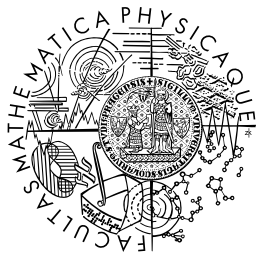

FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

DOCTORAL THESIS

Mgr. Radoslav Krivák

Prediction of ligand binding sites from protein structure

Department of Software Engineering

Supervisor of the doctoral thesis: doc. RNDr. David Hoksza, Ph.D.

Study programme: Computer Science

Specialization: Software Systems

Prague 2023

I declare that I carried out this doctoral thesis independently, and only with the cited sources, literature and other professional sources.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In ..... date ....

Author's signature

First of all, I would like to thank David Hoksza for supervising my work, for providing constant support and inspiration and ultimately for introducing me to the exciting field of structural bioinformatics. I am also thankful to all my friends, collaborators and colleagues that helped in any way to improve my work, either by direct help or by providing helpful suggestions and inspiration.

Dedicated to my grandparents: Brigita, Sabína, Andrej and František.

Title: Prediction of ligand binding sites from protein structure

Author: Mgr. Radoslav Krivák

Department: Department of Software Engineering

Supervisor: doc. RNDr. David Hoksza, Ph.D., Department of Software Engineering

Abstract:

Ligand binding site prediction from protein structure is a fundamental problem in the field of structural bioinformatics that has many applications related to the elucidation of protein function and structure-based drug discovery. The first focus of this thesis was the application of machine learning to this and related problems. The second focus was the development of practically usable tools based on our research. The machine learning based tools produced as a result of the work on this thesis include the pocket re-scoring method PRANK, a stand-alone ligand binding site prediction method P2Rank (together with its extended web interface PrankWeb) and the peptide binding prediction method P2Rank-Pept. We have shown that our methods outperformed available state-of-the-art tools while providing other benefits like prediction speed and stability. Furthermore, we have developed AHoJ, a flexible tool for the search and alignment of Apo-Holo protein pairs in the PDB. AHoJ that is ideal for creating Apo-Holo datasets which can in turn help to better evaluate binding site prediction methods in the future.

Keywords: Structural Bioinformatics, Protein-ligand binding sites, Machine learning

Contents

Part I

Commentary

Chapter 1

Introduction

1.1 Structure of the thesis

The thesis is structured in the following way. Part I: Commentary summarizes my work and contribution and puts it in the context while Part II: Publications contains the full text of ten peer-reviewed co-authored publications that constitute the core of the contribution and were published during my PhD study.

In Part II, some of the publications are introduced by a concise "Author's highlights". These are not necessarily summaries of the articles and are not meant to replace abstracts but rather highlight points that might be relevant to the readers of the thesis.

Instead of trying to be just a summary of included publications, the text of the thesis is intended to be an accompanying commentary to my work as a whole with some added value. The thesis contains some of my personal opinions and experiences and better explains the motivation behind some efforts and decisions. This includes some points that did not find a way to the original publications or could be only said with hindsight. Furthermore, while the included papers describe the software as it was when it was initially released, this thesis describes the software as it is now, with all accumulated improvements and changes.

1.2 Binding site prediction and related problems

Ligand binding site prediction is a fundamental problem in the field of computational biology that seeks to identify the location and shape of binding sites on protein structures that can interact with small molecules. This section contains a concise introduction to the problem and its context. The main goal is, however, to highlight inherent complications with the problem definition and bring up considerations that shaped the work presented in this thesis.

1.2.1 Motivation

Prediction of ligand binding sites from protein structure has many applications in elucidation of protein function [KJ14] and rational drug design [ZGWW12, PSM*10, TBNT16]. It has been employed in drug side-effects prediction [XXB11], fragment-based drug discovery [LEG16], docking prioritization [LJ06, FB15], structure based virtual screening [LSCZ14] and structure-based target prediction (or so called inverse virtual screening) [SBB*14]. Increasingly it is being used in large-scale structural studies that try to analyze and compare all known and putative binding sites on a genome-wide or PDB-wide level [DWH15, MBB16, MZF*17, SCS*17, BSSC18].

In practice, it is often the case that predicting ligand binding sites is not an end in itself but it represents only a step in a larger automated solution or pipeline. For instance, a druggability prediction server PockDrug-Server [HBG*15] relies on ligand binding site prediction internally. Similarly, allosteric site prediction tools Allosite [HLH*13] and AlloPred [HLH*13] both internally employ a ligand binding site prediction tool Fpocket [LGST09] as the first step of their workflows.

1.2.2 The problem statement

The problem of ligand binding site prediction from protein structure can be defined in the following way: given a protein structure, produce a list of putative binding sites and score/order them according to the likelihood of binding relevant ligands.

This definition is rather technical but still leads to several questions:

How can/should be predicted binding sites represented? It turns out that in whatever way possible and that the existing methods represent binding sites in various ways, which include but are not limited to: a set of protein surface atoms, a set of residues or a set of points around the surface of the protein (points on a regular 3D grid, alpha sphere centers or points on protein's solvent accessible surface). To evaluate a prediction method we need a binding site to be represented at least as a single point, i.e. center/centroid of a binding site.

Why it is important to score/order predicted binding sites? To meaningfully evaluate prediction methods and to determine their identification success rate it is necessary to consider only predicted sites with the highest score (e.g. Top-1/Top3 or better Top-n/Top-(n+2) where n is the number of known ligands on a given protein). If we were to consider all predicted pockets, an obviously useless method that would cover the whole surface of the protein with predicted binding sites would achieve 100% success rate.

Which types of ligands are relevant? This is often only implicitly defined by the datasets on which are particular methods trained and/or benchmarked. For a detailed discussion see Supplementary Materials to [KH18].

1.2.3 Related problems

Proteins can interact with a variety of binding partners: small molecule ligands, ions, peptides, other proteins and nucleic acids. For each type of binding partner, we can consider the problem of predicting its binding locations. Developing a prediction method for each of those molecular types presents distinct challenges and also offers specific clues that can be best utilized by specialized methods.

In contrast with the task of binding site prediction, there is a closely related task of binding residue prediction. Although the difference may seem only technical, it is important to distinguish between the two. The task of binding site prediction involves the prediction of binding sites as such, i.e. a binding site is considered an entity which shape and location (represented at least as a center point) needs to be determined. On the other hand, the task of binding residue prediction can be viewed as a task of labeling residues by a binary label (binding vs. non-binding), or by a binding probability score from the range of $[0, 1]$. One way to look at is that the task of binding residue prediction does not include the final step of clustering binding residues into binding sites.

An important variation of the problem is predicting binding residues from the sequence alone.

1.2.4 Existing methods and tools

Ligand binding site prediction methods have been in development for almost 40 years now (the first known method, to my knowledge, was published in 1985). During this time more than 50 different algorithms or improvements have been published.

Existing methods for ligand binding site prediction are based on a variety of algorithmic approaches. Traditionally, methods have been categorized based on their main algorithmic strategy into geometric, energetic, conservation based, template based, knowledge based and machine learning based. In reality, many of the existing tools are based on some combination of the mentioned approaches. Methods based on a consensus of results of other algorithms have also emerged.

More details on existing methods and tools can be found in numerous reviews and surveys [LJ06, HOH*10, PSM*10, LSZ10, CMGK11, FRH11, RBJ15, BS17, SLD*]. In the introduction to the paper [KH18] I have provided another comprehensive survey of existing tools with a focus on their practical usability. In it I have highlighted the importance of the categorization of the tools along several lines: template based / template-free methods, web servers / stand-alone tools, and residue-centric / pocket-centric methods and I have argued that there is a strong case for a new fast stand-alone user-friendly and template-free tool.

Studies that introduced existing methods reported relatively high prediction accuracy, usually on traditional small datasets. However, the results of the only independent benchmark [CMGK11] suggested that existing methods may not be as accurate as previously believed when applied to new datasets.

When I started working on the problem at the first sight it might seem that the field is crowded with tools available for researchers. However, after a closer survey [KH18] I found that only a few of the published methods were available as a stand-alone software that can be used locally (in contrast with web-based methods). Furthermore, most of those stand-alone tools were unnecessarily complicated to use (users were required to perform preprocessing tasks that could have been automated by the authors of the software). Even fewer of the tools were available as open-source software.

1.3 Goals

There is no reason to pretend that the work presented in this thesis was a liner process of first setting some fixed set of goals and then gradually accomplishing them. Indeed, what is included in the thesis is mostly only the work that led to in some way successful results. With that in mind, the following list is included here mainly to clarify my intentions and motivations and highlight the issues of existing tools I decided to focus on improving.

- Explore the possibility of improving existing ligand binding site prediction methods by replacing their scoring function.
- Develop a stand-alone ligand binding site prediction method based on machine learning. Although machine learning has been applied to the problem before and some studies have been published, their focus was on predicting binding residues rather than on predicting binding sites as such [KK09, QW00, CHG14].
- Produce command line tools that can be used locally and are easy to set up and use and therefore can be easily employed in larger bioinformatics pipelines.
- Produce intuitive web based tools with integrated visualizations that have documented REST APIs.
- Work towards a better evaluation of ligand binding site prediction methods on Apo-Holo datasets.

Chapter 2

Overview of the contribution

2.1 List of Publications

The following peer-reviewed publications and associated structural bioinformatics software constitute the core contribution presented in this thesis. Full texts of these publications (except [con19, con21]) including relevant supplementary materials are included in Part II.

[KH15a] KRIVÁK R., HOKSZA D.: **Improving protein-ligand binding site prediction accuracy by classification of inner pocket points using local features**. *Journal of Cheminformatics 7*, 1 (Apr 2015), 12. doi:10.1186/s13321-015-0059-5

[KH15b] KRIVÁK R., HOKSZA D.: **P2RANK: Knowledge-Based Ligand Binding Site Prediction Using Aggregated Local Features**. In *International Conference on Algorithms for Computational Biology* (2015), Springer, pp. 41–52. doi:10.1007/978-3-319-21233-3_4

[KH18] KRIVÁK R., HOKSZA D.: **P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure**. *Journal of cheminformatics 10*, 1 (2018), 39. doi:10.1186/s13321-018-0285-8

[KH7] KRIVÁK R., HOKSZA D., ŠKODA P.: **Improving quality of ligand-binding site prediction with Bayesian optimization**. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (2017), pp. 2278–2279. doi:10.1109/BIBM.2017.8218024

[KJH18] KRIVÁK R., JENDELE L., HOKSZA D.: **Peptide-Binding Site Prediction From Protein Structure via Points on the Solvent Accessible Surface**. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics* (New York, NY, USA, 2018), BCB '18, Association for Computing Machinery, p. 645–650. doi:10.1145/3233547.3233708

[JKS*19] JENDELE L., KRIVAK R., SKODA P., NOVOTNY M., HOKSZA D.: **PrankWeb: a web server for ligand binding site prediction and visualization**. *Nucleic Acids Res. 47*, W1 (Jul 2019), W345–W349. doi:10.1093/nar/gkz424

[JSK*22] JAKUBEC D., SKODA P., KRIVAK R., NOVOTNY M., HOKSZA D.: **PrankWeb 3: accelerated ligand-binding site predictions for experimental and modelled protein structures**. *Nucleic Acids Research 50*, W1 (05 2022), W593–W597. doi:10.1093/nar/gkac389

[con19] CONSORTIUM P.-K.: **PDBe-KB: a community-driven resource for structural and functional annotations**. *Nucleic Acids Research 48*, D1 (10 2019), D344–D353. doi:10.1093/nar/gkz853

[con21] CONSORTIUM P.-K.: **PDBe-KB: collaboratively defining the biological context of structural data**. *Nucleic Acids Research 50*, D1 (11 2021), D534–D542. doi:10.1093/nar/gkab988

[FKHN22] FEIDAKIS C. P., KRIVAK R., HOKSZA D., NOVOTNY M.: **AHoJ: rapid, tailored search and retrieval of apo and holo protein structures for user-defined ligands**. *Bioinformatics 38*, 24 (10 2022), 5452–5453. doi:10.1093/bioinformatics/btac701

2.1.1 Autorship notes

In the publications where I am the first author [KH15a, KH15b, KH18, KH7, KJH18] I have contributed most of the research ideas and software development, performed the experiments and I have written most of the text of the manuscripts (all under the supervision and with consultation with my supervisor David Hoksza following his initial ideas about an aggregated representation of protein physico-chemical features).

In the publications related to the web interface (PrankWeb) [JKS*19, JSK*22] I have contributed some of the development, performed the experi-

ments [JKS*19] or helped with their design [JSK*22] and written parts of the manuscript [JKS*19].

Publications related to PDB-KB [con19, con21] were written by a consortium of authors and P2Rank is only one of the tools integrated with PDB-KB. I have helped to develop data transformation of P2Rank output to PDB-KB input format, contributed to the validator of PDB-KB input data and performed predictions on all proteins in the PDB.

In [FKHN22] I have developed the web interface and contributed to the development of the command line version of the software.

2.2 Summary of the contribution

This section summarizes the most important contributions of the work presented in this thesis. Most of the work was produced in cooperation with co-authors of respective publications.

A list of released bioinformatics software and practical/usable contributions follows.

1. We have developed PRANK, a machine learning based method that allows to re-score (re-rank) ligand binding sites predicted produced by other methods. Since it helps true binding sites to be ranked higher, it improves the applicability and usefulness of their predictions. PRANK is useful especially in combination with methods like Fpocket, which produce a large amount of predicted binding sites for each protein but do not always score true binding sites at the top. PRANK was made available as a free command line tool with source code available upon request. Later it became part of the P2Rank codebase and was released as open-source software.

2. We have developed P2Rank, a fully independent method for ligand binding site prediction based on machine learning. Although some machine learning based methods for a given problem were described in the literature before, to our knowledge P2Rank was the first pragmatically usable tool for ligand binding site prediction based on machine learning. P2Rank makes predictions by scoring and clustering points on the protein's solvent accessible surface. The ligandability score of individual points is determined by a Random Forest model trained on the dataset of known protein-ligand complexes. P2Rank

is released as open-source software (under MIT license) on GitHub (https://github.com/rdk/p2rank).

3. We have developed PrankWeb, a web application interface for P2Rank [JKS*19]. In addition to a standalone version of P2Rank, PrankWeb employs a custom-made conservation pipeline and improved prediction models trained using conservation as one of the features (i.e. descriptors). Unlike many similar tools at the time of the release, PrankWeb came with a documented REST API. The later version introduced the support for mmCIF format and prediction model specialized for AlphaFold structures [JSK*22]. PrankWeb is freely available at https://prankweb.cz/ and open-sourced (under Apache License 2.0) on GitHub (https://github.com/cusbg/prankweb).

4. We have integrated P2Rank/PrankWeb with EBI's Protein Data Bank in Europe – Knowledge Base (PDBe-KB), the new PDBe's major resource of integrated protein data [con19, con21]. PDB-KB now contains annotations based on P2Rank predictions precomputed for almost every protein in the PDB and it is being periodically updated with predictions on new proteins. PDBe-KB is available at https://pdbe-kb.org.

5. We have developed AHoJ, a highly-configurable tool for the search and alignment of Apo-Holo protein pairs in the PDB [FKHN22]. AHoJ is available as an open-source command line program and a web application that allows running searches for multiple queries at the same time (and thus produce Apo-Holo datasets) and includes integrated web-based visualization. The web application is freely available at http://apoholo.cz/ and the command line tool is open-sourced (under Apache License 2.0) on GitHub (https://github.com/cusbg/AHoJ-project).

6. I have developed FasterForest, a Java library that contains two highly optimized Random Forest implementations. These implementations represent mainly technical optimizations of the previous original open-source work [Sup13, Sel17] and require roughly 75% time and 50% space compared to the original implementations. The library was used during the development and optimization of our later methods [KJH18, JKS*19]. FasterForest library is available as open source under GNU GPL v2 (https://github.com/rdk/FasterForest).

The following list summarizes my research contributions, i.e. theoretically interesting results or novel contributions to the discussion in the field of binding site prediction.

1. P2Rank was the first machine learning based method related to a protein structure that internally used points on the solvent accessible surface of the protein instead of a typical approach of using points on a regular 3D grid.

2. In publication [KH18] I introduced some points that I believe were missing from the discussion in the field. These include the following: running times (i.e. speed) of prediction methods, we highlighted the difference between pocket-centric and residue-centric methods and respective evaluation methodologies, and included a discussion of the possibility of reaching Bayes optimal rate on inherently noisy datasets.

3. During the development of the prediction methods, I used the technique of Bayesian optimization [BCdF09] that allowed me to optimize several arbitrary parameters simultaneously.

4. We have developed and published the results of P2Rank-Pept, a method specialized for the prediction of peptide binding sites from protein structure. This demonstrated the applicability of our general approach to different related tasks P2Rank-Pept is a part of the P2Rank codebase, but up to this date it has not been released with a pre-trained model.

Chapter 3

Tools for ligand binding site prediction

3.1 PRANK: replacing the scoring function of existing methods

Most of the existing ligand binding site prediction methods find much more pockets on a given structure than there are actual true binding sites. At the same time, they employ a fairly simple ranking function leading to sub-optimal prediction results[1].

To address this problem, we introduced a novel machine learning-based pocket ranking algorithm called PRANK (Pocket RANKing) that can be used post-processing step which improves the performance of existing ligand binding site prediction methods. The outline of the algorithm is shown in Figure 3.1 and further described in Figure 3.2 which shows an internal pocket representation used by PRANK. A detailed description of the algorithm can be found in [KH15a].

Our benchmarks showed that our new scoring function considerably outperformed the native scoring functions of Fpocket [LGST09] and ConCavity [CLT*09] on all evaluated datasets. Furthermore, we showed that it outperformed two simpler scoring functions: PLB index, which is based on amino acid composition [SSKH07] and a simple ordering by pocket volume. Improvements in the prediction success rate achieved by PRANK

[1] measured as binding site prediction success rate considering Top-k predicted pockets with the highest score

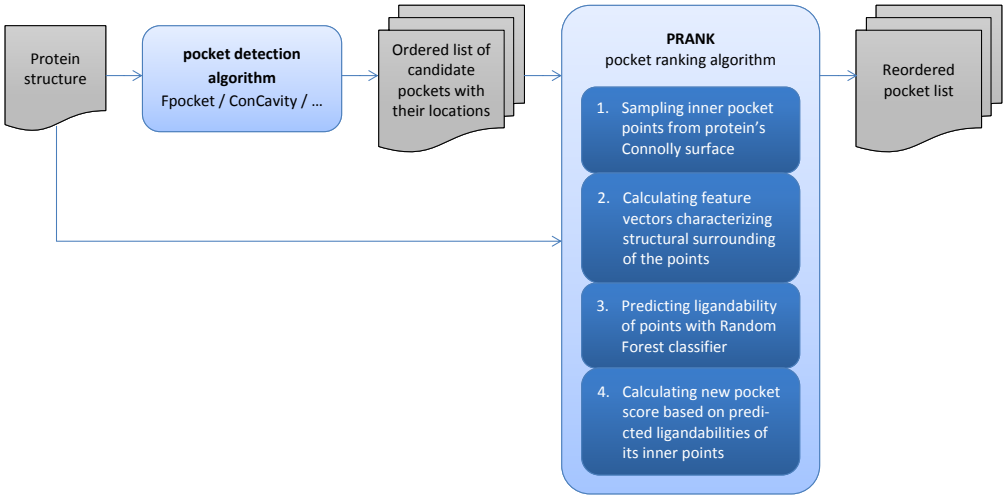

FIGURE 3.1: **Flowchart that outlines PRANK algorithm.**

when applied to Fpocket predictions can be seen in Figure 3.3.

PRANK takes a protein structure and the output of a third-party prediction method on the input and produces a list of re-scored and re-ranked pockets on the output. PRANK can currently process the output of the following methods: Fpocket , ConCavity , SiteHound [GS09], MetaPocket 2.0 [ZLL*11], LISE [XH12] and DeepSite [JDMR*17]. Furthermore, a clean internal API allows parsers for new methods to be easily implemented.

PRANK was originally developed and distributed as a set of scripts written in Groovy programming language and later integrated into the codebase and distribution of P2Rank as a standalone command line application running on Java Virtual Machine.

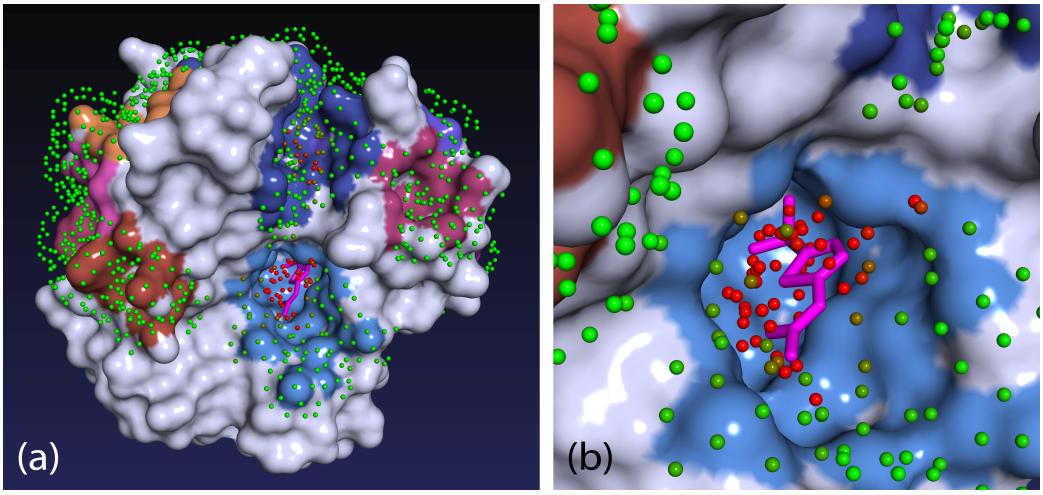

FIGURE 3.2: **PRANK: Visualization of inner pocket points. (a)** Displayed is the protein 1AZM bound to one ligand (magenta). Fpocket predicted 13 pockets that are depicted as colored areas on the protein surface. To rank these pockets, the protein was first covered with evenly spaced points on a solvent accessible surface (probe radius 1.6 Å) and only the points adjacent to one of the pockets were retained. The colour of the points reflects their ligandability (green = 0...red = 0.7) predicted by Random Forest classifier. PRANK algorithm rescores pockets according to the cumulative ligandability of their corresponding points (calculated as a sum of squares). Note that there are two clusters of ligandable (red) points in the picture, one located in the upper dark-blue pocket and the other in the light-blue pocket in the middle. The light-blue pocket, which is, in fact, the true binding site, contains more strongly ligandable points and therefore will be ranked higher. **(b)** Detailed view of the binding site with the ligand and the inner pocket points.

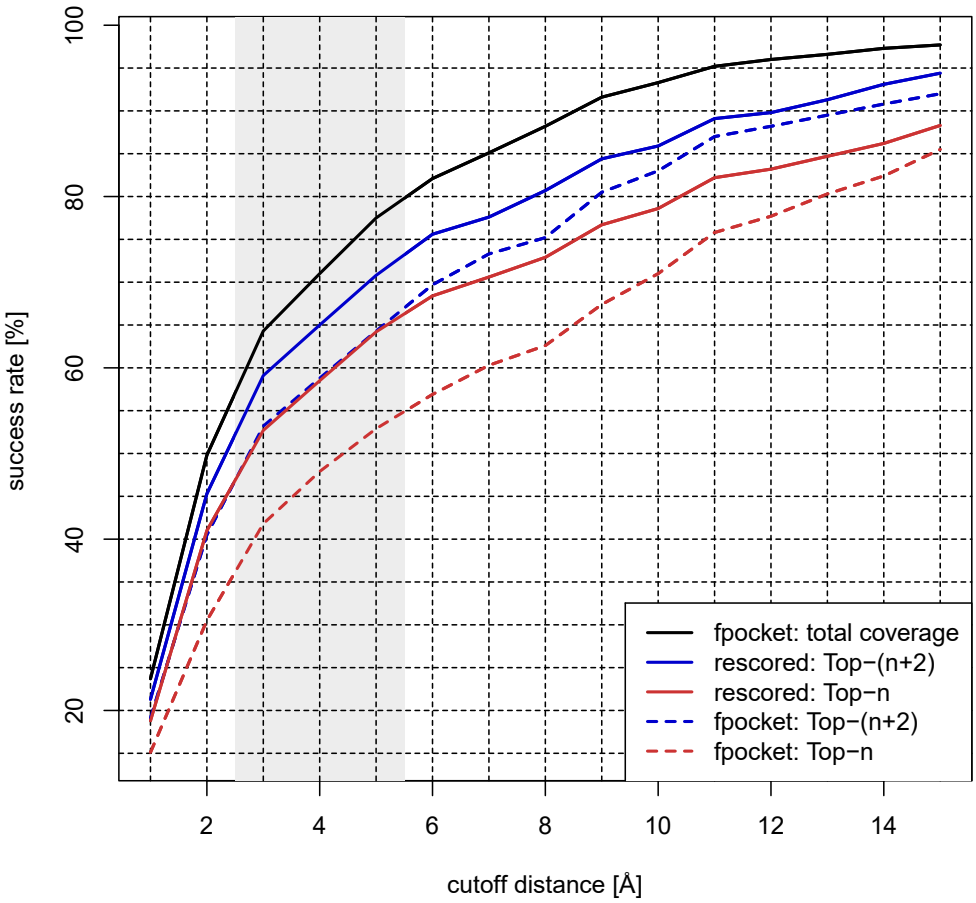

FIGURE 3.3: **PRANK: Results of rescoring Fpocket predictions on CHEN11 dataset.** Chart showing prediction success rates of Fpocket compared with results rescored by PRANK on CHEN11 dataset considering Top-n, Top-(n+2) and all pockets (total coverage). The success rate is measured by D_{CA} criterion for the range of integer cutoff distances (i.e. distance between the center of a predicted pocket and any atom of the ligand). Displayed results for rescored pockets are averaged from ten independent 5-fold cross-validation runs.

3.2 P2Rank: machine learning based method

Building on PRANK we have developed P2Rank a stand-alone independent ligand binding site prediction method. We have realized that relying on third-party methods for making predictions and then rescoring them is actually limiting and that our machine learning based approach can predict that the other methods are not able to identify at all. Compared to PRANK, P2Rank is looking at the whole surface of the protein. It covers it with points on a solvent accessible surface, predicts their ligandability and then clusters points with high ligandability into predicted binding sites. The working of the algorithm is illustrated in Figure 3.4 which shows an entire surface of the protein covered with points with predicted ligandability. A detailed description of the algorithm can be found in [KH18].

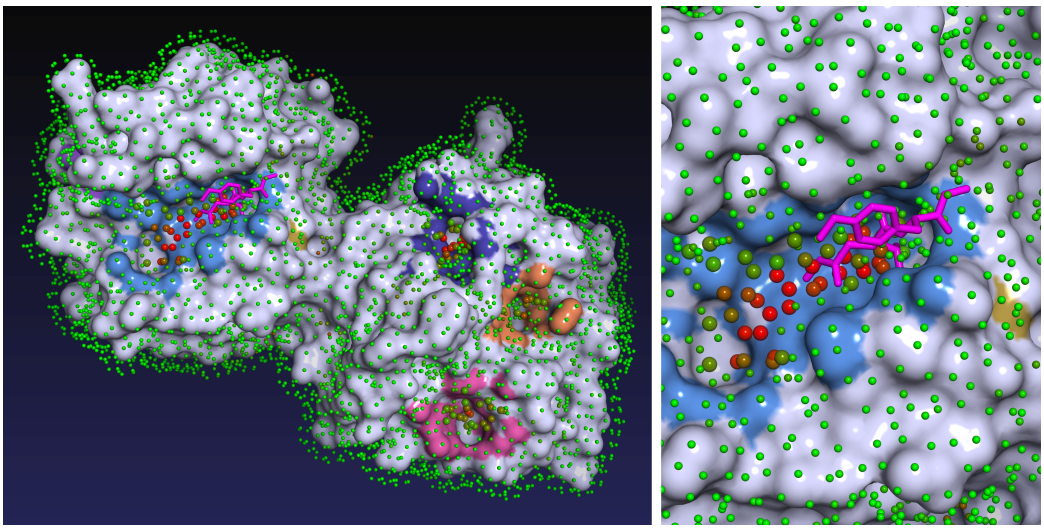

FIGURE 3.4: **P2Rank: Visualization of ligand binding sites predicted by for structure 1FBL.** Protein is covered by a layer of points lying on the Solvent Accessible Surface of the protein. Each point represents its local chemical neighborhood and is colored according to its predicted ligandability score (from 0=green to 1=red). Points with high ligandablity score are clustered to form predicted binding sites (marked by coloring adjacent protein surface). In this case, the largest predicted pocket (shown in the close-up) is indeed a correctly predicted true binding site that binds a known ligand (magenta). Visualization is based on a PyMOL script produced by P2Rank.

3.2.1 Features

This section contains a summary of the features and characteristics of the software from the point of view of a user and from the point of view of a new model/method developer. The current version of the software is described (P2Rank 2.4).

User facing features

- Ease of setup. P2Rank is distributed as a precompiled binary package with pre-trained prediction models that requires no compilation or installation. P2Rank does not depend on any third-party bioinformatics software and the only dependency is Java Virtual Machine.
- Ease of use. Given any protein structure, P2Rank is able to produce prediction by running a single command (i.e. no preprocessing steps or multiple-step procedures are needed). This is still quite rare among available methods.
- High prediction accuracy, especially when compared to methods that are comparably fast.
- PyMol visualisation. P2Rank optionally produces PyMol visualizations such as the one that can be seen in Figure 3.4.
- Optimized multi-threaded implementation. P2Rank is only one of two methods that need under one second to generate a prediction on a single protein of average size [KH18].
- Support for both PDB and mmCIF formats. P2Rank is one of the few existing ligand binding site prediction methods that are currently able to process mmCIF format and produce predictions on proteins of unlimited size as well as on AlphaFold models.
- Stability. Great care has been taken so that P2Rank finishes successfully (without crashing) on any valid PDB or mmCif input that contains protein structure. It is admittedly a moving target. P2Rank has been therefore evaluated by running it on the whole PDB and is regularly automatically run on new PDB entries. This stands in contrast with many available tools, some of which have a failure rate that can be as high as 20-80% (see supplementary materials to [KH18]).

- Interpretability. For each pocket and each residue, P2Rank produces a probability score, which is a number from the $[0, 1]$. Transformations from raw scores to probability scores are trained/fitted for each prediction model on a calibration dataset.

Features related to training new models and development of new methods

P2Rank can be also seen as a framework and a workbench for training new prediction models and developing new prediction methods. The following list summarizes the features that are relevant for advanced users/developers that want to do one of the following: train new models on specific datasets, develop methods for new prediction tasks, or develop new local prtoein descriptors and compare their contribution to predictive performance.

- Java API for predictions. P2Rank can be used as a library by the programs running on JVM.
- Training and evaluation of new models. P2Rank is able to train and evaluate new models on different dataset running single command.
- Configurability. P2Rank has more than 100 documented configurable parameters. Configuration can be stored in a config file and overridden in the command line.
- Different evaluation modes and metrics. P2Rank implements pocket-centric and also residue-centric evaluation and within them calculates various prediction performance metrics.
- Grid optimization with visualization. P2Rank implements an internal optimization loop for grid optimization based on a list of parameter values. If only one or two parameters are optimized at the same time P2Rank can produce bar charts or heatmaps for every calculated metric.
- Integration with external optimizers. P2Rank implements an internal optimization loop that can make use of third-party optimizers. Two optimizers that implement Bayesian optimization are currently integrated [SLA12, JG17].
- Easy development of new features/descriptors. P2Rank contains a clean internal API for the development of new features. New features

can be calculated either for protein atoms or residues (those are then projected onto solvent accessible surface points) or for solvent accessible surface points directly, depending on what comes most naturally.

- Ability to use externally calculated features/descriptors via CSV files which contain features calculated for every residue in the dataset.

3.2.2 Results

Results in Table 3.1 show that P2Rank clearly outperforms other evaluated tools in Top-n and Top-(n+2) categories on two datasets. P2Rank also achieves higher success rates than were possible to achieve just by re-scoring predictions of Fpocket using PRANK algorithm. Still, Fpocket+PRANK performed better than any of the other tools except for P2Rank. We have also evaluated the performance of a reduced version of P2Rank that uses only a single geometric feature (descriptor): protrusion. Surprisingly, even this simplified, purely geometric version of P2Rank slightly outperforms other tools in most cases (except for MetaPocket 2.0 in Top-(n+2) category).

TABLE 3.1: **P2Rank: Comparison of predictive performance on COACH420 and HOLO4K datasets.**

	COACH420		HOLO4K	
	Top-n	Top-(n+2)	Top-n	Top-(n+2)
Fpocket	56.4	68.9	52.4	63.1
Fpocket+PRANK[a]	63.6	76.5	62.0	71.0
SiteHound[†]	53.0	69.3	50.1	62.1
MetaPocket 2.0[†]	63.4	74.6	57.9	68.6
DeepSite[†]	56.4	63.4	45.6	48.2
P2Rank[protrusion][b]	64.2	73.0	59.3	67.7
P2Rank	**72.0**	**78.3**	**68.6**	**74.0**

The numbers represent identification success rate [%] measured by D_{CA} criterion (distance from pocket center to closest ligand atom) with 4 Å threshold considering only pockets ranked at the top of the list (n is the number of ligands in considered structure).
[†]These methods failed to produce predictions for some portion of input proteins. Here are displayed success rates calculated only based on subsets of proteins, on which they finished successfully. Detailed, pairwise comparison with P2Rank on the exact subsets can be found in the Supplementary Information of [KH18].
[a]predictions of Fpocket re-scored by PRANK algorithm
[b]reduced version of P2Rank that uses only single geometric feature: protrusion

3.3 PrankWeb: more than a web interface for P2Rank

We have developed PrankWeb, a web application for the prediction of ligand binding sites [JKS*19]. While PrankWeb uses P2Rank in the backend, it is not just a simple web interface for P2Rank. It additionally employs a custom-made conservation pipeline and improved prediction models trained using conservation as one of the features (i.e. descriptors). The new version [JSK*22] introduced the support for mmCIF format and prediction model specialized for AlphaFold structures [TAW*21].

Note: the pre-trained models that use conservation are included in the standalone command line distribution of P2Rank, but the conservation pipeline is not. To use these models in command line mode users can make use of PrankWeb's docker images.

3.3.1 Features

- PrankWeb is able to predict binding sites on experimental structures (PDB), AlphaFold models or any valid sructure uploaed by the user.
- Conservation pipeline. PrankWeb can calculate sequence conservation scores and employ this information in binding site prediction.
- Customizable web-based visualization of prediction results that integrates sequence and structural visualization. Visualization includes conservation score and AlphaFold score (pLDDT) if available.
- Precomputed predictions. We have computed the ligand binding site predictions for two components of the AlphaFold DB, the "model organism proteomes" and "Swiss-Prot", as well as for the whole PDB. For each database, AlphaFold DB and PDB, we computed the prediction with and without using conservation. Results precomputed for PDB are being automatically periodically updated by running predictions with the structures newly added to PDB. PrankWeb can serve the predictions on those structures to users instantaneously via its web interface. Moreover, precomputed predictions on individual databases are available for bulk download on PrankWeb's website.
- Documented REST API.

3.3.2 Results

Table 3.2 presents the evaluation of all new P2Rank models used for PrankWeb 3, as well as their comparison with the former models user by the original version or PrankWeb. It can be seen that the new Default models exceed the performance of the corresponding old models when evaluated on the representative HOLO4K dataset.

3.3.3 Implementation details

The original version of PrankWeb [JKS*19] was developed as a Java web application that was using P2Rank internally as a library via P2Rank's Java API. The advantage of this approach was that it avoided repeated JVM and model loading cost on each prediction run (which is measured in order of seconds).

TABLE 3.2: **PrankWeb: Results of four prediction models employed by PrankWeb 3** and comparison with two previously used models

	COACH420		HOLO4K	
	Top-n	Top-(n+2)	Top-n	Top-(n+2)
Default (old)	72.0	78.3	68.6	74.0
Default + conservation (old)	73.2	77.9	72.1	76.7
Default	71.6	76.8	72.7	78.0
Default + conservation	74.3	77.2	74.5	78.4
B-factor-free	71.2	77.5	72.1	77.2
B-factor-free + conservation	74.9	78.5	73.9	77.7

The numbers represent identification success rates (in %) measured using the DCA criterion utilizing a 4.0 Å threshold for the distance between the center of the predicted LBS and any ligand atom; only the n or (n+2), respectively, top-ranking predicted sites are considered in the evaluation, where n is the number of ligands in the respective 3D structure. Values for Default (old) and Default + conservation (old) represent results of old models used by the original version of PrankWeb. B-factor-free are used with AlphaFold predictions which utilize the B-factor field for confidence scores. Please note that old models were generated by the older version of P2Rank, which used older versions of BioJava and CDK. Using newer versions changed how certain PDB files are parsed, and an upgrade of the CDK library fixed a bug in the algorithm that generates SAS points. This, together with bug fixes in P2Rank itself, causes the scores for the Default (old) and Default models to differ.

With the new release, PrankWeb's architecture has been completely redesigned [JSK*22]. PrankWeb is now developed as a modern Python web application with modular architecture that strictly separates web-based user interface, data storage, and an execution component. Each component corresponds to a Docker image. Combined with docker-compose, it is easy to deploy and update PrankWeb instances, or using just the execution component run predictions on private data without exposing them to third-party servers. Each new prediction is now executed as a separate P2Rank process. This brings higher flexibility but also brings back JVM and model loading cost. This fact is now offset by faster startup times on newer JVMs and by the fact that predictions for many available structures are automatically precomputed by PrankWeb.

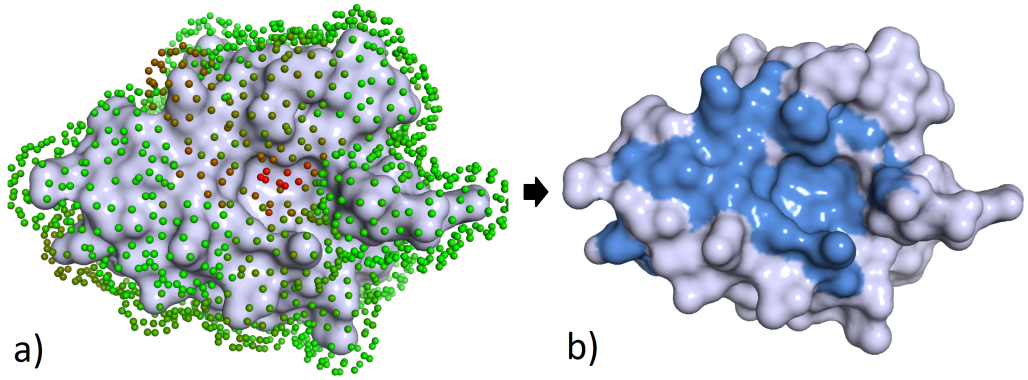

FIGURE 3.5: **Peptide-binding residue prediction based on points on the Solvent Accessible Surface**. **a)** Protein (3NFK/A) is covered in a layer of points lying on the solvent accessible surface. Each point represents its local chemical neighborhood and is described by a feature vector calculated from its surroundings. Points are colored according to the peptide-binding score ($\in$ [0,1]) predicted by a Random Forest classifier (*green=0/red=1*). **b)** Peptide-binding score of any given solvent exposed residue is based on the score of its adjacent points (radius of the cutoff and the form of aggregation function were subject to optimization). Residues with the score above a certain threshold are labeled as predicted positives (*blue*).

3.4 P2Rank-Pept: prediction of peptide binding sites

We have applied our approach to the task of peptide binding site prediction. Compared to P2Rank we had to develop and employ a variety of new features to achieve top performance. Among them were features related to protein geometry, secondary structure and sequence conservation. Figure 3.6 shows the outline of the algorithm i.e. the steps that P2Rank-Pept follows to predict peptide-binding residues using previously trained classification model. Prediction on a particular protein is further illustrated in Figure 3.5. P2Rank-Pept is a part of the P2Rank codebase, but up to this date it has not been released with a pre-trained model. Although we achieved predictive performance that was significantly higher than the competition, I was not convinced that the method is practically useful in its current state.

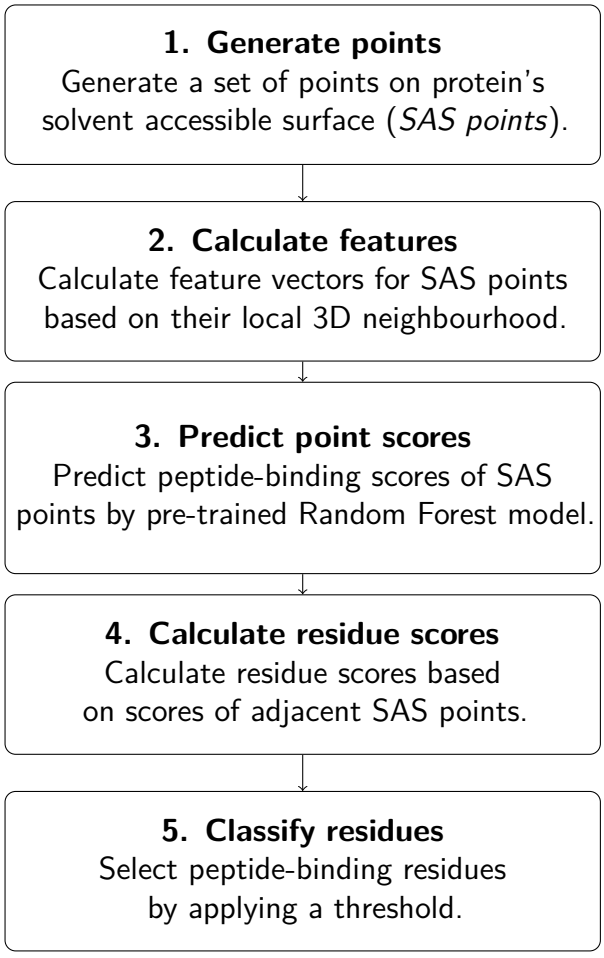

FIGURE 3.6: **P2Rank-Pept algorithm outline**

3.5 Integration with PDB-KB

We have integrated P2Rank/PrankWeb with EBI's Protein Data Bank in Europe – Knowledge Base (PDBe-KB), the new PDBe's major resource of integrated protein data [con19, con21]. PDB-KB now contains annotations based on P2Rank predictions precomputed for almost every protein in the PDB and it is being periodically updated with predictions on new proteins.

Chapter 4

Apo-Holo protein search

4.1 Introduction

Ligand-binding proteins exist in a bound (Holo) and an unbound (Apo) state. Structurally those states are almost always, to some extent, conformationally different due to the binding-induced conformational changes. For many proteins, both of these states can be found in the PDB, often in multiple entries.

This picture gets further complicated when we consider proteins that can bind multiple ligands on multiple binding sites (which is probably a majority of ligand-binding proteins). One particular protein with two binding sites can thus exist in a few different versions in the PDB: not binding any ligand, binding a ligand in one of the binding sites but not in the other, and binding ligands in both sites. The generally accepted definition is that a protein in the Apo state does not bind any ligands at all and Holo state covers the situations where it binds one or multiple ligands. However, when we talk about Apo-Holo protein pairs and their search, it is more useful to think about a pair of Apo-Holo structures with respect to: (a) a specific binding site, (b) a set of specific binding sites, (c) all known binding sites.

The Apo-Holo protein pairing is not readily available in the PDB and the consideration about multiple binding sites just illustrates one of the reasons. The process of Apo-Holo pairing is further complicated by sequence irregularities in the PDB, a consideration of which type of molecules should be considered as relevant ligands and a specific way how the binding site occupancy is determined (which is a process that necessarily involves some arbitrary thresholds). Apo-Holo protein pairing should thus not be seen as

a static link between PDB entries, but rather as a qualified search process, which results depend on a user query that can specify various arbitrary search options.

4.2 Motivation

Our motivation for developing Apo-Holo protein search tool was the need to create Apo-Holo datasets for better evaluation of binding site prediction methods. The general problem in the field of ligand binding site prediction (and arguably a shortcoming of my own work) is the fact that methods are typically being evaluated only on Holo datasets. Evaluating binding site prediction methods on Holo datasets means that the prediction method can "see" the protein structure as it is after the ligand-induced conformational changes. A prediction method can then use the information encoded in the conformational change in the Holo structure to predict a binding site that it would not be able to predict on the Apo structure. The consequence is that the reported results of success rates of binding site prediction methods can be overly optimistic and may not represent expected results when we apply them to Apo structures (which is almost always what we are looking for when running binding site prediction).

Many other bioinformatics tasks also require access to several conformations (preferably Apo and Holo) and can benefit from the existence of a flexible Apo-Holo search tool. These include observing the effects of ligand binding [BS08], exploring the specificity of a binding site [MSWN02], unveiling cryptic binding sites [CWR*16] and assessing the importance and consistency of water molecules [WDP*18].

4.3 Existing resources

Some resources to address the need of Apo-Holo protein pairing have been built previously. These can be divided into pre-calculated datasets or databases [LSG*10, CYF*12, DLOW07], and one search tool [MTNS11]. However, all the available resources seem to be either not actively updated or are not available at all at the time of writing.

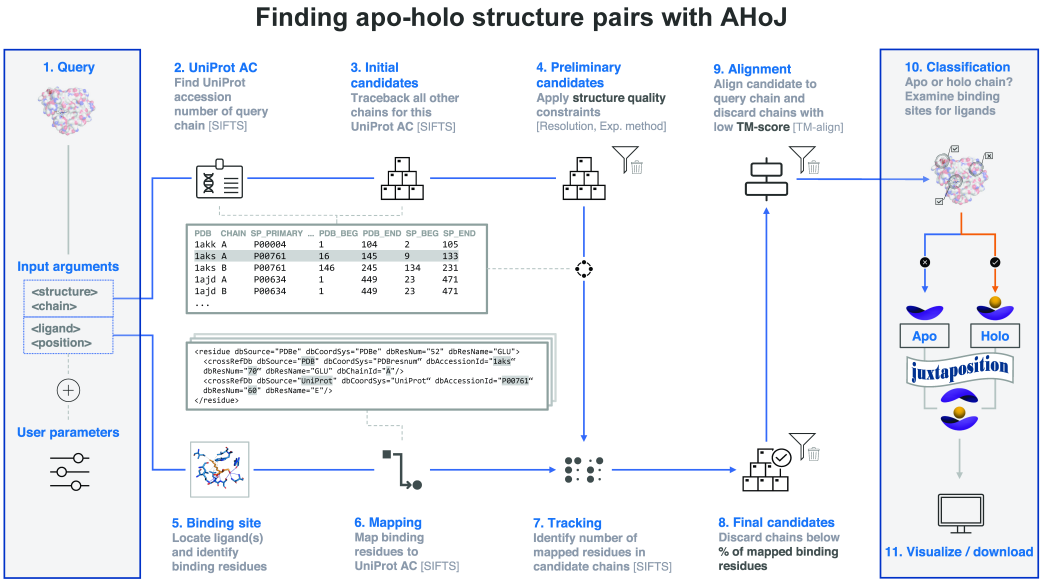

FIGURE 4.1: **Flowchart depicting the workflow in AHoJ**

4.4 Our solution

We have developed AHoJ, a command line tool and a web application that enables the user to conduct easy, fast and parameterizable searches for Apo-Holo structural pairs in the PDB against a query structure [FKHN22]. The user is allowed to specify one or more ligands or binding sites of interest as a part of a query, or can let the application detect the ligands instead. The query structure itself can be Holo or Apo and the result consists of two lists of found structures: those that are Apo with respect to specified binding sites and those that are Holo. All structures are furthermore aligned to the query structure and various metrics for each structure are calculated (including a sequence overlap with the query, RMSD and TM-score). The search process is illustrated in Figure 4.1.

Both the command line tool and the web application can process multiple queries in one run and thus allow to easily create custom Apo-Holo datasets or allow researchers to work in a batch mode without any further programming. The web application allows downloading the results of individual queries or the results of all the queries in a job together. The command line tool produces PyMol visualization and the web application additionally contains an integrated Mol* [SBD*21] visualization of the results (see Figure 4.2). Both applications are freely available and the command line tool is open-sourced.

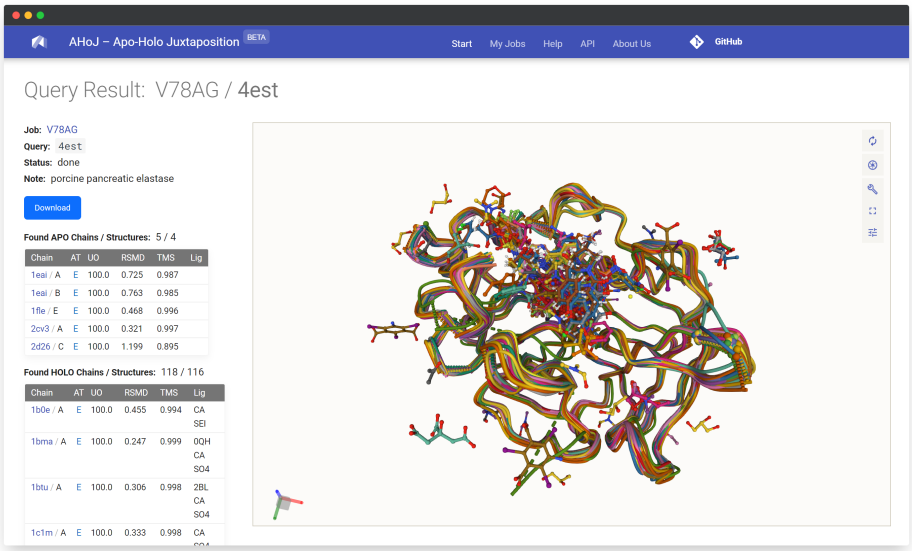

FIGURE 4.2: **AHoJ web application:** screenshot of a page that displays the result of a single search query.

Chapter 5

Conclusion

The main focus of my Ph.D. study was the application of machine learning to the problem of ligand binding site prediction from protein structure and related problems.

I have developed or contributed to the development of several novel methods which include the pocket re-scoring method PRANK, a stand-alone ligand binding site prediction method P2Rank (together with its extended web interface PrankWeb) and the peptide binding prediction method P2Rank-Pept.

The emphasis was always put also on producing pragmatically usable and user-friendly tools, not just on the publication of the methods. This seems to have been a largely successful approach which can be seen in the adoption data. To this date, a binary distribution of P2Rank has been downloaded more than 6500 times while PrankWeb is currently being used by more than 1300 unique users a month.

Furthermore, I have helped to develop AHoJ, a flexible tool for the search and alignment of Apo-Holo protein pairs in the PDB. The main motivation behind it was the need to create Apo-Holo datasets for better evaluation of binding site prediction methods. The existence of this tool will hopefully contribute to binding site prediction methods being again more commonly evaluated on Apo-Holo datasets.

Part II

Publications

Improving protein-ligand binding site prediction accuracy by classification of inner pocket points using local features

Reference

KRIVÁK R., HOKSZA D.: **Improving protein-ligand binding site prediction accuracy by classification of inner pocket points using local features**. *Journal of Cheminformatics 7*, 1 (Apr 2015), 12. doi:10.1186/s13321-015-0059-5

Author's highlights

We have developed PRANK, a machine learning based method that allows to re-score (re-rank) ligand binding sites predicted produced by other methods. Since it helps true binding sites to be ranked higher, it improves the applicability and usefulness of their predictions. PRANK was made available as a free command line tool with source code available upon request.

Note: in this paper we have used the term Connolly surface referring to the surface which would be more precisely described as solvent accessible surface.

P2RANK: Knowledge-Based Ligand Binding Site Prediction Using Aggregated Local Features

Reference

KRIVÁK R., HOKSZA D.: **P2RANK: Knowledge-Based Ligand Binding Site Prediction Using Aggregated Local Features**. In *International Conference on Algorithms for Computational Biology* (2015), Springer, pp. 41–52. doi:10.1007/978-3-319-21233-3_4

Author's highlights

Building on PRANK method we have deceloped P2Rank: a method for prediction of ligand binding sites. This conference article contains the cleanest exposition of P2Rank algorithm itself.

Note: in this paper we have used the term Connolly surface referring to the surface which would be more precisely described as solvent accessible surface.

P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure

Reference

KRIVÁK R., HOKSZA D.: **P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure**. *Journal of cheminformatics 10*, 1 (2018), 39. doi:10.1186/s13321-018-0285-8

Author's highlights

An expanded version of the previous conference contribution that introduced P2Rank as a freely available open-source tool. P2Rank was extensively tested against 5 other state-of-the-art methods. Notable is a longer introduction discussing the state of the field of LBS prediction methods.

The supplementary material contains a discussion about which ligands are considered biologically relevant. Furthermore, there is a detailed pairwise comparison with every other method from our evaluation.

Improving quality of ligand-binding site prediction with Bayesian optimization

Reference

KRIVÁK R., HOKSZA D., ŠKODA P.: **Improving quality of ligand-binding site prediction with Bayesian optimization**. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (2017), pp. 2278–2279. doi:10.1109/BIBM.2017.8218024

Author's highlights

Extended conference abstract summarising several updates to the algorithm. Most notable of them is the implementation of a framework for using Bayesian optimization to optimize several arbitrary parameters of the algorithm at the same time.

Peptide-Binding Site Prediction From Protein Structure via Points on the Solvent Accessible Surface

Reference

KRIVÁK R., JENDELE L., HOKSZA D.: **Peptide-Binding Site Prediction From Protein Structure via Points on the Solvent Accessible Surface**. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics* (New York, NY, USA, 2018), BCB '18, Association for Computing Machinery, p. 645–650. doi:10.1145/3233547.3233708

Author's highlights

P2Rank was modified for the prediction of peptide binding Sites. Apart from working with different datasets the method had to be redesigned to include a residue-centric prediction mode. To achieve performance better than other existing methods several new descriptors were introduced, including geometrical descriptors and sequence conservation score.

PrankWeb: a web server for ligand binding site prediction and visualization

Reference

JENDELE L., KRIVAK R., SKODA P., NOVOTNY M., HOKSZA D.: **PrankWeb: a web server for ligand binding site prediction and visualization**. *Nucleic Acids Res. 47*, W1 (Jul 2019), W345–W349. doi:10.1093/nar/gkz424

Author's highlights

We have developed easy to use web interface for P2Rank with web based visualization, the ability to download the results and documented REST API. A custom pipeline for calculating sequence conservation scores was developed as part of the project and a new default model for P2Rank was trained (the one using sequence conservation among features). The performance of the model using conservation was compared to the model without conservation with the result that conservation contributes to a slightly better prediction success rate and results in producing a lower number of more relevant pockets. At the same time P2Rank introduced Java API which allowed it to be used as a library by programs running on JVM.

PrankWeb 3: accelerated ligand-binding site predictions for experimental and modelled protein structures

Reference

JAKUBEC D., SKODA P., KRIVAK R., NOVOTNY M., HOKSZA D.: **PrankWeb 3: accelerated ligand-binding site predictions for experimental and modelled protein structures**. *Nucleic Acids Research 50*, W1 (05 2022), W593–W597. doi:10.1093/nar/gkac389

Author's highlights

Complete rewrite of PrankWeb as a modern modular Python web application. The conservation pipeline was completely redesigned. The new version is more efficient and consistent with regard to the required time for calculation for any single sequence. Four new prediction models were trained.

PDBe-KB: a community-driven resource for structural and functional annotations

Reference

CONSORTIUM P.-K.: **PDBe-KB: a community-driven resource for structural and functional annotations**. *Nucleic Acids Research 48*, D1 (10 2019), D344–D353. doi:10.1093/nar/gkz853

Abstract

The Protein Data Bank in Europe-Knowledge Base (PDBe-KB, https://pdbe-kb.org) is a community-driven, collaborative resource for literature-derived, manually curated and computationally predicted structural and functional annotations of macromolecular structure data, contained in the Protein Data Bank (PDB). The goal of PDBe-KB is two-fold: (i) to increase the visibility and reduce the fragmentation of annotations contributed by specialist data resources, and to make these data more findable, accessible, interoperable and reusable (FAIR) and (ii) to place macromolecular structure data in their biological context, thus facilitating their use by the broader scientific community in fundamental and applied research. Here, we describe the guidelines of this collaborative effort, the current status of contributed data, and the PDBe-KB infrastructure, which includes the data exchange format, the deposition system for added value annotations, the distributable database containing the assembled data, and programmatic access endpoints. We also describe a series of novel web-pages—the PDBe-KB aggregated views of structure data—which combine information on macromolecular structures from many PDB entries. We have recently

released the first set of pages in this series, which provide an overview of available structural and functional information for a protein of interest, referenced by a UniProtKB accession.

PDBe-KB: collaboratively defining the biological context of structural data

Reference

CONSORTIUM P.-K.: **PDBe-KB: collaboratively defining the biological context of structural data**. *Nucleic Acids Research 50*, D1 (11 2021), D534–D542. doi:10.1093/nar/gkab988

Abstract

The Protein Data Bank in Europe – Knowledge Base (PDBe-KB, https://pdbe-kb.org) is an open collaboration between world-leading specialist data resources contributing functional and biophysical annotations derived from or relevant to the Protein Data Bank (PDB). The goal of PDBe-KB is to place macromolecular structure data in their biological context by developing standardised data exchange formats and integrating functional annotations from the contributing partner resources into a knowledge graph that can provide valuable biological insights. Since we described PDBe-KB in 2019, there have been significant improvements in the variety of available annotation data sets and user functionality. Here, we provide an overview of the consortium, highlighting the addition of annotations such as predicted covalent binders, phosphorylation sites, effects of mutations on the protein structure and energetic local frustration. In addition, we describe a library of reusable web-based visualisation components and introduce new features such as a bulk download data service and a novel superposition service that generates clusters of superposed protein chains weekly for the whole PDB archive.

AHoJ: rapid, tailored search and retrieval of apo and holo protein structures for user-defined ligands

Reference

FEIDAKIS C. P., KRIVAK R., HOKSZA D., NOVOTNY M.: **AHoJ: rapid, tailored search and retrieval of apo and holo protein structures for user-defined ligands**. *Bioinformatics 38*, 24 (10 2022), 5452–5453. doi:10.1093/bioinformatics/btac701

Author's highlights

We have developed AHoJ, a highly-configurable tool for the search and alignment of Apo-Holo protein pairs in the PDB. AHoJ is available as an open-source command line program and a web application that allows running searches for multiple queries at the same time (and thus produce Apo-Holo datasets) and includes integrated web-based visualization.

Bibliography

[BCdF09] BROCHU E., CORA V. M., DE FREITAS N.: **A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning**. *CoRR abs/1012.2599* (2009).

[BS08] BRYLINSKI M., SKOLNICK J.: **What is the relationship between the global structures of apo and holo proteins?** *Proteins: Structure, Function, and Bioinformatics 70*, 2 (2008), 363–377. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.21510, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.21510, doi:https://doi.org/10.1002/prot.21510.

[BS17] BROOMHEAD N. K., SOLIMAN M. E.: **Can We Rely on Computational Predictions To Correctly Identify Ligand Binding Sites on Novel Protein Drug Targets? Assessment of Binding Site Prediction Methods and a Protocol for Validation of Predicted Binding Sites**. *Cell Biochemistry and Biophysics 75*, 1 (Mar 2017), 15–23. URL: https://doi.org/10.1007/s12013-016-0769-y, doi:10.1007/s12013-016-0769-y.

[BSSC18] BHAGAVAT R., SANKAR S., SRINIVASAN N., CHANDRA N.: **An Augmented Pocketome: Detection and Analysis of Small-Molecule Binding Pockets in Proteins of Known 3D Structure**. *Structure 26*, 3 (2018), 499 – 512.e2. doi:https://doi.org/10.1016/j.str.2018.02.001.

[CHG14] CHEN P., HUANG J. Z., GAO X.: **LigandRFs: random forest ensemble to identify ligand-binding residues from sequence information alone.** *BMC bioinformatics 15 Suppl 15* (Jan 2014), S4. doi:10.1186/1471-2105-15-S15-S4.

[CLT*09] CAPRA J. A., LASKOWSKI R. A., THORNTON J. M., SINGH M., FUNKHOUSER T. A.: **Predicting Protein Ligand Binding Sites by Combining Evolutionary Sequence Conservation and 3D Structure**. *PLoS Comput Biol 5*, 12 (12 2009), e1000585.

[CMGK11] CHEN K., MIZIANTY M., GAO J., KURGAN L.: **A critical comparative assessment of predictions of protein-binding sites for biologically relevant organic compounds.** *Structure (London, England : 1993) 19*, 5 (2011), 613–621. URL: http://dx.doi.org/10.1016/j.str.2011.02.015, doi:10.1016/j.str.2011.02.015.

[con19] CONSORTIUM P.-K.: **PDBe-KB: a community-driven resource for structural and functional annotations**. *Nucleic Acids Research 48*, D1 (10 2019), D344–D353. doi:10.1093/nar/gkz853.

[con21] CONSORTIUM P.-K.: **PDBe-KB: collaboratively defining the biological context of structural data**. *Nucleic Acids Research 50*, D1 (11 2021), D534–D542. doi:10.1093/nar/gkab988.

[CWR*16] CIMERMANCIC P., WEINKAM P., RETTENMAIER T. J., BICHMANN L., KEEDY D. A., WOLDEYES R. A., SCHNEIDMAN-DUHOVNY D., DEMERDASH O. N., MITCHELL J. C., WELLS J. A., ET AL.: **CryptoSite: expanding the druggable proteome by characterization and prediction of cryptic binding sites**. *Journal of molecular biology 428*, 4 (2016), 709–719.

[CYF*12] CHANG D. T.-H., YAO T.-J., FAN C.-Y., CHIANG C.-Y., BAI Y.-H.: **AH-DB: collecting protein structure pairs before and after binding**. *Nucleic acids research 40*, D1 (2012), D472–D478.

[DLOW07] DESSAILLY B. H., LENSINK M. F., ORENGO C. A., WODAK S. J.: **LigASite—a database of biologically relevant binding sites in proteins with known apo-structures**. *Nucleic acids research 36*, suppl_1 (2007), D667–D673.

[DWH15] DEGAC J., WINTER U., HELMS V.: **Graph-Based Clustering of Predicted Ligand-Binding Pockets on Protein Surfaces**. *Journal of Chemical Information and Modeling 55*, 9 (2015), 1944–1952. PMID: 26325445. doi:10.1021/acs.jcim.5b00045.

[FB15] FEINSTEIN W. P., BRYLINSKI M.: **Calculating an optimal box size for ligand docking and virtual screening against experimental and predicted binding pockets**. *Journal of Cheminformatics 7*, 1 (2015), 1–10. URL: http://dx.doi.org/10.1186/s13321-015-0067-5, doi:10.1186/s13321-015-0067-5.

[FKHN22] FEIDAKIS C. P., KRIVAK R., HOKSZA D., NOVOTNY M.: **AHoJ: rapid, tailored search and retrieval of apo and holo protein structures for user-defined ligands**. *Bioinformatics 38*, 24 (10 2022), 5452–5453. doi:10.1093/bioinformatics/btac701.

[FRH11] FAUMAN E. B., RAI B. K., HUANG E. S.: **Structure-based druggability assessment—identifying suitable targets for small molecule therapeutics**. *Current Opinion in Chemical Biology 15*, 4 (2011), 463 – 468. Next Generation Therapeutics. doi:https://doi.org/10.1016/j.cbpa.2011.05.020.

[GS09] GHERSI D., SANCHEZ R.: **EasyMIFS and SiteHound: a toolkit for the identification of ligand-binding sites in protein structures.** *Bioinformatics (Oxford, England) 25*, 23 (2009), 3185–3186. URL: http://dx.doi.org/10.1093/bioinformatics/btp562, doi:10.1093/bioinformatics/btp562.

[HBG*15] HUSSEIN H., BORREL A., GENEIX C., PETITJEAN M., REGAD L., CAMPROUX A.: **PockDrug-Server: a new web server for predicting pocket druggability on holo and apo proteins**. W436–W442. doi:10.1093/nar/gkv462.

[HLH*13] HUANG W., LU S., HUANG Z., LIU X., MOU L., LUO Y., ZHAO Y., LIU Y., CHEN Z., HOU T., ZHANG J.: **Allosite: a method for predicting allosteric sites**. *Bioinformatics 29*, 18 (2013), 2357–2359. doi:10.1093/bioinformatics/btt399.

[HOH*10] HENRICH S., OUTI S., HUANG B., RIPPMANN F., CRUCIANI G., WADE R.: **Computational approaches to identifying and characterizing protein binding sites for ligand design.** *Journal of molecular recognition : JMR 23*, 2 (2010), 209–219. URL: http://dx.doi.org/10.1002/jmr.984, doi:10.1002/jmr.984.

[JDMR*17] JIMÉNEZ J., DOERR S., MARTÍNEZ-ROSELL G., ROSE A. S., DE FABRITIIS G.: **DeepSite: protein-binding site predictor**

using 3D-convolutional neural networks. *Bioinformatics 33*, 19 (2017), 3036–3042. doi:10.1093/bioinformatics/btx350.

[JG17] JIMÉNEZ J., GINEBRA J.: **pyGPGO: Bayesian Optimization for Python**. *Journal of Open Source Software 2*, 19 (2017), 431. URL: https://doi.org/10.21105/joss.00431, doi:10.21105/joss.00431.

[JKS*19] JENDELE L., KRIVAK R., SKODA P., NOVOTNY M., HOKSZA D.: **PrankWeb: a web server for ligand binding site prediction and visualization**. *Nucleic Acids Res. 47*, W1 (Jul 2019), W345–W349. doi:10.1093/nar/gkz424.

[JSK*22] JAKUBEC D., SKODA P., KRIVAK R., NOVOTNY M., HOKSZA D.: **PrankWeb 3: accelerated ligand-binding site predictions for experimental and modelled protein structures**. *Nucleic Acids Research 50*, W1 (05 2022), W593–W597. doi:10.1093/nar/gkac389.

[KH15a] KRIVÁK R., HOKSZA D.: **Improving protein-ligand binding site prediction accuracy by classification of inner pocket points using local features**. *Journal of Cheminformatics 7*, 1 (Apr 2015), 12. doi:10.1186/s13321-015-0059-5.

[KH15b] KRIVÁK R., HOKSZA D.: **P2RANK: Knowledge-Based Ligand Binding Site Prediction Using Aggregated Local Features**. In *International Conference on Algorithms for Computational Biology* (2015), Springer, pp. 41–52. doi:10.1007/978-3-319-21233-3_4.

[KH18] KRIVÁK R., HOKSZA D.: **P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure**. *Journal of cheminformatics 10*, 1 (2018), 39. doi:10.1186/s13321-018-0285-8.

[KH7] KRIVÁK R., HOKSZA D., ŠKODA P.: **Improving quality of ligand-binding site prediction with Bayesian optimization**. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (2017), pp. 2278–2279. doi:10.1109/BIBM.2017.8218024.

[KJ14] KONC J., JANEŽIČ D.: **Binding site comparison for function prediction and pharmaceutical discovery.** *Current opinion in*

structural biology 25 (Apr 2014), 34–9. doi:10.1016/j.sbi.2013.11.012.

[KJH18] KRIVÁK R., JENDELE L., HOKSZA D.: **Peptide-Binding Site Prediction From Protein Structure via Points on the Solvent Accessible Surface**. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics* (New York, NY, USA, 2018), BCB '18, Association for Computing Machinery, p. 645–650. doi:10.1145/3233547.3233708.

[KK09] KAUFFMAN C., KARYPIS G.: **LIBRUS: combined machine learning and homology information for sequence-based ligand-binding residue prediction.** *Bioinformatics (Oxford, England) 25*, 23 (Dec 2009), 3099–107. URL: http://bioinformatics.oxfordjournals.org/cgi/pmidlookup?view=long&pmid=19786483, doi:10.1093/bioinformatics/btp561.

[LEG16] LAURIE E. GROVE SANDOR VAJDA D. K.: **Computational Methods to Support Fragment-based Drug Discovery**. In *Fragment-based Drug Discovery: Lessons and Outlook*, Fagerberg J., Mowery D. C., Nelson R. R., (Eds.). Wiley, Weinheim, 2016, ch. 9, pp. 197–222.

[LGST09] LE GUILLOUX V., SCHMIDTKE P., TUFFERY P.: **Fpocket: an open source platform for ligand pocket detection.** *BMC bioinformatics 10* (2009). URL: http://dx.doi.org/10.1186/1471-2105-10-168, doi:10.1186/1471-2105-10-168.

[LJ06] LAURIE A., JACKSON R.: **Methods for the prediction of protein-ligand binding sites for structure-based drug design and virtual ligand screening.** *Current protein & peptide science 7*, 5 (2006), 395–406.

[LSCZ14] LIONTA E., SPYROU G., COURNIA D. K. V., ZOE: **Structure-Based Virtual Screening for Drug Discovery: Principles, Applications and Recent Advances**. *Current Topics in Medicinal Chemistry 14*, 16 (2014), 1923–1938. URL: http://www.eurekaselect.com/node/124979/article.

[LSG*10] LOBANOV M. Y., SHOEMAKER B. A., GARBUZYNSKIY S. O., FONG J. H., PANCHENKO A. R., GALZITSKAYA O. V.: **ComSin:**

database of protein structures in bound (complex) and unbound (single) states in relation to their intrinsic disorder. *Nucleic Acids Research 38*, suppl_1 (2010), D283–D287.

[LSZ10] LEIS S., SCHNEIDER S., ZACHARIAS M.: **In silico prediction of binding sites on proteins.** *Current medicinal chemistry 17*, 15 (2010), 1550–1562.

[MBB16] MEYERS J., BROWN N., BLAGG J.: **Mapping the 3D structures of small molecule binding sites**. *Journal of Cheminformatics 8*, 1 (2016), 70. doi:10.1186/s13321-016-0180-0.

[MSWN02] MA B., SHATSKY M., WOLFSON H. J., NUSSINOV R.: **Multiple diverse ligands binding at a single protein site: a matter of pre-existing populations**. *Protein science 11*, 2 (2002), 184–197.

[MTNS11] MORITA M., TERADA T., NAKAMURA S., SHIMIZU K.: **BUDDY-system: A web site for constructing a dataset of protein pairs between ligand-bound and unbound states**. *BMC Research Notes 4* (2011), 1–4.

[MZF*17] MONZON A. M., ZEA D. J., FORNASARI M. S., SALDAÑO T. E., FERNANDEZ-ALBERTI S., TOSATTO S. C. E., PARISI G.: **Conformational diversity analysis reveals three functional mechanisms in proteins**. *PLOS Computational Biology 13*, 2 (02 2017), 1–18. URL: https://doi.org/10.1371/journal.pcbi.1005398, doi:10.1371/journal.pcbi.1005398.

[PSM*10] PÉROT S., SPERANDIO O., MITEVA M., CAMPROUX A., VILLOUTREIX B.: **Druggable pockets and binding site centric chemical space: a paradigm shift in drug discovery.** *Drug discovery today 15*, 15-16 (2010), 656–667. URL: http://dx.doi.org/10.1016/j.drudis.2010.05.015, doi:10.1016/j.drudis.2010.05.015.

[QW00] QIU Z., WANG X.: **Improved Prediction of Protein Ligand-Binding Sites Using Random Forests**. *Protein and Peptide Letters 18*, 12 (2011-12-01T00:00:00), 1212–1218. URL: http://www.ingentaconnect.com/content/ben/ppl/2011/00000018/00000012/art00005, doi:doi:10.2174/092986611797642788.

[RBJ15] ROCHE D. B., BRACKENRIDGE D. A., J M. L.: **Proteins and Their Interacting Partners: An Introduction to Protein-Ligand Binding Site Prediction Methods.** *Int J Mol Sci 16*, 12 (2015), 29829–42. doi:10.3390/ijms161226202.

[SBB*14] SCHOMBURG K., BIETZ S., BRIEM H., HENZLER A., URBACZEK S., RAREY M.: **Facing the challenges of structure-based target prediction by inverse virtual screening.** *Journal of chemical information and modeling 54*, 6 (2014), 1676–86. doi:10.1021/ci500130e.

[SBD*21] SEHNAL D., BITTRICH S., DESHPANDE M., SVOBODOVÁ R., BERKA K., BAZGIER V., VELANKAR S., BURLEY S. K., KOČA J., ROSE A. S.: **Mol* Viewer: modern web app for 3D visualization and analysis of large biomolecular structures**. *Nucleic Acids Research 49*, W1 (05 2021), W431–W437. URL: https://doi.org/10.1093/nar/gkab314, arXiv:https://academic.oup.com/nar/article-pdf/49/W1/W431/38842088/gkab314.pdf, doi:10.1093/nar/gkab314.

[SCS*17] SHEN Q., CHENG F., SONG H., LU W., ZHAO J., AN X., LIU M., CHEN G., ZHAO Z., ZHANG J.: **Proteome-Scale Investigation of Protein Allosteric Regulation Perturbed by Somatic Mutations in 7,000 Cancer Genomes**. *The American Journal of Human Genetics 100*, 1 (2017), 5 – 20. URL: http://www.sciencedirect.com/science/article/pii/S0002929716304013, doi:https://doi.org/10.1016/j.ajhg.2016.09.020.

[Sel17] SELLÉS J. P.: **FastRandomForest 2.0**, 2017. URL: https://github.com/GenomeDataScience/FastRandomForest.

[SLA12] SNOEK J., LAROCHELLE H., ADAMS R. P.: **Practical Bayesian Optimization of Machine Learning Algorithms**. In *Advances in Neural Information Processing Systems 25*. 2012, pp. 2951–2959.

[SLD*] SIMÕES T., LOPES D., DIAS S., FERNANDES F., PEREIRA J., JORGE J., BAJAJ C., GOMES A.: **Geometric Detection Algorithms for Cavities on Protein Surfaces in Molecular Graphics: A Survey**. *Computer Graphics Forum*. doi:10.1111/cgf.13158.

[SSKH07] SOGA S., SHIRAI H., KOBORI M., HIRAYAMA N.: **Use of Amino Acid Composition to Predict Ligand-Binding Sites**. *Journal of Chemical Information and Modeling 47*, 2 (2007), 400–406. PMID: 17243757. doi:10.1021/ci6002202.

[Sup13] SUPEK F.: **FastRandomForest**, 2013. URL: https://code.google.com/archive/p/fast-random-forest/.

[TAW*21] TUNYASUVUNAKOOL K., ADLER J., WU Z., GREEN T., ZIELINSKI M., ŽÍDEK A., BRIDGLAND A., COWIE A., MEYER C., LAYDON A., ET AL.: **Highly accurate protein structure prediction for the human proteome**. *Nature 596*, 7873 (2021), 590–596.

[TBNT16] TIBAUT T., BORIŠEK J., NOVIČ M., TURK D.: **Comparison of in silico tools for binding site prediction applied for structure-based design of autolysin inhibitors**. *SAR and QSAR in Environmental Research 27*, 7 (2016), 573–587. PMID: 27686112. doi:10.1080/1062936X.2016.1217271.

[WDP*18] WLODAWER A., DAUTER Z., POREBSKI P. J., MINOR W., STANFIELD R., JASKOLSKI M., POZHARSKI E., WEICHENBERGER C. X., RUPP B.: **Detect, correct, retract: How to manage incorrect structural models**. *The FEBS journal 285*, 3 (2018), 444–466.

[XH12] XIE Z., HWANG M.: **Ligand-binding site prediction using ligand-interacting and binding site-enriched protein triangles**. *Bioinformatics 28*, 12 (2012), 1579–1585. doi:10.1093/bioinformatics/bts182.

[XXB11] XIE L., XIE L., BOURNE P. E.: **Structure-based systems biology for analyzing off-target binding.** *Current opinion in structural biology 21*, 2 (Apr 2011), 189–99. doi:10.1016/j.sbi.2011.01.004.

[ZGWW12] ZHENG X., GAN L., WANG E., WANG J.: **Pocket-Based Drug Design: Exploring Pocket Space**. *The AAPS Journal* (2012). doi:10.1208/s12248-012-9426-6.

[ZLL*11] ZHANG Z., LI Y., LIN B., SCHROEDER M., HUANG B.: **Identification of cavities on protein surface using multiple computational approaches for drug binding site prediction.** *Bioinformatics (Oxford, England) 27*, 15 (2011), 2083–2088. URL: http://dx.doi.org/10.1093/bioinformatics/btr331, doi:10.1093/bioinformatics/btr331.

List of Figures

List of Tables