu.2023.101162
38. Flachsenberg F, Meyder A, Sommer K, Penner P, Rarey M (2020) A consistent Scheme for gradient-based optimization of protein–ligand poses. J Chem Inf Model 60:6502–6522. https://doi.org/10.1021/acs.jcim.0c01095
39. Flachsenberg F, Ehrt C, Gutermuth T, Rarey M (2024) Redocking the PDB. J Chem Inf Model 64:219–237. https://doi.org/10.1021/acs.jcim.3c01573
40. Holcomb M, Chang Y, Goodsell DS, Forli S (2022) Evaluation of AlphaFold2 structures as docking targets. Protein Sci 32:e4530. https://doi.org/10.1002/pro.4530
41. Rudenko AY, Mariasina SS, Sergiev PV, Polshakov VI (2022) Analogs of S-Adenosyl-L-Methionine in studies of Methyltransferases. Mol Biol 56:229–250. https://doi.org/10.1134/S002689332202011X
42. Kozbial PZ, Mushegian AR (2005) Natural history of S-adenosylmethionine-binding proteins. BMC Struct Biol 5:19. https://doi.org/10.1186/1472-6807-5-19
43. Bietz S, Rarey M (2016) SIENA: efficient compilation of selective protein binding site ensembles. J Chem Inf Model 56:248–259. https://doi.org/10.1021/acs.jcim.5b00588

## C.6 ProteinsPlus: interactive analysis of protein-ligand binding interfaces

[D6] K. Schöning-Stierand, K. Diedrich, R. Fährrolfes, F. Flachsenberg, A. Meyder, E. Nittinger, R. Steinegger, and M. Rarey. Nucleic Acids Research 48 (2020), pp. 48–53. Available: https://doi.org/10.1093/nar/gkaa235. 

*W48–W53* *Nucleic Acids Research, 2020, Vol. 48, Web Server issue* *Published online 16 April 2020*
*doi: 10.1093/nar/gkaa235*

# ProteinsPlus: interactive analysis of protein–ligand binding interfaces

**Katrin Schöning-Stierand*, Konrad Diedrich, Rainer Fährrolfes, Florian Flachsenberg, Agnes Meyder, Eva Nittinger, Ruben Steinegger and Matthias Rarey***

Universität Hamburg, ZBH - Center for Bioinformatics (ZBH), 20146 Hamburg, Germany

Received February 12, 2020; Revised March 19, 2020; Editorial Decision March 30, 2020; Accepted April 14, 2020

## ABSTRACT

**Due to the increasing amount of publicly available protein structures searching, enriching and investigating these data still poses a challenging task. The ProteinsPlus web service (https://proteins.plus) offers a broad range of tools addressing these challenges. The web interface to the tool collection focusing on protein–ligand interactions has been geared towards easy and intuitive access to a large variety of functionality for life scientists. Since our last publication, the ProteinsPlus web service has been extended by additional services as well as it has undergone substantial infrastructural improvements. A keyword search functionality was added on the start page of ProteinsPlus enabling users to work on structures without knowing their PDB code. The tool collection has been augmented by three tools: StructureProfiler validates ligands and active sites using selection criteria of well-established protein–ligand benchmark data sets, WarPP places water molecules in the ligand binding sites of a protein, and METALizer calculates, predicts and scores coordination geometries of metal ions based on surrounding complex atoms. Additionally, all tools provided by ProteinsPlus are available through a REST service enabling the automated integration in structure processing and modeling pipelines.**

## INTRODUCTION

Available structural data of macromolecular complexes in the Protein Data Bank (PDB) (1) are often used as starting point for the successful development of new drugs (2). Although data quality and resolution increase with continuous improvement of methods, structure quality assessment, data enrichment and investigation are a prerequisite for successful structure-driven life science research. Selecting an appropriate macromolecular complex as starting structure poses a great challenge with regard to the growing number of available data and the great differences in quality and applied structure determination methods. Manually curated benchmark datasets like the Astex Diverse Set (3) or the Iridium HT (4) are outdated by now, but the selection criteria used for the generation of these sets are still applicable to the search for new reliable structures. In order to keep pace with the rate of data generation, there is a need for fully automated structure validation methods. The data selection step is followed by structure enrichment consisting of adding computed properties, which cannot be derived directly from the structure determination. A prominent example for an essential enrichment step is the addition of hydrogens to X-ray or Cryo-EM determined structures. The estimation of the formation of hydrogen bonds between protein and ligand directly depends on the calculated positions of hydrogens, protonation state, and the tautomeric state of the amino acid side chains and bound ligands. Also the correctness in prediction of water molecule positions plus the orientation of the water hydrogens and the assignment of metal coordination geometries are essential for a functional understanding of binding and influences the prospects of a design process.

Finding answers to the various questions emerging in a modeling process poses a great challenge for scientists. Many web services addressing specific topics like pocket detection (5), protein–ligand interaction visualization (6,7), protein–protein interface analysis (8,9) and metal interactions (10–12) exist. But there is a lack for comprehensive solutions offering different tools in a unified interface that facilitates the reuse of intermediate results and provides interoperability between tools. The members of the Worldwide PDB partnership (wwPDB; wwPDB.org) (13), for example, provide numerous tools and services to access and explore PDB content (14–16) on their own web pages. Additional to the web services, many software tools for protein struc-

*To whom correspondence should be addressed. Tel: +49 40 42838 7350; Fax: +49 40 42838 7352; Email: rarey@zbh.uni-hamburg.de
Correspondence may also be addressed to Katrin Schöning-Stierand. Tel: +49 40 42838 7372; Fax: +49 40 42838 7352; Email: stierand@zbh.uni-hamburg.de
Present address: Eva Nittinger, Medicinal Chemistry, Respiratory Inflammation and Autoimmune (RIA), BioPharmaceutical R&D, AstraZeneca, Gothenburg, Sweden; Ruben Steinegger, IT, Perfood GmbH, Lübeck 23564, Germany.

tures and their complexes have been developed both as open source (17) and commercial solutions. Often, software usage is restricted by platform dependencies and installation obstacles. Web servers circumvent these issues, however, interoperability between tools and command line based applications remain problematic in practice.

Here, we present an extended version of Proteins*Plus* that addresses a large variety of molecular modelling tasks covering the following areas: structure quality assessment by EDIA (18) and StructureProfiler (19), structure enrichment by Protoss (20), WarPP (21), METALizer, 2D visualization by PoseView (22), binding site ensemble generation by SIENA (23), protein–protein interface classification by HyPPI and pocket detection and druggability estimation by DoGSiteScorer (24). The web interface to our tool collection focusing on protein–ligand interactions has been geared towards easy and intuitive access for life scientists. This includes the visualization of the 3D structure in the embedded NGL viewer (25,26) and the 2D structure diagrams of all ions and small molecules. The layout of the start page with only a text field and two upload buttons is similar to the start page of popular internet search engines and therefore self-explanatory. Once the desired structure has been selected or uploaded, the default layout of the page consisting of the 3D view of the complex on the left hand side, a column containing the aforementioned structure diagrams in the middle of the page, and a tool panel on the right hand side is loaded, see Figure 1. The textual or tabular results of the different tools are presented on tool panel while the 3D view is updated accordingly in order to visualize the calculated result. Structure selections for the different calculations, e.g. a metal ion for running the METALizer (see below), can be done by clicking on the structures of interest in either the 2D or 3D representation. Clicking on results in the tool panel highlights or toggles the corresponding structure in the NGL viewer.

## MATERIALS AND METHODS—ENHANCEMENTS OF THE ProteinsPlus WEB SERVER

Since our last publication in 2017, the Proteins*Plus* web service has been extended by additional tools (WarPP, METALizer and StructureProfiler) as well as it has undergone substantial infrastructural improvements. Most noteworthy, a keyword search, interactive pocket definition and handling, and a REST API have been implemented. The keyword search enables the user to start Proteins*Plus* without knowing the PDB code of the structure of interest. StructureProfiler screens structures based on selection criteria typically used upon data set assembly for structure-based design methods. Combined with EDIA, a comprehensive structure validation is enabled within Proteins*Plus*. WarPP places water molecules for the active site of a given PDB file. METALizer predicts the metal coordination geometry and provides statistical information about the coordination distribution of metal ions in the PDB. The usage of the tools is visually supported within Proteins*Plus*. All tools can be used in an automated way via a REST service. The newly added functionality of Proteins*Plus* will be described in detail below.

### Data handling

*Keyword search.* The entry point to the Proteins*Plus* service is in many cases a publicly available structure from the PDB whose PDB code is not necessarily known to the user. To overcome this issue, a keyword search combined with a small number of quality filters was introduced. Searching with a keyword enables the user to find structures by e.g. a protein or author name, a ligand id or a SMILES string. The obtained results can be further filtered by the deposition date, the experimental method and resolution, and the organism. The keyword search is performed directly on the PDB via its RESTful Web Service APIs (https://www.rcsb.org/pdb/software/rest.do). The service provides different query types controlling the data fields considered. The initially used query type in Proteins*Plus* is 'Text Search' that searches all fields in each entry and can be refined afterwards by additional keyword searches in user-selected fields. All results are presented in a list sorted by the Match Score for the keyword. Interactive histograms give an overview on key data elements like resolution and deposition time and enable easy filtering. Besides summarizing textual information, a 3D picture of the whole structure and 2D structure diagrams of the ligands are provided. Next, with the selected structure the user can decide to start Proteins*Plus* with the default tool overview or directly with a specific tool.

*Pocket handling.* Most of the provided tools within the ProteinsPlus service perform their calculations on the binding site of the protein–ligand complex. Hence, a pocket definition functionality has been added. Pockets can be generated automatically from ligands, manually created by the user by selecting individual amino acids, or extracted from DoGSiteScorer (24) calculations. If a pocket is derived from a ligand, all amino acids in a radius of 6.5 Å to any ligand atom are selected following the recommendations concerning pocket sizes given in former publications (27,28). All pockets can be modified interactively by adding or removing amino acids. Several pocket definitions can be generated for the starting structure. Pockets can be visualized and used as input for binding site ensemble calculations with SIENA (23).

### Structure validation and selection

Up to date, high quality data sets for the validation of structure-based design methods are often manually curated. The growing amount of available structures and the need for specially tailored data sets requires an automated generation of such data sets. StructureProfiler (19) was developed as an all-in-one tool to screen structures based on selection criteria typically used in data set assembly for structure-based design methods. Combined with EDIA (18), which calculates an electron density score for individual atoms and was presented in our previous publication (29), a comprehensive structure validation is enabled within the Proteins*Plus* web service. The analysis performed by StructureProfiler can be divided into four different areas: First, the quality of the experimental data is evaluated using the resolution of the protein structure, its diffraction precision index, R and R_free factor, their difference, and the model

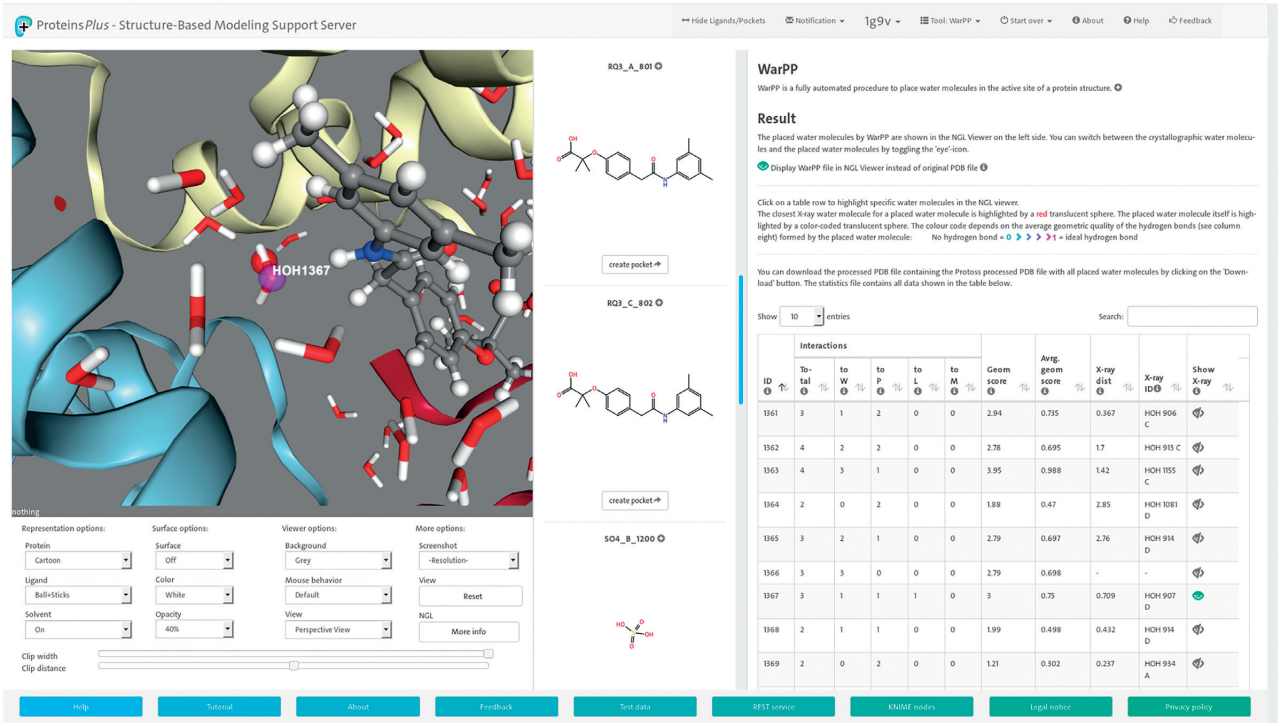



**Figure 1.** The Proteins*Plus* GUI. The 3D representation of human deoxy hemoglobin (Hb) complexed with RSR-13 (PDB code: 1G9V (41)) is shown on the left hand side together with the control panel for the NGL viewer options. The central panel contains a scrollable list of structure diagrams of all ligands contained in the PDB file. On the tool panel on the right hand side, the calculation results of WarPP are shown in a table. In the NGL viewer, a water molecule corresponding to one line in the table is shown. The red translucent sphere shows the position of the closest X-ray water molecule and the pink sphere denotes a good average hydrogen bond quality for this water molecule.

significance. Secondly, the pocket around a ligand is analyzed for its occupancies, intramolecular clashes, $EDIA_m$ per residue, deviations from standard VSEPR bond angles and usual bond lengths. Thirdly, the pocket to ligand B-factor ratio and their intermolecular clash is inspected. Last, small molecules in up to 8.0 Å distance to the protein complex are analyzed as ligands by StructureProfiler. With 21 tests ranging from $EDIA_m$ over torsion angle analysis to their possible exclusion through a SMARTS and a ligand id (PDB HET code) filter, the features of the ligand can be well profiled. Overall, a thorough, objective, transparent, and automatic analysis of any complex available in the PDB can be performed with the help of StructureProfiler.

## Water molecules and metal ions

Water molecules and metal ions play a key role in the mediation of protein–ligand interactions. Therefore, the Proteins*Plus* tool suite has been augmented by a water placement procedure (WarPP) (21) and a metal complex geometry prediction tool (METALizer). WarPP, validated on ten thousands of crystallographic waters, places water molecules in the binding sites of a given PDB structure. METALizer predicts the metal coordination geometry and provides statistical information about metal coordination type distribution in the PDB.

WarPP predicts the energetically favorable, stable, positions of water molecules in protein–ligand binding sites. In a first step, free interaction directions are identified, which include nitrogen or oxygen atoms with an unsaturated hydrogen bond function. Additionally, hydrogen bond acceptors and donors with a bad geometry are considered (geometric score $< 0.85$, see (21)). Based on interaction geometries, previously derived from a large scale analysis of interactions in high resolution protein structures using NAOMInova (30,31), potential water positions are generated in ideal hydrogen bond distances (2.6 and 2.8 Å). These discrete points receive a geometric score based on their deviation angle to the ideal interaction direction. Next, the availability of these potential water positions needs to be determined. Due to close contact with other ligand or protein atoms, some of the interaction surface may not be available and thus cannot be converted into potential water positions. Finally, all potential water positions that are position-optimized and merged in a self-assembling procedure. Herein, based on the individual geometric score, the potential water positions are shifted towards each other until clusters are generated. These clusters are then used to predict water molecules whose location undergo a final numerical optimization.

In a second iteration, further water placement identifies water-water interactions in binding sites, which otherwise might not be identified and can contribute to important water networks. For more details on the WarPP method and its parametrization, please refer to our publication (21).

The web service displays the placed water molecules in the protein–ligand binding site. Additionally, important information regarding the formed hydrogen bonds to water molecules are summarized. If crystallographic water

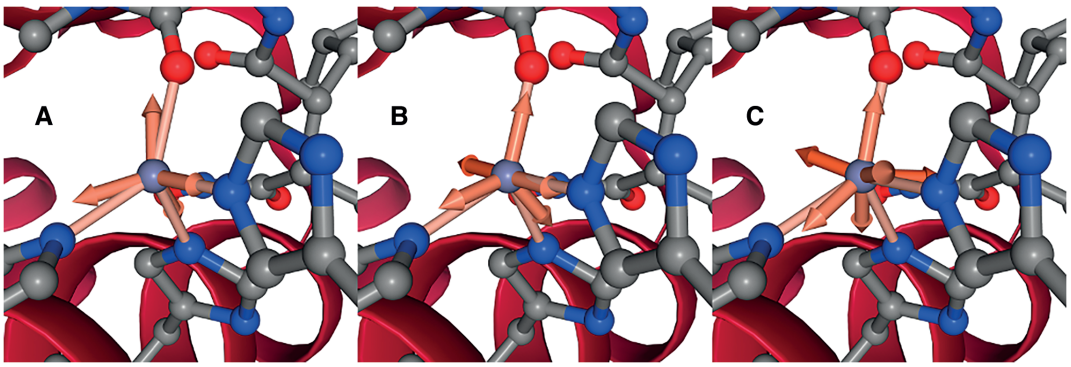



**Figure 2.** Visualization of METALizer results for Atrolysin C with Batimastat (PDB code: 1DTH (42)). The connection between the coordinating atoms and the metal ion are denoted by solid lines, the optimal geometry is denoted as arrows outgoing from the ion. METALizer predicts three different coordination geometries for the zinc ion bound to chain A of the protein: (**A**) tetrahedral (Free Sites: 0, Geometry RMSD: 0.190, Overlap Penalty: 0.0, Score: 9.51), (**B**) trigonal bipyramid (Free Sites: 1, Geometry RMSD: 0.173, Overlap Penalty: 0.0, Score: 12.63) and (**C**) trigonal prismatic (Free Sites: 2, Geometry RMSD: 0.246, Overlap Penalty: 0.001, Score: 20.29). The tetrahedral geometry is considered to be the best one due to the lowest calculated score.

molecules were available in the starting structure, these water molecules will be used as a reference for the placed water molecules. The closest water molecule to each predicted one is available in a tabular representation and can also be displayed in the 3D view, see Figure 1.

METALizer is a tool to analyze the coordination geometry in protein–ligand complexes. In the Proteins*Plus* server METALizer is combined with EDIA (18) for additional quality assessments and SIENA (23) for the search for similar metal sites. Initially, METALizer identifies the coordinating atoms in the metal's coordination sphere; Supplementary Table S1 in the Supporting Information contains a list with element-specific radii of the coordination spheres. All oxygen, nitrogen, sulfur, and chlorine atoms are used as coordinating atoms; carboxylate groups are treated as potential bidentates (32). METALizer identifies the best fitting metal coordination geometries by superposing the geometric arrangement of the coordinating atoms in the binding site to ideal reference geometries (see Supporting Information, Supplementary Table S2 for a list). First, the RMSDs between the angle list of the query site and the angle lists of the reference geometries are calculated (32). For the selected reference geometries the actual superposition is calculated and the resulting distance RMSD is then used for scoring (33). The resulting coordination geometries are scored with a function that includes - besides the superposition RMSD—also the number of free coordination sites (preferring simple geometries) as well as the overlap that a potential binding partner at the free sites would have with other atoms in the protein–ligand complex (this parameter is also used, e.g. by UCSF Chimera (34)). The superposed coordination geometries are supplemented with statistics calculated on the PDB on the frequency of different coordination geometries for the given metal ion and the distribution of metal–partner distances. As an example, the superposition of a calculated zinc geometry and the three closest reference coordination geometries is shown in Figure 2.

Using EDIA (18) it can be checked how well each coordinating atom is supported by the electron density providing an additional quality assessment of the metal coordination site. SIENA (23) allows the fast retrieval, structural superposition and analysis of similar metal sites (with a sequence identity of $\geq$70% within the metal site) from the PDB. Within seconds to minutes similar metal sites can be retrieved using SIENA, analyzed and compared with METALizer, finding at least one similar site in another PDB structure for 75% of our test queries (for details, see Supporting Information). Additionally, the very same statistics as for the PDB (coordination geometry frequency and metal-partner distances) are calculated for the SIENA ensemble of similar metal sites.

METALizer provides the same basic functionalities (metal coordination geometry identification and statistics) as other—still maintained—web servers with a focus on metal ions in biological complexes like the MetalPDB (12) or the CheckMyMetal server (11) do, however, has some unique features making it complementary to existing tools: The integration of our EDIA score adds valuable information to the quality assessments given by the CheckMyMetal server. Our SIENA-based search for similar metal binding sites has a different focus than the MetalS$^3$ (10) database search tool within the MetalPDB: The SIENA-based search together with METALizer is able to find and analyze metal sites with a similar amino acid sequence to the query metal site within seconds to minutes. On the other hand, the MetalS$^3$ tool searches for metal sites that are structurally similar, however, can take hours to run for user-provided PDB files (10). For more information about computing times and search results of METALizer in combination with SIENA, see last paragraph and Supplementary Figure S1 in the Supporting Information.

## Accessibility

Additional to the graphical user interface, a REST API for each of the Proteins*Plus* tools has been made available. API requests can be sent with the command line tool curl or with a browser rest client plugin. The API allows the user to create jobs for the respective tools, each requiring a different set of parameters. Calculation results can then be

accessed and downloaded. The base URL for version 1 is https://proteins.plus/api. The REST service usage and output is documented in detail for each individual tool on the Proteins*Plus* website together with a sample call for both the POST and the GET method. Providing a REST API makes the different tools available for an automated integration in modelling pipelines and software libraries. As an application example, a KNIME node (https://www.knime.com) has been developed for each tool and made available on the Proteins*Plus* website showcasing the usage of the respective APIs.

## SUMMARY AND OUTLOOK

Together with the additional functionality described above, Proteins*Plus* evolved into a versatile instrument for molecular modeling processes. The analysis and processing of binding sites and ligands on atomic level give comprehensive insights in the binding mode of the interacting molecules. In 2019, the server received 61,830 unique page view requests from 21,217 users. Further usability improvement of Proteins*Plus* workflows could be reached by an increase of tool interoperability: using results from calculations of other tools as input without needing intermediate formatting steps would enable the implementation of automated workflows.

Proteins*Plus* combines the advantages of a web service and a molecular modeling desktop application: the unified graphical user interface makes the usage of new or unfamiliar tools possible without a tedious learning effort, calculation results can be interconnected or reused for further calculations, and no local installation is needed. Connecting Proteins*Plus* to other web services could lead to deeper knowledge of a PDB structure. So far, a connection to the enzyme database BRENDA (35) already exists. A tool that searches for related bioactivity data of a complex in ChEMBL (36) is already included as alpha version in Proteins*Plus*. Currently, we investigate methods to include alternative structure files, for example from PDB-REDO (37).

For the near future, we plan to extend Proteins*Plus* by a search functionality that performs a textual, numerical and 3D search with full chemical awareness in protein–ligand interfaces (38). Additionally, we intend to incorporate docking and virtual screening methods (39,40). Thus, Proteins*Plus* opens the way to a large range of functionality from the analysis of protein structure and function to molecular design techniques for every life scientist.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## FUNDING

Development of Proteins*Plus* by de.NBI (in part); German Federal Ministry of Education and Research (BMBF) [031L0105]. Funding for open access charge: German Federal Ministry of Education and Research (BMBF) [031L0105].

*Conflict of interest statement*. Proteins*Plus* is a web service offered to the whole scientific community free of charge. Some tools within Proteins*Plus* are available for download within the NAOMI ChemBio Suite. The suite is available free of charge for academic use only. The authors declare financial conflict of interest in case that the NAOMI ChemBio Suite is licensed to commercial users for charge. Some methods used in Proteins*Plus* and in the NAOMI ChemBio Suite are jointly owned and/or licensed to BioSolveIT GmbH, Germany, M.R. is a shareholder of BioSolveIT GmbH.

## REFERENCES

1. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H. and Shindyalov,I.N. (2000) The Protein Data Bank (www.rcsb.org). *Nucleic Acids Res.*, **28**, 235–242.
2. Goodsell,D.S., Zardecki,C., Di Costanzo,L., Duarte,J.M., Hudson,B.P., Persikova,I., Segura,J., Shao,C., Voigt,M., Westbrook,J.D. *et al.* (2020) RCSB Protein Data Bank: enabling biomedical research and drug discovery. *Protein Sci.*, **29**, 52–65.
3. Warren,G.L., Do,T.D., Kelley,B.P., Nicholls,A. and Warren,S.D. (2012) Essential considerations for using protein–ligand structures in drug discovery. *Drug Discov. Today*, **17**, 1270–1281.
4. Hartshorn,M.J., Verdonk,M.L., Chessari,G., Brewerton,S.C., Mooij,W.T.M., Mortenson,P.N. and Murray,C.W. (2007) Diverse, high-quality test set for the validation of protein–ligand docking performance. *J. Med. Chem.*, **50**, 726–741.
5. Jendele,L., Krivak,R., Skoda,P., Novotny,M. and Hoksza,D. (2019) PrankWeb: a web server for ligand binding site prediction and visualization. *Nucleic Acids Res.*, **47**, W345–W349.
6. Salentin,S., Schreiber,S., Haupt,V.J., Adasme,M.F. and Schroeder,M. (2015) PLIP: fully automated protein–ligand interaction profiler. *Nucleic Acids Res.*, **43**, W443–W447.
7. Jubb,H.C., Higueruelo,A.P., Ochoa-Montaño,B., Pitt,W.R., Ascher,D.B. and Blundell,T.L. (2017) Arpeggio: a web server for calculating and visualising interatomic interactions in protein structures. *J. Mol. Biol.*, **429**, 365–371.
8. Roy,A.A., Dhawanjewar,A.S., Sharma,P., Singh,G. and Madhusudhan,M.S. (2019) Protein Interaction Z Score Assessment (PIZSA): an empirical scoring scheme for evaluation of protein–protein interactions. *Nucleic Acids Res.*, **47**, W331–W337.
9. Evgeny,K. (2010) Crystal contacts as nature's docking solutions. *J. Comput. Chem.*, **31**, 133–143.
10. Valasatava,Y., Rosato,A., Cavallaro,G. and Andreini,C. (2014) MetalS3, a database-mining tool for the identification of structurally similar metal sites. *J. Biol. Inorg. Chem.*, **19**, 937–945.
11. Zheng,H., Cooper,D.R., Porebski,P.J., Shabalin,I.G., Handing,K.B. and Minor,W. (2017) CheckMyMetal: a macromolecular metal-binding validation tool. *Acta Crystallogr. Sect. D Struct. Biol.*, **73**, 223–233.
12. Putignano,V., Rosato,A., Banci,L. and Andreini,C. (2018) MetalPDB in 2018: A database of metal sites in biological macromolecular structures. *Nucleic Acids Res.*, **46**, D459–D464.
13. Burley,S.K., Berman,H.M., Bhikadiya,C., Bi,C., Chen,L., Costanzo,L. Di, Christie,C., Duarte,J.M., Dutta,S., Feng,Z. *et al.* (2019) Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res.*, **47**, D520–D528.
14. PDBe-KB consortium (2020) PDBe-KB: a community-driven resource for structural and functional annotations. *Nucleic Acids Res.*, **48**, D344–D353.
15. Burley,S.K., Berman,H.M., Bhikadiya,C., Bi,C., Chen,L., Di Costanzo,L., Christie,C., Dalenberg,K., Duarte,J.M., Dutta,S. *et al.* (2019) RCSB Protein Data Bank: Biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Res.*, **47**, D464–D474.
16. Dana,J.M., Gutmanas,A., Tyagi,N., Qi,G., O'Donovan,C., Martin,M. and Velankar,S. (2019) SIFTS: Updated Structure Integration with Function, Taxonomy and Sequences resource allows

40-fold increase in coverage of structure-based annotations for proteins. *Nucleic Acids Res.*, **47**, D482–D489.
17. Pirhadi,S., Sunseri,J. and Koes,D.R. (2016) Open source molecular modeling. *J. Mol. Graph. Model.*, **69**, 127–143.
18. Meyder,A., Nittinger,E., Lange,G., Klein,R. and Rarey,M. (2017) Estimating electron density support for individual atoms and molecular fragments in X-ray structures. *J. Chem. Inf. Model.*, **57**, 2437–2447.
19. Meyder,A., Kampen,S., Sieg,J., Fährrolfes,R., Friedrich,N.O., Flachsenberg,F. and Rarey,M. (2019) StructureProfiler: an all-in-one tool for 3D protein structure profiling. *Bioinformatics*, **35**, 874–876.
20. Bietz,S., Urbaczek,S., Schulz,B. and Rarey,M. (2014) Protoss: a holistic approach to predict tautomers and protonation states in protein–ligand complexes. *J. Cheminform.*, **6**, 12.
21. Nittinger,E., Flachsenberg,F., Bietz,S., Lange,G., Klein,R. and Rarey,M. (2018) Placement of water molecules in protein structures: from large-scale evaluations to single-case examples. *J. Chem. Inf. Model.*, **58**, 1625–1637.
22. Stierand,K., Maaß,P.C. and Rarey,M. (2006) Molecular complexes at a glance: automated generation of two-dimensional complex diagrams. *Bioinformatics*, **22**, 1710–1716.
23. Bietz,S. and Rarey,M. (2016) SIENA: efficient compilation of selective protein binding site ensembles. *J. Chem. Inf. Model.*, **56**, 248–259.
24. Volkamer,A., Kuhn,D., Grombacher,T., Rippmann,F. and Rarey,M. (2012) Combining global and local measures for structure-based druggability predictions. *J. Chem. Inf. Model.*, **52**, 360–372.
25. Rose,A.S., Bradley,A.R., Valasatava,Y., Duarte,J.M., Prlic,A. and Rose,P.W. (2018) NGL viewer: web-based molecular graphics for large complexes. *Bioinformatics*, **34**, 3755–3758.
26. Rose,A.S., Hildebrand, and,P.W. (2015) NGL Viewer: a web application for molecular visualization. *Nucleic Acids Res.*, **43**, W576–W579.
27. Kramer,B., Rarey,M. and Lengauer,T. (1999) Evaluation of the FlexX incremental construction algorithm for protein- ligand docking. *Proteins Struct. Funct. Genet.*, **37**, 228–241.
28. Rarey,M., Kramer,B., Lengauer,T. and Klebe,G. (1996) A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.*, **261**, 470–489.
29. Fährrolfes,R., Bietz,S., Flachsenberg,F., Meyder,A., Nittinger,E., Otto,T., Volkamer,A. and Rarey,M. (2017) ProteinsPlus: a web portal for structure analysis of macromolecules. *Nucleic Acids Res.*, **45**, W337–W343.
30. Nittinger,E., Inhester,T., Bietz,S., Meyder,A., Schomburg,K.T., Lange,G., Klein,R. and Rarey,M. (2017) Large-scale analysis of hydrogen bond interaction patterns in protein-ligand interfaces. *J. Med. Chem.*, **60**, 4245–4257.
31. Inhester,T., Nittinger,E., Sommer,K., Schmidt,P., Bietz,S. and Rarey,M. (2017) NAOMInova: interactive geometric analysis of noncovalent interactions in macromolecular structures. *J. Chem. Inf. Model.*, **57**, 2132–2142.
32. Seebeck,B., Reulecke,I., Kämper,A. and Rarey,M. (2008) Modeling of metal interaction geometries for protein–ligand docking. *Proteins Struct. Funct. Genet.*, **71**, 1237–1254.
33. Andreini,C., Cavallaro,G. and Lorenzini,S. (2012) FindGeo: a tool for determining metal coordination geometry. *Bioinformatics*, **28**, 1658–1660.
34. Pettersen,E.F., Goddard,T.D., Huang,C.C., Couch,G.S., Greenblatt,D.M., Meng,E.C. and Ferrin,T.E. (2004) UCSF chimera - a visualization system for exploratory research and analysis. *J. Comput. Chem.*, **25**, 1605–1612.
35. Jeske,L., Placzek,S., Schomburg,I., Chang,A. and Schomburg,D. (2019) BRENDA in 2019: a European ELIXIR core data resource. *Nucleic Acids Res.*, **47**, D542–D549.
36. Mendez,D., Gaulton,A., Bento,A.P., Chambers,J., De Veij,M., Félix,E., Magariños,M.P., Mosquera,J.F., Mutowo,P., Nowotka,M. *et al.* (2019) ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.*, **47**, D930–D940.
37. Joosten,R.P., Long,F., Murshudov,G.N. and Perrakis,A. (2014) The PDB-REDO server for macromolecular structure model optimization. *IUCrJ*, **1**, 213–220.
38. Inhester,T., Bietz,S., Hilbig,M., Schmidt,R. and Rarey,M. (2017) Index-based searching of interaction patterns in large collections of protein-ligand interfaces. *J. Chem. Inf. Model.*, **57**, 148–158.
39. Schomburg,K.T., Bietz,S., Briem,H., Henzler,A.M., Urbaczek,S. and Rarey,M. (2014) Facing the challenges of structure-based target prediction by inverse virtual screening. *J. Chem. Inf. Model.*, **54**, 1676–1686.
40. Von Behren,M.M., Volkamer,A., Henzler,A.M., Schomburg,K.T., Urbaczek,S. and Rarey,M. (2013) Fast protein binding site comparison via an index-based screening technology. *J. Chem. Inf. Model.*, **53**, 411–422.
41. Safo,M.K., Moure,C.M., Burnett,J.C., Joshi,G.S. and Abraham,D.J. (2001) High-resolution crystal structure of deoxy hemoglobin complexed with a potent allosteric effector. *Protein Sci.*, **10**, 951–957.
42. Botos,I., Scapozza,L., Zhang,D., Liotta,L.A. and Meyer,E.F. (1996) Batimastat, a potent matrix metalloproteinase inhibitor, exhibits an unexpected mode of binding. *Proc. Natl. Acad. Sci. U.S.A.*, **93**, 2749–2754.

## C.7 ProteinsPlus: a comprehensive collection of web-based molecular modeling tools

[D7] K. Schöning-Stierand, K. Diedrich, C. Ehrt, F. Flachsenberg, J. Graef, J. Sieg, P. Penner, M. Poppinga, A. Ungethüm, and M. Rarey. Nucleic Acids Research 50 (2022), pp. 611–615. Available: https://doi.org/10.1093/nar/gkac305. Reprinted with permission from [D7] and Oxford University Press.

*Published online 30 April 2022* *Nucleic Acids Research, 2022, Vol. 50, Web Server issue* ***W611–W615***
*https://doi.org/10.1093/nar/gkac305*

# Proteins*Plus*: a comprehensive collection of web-based molecular modeling tools

**Katrin Schöning-Stierand [1,†], Konrad Diedrich [1,†], Christiane Ehrt [1,†], Florian Flachsenberg [1,†], Joel Graef [1,†], Jochen Sieg [1,†], Patrick Penner [1], Martin Poppinga [1,2], Annett Ungethüm [3] and Matthias Rarey [1,*]**

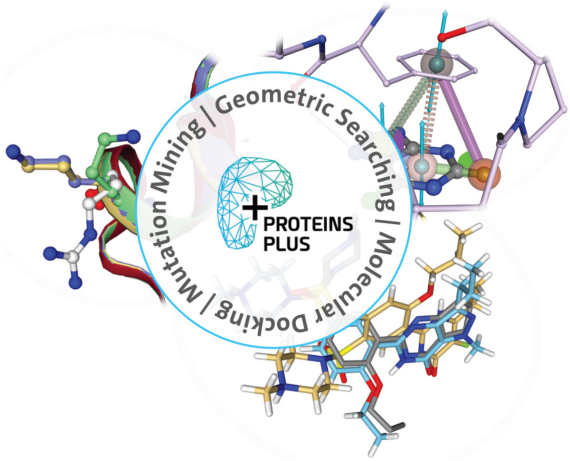
[1]Universität Hamburg, ZBH - Center for Bioinformatics, Bundesstraße 43, 20146 Hamburg, Germany, [2]Universität Hamburg, Department of Informatics, Vogt-Kölln-Straße 30, 22527 Hamburg, Germany and [3]Universität Hamburg, Center for Data and Computing in Natural Sciences (CDCS), Notkestraße 11, 22607 Hamburg, Germany

Received February 26, 2022; Revised April 05, 2022; Editorial Decision April 10, 2022; Accepted April 19, 2022

## ABSTRACT

**Upon the ever-increasing number of publicly available experimentally determined and predicted protein and nucleic acid structures, the demand for easy-to-use tools to investigate these structural models is higher than ever before. The Proteins*Plus* web server (https://proteins.plus) comprises a growing collection of molecular modeling tools focusing on protein–ligand interactions. It enables quick access to structural investigations ranging from structure analytics and search methods to molecular docking. It is by now well-established in the community and constantly extended. The server gives easy access not only to experts but also to students and occasional users from the field of life sciences. Here, we describe its recently added new features and tools, beyond them a novel method for on-the-fly molecular docking and a search method for single-residue substitutions in local regions of a protein structure throughout the whole Protein Data Bank. Finally, we provide a glimpse into new avenues for the annotation of AlphaFold structures which are directly accessible via a RESTful service on the Proteins*Plus* web server.**

## GRAPHICAL ABSTRACT




## INTRODUCTION

The Proteins*Plus* (1,2) web server, openly available at https://proteins.plus, offers molecular modeling support for all protein structures that are publicly available as PDB files in the Protein Data Bank (PDB) (3). Usually, workflows for structure-based design necessitate a comprehensive user knowledge of different molecular modeling tools. For example, predicting potential binding sites, finding similar binding sites for ensemble docking, and molecular docking of small molecules of interest into a binding site requires access to and knowledge of a high number of tools with a multitude of parameters. Furthermore, researchers must rely on their computational resources. With the Proteins*Plus* server, these shortcomings are overcome by enabling users to perform all these steps via one unique and easily accessible interface. The server is under constant development including

*To whom correspondence should be addressed. Tel: +49 40 428387350; Fax: +49 40 428387352; Email: matthias.rarey@uni-hamburg.de
†The authors wish it to be known that, in their opinion, the first six authors should be regarded as Joint First Authors.
Present address: Florian Flachsenberg, BioSolveIT GmbH, An der Ziegelei 79, 53757 St. Augustin, Germany.

fine-tuning, feature extensions, and the integration of additional modeling tools.

Here, we offer insights into feature extensions for the structural multi-purpose comparison tool GeoMine, the newly integrated molecular docking tool JAMDA and MicroMiner - a method that can be used to screen for single-residue substitutions in local protein environments in the whole PDB.

Finally, the artificial intelligence-based protein structure predictions by AlphaFold (currently predicted by AlphaFold Monomer v2.0) enable unprecedented access to high-quality models of proteins of yet unknown structure (4). These models are now readily accessible via the AlphaFold Protein Structure Database (https://alphafold.ebi.ac.uk/) and can be directly imported via the provided REST API.

## MATERIAL AND METHODS: EXTENSIONS AND NOVEL TOOLS

### GeoMine

From the analysis of binding sites to investigations of geometric preferences for interactions, the ever-increasing number of molecular structures in the PDB offers a multitude of possibilities for in-depth studies of binding sites, their properties and their similarities. This requires comprehensive search capabilities. With GeoMine (5,6), we have developed a search engine that allows for the generation of and the search for atom-based geometric query patterns and an extensive textual and numerical filtering of the PDB. The query atoms can be described manually or automatically with varying degrees of detail, from major properties like the corresponding molecule type, i.e. nucleic acid, protein, ligand, water, or metal, to more restrictive ones, e.g. the molecular surface contribution of a protein or nucleic acid atom. Further feature points like aromatic ring centers can be added to the query and described equally. Distance ranges or hydrogen bond, pi–pi, pi stacking, pi–cation, metal and ionic interactions between atoms and feature points can be introduced into the query, and angle ranges between those can be specified. With the combination of all these features, almost any 3D pattern can be designed and searched in the entire PDB.

In the Proteins*Plus* user interface, the query can be created in a 3D viewer from scratch by the placement of new atoms and feature points or by selecting those in a visualized binding site of a PDB/AlphaFold structure or any uploaded structure file. For this structure, GeoMine predicts binding pockets with interactions and hydrogen atoms using the tools DoGSiteScorer (7) and Protoss (8,9), respectively. If a ligand is present but no pocket has been calculated, a pocket is defined using a radius of 6.5 Å of any ligand atom. The computing times for the iterative search of over one million preprocessed bindings sites depends on the specificity of the query. Most requests can be processed in the range of minutes. For each detected hit, the root-mean-square deviation (RMSD) between the query and the part of the site matching the query is calculated enabling a ranking of the results by geometric fit. The 150 best results are listed in a table and can be visually inspected superimposed to the query in the NGL viewer. Different visualization options are available, for example, choice of residues (complete pocket or only of the residues that match the query). The 150 best-matching pockets can be downloaded in PDB format together with a report containing the statistical overview of all results. The statistics report lists the PDB IDs and ligand names of all found pockets, the distributions of the RMSD values, and the properties of all matched atoms, feature points, distances, interactions, and angles of the query, e.g. the functional group distribution for a matched ligand atom. The user interface with a query history allows a continuous refinement of the results providing an interactive workflow of query modification and subsequent searching in the results. With this tool, protein function or ligand off-targets can be discovered by searching similar binding site properties in 3D space. GeoMine has recently been applied for a detailed analysis of structural features in protein kinase structures (5).

### JAMDA

Protein–ligand docking is one of the core tasks in structure-based drug design. With JAMDA, we aimed for the implementation of a fully-automated docking workflow in the Proteins*Plus* server that does not only provide the actual docking algorithm but also encompasses all necessary preprocessing steps, including protonation state assignment and calculation of hydrogen coordinates for the protein (8), prediction of protonation and tautomeric states of the molecules to be docked (10), as well as the generation of 3D coordinates/conformations (11). While a certain degree of manual intervention is possible, our goal was to provide a fully automated workflow with optimized default parameters. This enables even less experienced users to derive potential binding modes of small molecules in the binding site of interest. From the analysis of structure-activity relationships to the test of new binding hypotheses, the established pipeline offers unlimited access to predicted binding modes.

JAMDA docking combines the TrixX docking algorithm (12,13) for initial pose generation with the JAMDA scoring function (14), and our novel LSL-BFGS optimization algorithm(14,15) for scoring and pose optimization. Initially, conformers for the molecule to be docked are generated with the Conformator (11). The raw poses are subjected to a scoring and optimization cascade using the JAMDA scoring function to refine and rank the docking poses.

On Proteins*Plus*, JAMDA allows for a fully automated docking: Only the protein, the binding site, and the molecules to be docked must be provided by the user. The binding site can be defined based on a known ligand or selected from the pocket definitions in Proteins*Plus* (1) (e.g. predicted by DoGSiteScorer (16)). To enable the user to manually adjust the binding sites, all ligand-based and predicted binding sites which do not originate from GeoMine are editable by the user in the pockets tab by clicking on the pencil symbol of the pocket of interest in the upper right corner. Neither the protein nor the molecules to be docked must be manually prepared by the user because this is an integral part of the JAMDA docking workflow: The protein is prepared by assigning likely protonation states using Protoss (8). Furthermore, only structurally relevant water molecules and small molecules that are common cofac-

tors are kept. The molecules to be docked can be provided by picking a ligand from the NGL viewer for redocking studies or by uploading molecules in any common molecular file format (including SMILES without coordinates). Their predominant protonation and tautomeric states are predicted with UNICON (10) prior to docking. Most of these preprocessing steps can optionally be customized by the user.

The preprocessing and docking are performed on the server and, currently, up to five molecules may be docked simultaneously. In the Proteins*Plus* web interface, the resulting docking poses are shown in a table (with JAMDA score and the RMSD if a redocking was performed) and visualized in the NGL viewer panel for interactive analyses. They can also be downloaded for alternative visualizations and further processing. In consequence, JAMDA offers a pipeline for molecular docking that provides reliable results even in the absence of substantial knowledge regarding molecular modeling tools.

## MicroMiner

MicroMiner searches for mutations in protein structure databases. On Proteins*Plus*, it screens for single-residue substitutions in the experimental structures of the entire PDB. Retrieved mutant structures can be easily analyzed and compared to the wildtype through automatically generated superpositions in the NGL viewer. The tool focuses on the local 3D micro-environment of single residues in a query protein. It searches the protein structure database for similar local environments with a mutated central residue. For reasonably large wildtype protein structures it is feasible to search for substitutions of all residues in the query at once. In this way, a user can comprehensively explore the wealth of experimental protein structures that exemplify the local effects of mutations through the interactive web interface.

MicroMiner originates from the ASCONA (17) and SIENA (18) technology for binding site similarity search and ensemble compilation. However, instead of focusing on the protein environment of ligands, MicroMiner uses the local 3D micro-environment of any individual residue as the query to search for residues embedded in similar local arrangements. A database search starts by selecting a query residue from which the local 3D protein neighborhood within a distance cutoff (default 6.5 Å) represents the query micro-environment. The connected sequence fragments of this environment are used to identify candidate protein structures with similar sequence fragments in the database. Second, all potential matches are identified by residue-wise sequence alignments. A subsequent fuzzy geometric filter based on the Cα atom orientation and distances of the matching sequence fragments ensures a reasonably similar structural arrangement while tolerating structural changes upon mutation. Thus, we identify local micro-environments with a high sequence and structural similarity. Figure 1 shows the MicroMiner workflow.

Within the Proteins*Plus* server, the user can select single residues of interest or all residues in the input structure to be searched against the PDB. Searching for all residues is feasible within one minute or less on average, depending on the size of the input protein and the number of similar

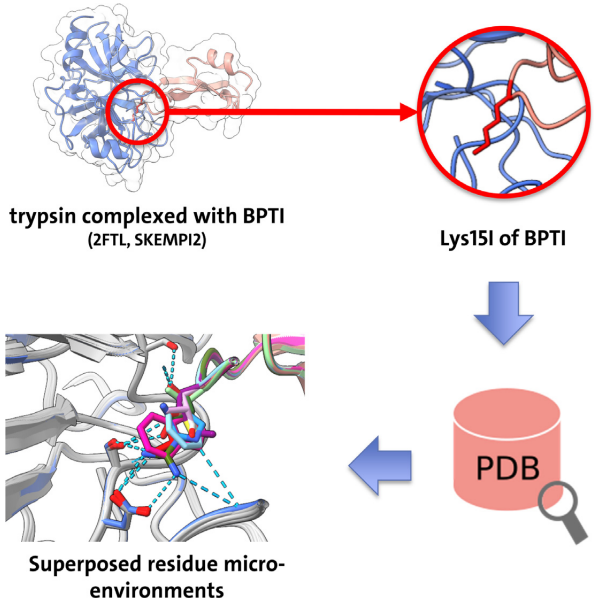



**Figure 1.** MicroMiner workflow. With the local 3D micro-environment of a selected query residue, the PDB is searched. Structures from the database containing a similar micro-environment identical in sequence except for the query residue position are retrieved and superposed for analysis. In this way, MicroMiner yields structure ensembles exemplifying the local effects of mutations.

micro-environments in the PDB. The protein structures of retrieved micro-environments can be explored interactively as a structure ensemble in the 3D viewer and sorted by properties of interest, for example, the RMSD of the local environments to investigate the structural effects of mutations. Further applications are the search for highly conserved regions in protein structures, comparisons of the impact of conservative and radical substitutions, or the investigation of structural effects upon substitution for evaluating the reliability and accuracy of computationally generated models of single-residue substitutions.

## Integration of AlphaFold structures

The inclusion of AlphaFold protein structure models (4) (https://alphafold.ebi.ac.uk/) in the Proteins*Plus* web server enables easy access to machine learning-based predictions of previously unknown structures. The models are accessible on our web server by entering the UniProt Accession Number on the landing page or uploading a preprocessed structure. The user can analyze these structures in the same way publicly available PDB structures can be analyzed by making use of all applicable capabilities of the Proteins*Plus* tools.

Besides the structural uncertainty of AlphaFold structures (19), the missing ligand annotations are a major drawback. This led to the development of the database AlphaFill (20) which annotates the 3D models with cofactors and metal ions and transfers them into the structure assisting in the functional annotation of the models. However, this annotation procedure was only followed for structures that show an identity of at least 35% to known 3D structures

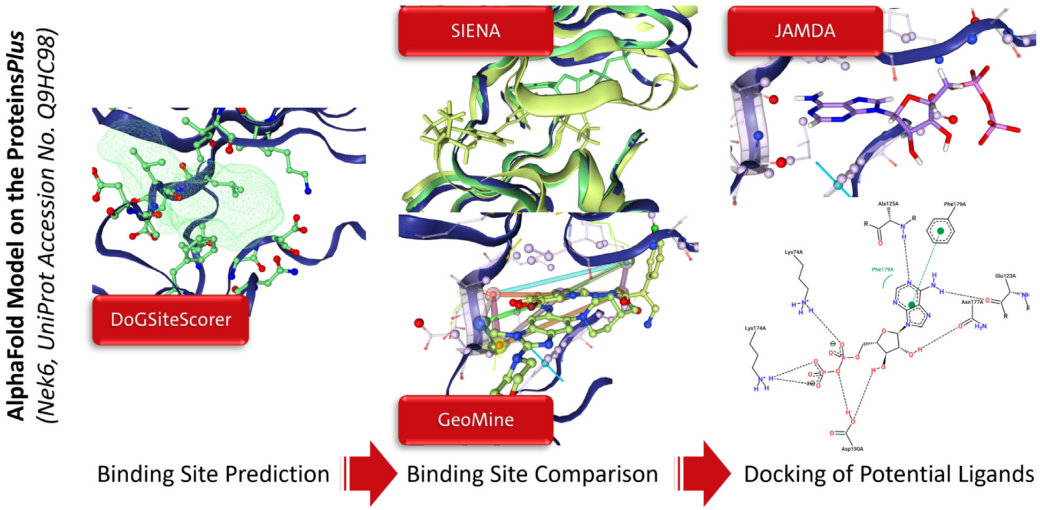



**Figure 2.** This workflow shows exemplary results for structural investigations of the AlphaFold model for the Nek6 (UniProt Accession Number Q9HC98). First, the user can detect druggable binding sites with DoGSiteScorer. Pocket 'P_2' which was predicted as druggable is depicted in green on the right. Next, the pocket can be used for a SIENA search for similar binding sites. Shown are two matches from this analysis with Nek7 structures: 2WQN with ADP and 6S73 in complex with the ligand with the ID F9N in the PDB. GeoMine can be applied for more specific user-defined searches in the binding sites of the PDB. Using a geometric query annotating solvent-exposed potentially interacting atoms and their distances, we found 116 pockets with a similar geometry in the PDB (e.g. cAMP-dependent protein kinase A with the PDB ID 7BAQ, PDB ligand ID T82 or interleukin-1 receptor-associated kinase 4 with the PDB ID 6O94, PDB ligand ID LRS). The corresponding query can be found in the Supplementary Data for upload to the GeoMine tool on the Proteins*Plus* for this structure. Interesting small molecules from the identified similar sites can be downloaded and subsequently be used for molecular docking with JAMDA. The figures on the right show the second highest-scoring predicted binding mode for ADP in the binding site of Nek6 and its 2D interaction visualization with PoseView (21).

stored in the PDB and restricted to common cofactors and ions with potentially functional roles. For researchers interested in the structural annotation of structures that have no known homologs in the PDB, the Proteins*Plus* web service comes in handy. It enables on-the-fly prediction of binding sites with DoGSiteScorer, retrieval of similar binding sites with SIENA, the identification of further potentially interesting ligands by user-defined GeoMine queries, and the molecular docking of these ligands into the AlphaFold model with JAMDA, see Figure 2.

### Ligand annotation for AlphaFold models

Given a protein of interest, e.g. the human protein kinase NIMA-related kinase 6 (Nek6), we can start our Proteins*Plus* investigations by providing its UniProt Accession Number Q9HC98 and entering the structural analysis mode of the web service. Next, we can predict potential binding sites using DoGSiteScorer. These predicted sites can be used to search for potential ligands with SIENA. By selecting, for example, the pocket named 'P_2' and performing a SIENA search for this predicted binding site, we can retrieve similar sites in complex with various ligands. Besides ADP (the annotation which was also found by AlphaFill), we find similar kinase binding sites in complex with further ligands, in this case, the inhibitor with the PDB ligand ID F9N in complex with Nek2 and Nek7. The active site sequence identity is 94%. The retrieved aligned complexes can be downloaded, together with the corresponding ligand SDF files. The results also enable the exploration of structural flexibility of similar binding sites that can be used, e.g. for the generation of other conformational states that are not covered in the AlphaFold database by homology modeling based on the identified structures.

The ligands retrieved from the SIENA run can either be transferred into the binding site based on the resulting alignment or using the on-the-fly docking tool JAMDA. It can be applied to find whether the found ligands from similar sites can be accommodated in the model's binding site. However, care should be taken regarding the model quality of the binding site residues as this can have a huge impact on the docking performance. Some preprocessing steps of the original AlphaFold structure might be necessary to obtain reliable ligand binding modes (22).

The search for similar binding sites using the Proteins*Plus*, however, is not restricted to binding sites with a high sequence identity. GeoMine can be applied to generate user-defined queries that search for geometric patterns of interacting binding site residues in nearly one million binding sites (predicted or ligand-annotated) in the PDB. For our example protein kinase, additional GeoMine queries result in the identification of further protein kinases in complex with inhibitors which can be used as idea generators for *in silico* drug design.

## SUMMARY AND OUTLOOK

The Proteins*Plus* web server offers a unique access point to protein structure and protein–ligand complex data processing on the worldwide web. Current developments with only conservative extensions of the user interface enable even broader access to molecular modeling tools which usually require comprehensive user knowledge. Furthermore, steady improvements and feature extensions based

on suggestions of users render it a lively and well-kept platform. To support users in getting started with the web server, we offer comprehensive documentation of the provided services (https://proteins.plus/help/index) and hands-on tutorials (https://proteins.plus/help/tutorial). As with all computational modeling approaches, the tools behind Proteins*Plus* have their limitations. All users are asked to consult the corresponding methods' publication for more details on the respective restrictions and application domains.

Besides the introduction of new features for GeoMine and the integration of the novel methods JAMDA and MicroMiner, we are in a constant process of elaborating the web server, its tool base, and its potential use cases. The first inclusion of AlphaFold structures in the web server opens new avenues for structural explorations that have not yet been fully explored. With numerous extensions in mind, including 2D and automated query generation in GeoMine or multiple mutations search in MicroMiner, we hope to create a steadily growing, easy-to-use modeling infrastructure for the life science community.

## DATA AVAILABILITY

Proteins*Plus* is a publicly available web-based protein structure analysis service, available at https://proteins.plus.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## FUNDING

Development of Proteins*Plus* was supported by de.NBI (in part); German Federal Ministry of Education and Research (BMBF) [031L0105 to K.S. and J.S.]; Development of GeoMine was supported by the German Federal Ministry of Education and Research as part of CompLS and de.NBI [031L0172, 031L0105 to K.D. and J.G.]; Development of MicroMiner was supported by the German Federal Ministry of Education and Research (BMBF) as part of protPSI [031B0405B to J.S.]; DASHH: Data Science in Hamburg - Helmholtz Graduate School for the Structure of Matter, funded by the Helmholtz Association [HIDSS-0002 to C.E.]; Center for Data and Computing in Natural Sciences (CDCS), funded by Authority for Science, Research and Equality of the Free and Hanseatic City of Hamburg (BWFGB) [LFF-HHX-03 to A.U.]. Funding for open access charge: Internal university funds.

*Conflict of interest statement.* Proteins*Plus* uses some methods that are jointly owned and/or licensed to BioSolveIT GmbH, Germany, M.R. is a shareholder of BioSolveIT GmbH.

## REFERENCES

1. Schöning-Stierand,K., Diedrich,K., Fährrolfes,R., Flachsenberg,F., Meyder,A., Nittinger,E., Steinegger,R. and Rarey,M. (2020) ProteinsPlus: interactive analysis of protein–ligand binding interfaces. *Nucleic Acids Res.*, **48**, W48–W53.
2. Fährrolfes,R., Bietz,S., Flachsenberg,F., Meyder,A., Nittinger,E., Otto,T., Volkamer,A. and Rarey,M. (2017) Proteins plus: a web portal for structure analysis of macromolecules. *Nucleic Acids Res.*, **45**, W337–W343.
3. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H. and Shindyalov,I.N. (2000) The protein data bank (www.rcsb.org). *Nucleic Acids Res.*, **28**, 235–242.
4. Jumper,J., Evans,R., Pritzel,A., Green,T., Figurnov,M., Ronneberger,O., Tunyasuvunakool,K., Bates,R., Žídek,A., Potapenko,A. *et al.* (2021) Highly accurate protein structure prediction with alphafold. *Nature*, **596**, 583–589.
5. Graef,J., Ehrt,C., Diedrich,K., Poppinga,M., Ritter,N. and Rarey,M. (2022) Searching geometric patterns in protein binding sites and their application to data mining in protein kinase structures. *J. Med. Chem.*, **65**, 1384–1395.
6. Diedrich,K., Graef,J., Schöning-Stierand,K. and Rarey,M. (2021) GeoMine: interactive pattern mining of protein–ligand interfaces in the protein data bank. *Bioinformatics*, **37**, 424–425.
7. Volkamer,A., Kuhn,D., Rippmann,F. and Rarey,M. (2012) DoGSiteScorer: a web server for automatic binding site prediction, analysis and druggability assessment. *Bioinformatics*, **28**, 2074–2075.
8. Bietz,S., Urbaczek,S., Schulz,B. and Rarey,M. (2014) Protoss: a holistic approach to predict tautomers and protonation states in protein–ligand complexes. *J. Cheminform.*, **6**, 12.
9. Lippert,T. and Rarey,M. (2009) Fast automated placement of polar hydrogen atoms in protein–ligand complexes. *J. Cheminform.*, **1**, 13.
10. Sommer,K., Friedrich,N.-O., Bietz,S., Hilbig,M., Inhester,T. and Rarey,M. (2016) UNICON: a powerful and Easy-to-Use compound library converter. *J. Chem. Inf. Model.*, **56**, 1105–1111.
11. Friedrich,N.-O., Flachsenberg,F., Meyder,A., Sommer,K., Kirchmair,J. and Rarey,M. (2019) Conformator: a novel method for the generation of conformer ensembles. *J. Chem. Inf. Model.*, **59**, 731–742.
12. Schlosser,J. and Rarey,M. (2009) Beyond the virtual screening paradigm: structure-based searching for new lead compounds. *J. Chem. Inf. Model.*, **49**, 800–809.
13. Henzler,A.M., Urbaczek,S., Hilbig,M. and Rarey,M. (2014) An integrated approach to knowledge-driven structure-based virtual screening. *J. Comput. Aided. Mol. Des.*, **28**, 927–939.
14. Flachsenberg,F., Meyder,A., Sommer,K., Penner,P. and Rarey,M. (2020) A consistent scheme for gradient-based optimization of protein–ligand poses. *J. Chem. Inf. Model.*, **60**, 6502–6522.
15. Flachsenberg,F. and Rarey,M. (2021) LSLOpt: an open-source implementation of the step-length controlled LSL-BFGS algorithm. *J. Comput. Chem.*, **42**, 1095–1100.
16. Volkamer,A., Griewel,A., Grombacher,T. and Rarey,M. (2010) Analyzing the topology of active sites: on the prediction of pockets and subpockets. *J. Chem. Inf. Model.*, **50**, 2041–2052.
17. Bietz,S. and Rarey,M. (2015) ASCONA: rapid detection and alignment of protein binding site conformations. *J. Chem. Inf. Model.*, **55**, 1747–1756.
18. Bietz,S. and Rarey,M. (2016) SIENA: efficient compilation of selective protein binding site ensembles. *J. Chem. Inf. Model.*, **56**, 248–259.
19. Perrakis,A. and Sixma,T.K. (2021) AI revolutions in biology. *EMBO Rep.*, **22**, e54046.
20. Hekkelman,M.L., de Vries,I., Joosten,R.P. and Perrakis,A. (2021) AlphaFill: enriching the alphafold models with ligands and co-factors. bioRxiv doi: https://doi.org/10.1101/2021.11.26.470110, 27 November 2021, preprint: not peer reviewed.
21. Stierand,K., Maass,P.C. and Rarey,M. (2006) Molecular complexes at a glance: automated generation of two-dimensional complex diagrams. *Bioinformatics*, **22**, 1710–1716.
22. Skolnick,J., Gao,M., Zhou,H. and Singh,S. (2021) AlphaFold 2: why it works and its implications for understanding the relationships of protein sequence, structure, and function. *J. Chem. Inf. Model.*, **61**, 4827–4831.

# Eidesstattliche Versicherung

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

Hamburg, den November 29, 2024

---

Konrad Diedrich