Charles University
Faculty of Science

Study programme: Bioinformatics

Branch of study: Bioinformatics

Kateřina Břicháčková

Use of residue-level annotations for structural prediction of protein-ligand binding sites

Využití anotací primární struktury pro strukturní predikci protein-ligand aktivních míst

Master thesis

Supervisor: RNDr. David Hoksza, Ph.D.

Prague, 2020

Prohlášení

Prohlašuji, že jsem závěrečnou práci zpracovala samostatně a že jsem uvedla všechny použité informační zdroje a literaturu. Tato práce ani její podstatná část nebyla předložena k získání jiného nebo stejného akademického titulu.

V Praze, XXX Kateřina Břicháčková

Poděkování

Acknowledgement

Abstract

abstract

Keywords: keyword1, keyword2

Abstrakt

abstract

Klíčová slova: keyword1, keyword2

Contents

1. Introduction

TODO biologicke background, struktura proteinu, interakce proteinu (protein-protein, protein-ligand, DNA, RNA...)

TODO co jsou active sites, jakou maji funkci, proc je jejich predikce dulezita, drug design

TODO cile prace

TODO napsat jak by bylo slozity zjistovat jestli mi nejaka vlastnost zmeni predikci, ze je tam problem s mapovanim, parsovanim atd.

The aims of the thesis are following:

- To implement a pipeline for the statistical analysis of residue-level annotations. The pipeline should be designed to be easily extensible by new annotations defined by the user. It should be possible to run the statistical analysis with data supplied by the user, as well as to run the whole process and obtaining the data automatically.
- To use the pipeline for the analysis of existing experimental and predicted residue-level annotations of protein structure. According to the results, select the ligand binding sites properties which appear to have different values for binding and non-binding sites.
- To use existing method P2Rank [32] for prediction of ligand binding sites to test the practical significance of the results.

2. Ligand binding sites prediction

2.1 Existing approaches

todo v cem se to lisi od protein-peptide binding sites

2.1.1 P2Rank

2.2 Evaluation of success rates

the cutoff D = 4 Å, which was suggested by Skolnick and Brylinski (2008), ...

3. Methodology

One of the main aims of the thesis was to develop a pipeline for statistical analysis of available protein structure annotations (hereinafter referred to as features), and to prepare this pipeline for adding user-defined features.

The process starts with downloading FASTA and PDB files for input proteins from databases. Residue-level mappings are downloaded as well, to allow cross-referencing the protein tertiary structure with the sequence annotations. After that, values for all features are computed or downloaded and assigned to each residue. Residues are labeled as binding or non-binding according to the ligands defined in the PDB file. As the next step, we perform statistical analysis of the features, using computed ligand binding sites labels and feature values. After examining the results, we can decide which features could be potentially interesting for the ligand binding sites prediction. Finally, we can train new P2Rank models with these new features and see if the performance has improved.

The analysis pipeline covers most steps of the process, from downloading the files from databases, to computing the statistical analysis of individual features. The only needed input is a dataset file with listed protein identifiers. Moreover, there are two scripts that further extend the pipeline and can be used to train and evaluate P2Rank models with new features. The structure of the pipeline is depicted in Diagram 3.1. The details about individual parts are described in this Chapter.

The pipeline is implemented in Python, utilizing several Python packages, such as BioPython [18], NumPy [**?**] or SciPy [**?**]. BioPython is an open-source collection of Python tools for computational biology and it was very useful in this work, especially for parsing PDB and FASTA files.

The pipeline comprises a set of Python scripts which are connected together by the main script `analysis_pipeline.py`. The main script should be used to run the pipeline. It defines the user API, parses and checks arguments, takes care of logging and runs individual parts of the pipeline. See `https://github.com/katebrich/LBS_analysis_pipeline` for more details about options, examples of usage, setup, requirements, input and output.

The features can be defined in 'config file'. It is a file in JSON format that lists names of features, their type (binary, categorical, ordinal or continuous) and a path to the class with implementation. Custom config file with user-defined features or with subsets of features can be created and passed as argument `-c new_config_path`.

TODO vic popsat usage? Nebo usage a setup v attachments? Příklady?

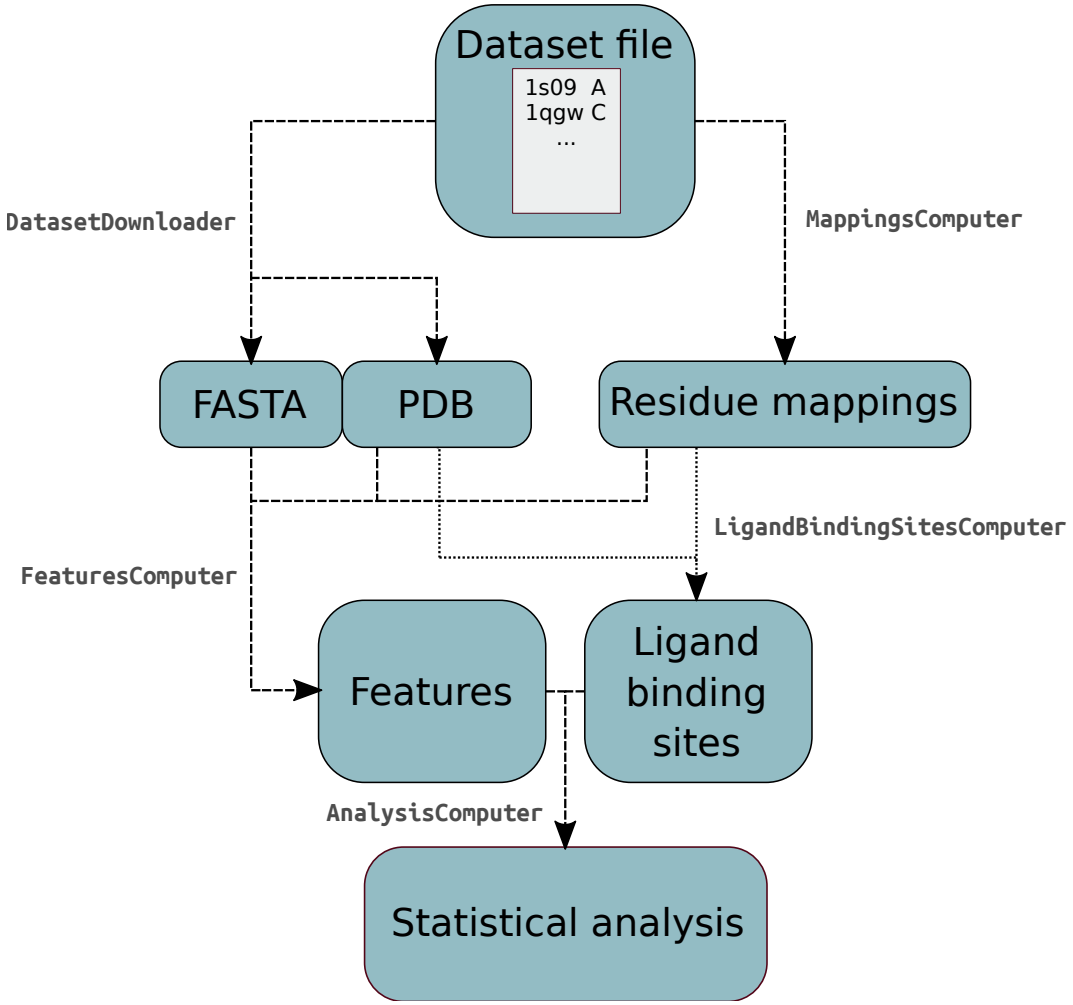


Figure 3.1: Diagram of the pipeline structure.

The information about versions of used software and databases can be found in Appendix TODO.

3.1 Dataset file

Dataset file is a mandatory input for the pipeline. It is a plain-text file with each row representing one structure. It has several columns separated by whitespace. The first two columns are mandatory and they contain PDB ID and chain ID (pipeline can only work with single-chain structures). The third column is optional and it can define a comma-separated list of ligands which will be used for the ligand binding sites computation.

3.2 FASTA and PDB download

For each structure, FASTA and PDB files are downloaded from PDBe, via Entry-based REST API [8], which is one of the possibilities to access large amouts of data about individual PDB entries programatically.

3.3 Residue mappings

The residue-level mappings are needed for cross-referencing the protein tertiary structure with the sequence annotations and UniProt records. The main reason is that the PDB entry may cover only a segment of the full-length protein and the segment does not have to be continuous.

Furthermore, the numbering of residues in the PDB file can differ from the primary sequence numbering. The residues in the PDB file are assigned identifiers by the author, in order to match the identification used in the publication. The identifier of the residue is composed of two parts. The first one is a residue number, and the second one, called 'insertion code', is usually a character and it is left empty for most residues. Typically, it is used to label insertions relative to the reference sequence. The author can assign the numbers how he or she desires; they do not have to start with one or zero, do not have to be in consecutive order and can even be negative.

For those reasons, the residue-level mappings are downloaded from PDBe REST API [8] early in the pipeline process, they are cached in files and can be used whenever sequence annotations need to be mapped on the structure described in the PDB file.

For the cross-referencing between protein structures and protein sequences in UniProtKB [51], we used UniProt segments mapping implemented in PDBe REST API. The implementation is based on SIFTS [**?**], a resource for the transfer of annotations between protein structure and sequence. The mappings are assigned by the SIFTS process with the UniProt sequence as reference and thus, the output is a set of segments that reflects discontinuities in the UniProt sequence.

There is a presumable issue with UniProt segments - sometimes the returned segment does not have the same length in UniProt coordinates and PDB coordinates. We reported it and it was recognized as a bug, but unfortunately has not been corrected before running the experiments in this work; however, it is very rare and it concerns only a few structures; those were removed from the datasets.

3.4 Ligand binding sites

Each residue is assigned a label *non-binding/binding* (0/1) according to positions of ligands in the PDB file. A residue is labeled as *binding* if it has at least one non-hydrogen surface atom within distance 4.0 Å of a non-hydrogen atom of any ligand. The distance 4.0 Å can be changed in the pipeline with `-l` argument. This ligand-based definition of binding sites was used in previous large-scale study exploring the composition of binding sites [31].

Only residues on the surface of the protein were taken into consideration in the analysis. The non-surface residues cannot be binding anyway, and excluding them decreases the imbalance between binding and non-binding residues counts. Furthermore, excluding the inner residues helps to reduce potential influence of difference of feature values in surface vs. non-surface residues. For example, inner residues tend to be more hydrophobic in general. Thus, binding sites could seem to be more hydrophilic than non-binding sites, but it would not be clear whether it is not simply the effect of being on the surface of the protein.

To decide which residues are located on the surface, solvent-accessible surface area of each residue was computed. We defined the surface residues as residues that have less than 5% of their surface accessible to the solvent. This cut-off was proposed by Miller *et al.* [38] and used in other studies[29, 31].

The solvent-accessible surface was computed using `Bio.PDB.SASA` module in BioPython [9]. It implements Shrake & Rupley algorithm [48] which uses a sphere of particular radius to probe the surface of the protein. It can be imagined as 'rolling a ball' along the surface (see Figure 3.2). The smaller the sphere radius, the more surface details it can detect. For this work, we used the default radius of 1.4 Å, which approximates the radius of a water molecule.

3.5 Features

This section describes all implemented features and provides information on how to add user-defined features. The features names and types are summarized in Table 3.1

The pipeline can run only with a subset of implemented features, by listing them and passing as argument (e.g. `-f hydropathy,aromaticity`). If argument `-f` is not stated, all features defined in the config file are computed.

The individual features are described in detail in following sections, categorized by the resource that was used for their retrieval.

Name	Type	Source
PTM	binary	UniProtKB
lipidation	binary	UniProtKB
glycosylation	binary	UniProtKB
mod_res	binary	UniProtKB
disulfid	binary	UniProtKB
non_standard	binary	UniProtKB
sec_str	categorical	UniProtKB
helix	binary	UniProtKB
turn	binary	UniProtKB
strand	binary	UniProtKB
natural_variant	binary	UniProtKB
variation	binary	UniProtKB
compbias	binary	UniProtKB
pdbekb_conservation	ordinal	PDBe-KB
dynamine	continuous	PDBe-KB
efoldmine	continuous	PDBe-KB
depth	continuous	PDBe-KB
bfactor	continuous	PDBe-KB
exposure_CN	continuous	PDB
HSE_up	continuous	PDB
HSE_down	continuous	PDB
phi_angle	continuous	PDBe
psi_angle	continuous	PDBe
cis_peptide	binary	PDBe
aa	categorical	FASTA
hydropathy	ordinal	FASTA
mol_weight	ordinal	FASTA
polarity	categorical	FASTA
charge	binary	FASTA
aromaticity	binary	FASTA
H_bond_atoms	ordinal	FASTA
mobiDB	continuous	MobiDB
conservation	continuous	P2Rank

Table 3.1: Summary of analysed features.

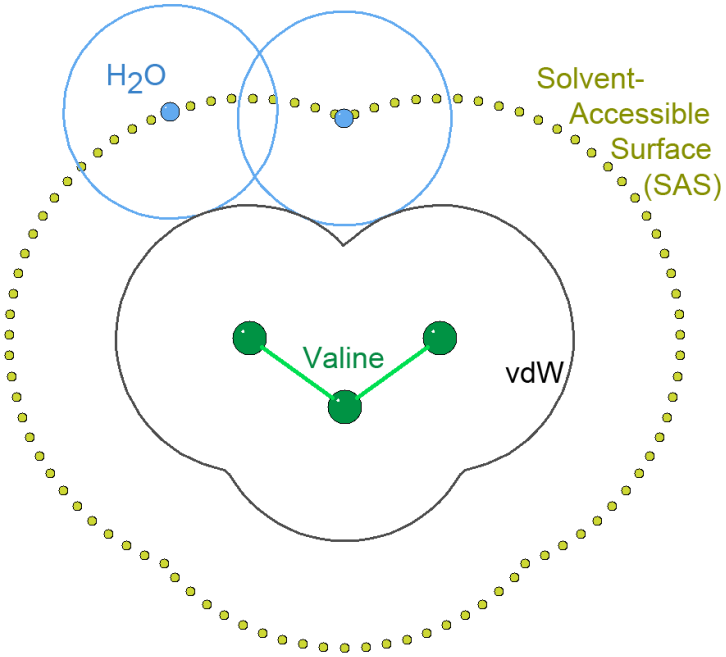


Figure 3.2: Illustration of the solvent accessible surface. It was created by rolling the probe (in blue) along the molecule surface and tracing the center of the probe. Retrieved 02-01-2020 from `https://commons.wikimedia.org/wiki/File:Surfacetype_Solvent-Accessible.png`

3.5.1 UniProtKB

The UniProt Knowledgebase (UniProtKB) is a large database of well-annotated protein sequence data. It tries to achieve the minimal redundancy of proteomes and it provides detailed, accurate and consistent annotations of the sequences [51].

Sequence annotations (called 'features') are available for every UniProtKB entry. They describe interesting sites and regions on the protein sequence and every feature has an associated description with available evidence, source and related publications. The features are arranged in a well-organized manner on the UniProt website [**?**], in so called 'Features viewer' with many overlapping tracks for different features. Nonetheless, for the purpose of this work, the best way to obtain the features was via the Proteins REST API [41]. It provides the interface to access the sequence annotation data as well as mapped variation data programmatically. The API is available at (`http://www.ebi.ac.uk/proteins/api/doc`) [**?**].

Features are classified into eight categories which are further subdivided into types. For example, the category 'STRUCTURAL' comprises the types 'HELIX', 'TURN' and 'STRAND'.

The types and categories that were chosen as potentially relevant for ligand

binding sites prediction are described below.

3.5.1.1 PTM

Post-translational modifications are covalent chemical modifications of polypeptide chains after translation, usually modifying the functional group of the standard amino acids, or introducing a new group. They extend the set of the 20 standard amino acids and they can be important for the function of many proteins,as they can alter the interactions with other proteins, localization, activity, signal transduction, cell-cell interactions and other properties. Their enrichment in binding sites is very interesting to examine.

Three UniProtKB feature types were analysed: lipidation, glycosylation and type 'MOD_RES' which comprises phosphorylation, methylation, acetylation, amidation, formation of pyrrolidone carboxylic acid, isomerization, hydroxylation, sulfation, flavin-binding, cysteine oxidation and nitrosylation. Only experimentally determined modification sites are annotated, and they are further propagated to related orthologs when specific criteria are met [5].

Since lipidaton and glycosylation data were very sparse (e.g. there were only 15 lipidation sites in the whole holo4k dataset composed of 3973 proteins), the fourth feature called 'PTM' including all three types was added to the analysis.

3.5.1.2 Disulfide bonds

Another type of post-translational modifications are disulfide bonds formed between two cysteine residues. Both intrachain and interchain bonds are annotated by UniProtKB. The disulfide bonds may be either experimentally determined or predicted (occuring in specific protein families) [3].

3.5.1.3 Non-standard residues

Describes the occurence of non-standard amino acids (selenocysteine and pyrrolysine). There must be experimental evidence for this occurence; however, it can be propagated to close homologs [7].

3.5.1.4 Secondary structure

This feature category annotates three types of secondary structures: helices, beta sheets and hydrogen-bonded turns. Residues not belonging to any of the classes are in a random-coil structure. The 'helix' class comprises alpha-helices, pi-helices and 3_{10} helices.

The secondary structure assignment is made by DSSP algorithm [30] based on the coordinate data sets extracted from the Protein Data Bank (PDB). They are neither predicted computationally, nor propagated to related species [10].

3.5.1.5 Natural variant

This feature includes naturally occuring polymorphisms, variations between strains or RNA editing events [6].

3.5.1.6 Variation

Variation service is a utility that can retrieve variation data from UniProtKB. The variants are either extracted from the scientific literature and manually reviewed, or mapped from large scale studies, such as 1000 Genomes [**?**], COSMIC [**?**], ClinVar [**?**] or ExAC [**?**]. The Proteins REST API provides various options for variants retrieval, such as to filter by the consequence type, associated disease name, cross reference database type (e.g. ClinVar) or by the source type [41].

3.5.1.7 Compositional bias

The regions of compositional bias are parts of the polypeptide chain where some of the amino acids are over-represented, not following the standard frequencies. The regions can be enriched in one or more different amino acids [**?**].

3.5.2 PDBe-KB

PDBe-KB (Protein Data Bank in Europe - Knowledge Base) is managed by the PDBe team at the European Bioinformatics Institute. It is a collaborative resource that aims to bring together the annotations from various sources and to show the macromolecular structures in broader biological context.

One drawback of PDB is that every page represents only one entry that is based on a single experiment. There may be several PDB entries for the full-length protein, each covering only a segment of it. Nevertheless, the entries for the same protein are not interconnected. PDBe-KB has developed the *aggregated views of proteins*, displaying an overview of all the data related to the full-length protein defined by the UniProtKB accession.

The structures from the PDB are extensively used by scientific software and other resources. There exist many valuable annotations, such as ligand binding sites, post-translational modification sites, molecular channels or effects of mutations, that are created outside of the PDB. The problem is that the data

is fragmented and therefore it would require immense effort of a researcher to collect and make use of all available data for a structure of interest.

The aggregated views of proteins integrates the annotations from *PDBe-KB partners*, collaborating scientific software developers. It facilitates the retrieval of these annotations with a uniform data access mechanism (via FTP or REST API). The project is called 'FunPDBe'. A common data exchange scheme was defined to facilitate the transfer of data. [19]

The use of PDBe-KB was difficult because of the lack of documentation and a few bugs that were encountered during this work (some of them corrected by now after pointing them out). However, it is understandable since it was launched only two years ago and the constant improvements are done since then.

3.5.2.1 Conservation

PDBe-KB provides pre-calculated residue-level conservation scores, obtained by a pipeline using HMMER and Skylign web servers that was described by Jakubec et al. [27].

The values of the score are integers ranging from 0 to 9, with 9 being the most conserved. Since scores higher than 4 were very sparse and the feature would not meet the assumptions of the Chi-squared test, the scores 4 and higher were merged into one category (4). This does not deteriorate the prediction nor the hypothesis test, as vast majority (over 95%) of non-binding residues were scored 1 and lower.

3.5.2.2 DynaMine

DynaMine [15] was developed by the Bio2Byte group [1] and it is one of the PDBe-KB partner resources. It provides the annotations of the backbone dynamics predicted only from the FASTA sequence. DynaMine predicts backbone flexibility at the residue-level, using a linear regression model trained on a large dataset of curated NMR chemical shifts extracted from the Biological Magnetic Resonance Data Bank [52]. The predictor estimates the value of the 'order parameter' (S^2) which is related to the rotational freedom of the N-H bond vector of the backbone. The values range from 0 (highly dynamic) to 1 (complete order).

3.5.2.3 EFoldMine

EFoldMine [42] tool comes from the same group as DynaMine. It is a predictor of the early folding regions of proteins. It makes predictions at the residue-level derived only from the FASTA sequence. Internally it uses dynamics predictions and secondary structure propensities as features and the linear regression model

is trained on data from NMR pulsed labelling experiments. Unfortunatelly, the early stages of protein folding are not understood very well so far and experimental data is very difficult to obtain. The predictor was trained on the dataset of only 30 proteins and its performance is quite poor.

3.5.2.4 Depth

Depth [50] is a webserver that can measure residue burial within the protein. It is able to find small cavities in proteins and could be used as a ligand-binding sites predictor as such. The residue depth values are computed from the input PDB file.

The algorithm places input 3D structure in the box of model water, each residue with at least two hydration shells around itself. The water molecules in cavities are removed: the algorithm removes the water molecule if there are less than a given number of water molecules in its spherical volume of given size. The minimum number of neighbouring molecules and the spherical volume can be defined by the user. The removal is iterated until there are no more cavity waters. Residue depth is then computed as the distance to the closest water molecule.

3.5.3 PDB

The following features can be computed or obtained directly from the PDB file.

3.5.3.1 B factor

The B factor, also called the Debye-Waller factor or the temperature factor, describes "the attenuation of X-ray or neutron scattering caused by thermal motion" [49]. It can be used to identify and interpret flexibility of proteins, supposing that high B factors are indicators of higher flexibility, whereas atoms with low B factors generally belong to the well-ordered parts of the structure. B factors can be also view as indicators of the relative vibrational motion of atoms in a protein [49].

The values can be obtained directly from the PDB files: each ATOM record of a X-ray structure (except for hydrogens) deposited in PDB contains B factor value for the atom. B factor for a residue was computed be averaging B factors of all its atoms.

3.5.3.2 Contact number exposure

Contact number (CN) is a simple solvent exposure measure that can be computed directly from the 3D structure. The CN value for residue is number of Cα atoms within a sphere of chosen radius around the Cα of that residue [**?**].

The implementation in BioPython module `Bio.PDB.HSExposure` was used for computation, with default sphere radius 12 Å.

3.5.3.3 Half sphere exposure

Half sphere exposure (HSE) is a solvent exposure measure introduced by Hamelryck (2005) [24]. The CN sphere (defined above) around the Cα atom is split in two halves by the plane perpendicular to the Cα-Cβ vector, going through the Cα, as illustrated in Figure 3.3. Two different measures are obtained, HSE-up, which is number of Cα in 'upper' half sphere (containing Cβ), and HSE-down, number of Cα in the opposite sphere.

Class `HSExposureCB` from BioPython module `Bio.PDB.HSExposure` with default sphere radius 12 Å was used.

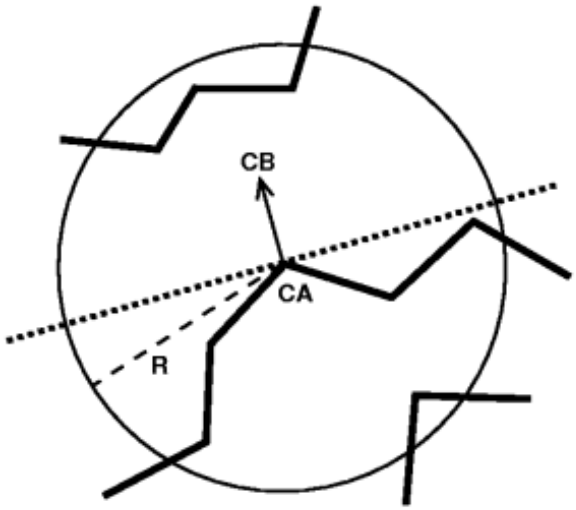


Figure 3.3: Half sphere exposure. Retrieved 02-01-2020 from `https://en.wikipedia.org/wiki/File:HSECa.png`

3.5.3.4 Phi and psi angles

Three dihedral angles of a popypeptide backbone phi (ϕ), psi (ψ) and omega (ω) are depicted on the Figure 3.4. While the ω angle is restricted due to the planar character of the peptide bond, the ϕ and ψ angles have high rotational freedom around the N-Cα (ϕ torsion) or Cα-C (ψ torsion) bonds. The Ramachandran plot provides good visualization of the whole ϕ

ψ space [43]. The angles sizes can be computed directly from the PDBe; for the purpose of this work, they were obtained from PDBe via REST API.

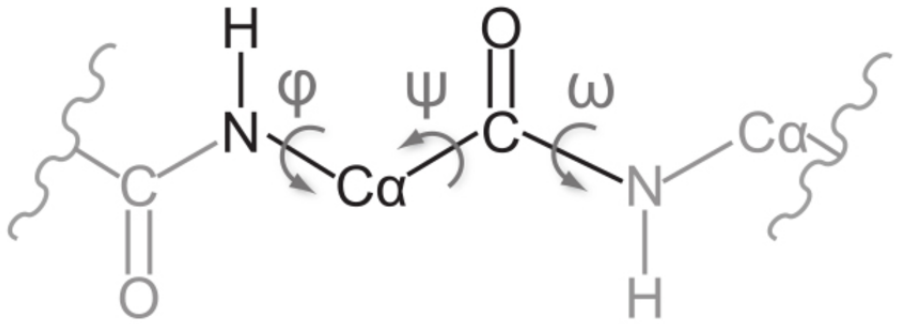


Figure 3.4: Polypeptide torsion angles phi, psi and omega. Retrieved 02-01-2020 from `https://www.researchgate.net/figure/Backbone-torsion-angles-of-a-prototypical-amino-acid-building-block-embedded-i fig2_284713304`

3.5.3.5 Cis peptide

The majority of protein bonds is found with torsion angle ω close to 180°, in so-called *trans* conformation. The *cis* isomer, having ω close to 0°, is rather rare. The *cis-trans* isomeration is involved in some biological processes, such as protein folding or membrane binding [20].

The residues with *trans* bond are obtained from PDBe via REST API.

3.5.4 FASTA

There are features that can be derived directly from the FASTA sequence. Every amino acid is assigned a value and the feature values are obtained according to the FASTA file. These features are:

- **Amino acid** - Categorical feature which is simply the amino acid letter.
- **Hydropathy** - The values of hydropathy index proposed by J. Kyte and R. F. Doolittle [33]. It takes into consideration hydrophilic and hydrophobic properties of the 20 amino acid side chains. It is based on experimental observations derived from the literature. It ranges from -4.5 (Arg) to 4.5 (Ile) and the larger the number is, the more hydrophobic the amino acid.
- **Molecular weight** - Residue mass in Daltons.
- **Polarity** - Classification of amino acids according to the side chain - categories Polar, Nonpolar and Polar uncharged.

- **Charge** - Binary feature indicating whether the side group is charged in physiological pH.
- **Aromaticity** - Binary feature labeling residues that contain aromatic ring.
- **Hydrogen bond atoms** - Number of atoms of the side chain that are either hydrogen donor or hydrogen acceptor.

The biochemical properties of amino acids (all features above except hydropathy) were obtained from TODO citovat biochemie voet

3.5.5 Other resources

3.5.5.1 MobiDB

MobiDB is a database of protein disorder and mobility annotations. It provides annotations and predictions for intrinsically disordered (ID) proteins. MobiDB-lite is a method for highly specific predictions of long (at least 20 residues) disorders. It is a consensus-based prediction, combining results of eight different predictors [40]. It has been integrated into the MobiDB and the web server provides programmatic access to retrieve single entries via REST API [4].

3.5.5.2 P2Rank

The sequence conservation scores for the feature `conservation` are computed through Conservation pipeline [2] implemented for P2Rank. The pipeline runs on local computer and needs to have SwissProt, UniRef90 and TrEMBL databases downloaded locally.

3.5.6 User-defined features

There are two ways to run the statistical analysis with user-defined features. The first, more time-demanding way is to compute all requited inputs (i.e. feature values and ligand binding sites labels) outside of the pipeline, and then run only the analysis task (with argument `-tasks A`). More details about the usage and a few examples are described in the README file in the GitHub repository of the project (`https://github.com/katebrich/LBS_analysis_pipeline`).

Another, more straightforward possibility is to implement the new feature directly inside the pipeline. Two steps need to be made:

- to add the feature to the config file. The class with feature implementation is loaded dynamically according to the feature name, for the easier definition of new features,

- to implement class with method `get_values` which computes the feature values and returns them in the required format.

3.6 Statistical analysis

To find the features which are possibly important for prediction of protein-ligand binding sites, statistical analysis has the crucial role. This section describes the method that was used to analyse the statistical significance of the features and to distinguish the ones that stand out in the known protein-ligand binding sites.

In this work, the problem is seen as a hypothesis testing problem. Two populations will be compared: we take values of a feature for all the residues across all the proteins in the dataset and then compare the values associated with the binding residues and non-binding residues.

The *null hypothesis* and the *alternative hypothesis*, denoted by H_0 and H_1, respectively, will be tested:

- H_0 - The feature values in binding sites do not significantly differ from the values in non-binding sites.
- H_1 - There is a significant difference of feature values in binding sites and non-binding sites.

To decide which one of two complementary hypotheses is true, we employ a suitable *hypothesis test*. Welch's test and Chi-squared test of indepence, both described below in more detail, will be used according to the feature type (binary, categorical or continuous).

As one may expect, the tests are not error-proof and a mistake can be made in the decision of whether to accept or reject the null hypothesis. There are two types of errors in hypothesis testing, commonly known as *Type I error* and *Type II error*. The test has made a Type I error if it incorrectly rejects a true null hypothesis. If, on the other hand, a null hypothesis is accepted and it is not true, a Type II error has been made. Both situations are depicted in the Table 3.2. The ideal test would have both error probabilities equal to zero. Nevertheless, in most cases it is not possible to make both error probabilities arbitrarily small for a fixed sample size [13].

To control statistical significance of the result, we define a *significance level*, a constant denoted by α. It represents the probability of making a Type I error, in other words, the probability that the study rejects the null hypothesis when it is true. The typical choices in practice are $\alpha = 0.01$, 0.05 or 0.10 [13]. One should be aware that by fixing the significance level of the test, the experimenter

		Prediction	
		Accept H_0	Reject H_0
Truth	H_0	Correct (true positive)	**Type I error** (false positive)
	H_1	**Type II error** (false negative)	Correct (true negative)

Table 3.2: Type I and II Error in hypothesis testing.

is controlling only the Type I error probabilities. The probability of the Type II error is subject to factors such as the accuracy and completeness of the data and most importantly, the true effect size [47]. Our choice will be $\alpha = 0.05$.

The *P value* is reported as a result of the statistical test. The P value is the probability that, under the assumption the null hypothesis is true, we observe the same or greater difference between groups. Smaller values of $p(X)$ give stronger evidence for rejecting the null hypothesis. The null hypothesis is rejected when $p(X) \leq \alpha$. P value gives an idea of how strongly the data contradict the null hypothesis; furthermore, it allows other researchers to make a decision according to the significance level of their choice [11, 23, 47].

3.6.1 Implementation

The following subsections rationalize the choice of Welch's and chi-squared test. The imlementation of these tests in module `scipy.stats` from the SciPy Python library [**?**] was used in the pipeline; namely `ttest_ind` with `equal_var=False` to perform Welch's test, and `chi2_contingency` for chi-squared test of independence.

By default, the pipeline computes the analysis for all the residues across all proteins in dataset. Neverthless, random sampling (without repetition) can be performed by specifying the sample size with argument `-s`. It is also possible to run more iterations of random sampling (argument `-i`). In this case, mean P values will be reported in summary. Individual P values from all iterations will be reported in separate files for each feature. Another possibility is to balance the number of binding and non-binding sites with argument `-b`. Same number of binding and non-binding residues will be sampled in that case.

The significance level is 0.05 by default and can be changed with `-a` argument.

The output of the whole analysis pipeline are folders with results for each feature, as well as several summary files:

- **p_values_means.csv** - Averaged P values obtained from all iterations.
- **p_values.csv** - List of P values from all iterations for all features. P values for individual features are also included in the results folder for each feature.
- **p_vals_perc.csv** - Summary of how many percent of iterations had P value below given significance level α for each feature.
- **means_difference.csv** - Summary only for continuous features. Lists the differences of means and variances in both populations (all binding vs. all non-binding).
- **binding_ratios.csv** - binding/non-binding sites ratios for all iterations and all features.
- **errors.txt** - Lists features that ended up with an error. This is often caused by lack of data (e.g. the sample size is bigger than number of rows or the data for a categorical feature are too sparse to meet the assumptions of Chi-squared test). Detailed information about the errors can be found in the log file.

The results folder for each feature contain file `pairs.txt` with paired ligand binding sites values and feature values, detailed information about all iterations, and various histograms and plots (according to the type of feature).

3.6.2 Welch's test

Welch's unequal variances t-test, or Welch's test in short, is a two-sample hypothesis test used to decide whether two populations have different central tendencies (means or medians). The decision is made based on the samples from the two populations. It is a more robust alteration of the widely-used Student's t-test [54].

Both Student's and Welch's t-test assume that the two examined populations follow a normal distribution [54]. Nevertheless, when testing for the equality of means of 'large enough samples', the normality assumption can be violated thanks to the large sample theory and the Central Limit Theorem [23]. It has been shown in previous studies that for large samples, the statistical significance level is protected not only for normally distributed data, but also for many non-normal distributions; moreover, in case of Welch's test, this is true even for unequal variances [36, 59, 60]. According to Lehmann and Romano [23], the Type II error is also relatively insensitive to non-normality. Many articles and textbooks mention that when the sample sizes are small, nonparametric tests (i.e. tests that

do not assume a specific distribution) such as the Mann-Whitney test [**?**] should be considered as an alternative to t-tests. However, t-tests become superior when sample sizes increase [36, 57]. The simulations made by Lumley *et al.* [36] show that 'sufficiently large sample size' means under 100 in most cases. Even for extremely non-normal data, the sufficient size is at most 500. This suggests that the choice of Welch's test is legitimate for this work.

The problem of the Student's t-test is that it performs badly when the variances of the two compared populations are unequal. Both Type I and Type II errors are negatively affected by violation of the equal variances assumption. The unequal variances can be less problematic if sample sizes are similar, but in practice, that is not always the case [45].

Unlike Student's t-test, Welch's test does not assume equal variances of the populations. It performs well when the samples have unequal variances; furthermore, it can be used even when the samples have unequal sizes [22].

Some researchers tend to pre-test for variance equality by a preliminary test of variances (such as Levene's [34], Bartlett's [**?**] or Brown-Forsythe test [**?**]) and then choose whether to use Student's or Welch's t-test. However, although this approach persists in some textbooks and software packages, it is not recommended by statisticians. As a preliminary test itself is subject to Type I and II errors, this two-stage procedure would not protect the significance level and could lead to incorrect decisions. One should be aware of the fact that even if the test suggested that the samples variances are nearly equal, it would not mean that the whole population variances could not differ to a larger extent [58]. Some researchers may try to make the significance level of a preliminary test more strict, so that they could be more confident about the choice of the subsequent test; however, as the significance level decreases, the performance of the compound test paradoxically gets worse. According to Zimmerman [58], "a higher Type I error rate of the preliminary test actually improves the performance of the compound test" [58]. This suggests that using the preliminary test is not correct in principle.

Welch's test should be used whenever the researcher is not sure that the variances are truly equal. Ruxton [45] even suggests the routine use of Welch's test. When the sample sizes and variances are equal, both tests perform similarly. When dealing with unequal variances and unequal sample sizes, Welch's test is more robust than Student's t-test and the Type I error rate does not deviate far from the nominal value [22]. Hence, Welch's test can be applied without any significant disadvantages to Student's t-test.

For all the reasons stated above, Welch's test seems to be the best choice for the purpose of this study. It has the best combination of performance and ease of use, the calculation is straightforward and it is available in commonly used

statistics packages. This test will be used for continuous features.

3.6.3 Chi-squared (χ^2) test of independence

A different kind of tests will be needed for the analysis of categorical and binary features. In this section, the χ^2 test will be compared to another well-known test for the analysis of data in contingency tables, the Fisher's exact test.

A *contigency table* is a table displayed in a form of a matrix where cells represent a frequency distribution of samples in the categories. An example of a contingency table can be seen in Table 3.3. The sums of frequencies in rows and columns are called *marginal totals*.

	Aromatic residue	Non-aromatic residue	Total
Binding sites	1016	4654	5670
Non-binding sites	4829	44545	49374
Total	5845	49199	55044

Table 3.3: A 2×2 contingency table for binary feature `aromaticity` computed on dataset Chen11.

The null hypothesis assumes independence of the groups; in our case, the assumption is that there is no difference in the proportions of the analysed feature between binding sites and non-binding sites.

Fisher's exact test belongs to a class of so-called *exact tests*; it means that the P value is calculated accurately, not approximately, as is the case of many tests including Welch's test and χ^2 test. Fisher's test is mostly used for 2×2 contigency tables, although the principle of the computation can be extended to a general $m \times n$ table [37]. The principle of the test lies in computing the probability of obtaining a table that is more or equally extreme in the departure from the null hypothesis than the analysed table and has identical marginal totals [12].

Chi-squared test if independence is able to decide whether the difference between the observed frequencies and the 'expected frequencies' is statistically significant. The expected frequencies are computed for every cell using this formula:

$$\frac{row\ total \times column\ total}{grand\ total}$$

It can be imagined as the average frequencies we would get in the long run with the same marginal totals, assuming the null hypothesis is true (i.e. there is no association between groups). The result of the test tells how likely are we to observe given data under the assumption of the true null hypothesis [12].

The biggest difference between the two mentioned tests is that the chi-squared test is based on a aproximation approach; therefore, it needs a 'large enough' sample. W. G. Cochran (1952, 1954) proposed a set of recommendations about the minimum expectations to be used in χ^2 tests and about the choice between Fisher's test and χ^2 test:

- **The 2 x 2 table** - Fisher's exact test should be used whenever the sample size is smaller than 20, or when the sample size is smaller than 40 and if the expected frequency in at least one cell is less than 5. For sample sizes bigger than 40, always use chi-squared test [16, 17].
- **More than 1 degree of freedom** - Chi-squared test can be used when at most 20% of cells have the expected frequency less than 5 and no cell have the expected frequencies less than 1 [**?**].

These recommendations are presented in several textbooks and articles as a rule of thumb [**?**] and recommended to be used in practice.

As the sample sizes in this work are very large, the number of binding and non-binding sites is unbalanced, and the data for some features can be sparse, chi-squared test should be better choice for both binary and categorical features.

3.7 P2Rank models training and evaluation

Two scripts further extend the analysis pipeline and can be used to train P2Rank models with obtained data. Two dataset files are needed as input - one for training and another for evaluation. The process is following:

- for both datasets, run the whole pipeline twice:
  - in the first run, download data, compute mappings, ligand binding sites, features and analysis with default parameters
  - in the second run, recompute the analysis with random sampling (sample size 500, 1000 iterations)
- convert both dataset files to the format accepted by P2Rank
- create .csv files with custom features using previously computed feature values
- train and evaluate new P2Rank model on given datasets with custom features

It is possible either to train one model with all given features at once with script `pipeline_P2Rank_allFeatures.sh`, or to train one model per feature with `pipeline_P2Rank_oneFeature.sh`

4. Evaluation and results

4.1 Datasets

The choice of datasets of protein-ligand complexes used for statistical analysis and P2Rank model training and evaluation was strongly inspired by the datasets described in the P2Rank article [32]. The structures were re-downloaded directly from PDBe, according to their PDB ID (four-character alphanumeric identifier) and chain ID (one-character identifier) used in the original datasets. It was not possible to take the original datasets as they were, since the structures were not up-to-date and the annotations downloaded from the databases (e.g. feature values) could not be mapped properly.

Downloaded datasets were further checked and filtered: Obsolete structures were replaced with their current entries, structures that do not have a corresponding UniProt record were removed, as well as structures with the incorrect segments mapping due to the bug in PDBe (mentioned in section TODO-ODKAZ). The pipeline can only work with single-chain structures, and the structures from Holo4k and a few structures from Joined were multi-chain; thus, only one chain was chosen from each such structure.

The resulting datasets were named identically with the original datasets:

- **Chen11** - a smaller non-redundant dataset that was originally designed for a comparative study of ligand binding sites predictors [14]. It comprises at most one representative chain for every SCOP family [39] to ensure the minimal sequence similarity and maximal variability in tertiary structure. The original dataset covers 6 structural classes, 148 protein folds, 184 superfamilies and 251 families [14]; after re-downloading and filtering, the numbers are slightly smaller. Although this dataset is rather small, it covers wide range of non-homologous proteins. Therefore, it serves as good training dataset (P2Rank default model was trained on this dataset as well).

- **Coach420** - a dataset that was originally taken from a benchmark study [44] and used in other studies [32, 55]. The non-redundant dataset harbor mix of natural and drug-like ligand molecules.

- **Joined** - smaller datasets from previous studies merged together in one larger dataset. It comprises a set of drug-target complexes extracted from DrugBank, DrugPort and PDB DT198[56], a benchmark set for the validation of protein-ligand docking performance [25], and a dataset with bound

and unbound structures used for evaluation of a ligand binding sites predictor [26].

- **Holo4k** - a large set of protein-ligand complexes used in a large-scale evaluation of four binding sites predictors [46].

4.1.1 Ligands filtering

The downloaded PDB files contain more ligands per structure, and not all of them are of interest for drug design and other applications. These non-relevant ligands can be ions, peptides, small molecules such as solvents, buffers, detergents and salts that are merely artifacts, and other specific types of ligands.

For each dataset, three variants were created by different filters of the relevant ligands:

- **No filter** - Only water molecules were filtered out.
- **P2Rank filter** - Relevant ligands were obtained according to the rules used by P2Rank software [32]. These rules are:
  - the ligand has at least 5 atoms
  - at least one atom of the ligand is in distance 4 Å from any protein atom
  - the center of the mass of the ligand is not farther than 5.5 Å from the closest protein atom
  - the name of ligand PDB group is not any of following: HOH, DOD, WAT, NAG, MAN, UNK, GLC, ABA, MPD, GOL, SO4, PO4
- **MOAD filter** - Biologically relevant ligands according to the Binding MOAD [**?**] database. It contains manually curated crystalography protein-ligand complexes with validated biologically relevant ligands. Only structures obtained by X-ray crystalography with resolution higher than 2.5 Å have entry in Binding MOAD; other structures were removed from the datasets.

The structures without any remaining ligands after applying the P2Rank or MOAD filters were removed.

The summary of datasets properties can be seen in Table 4.1. The more strict the filter is, the lower the Binding/Non-binding ratio; nevertheless, the information obtained from the relevant binding sites should be more valuable. As we can see, MOAD filter is more strict and filteres out more ligands than P2Rank filter.

Dataset	Proteins	Ligands	Lig./Pro.	Binding	Non-bind.	B/N ratio
chen11	241	1039	4.3112	5670	49374	0.1148
chen11_filter_p2rank	223	401	1.7982	4590	47073	0.0975
chen11_filter_MOAD	178	266	1.4944	3032	39006	0.0777
coach420	417	841	2.0168	5988	80575	0.0743
coach420_filter_p2rank	369	427	1.1572	5247	71498	0.0734
coach420_filter_MOAD	258	291	1.1279	3688	48485	0.0761
joined	527	1522	2.888	8260	108337	0.0762
joined_filter_p2rank	446	585	1.3117	6492	97158	0.0668
joined_filter_MOAD	348	417	1.1983	4614	72363	0.0638
holo4k	3973	10391	2.6154	69866	790091	0.0884
holo4k_filter_p2rank	3842	5049	1.3142	62483	784885	0.0796
holo4k_filter_MOAD	3308	4023	1.2161	50834	679918	0.0748

Table 4.1: Summary of dataset properties with and without ligands filtering. Chen11 dataset has the highest average number of ligands per protein, but when the ligands are filtered, the number is comparable to the other datasets. It indicates that Chen11 has the highest ratio of biologically irrelevant ligands.

4.2 Statistical analysis

TODO vymenit grafy za ty nove, hezci, lepci!

The statistical analysis of ligand binding sites properties was computed using the analysis pipeline described in TODO section 3 with default parameters. The results were collected for all the datasets, including the versions with filtered ligands. Let's set the significance level, denoted by α, to 0.05.

Some features had to be excluded from the analysis, since the data were very sparse and the assumptions of the hypothesis tests would not be met. For example, there were only 15 lipidation sites in the whole holo4k dataset containing 857,635 residues. The excluded features are: `lipidation`, `glycosylation`, `non_standard` and `compbias`

The `conservation` feature was computed only for the three smaller datasets and was omitted for holo4k. The computational time would be very high, as it takes 15-30 minutes on average per structure, and the dataset contains almost four thousand proteins. Nevertheless, the comparison on the other three datasets seems sufficient.

The problem with feature `variation` was that the data were missing for many structures (around 3/4) and downloading via REST API resulted in 404 Not Found error. Data are not available on the UniProt website either. This might be caused by lack of variation data from large-scale studies for some organisms. UniProt helpdesk was contacted to help to explain the issue, but unfortunately, the question was left without an answer. Nevertheless, the feature was analysed on the subset of structures where the data is available.

For some features downloaded from databases, such as `depth` or `dynamine`,

there were missing data for a few structures as well. These cases were not very frequent and they most likely could not affect the analysis, so they were omitted.

Three artificial features were added for comparison and to check the validity of the tool:

- `lbs` - Ligand binding sites labels (0/1). Should have the best performance of all the features, the P-value should be zero.
- `random_binary` - Random binary numbers. Should not be significant.
- `random_cont` - Random continuous feature with values from uniform distribution from 0 to 10.

The results for datasets without ligands filter are shown in Table 4.2. As we can see, most features appear to be statistically significant, having the P-value below the significance level α. The results for the test features `lbs`, `random_binary` and `random_cont` seem to be okay. However, when looking at the histograms and plots, some results are not as expected. Let's take a look at the histogram depicted in Figure 4.1: the distribution of `dynamine` values does not seem significantly different in binding and non-binding sites. Note that for better comparison of binding and non-binding sites (since their ratio is very unbalanced), the density is computed with respect to the number of binding or non-binding sites; the value in the histogram bin can be understood as conditional probability of getting that value when having a binding/non-binding residue.

One conspicuous thing about the Table 4.2 is that, in general, the P-values are getting smaller as the dataset size grows (the datasets in the table are sorted from the smallest on the left to the largest on the right). This is referred to as the *P-value problem*. For very large samples, the statistical power of hypothesis tests is higher, and causes P-value going to zero. When dealing with large samples, even the miniscule effects can become statistically significant. The test can detect subtler and more complex effects, which can be advantageous in some cases, but also misleading. It all depends on the purpose of the statistical testing. The question we should ask is not whether the results are statistically significant (which there almost always will be for large samples), but whether they are interesting for our research [35].

The P-value itself does not have an objective meaning and is not an unambiguous measure of evidence. The sample size hugely influences the significance, and relying only on the P-value can lead to acceptance of the hypothesis of no practical significance. Despite that, this appears to be a common practice. Lin et al. [35] reviewed articles in two leading Information System (IS) journals and

	Chen11	Coach420	Joined	Holo4k
lbs (test)	0	0	0	0
conservation	0	0	0	—
pdbekb_conservation	0	0	0	0
HSE_up	1.48E-266	0	0	0
exposure_CN	2.08E-240	0	0	0
depth	1.63E-225	3.37E-244	0	0
bfactor	6.57E-176	1.02E-172	9.03E-280	0
aa	2.43E-141	4.01E-118	2.09E-224	0
mol_weight	2.54E-141	1.77E-117	2.71E-225	0
HSE_down	4.23E-139	6.18E-225	0	0
hydropathy	9.39E-136	1.83E-118	9.79E-222	0
aromaticity	5.53E-79	3.97E-56	1.79E-102	0
H_bond_atoms	3.59E-44	2.89E-36	4.50E-72	0
strand	2.11E-17	7.37E-32	7.58E-36	2.02E-252
sec_str	2.88E-16	1.57E-45	4.91E-42	0
helix	5.59E-06	3.28E-29	1.19E-26	5.83E-279
phi_angle	9.89E-06	4.29E-05	1.07E-07	1.36E-42
mobiDB	0.0006394	0.007984	4.54E-06	5.98E-51
PTM	0.007131	5.29E-05	1.77E-15	1.15E-104
psi_angle	0.009603	7.71E-16	0.0009644	1.30E-27
charged	0.009871	4.38E-13	2.99E-05	2.53E-159
dynamine	0.0143	0.02082	0.1595	1.70E-05
efoldmine	0.01699	0.002727	1.06E-09	5.02E-07
polarity	0.02564	9.01E-13	3.11E-06	4.18E-159
variation*	0.1513	0.07348	0.698	0.05166
cis_peptide	0.2373	0.0001902	4.44E-06	3.46E-45
disulfid	0.2753	1.82E-06	0.5603	1.71E-33
natural_variant	0.2793	0.02171	2.14E-07	3.39E-24
mod_res	0.3116	0.002696	9.69E-05	2.72E-49
random_cont (test)	0.4707	0.706	0.99	0.1021
random_binary (test)	0.5429	0.922	0.3561	0.9322
turn	0.8949	0.003081	0.7883	0.006317

Table 4.2: P-values returned by hypothesis tests for individual features for all four datasets (without ligands filtering). Features are sorted according to the P-value in the first column. Values highlighted with red colour are higher that the chosen significance level $\alpha = 0.05$.
*`variation` is computed only on the subsets of proteins for which the data were available in databases.

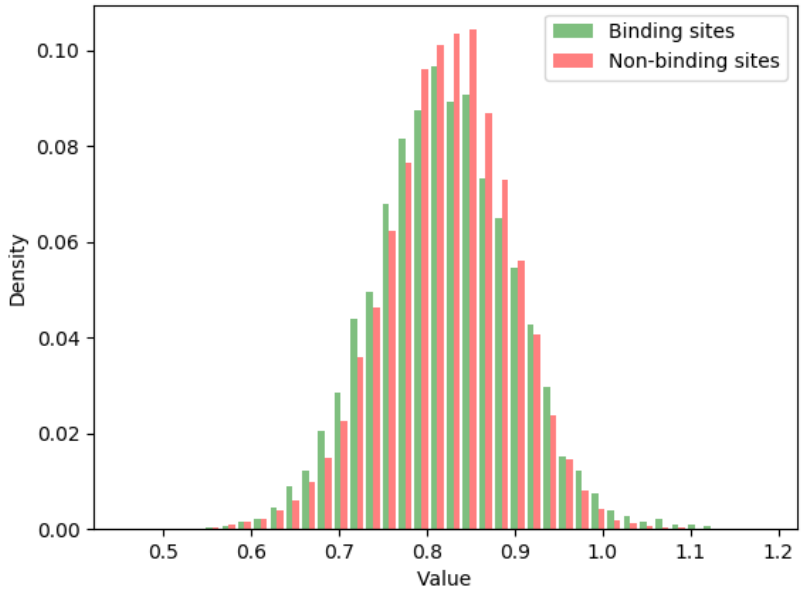


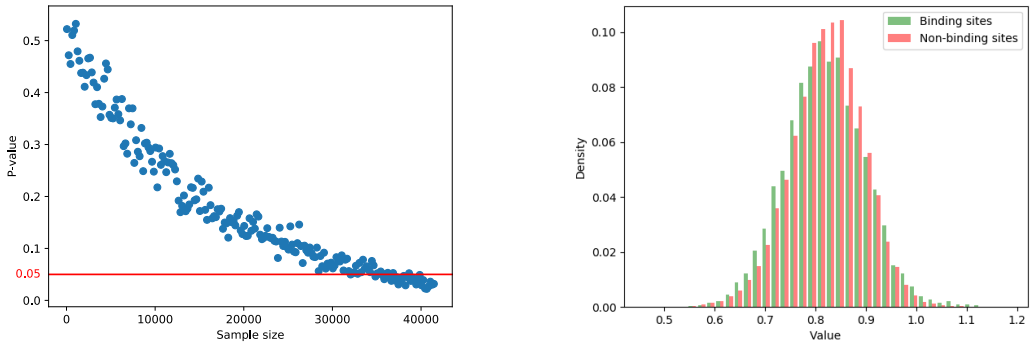
Figure 4.1: Histogram of feature `dynamine` computed on holo4k dataset. Density on the y-axis is computed with respect to the number of binding or non-binding sites. Difference in means: 0.0014; difference in variances: 0.0015.

reported that 50% of recent papers with sample sizes over 10,000 were relying on low P-values.

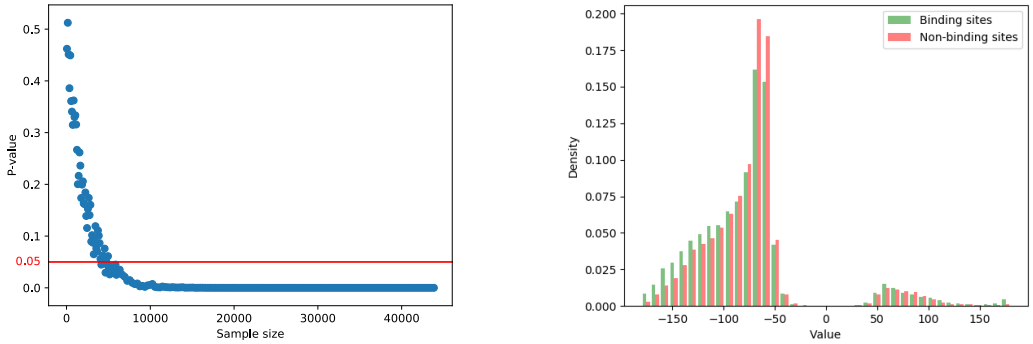
Let's see the P-value problem demonstated on our data. Figure 4.2 shows different speeds of P-value deflation for chosen features. At the first glance, the distributions of feature `exposure_CN` in binding and non-binding sites differ, and sample size 25 is sufficient to get the P-value below significance level 0.05. On the other hand, `dynamine` does not seem to be relevant for the binding sites recognition, and yet, if the sample size is large enough, we get the significant result.

Therefore, the low P-values reported in Table 4.2 are most likely mere artifacts of the large-sample sizes. Neverthleless, although P-value is not an objective measure of practical significance, it can be still used to compare the features relative to each other.

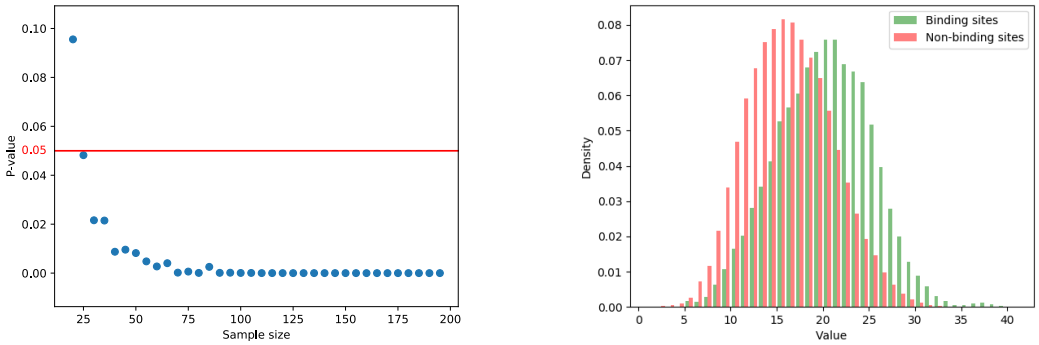
Another noticeable thing about Table 4.2 is that the results for some features vary across datasets. Let's take a look at feature `turn`, for example. The P-value is very high for datasets Chen11 and Joined - even higher than P-values for the random features; contrarily, it is low for Coach420 and Holo4k. It is not true that the P-value would decrease with the increasing sample size. This leads to a question of how the datasets are composed, and whether they are representative samples from the whole population of proteins. Taken into consideration the



(a) `dynamine`



(b) `phi_angle`



(c) `exposure_CN`

Figure 4.2: P-value deflation demonstrated on chosen features. The P-value decreases with increasing sample size. The speed of deflation is different for individual features. The y-axis shows mean P-values obtained from 100 iterations of random sampling with given sample size. The red line represents chosen significance level $\alpha = 0.05$.

way how the datasets were assembled, it is likely that some bias was introduced. The question is whether taking the whole PDB database would help to solve this issue. There probably would be the problem with redundancy of data, as close homologs and overlapping PDB entries would be included. Furthermore, the database itself is most likely a biased sample of the real world of proteins, as the tertiary structure is yet to be discovered for many of them. And most importantly, this approach would be computationally very demanding.

For the mentioned reasons, a different approach was implemented. Dataset 'mix' was created by merging all four datasets together, removing a few duplicates. Random sampling without replacement was applied on this dataset, in each iteration taking a sample of 500 binding and 500 non-binding sites. 1000 iterations were computed and mean P-values were reported. The results are shown in Table 4.3. Note that the mean P-value cannot be understood in the original meaning of P-value. Nevertheless, the numbers provide relative comparison of the features.

The sample size of 500 was chosen for two reasons: firstly, validity of the Central Limit Theorem needs to be assured, as described in section TODO. Lumley *et al.* [36] demonstrated that 500 is a sufficiently large sample even for extremely non-normal data. And secondly, the minimum sample size assuring the Central Limit Theorem validity should be chosen, to avoid the P-value problem. Smaller sample size would probably be sufficient for the Central Limit Theorem, as 500 is a very safe estimation. Nevertheless, the sample size could not be much smaller anyhow, since the data for some categorial features would be very sparse. Even with the sample size of 500, some features needed to be excluded from the analysis, as there was not sufficient number of positives in this smaller sample. TODO vyjmenovat je

TODO taky vyhoda, ze nebudou zavislosti mezi rezidui

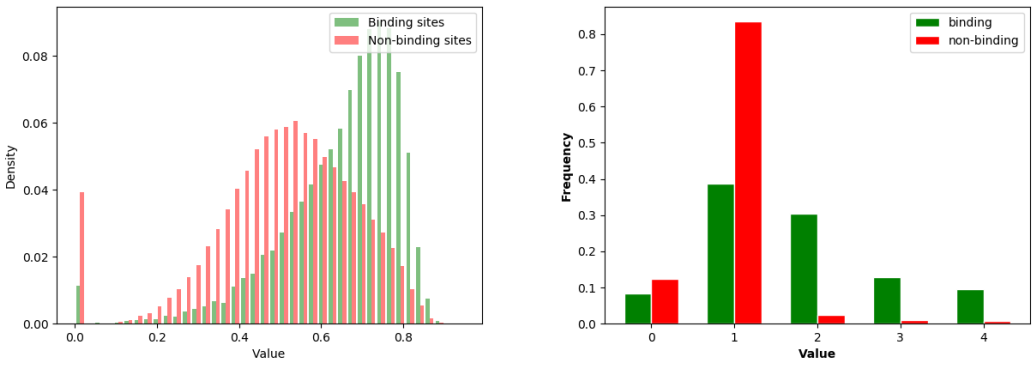
The results for the datasets with different ligands filters does not seem to differ widely and does not reveal any additional informations. For that reason, following features analysis and computed plots will be shown only for the dataset 'mix' with P2Rank ligands filter. The results obtained from the analysis pipeline for all the datasets and various sample sizes are included in the Attachments. TODO odkaz

The results for both conservation features `conservation` and `pdbekb_conservation` turned out as expected. Sequence conservation has been used previously in many approaches for protein-ligand binding sites prediction and its importance for the prediction has been repeatedly demonstrated [21, 26, 28, 53]. The higher values of conservation for binding residues are clearly visible from the Figure 4.3.

Features `HSE_up`, `HSE_down`, `exposure_CN` and `depth` are closely related to the

	no filter	P2Rank filter	MOAD filter
lbs (test)	1.33E-218	1.33E-218	1.33E-218
pdbekb_conservation	3.55E-27	1.35E-30	4.43E-36
conservation	1.11E-17	1.05E-27	7.86E-33
exposure_CN	4.72E-17	1.95E-21	1.30E-22
HSE_up	1.15E-14	2.15E-18	1.38E-18
depth	8.00E-14	9.13E-16	2.83E-16
HSE_down	1.48E-09	1.59E-11	2.28E-11
bfactor	2.56E-06	3.03E-08	3.97E-08
aa	0.006394	0.001172	—
mol_weight	—	0.00129	0.002037
hydropathy	0.00539	0.001376	0.001953
aromaticity	0.02027	0.01516	0.02523
H_bond_atoms	0.08081	0.02502	0.03019
charged	0.2683	0.06663	0.08965
polarity	0.2755	0.07131	0.1009
sec_str	0.133	0.0873	0.02696
strand	0.1491	0.1112	0.04838
helix	0.1361	0.1154	0.02435
mobiDB	0.3971	0.3844	0.3653
phi_angle	0.3973	0.399	0.3864
psi_angle	0.4213	0.4317	0.2875
efoldmine	0.4769	0.4373	0.4839
dynamine	0.4937	0.484	0.4208
random_cont (test)	0.5029	0.5021	0.4887
variation	0.5387	0.5283	0.5395
random_binary (test)	0.5374	0.5309	0.5223
turn	0.5785	0.5982	0.598

Table 4.3: Mean P-values computed from 1000 iterations of random sampling with sample size 500. Computed on dataset mix (4 datasets merged together) with three variations of ligands filtering. Features are sorted according to the P-value in the second column. Some values are missing because the assumptions of the test were not met.
*`variation` is computed only on the subsets of proteins for which the data were available in databases.



(a) `conservation` (b) `pdbekb_conservation`

Figure 4.3: Higher values of conservation for binding residues demonstrated on two features: (a) continuous feature computed by the conservation pipeline, and (b) ordinal feature downloaded from PDBe-KB database.

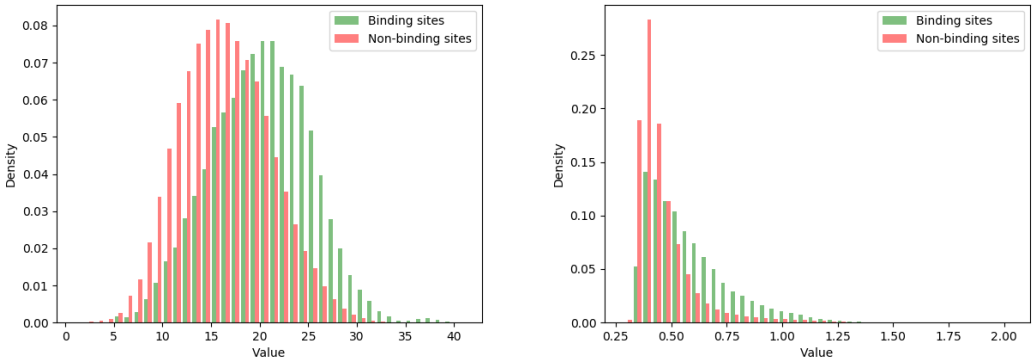
'buriedness' of the residue. Similarly as conservation, this feature was expected to be important for ligand binding sites recognition. Many binding sites are shaped as cavities, or concave pockets, on the surface of the 3D structure. The geometrical methods, such as LIGSITE [26] or PocketPicker [**?**], as well as other approached, make use of this property. The histograms for these continuous features are depicted in Figure 4.4

The analysis reveals that binding sites have on average slightly lower B factor values (see Figure 4.5). This suggests that binding sites are more well-ordered in general, whereas non-binding sites might have higher flexibility.

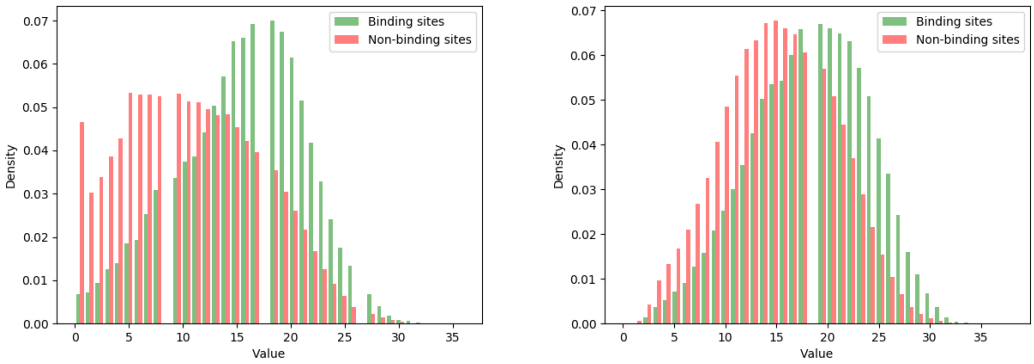
Binding and non-binding sites seem to have different residue composition. Let's take a look at Figure 4.6. Cys, Trp, Phe, Tyr, Gly, His, Met and Ile all have high binding/non-binding ratios, and thus, are more likely to occur in binding sites. On the other hand, Pro, Glu, Gln, Lys and Asp disfavour binding sites. Arg, Val, Ser and Leu are very frequent in binding sites; however, they have the ratios similar to the total binding/non-binding ratio, as they are very frequent on the whole protein surface, not only in binding sites. This result is in accordance with a large-scale study that explored the composition of protein-ligand binding sites [31]. This is an interesting result and higher propensities of some amino acids to appear in binding sites could be used for their prediction.

TODO analyza nejakych dalsich featur?

TODO !!!! It was discovered that high B-factor-characterized regions show a higher average flexibility index, more pronounced average hydrophilicity, and higher absolute net charge. - (11) Radivojac, P.; Obradovic, Z.; Smith, D. K.; Zhu, G.; Vucetic, S.; Brown, C. J.; Lawson, J. D.; Dunker, A. K. Protein flexibility and intrinsic disorder



(a) `exposure_CN` (b) `depth`



(c) `HSE_up` (d) `HSE_down`

Figure 4.4: The features related with buriedness of the residue have higher values in binding sites.

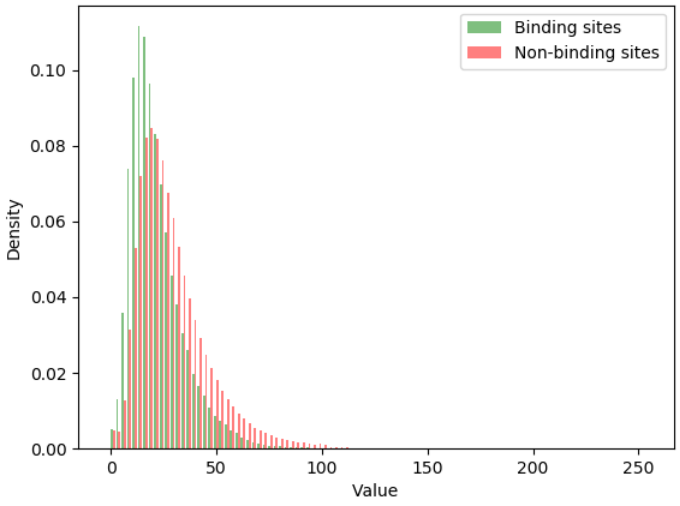
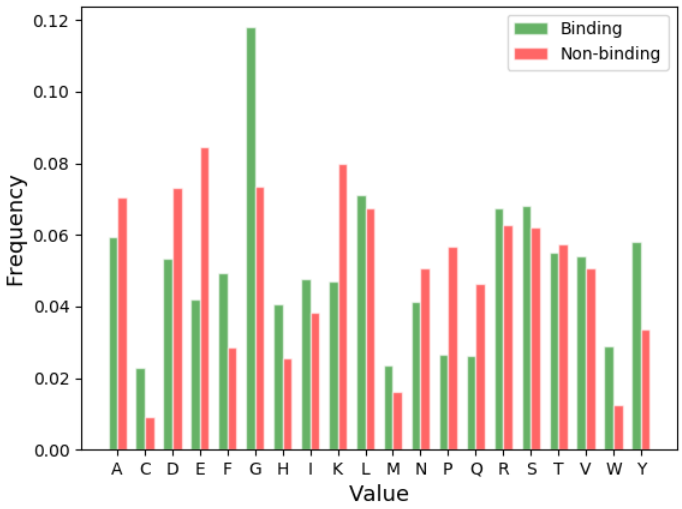
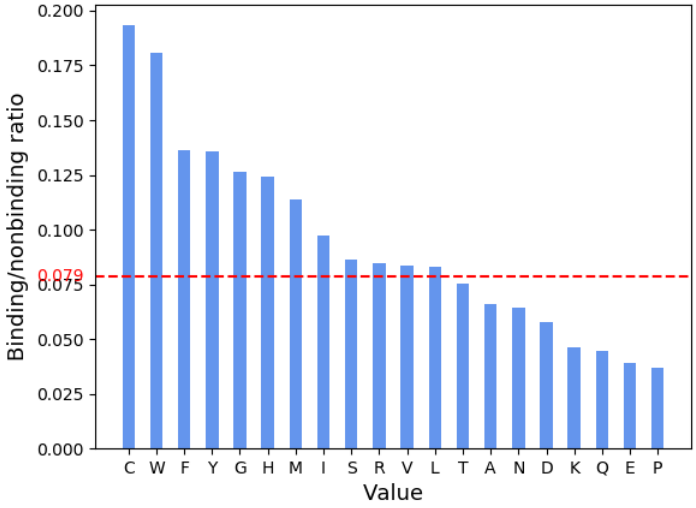


Figure 4.5: `bfactor`: Binding sites have lower B factor values on average.



(a) Frequencies of individual amino acids in binding and non-binding sites.



(b) Comparison of binding/non-binding ratios, computed as occurrences of the AA in binding sites divided by its occurrences in non-binding sites. The red line marks the total binding/non-binding ratio (total number of binding sites divided by total number of non-binding sites). High ratio means that the AA favours binding sites, and on the contrary, the low ratio indicates the tendency to occur in non-binding sites.

Figure 4.6: Feature `aa`.

TODO high-B-factor ordered regions have a higher average flexibility index, a higher average hydrophilicity, a higher average absolute net charge, and a higher total charge than do either short or long disordered regions. The low-B-factor ordered regions are significantly enriched in hydrophobic residues and depleted in the total number of charged residues compared to the other three classes.

4.3 P2Rank models

TODO evaluation metrics

The computed features were used to train new P2Rank models and analyze their practical significance. All models were trained with the same parameters as the default P2Rank model (100 trees, each grown with no depth limit using 6 features). The models were trained on the Chen11 dataset and evaluated on Coach420 dataset. The training was done with parameter `loop=10` which means that training was done 10 times, every time with different random seed, and the performance of the 10 resulting models was averaged at the end. This is important for the comparison of features importance so that the random behaviour of the random forest classifier does not have such big influence.

The features obtained by the analysis pipeline are called 'csv features' in the following section, in accordance with the terminology used by P2Rank for user-defined features in csv files.

TODO ze se nektere featury vynechaly a proc: aa, secstr, variation

The performance of the baseline model (all default P2Rank features and no csv feature) was compared with the model trained on all csv features (without test features `lbs`, `random_binary` and `random_cont`), and the model trained with all P2Rank features and all csv features. The results are summarized in Table 4.4. Our baseline model performs a little better in the Top-n category than the default P2Rank model described in the P2Rank article [32], which achieves the success rate of 72%. This can be caused by slightly different datasets, as described in Section TODO, or simply by the random behaviour of the classifier.

The performance of the model trained on the csv features is inferior to the baseline model, but surprisingly, the difference is very small. The reason probably is that many features used by P2Rank are identical or very similar to the new csv features. P2Rank already used B factor, amino acid properties, buriedness and other properties for training, and csv features evidently does not contribute with much new information. That is visible on the model with both csv and P2Rank features: the performance is superior only by 1.5%. Moreover, csv features are mutually correlated, as well as P2Rank features.

	Top-n	Top-(n+2)
P2Rank features (baseline model)	73.6	78
csv features	70.7	73.9
P2Rank + csv features	75.1	77.5

Table 4.4: Comparison of the performance of models with different sets of features.

	Top-n	Top-(n+2)
lbs (test)	90.2	90.5
pdbekb_conservation	78.9	81.6
conservation	76.8	79.8
HSE_down	74.1	78.6
helix	73.6	78.6
psi_angle	74.3	78.4
depth	73.2	78.4
turn	73.8	78.3
HSE_up	73.6	78.2
strand	73.6	78.2
mol_weight	73.5	78.2
aromaticity	73.6	78.1
dynamine	73.6	78.1
charged	73.5	78.1
H_bond_atoms	73.5	78.1
efoldmine	74.1	78
phi_angle	73.8	78
mobiDB	73.6	78
exposure_CN	73.5	78
hydropathy	73.3	77.8
bfactor	73	77.5

Table 4.5: Performance of models trained with all P2Rank features plus one extra csv feature at a time. The features are sorted according to the performance in Top-(n+2) category.

Let's take a look at how the csv features help to improve the performance when adding only one of them at a time. Table 4.5 summarizes the results of training one model per feature, with all P2Rank default features plus given csv feature. Models with features `pdbekb_conservation` and `conservation` are clearly superior to the baseline model. There are other features that perform better than the baseline model by tenths of percent; nevertheless, this difference is too small to proclaim the results significant. Although there are features that are enriched in binding sites, as has been shown previously, they do not help to improve P2Rank performance, probably due to the correlations and recurrence of the same features.

In another experiment, all P2Rank features were switched off and the models

	Top-n	Top-(n+2)
lbs (test)	88.3	90
pdbekb_conservation	57.1	71.9
HSE_up	36.7	52.1
conservation	34	49.7
exposure_CN	30.8	44.7
HSE_down	28.2	41.8
depth	25.5	37.5
aromaticity	14.3	25.5
dynamine	10.6	17.7
phi_angle	8.7	16.3
hydropathy	10.7	16
mol_weight	8.4	15.7
random_cont (test)	8.2	14.1
efoldmine	7.2	13.5
strand	6.3	10.9
random_binary (test)	7	10.7
psi_angle	5.8	9.8
bfactor	4.2	6.1
helix	3.9	6
charged	3.6	5
H_bond_atoms	3.3	4.9
mobiDB	2.4	4.7
turn	0	0

Table 4.6: Performance of models trained only on csv feature at a time (without the default P2Rank features). The features are sorted according to the performance in Top-(n+2) category.

were trained with csv features only, one at a time. The performances are of course wery poor, but it gives us the relative comparison of csv features, without the effect of correlation with existing P2Rank features. Table 4.6

Conclusion

Conclusion.

List of Abbreviations

AA	Amino acid
atd	a tak dale

Bibliography

[1] Bio2byte group. URL `https://bio2byte.be/info/`. Accessed: 20.11.2020.

[2] Sequence conservation pipeline. URL `https://github.com/cusbg/sequence-conservation`. Accessed: 5.12.2020.

[3] URL `https://www.uniprot.org/help/disulfid`. Accessed: 15.11.2020.

[4] Mobidb api documentation. URL `https://mobidb.bio.unipd.it/help/apidoc`. Accessed: 2.12.2020.

[5] Todo. URL `https://www.uniprot.org/help/mod_res`. Accessed: 15.11.2020.

[6] URL `https://www.uniprot.org/help/variant`.

[7] URL `https://www.uniprot.org/help/non_std`. Accessed: 15.11.2020.

[8] Pdbe rest api. URL `https://www.ebi.ac.uk/pdbe/pdbe-rest-api`. Accessed: 5.11.2020.

[9] Biopython api documentation. URL `https://biopython.org/docs/dev/api/Bio.PDB.SASA.html`. Accessed: 10.11.2020.

[10] URL `https://www.uniprot.org/docs/userman.htm`. Accessed: 15.12.2020.

[11] Chittaranjan Andrade. The p value and statistical significance: Misunderstandings, explanations, challenges, and alternatives. *Indian Journal of Psychological Medicine*, 41(3):210–215, may 2019. doi: 10.4103/ijpsym.ijpsym_193_19.

[12] Martin Bland. *An introduction to medical statistics*. Oxford University Press, Oxford New York, 1987. ISBN 0192615025.

[13] George Casella and Roger L Berger. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

[14] Ke Chen, Marcin J. Mizianty, Jianzhao Gao, and Lukasz Kurgan. A critical comparative assessment of predictions of protein-binding sites for biologically relevant organic compounds. *Structure (London, England : 1993)*, 19:613–621, May 2011. ISSN 1878-4186. doi: 10.1016/j.str.2011.02.015.

[15] Elisa Cilia, Rita Pancsa, Peter Tompa, Tom Lenaerts, and Wim F. Vranken. From protein sequence to dynamics and disorder with dynamine. *Nature communications*, 4:2741, 2013. ISSN 2041-1723. doi: 10.1038/ncomms3741.

[16] William G. Cochran. The χ^2 test of goodness of fit. *The Annals of Mathematical Statistics*, 23(3):315–345, sep 1952. doi: 10.1214/aoms/1177729380.

[17] William G. Cochran. The combination of estimates from different experiments. *Biometrics*, 10(1):101, mar 1954. doi: 10.2307/3001666.

[18] Peter J. A. Cock, Tiago Antao, Jeffrey T. Chang, Brad A. Chapman, Cymon J. Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, and Michiel J. L. de Hoon. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics (Oxford, England)*, 25:1422–1423, June 2009. ISSN 1367-4811. doi: 10.1093/bioinformatics/btp163.

[19] PDBe-K. B. consortium. Pdbe-kb: a community-driven resource for structural and functional annotations. *Nucleic acids research*, 48:D344–D353, January 2020. ISSN 1362-4962. doi: 10.1093/nar/gkz853.

[20] Pierrick Craveur, Agnel Praveen Joseph, Pierre Poulain, Alexandre G. de Brevern, and Joseph Rebehmed. Cis-trans isomerization of omega dihedrals in proteins. *Amino acids*, 45:279–289, August 2013. ISSN 1438-2199. doi: 10.1007/s00726-013-1511-3.

[21] Tianli Dai, Qi Liu, Jun Gao, Zhiwei Cao, and Ruixin Zhu. A new protein-ligand binding sites prediction method based on the integration of protein sequence conservation information. *BMC bioinformatics*, 12 Suppl 14:S9, December 2011. ISSN 1471-2105. doi: 10.1186/1471-2105-12-S14-S9.

[22] B. Derrick and P. White. Why welch's test is type i error robust. *The Quantitative Methods for Psychology*, 12(1):30–38, jan 2016. doi: 10.20982/tqmp.12.1.p030.

[23] Joseph P. Romano Erich L. Lehmann. *Testing Statistical Hypotheses*. Springer New York, 2008. ISBN 0387988645. URL `https://www.ebook.de/de/product/3186913/erich_l_lehmann_joseph_p_romano_testing_statistical_hypotheses.html`.

[24] Thomas Hamelryck. An amino acid has two sides: a new 2d measure provides a different view of solvent exposure. *Proteins*, 59:38–48, April 2005. ISSN 1097-0134. doi: 10.1002/prot.20379.

[25] Michael J. Hartshorn, Marcel L. Verdonk, Gianni Chessari, Suzanne C. Brewerton, Wijnand T. M. Mooij, Paul N. Mortenson, and Christopher W. Murray. Diverse, high-quality test set for the validation of protein-ligand docking performance. *Journal of medicinal chemistry*, 50:726–741, February 2007. ISSN 0022-2623. doi: 10.1021/jm061277y.

[26] Bingding Huang and Michael Schroeder. Ligsitecsc: predicting ligand binding sites using the connolly surface and degree of conservation. *BMC structural biology*, 6:19, September 2006. ISSN 1472-6807. doi: 10.1186/1472-6807-6-19.

[27] David Jakubec, Jirí Vondrášek, and Robert D. Finn. 3dpatch: fast 3d structure visualization with residue conservation. *Bioinformatics (Oxford, England)*, 35:332–334, January 2019. ISSN 1367-4811. doi: 10.1093/bioinformatics/bty464.

[28] Lukas Jendele, Radoslav Krivak, Petr Skoda, Marian Novotny, and David Hoksza. Prankweb: a web server for ligand binding site prediction and visualization. *Nucleic acids research*, 47:W345–W349, July 2019. ISSN 1362-4962. doi: 10.1093/nar/gkz424.

[29] S. Jones and J. M. Thornton. Analysis of protein-protein interaction sites using surface patches. *Journal of molecular biology*, 272:121–132, September 1997. ISSN 0022-2836. doi: 10.1006/jmbi.1997.1234.

[30] W. Kabsch and C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22:2577–2637, December 1983. ISSN 0006-3525. doi: 10.1002/bip.360221211.

[31] Nickolay A. Khazanov and Heather A. Carlson. Exploring the composition of protein-ligand binding sites on a large scale. *PLoS computational biology*, 9:e1003321, 2013. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1003321.

[32] Radoslav Krivák and David Hoksza. P2rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. *Journal of cheminformatics*, 10:39, August 2018. ISSN 1758-2946. doi: 10.1186/s13321-018-0285-8.

[33] J. Kyte and R. F. Doolittle. A simple method for displaying the hydropathic character of a protein. *Journal of molecular biology*, 157:105–132, May 1982. ISSN 0022-2836. doi: 10.1016/0022-2836(82)90515-0.

[34] Howard Levene. Robust tests for equality of variances. In Ingram Olkin, editor, *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, pages 278–292. Stanford University Press.

[35] Mingfeng Lin, Henry C. Lucas, and Galit Shmueli. Too big to fail: Large samples and thep-value problem. *Information Systems Research*, 24(4):906–917, dec 2013. doi: 10.1287/isre.2013.0480.

[36] Thomas Lumley, Paula Diehr, Scott Emerson, and Lu Chen. The importance of the normality assumption in large public health data sets. *Annual Review of Public Health*, 23(1):151–169, may 2002. doi: 10.1146/annurev.publhealth.23.100901.140546.

[37] Cyrus R. Mehta and Nitin R. Patel. A network algorithm for performing fisher's exact test in r × c contingency tables. *Journal of the American Statistical Association*, 78(382):427, jun 1983. doi: 10.2307/2288652.

[38] S. Miller, J. Janin, A. M. Lesk, and C. Chothia. Interior and surface of monomeric proteins. *Journal of molecular biology*, 196:641–656, August 1987. ISSN 0022-2836. doi: 10.1016/0022-2836(87)90038-6.

[39] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. Scop: a structural classification of proteins database for the investigation of sequences and structures. *Journal of molecular biology*, 247:536–540, April 1995. ISSN 0022-2836. doi: 10.1006/jmbi.1995.0159.

[40] Marco Necci, Damiano Piovesan, Zsuzsanna Dosztányi, and Silvio C. E. Tosatto. Mobidb-lite: fast and highly specific consensus prediction of intrinsic disorder in proteins. *Bioinformatics (Oxford, England)*, 33:1402–1404, May 2017. ISSN 1367-4811. doi: 10.1093/bioinformatics/btx015.

[41] Andrew Nightingale, Ricardo Antunes, Emanuele Alpi, Borisas Bursteinas, Leonardo Gonzales, Wudong Liu, Jie Luo, Guoying Qi, Edd Turner, and Maria Martin. The proteins api: accessing key integrated protein and genome information. *Nucleic acids research*, 45:W539–W544, July 2017. ISSN 1362-4962. doi: 10.1093/nar/gkx237.

[42] Daniele Raimondi, Gabriele Orlando, Rita Pancsa, Taushif Khan, and Wim F. Vranken. Exploring the sequence-based prediction of folding initiation sites in proteins. *Scientific reports*, 7:8826, August 2017. ISSN 2045-2322. doi: 10.1038/s41598-017-08366-3.

[43] G. N. RAMACHANDRAN, C. RAMAKRISHNAN, and V. SASISEKHARAN. Stereochemistry of polypeptide chain configurations. *Journal of molecular biology*, 7:95–99, July 1963. ISSN 0022-2836. doi: 10.1016/s0022-2836(63)80023-6.

[44] Ambrish Roy, Jianyi Yang, and Yang Zhang. Cofactor: an accurate comparative algorithm for structure-based protein function annotation. *Nucleic acids research*, 40:W471–W477, July 2012. ISSN 1362-4962. doi: 10.1093/nar/gks372.

[45] Graeme D. Ruxton. The unequal variance t-test is an underused alternative to student's t-test and the mann–whitney u test. *Behavioral Ecology*, 17(4): 688–690, may 2006. doi: 10.1093/beheco/ark016.

[46] Peter Schmidtke, Catherine Souaille, Frédéric Estienne, Nicolas Baurin, and Romano T. Kroemer. Large-scale comparison of four binding site detection algorithms. *Journal of chemical information and modeling*, 50:2191–2200, December 2010. ISSN 1549-960X. doi: 10.1021/ci1000289.

[47] Pak C. Sham and Shaun M. Purcell. Statistical power and significance testing in large-scale genetic studies. *Nature Reviews Genetics*, 15(5):335–346, apr 2014. doi: 10.1038/nrg3706.

[48] A. Shrake and J. A. Rupley. Environment and exposure to solvent of protein atoms. lysozyme and insulin. *Journal of molecular biology*, 79:351–371, September 1973. ISSN 0022-2836. doi: 10.1016/0022-2836(73)90011-9.

[49] Zhoutong Sun, Qian Liu, Ge Qu, Yan Feng, and Manfred T. Reetz. Utility of b-factors in protein science: Interpreting rigidity, flexibility, and internal motion and engineering thermostability. *Chemical reviews*, 119:1626–1665, February 2019. ISSN 1520-6890. doi: 10.1021/acs.chemrev.8b00290.

[50] Kuan Pern Tan, Raghavan Varadarajan, and M. S. Madhusudhan. Depth: a web server to compute depth and predict small-molecule binding cavities in proteins. *Nucleic acids research*, 39:W242–W248, July 2011. ISSN 1362-4962. doi: 10.1093/nar/gkr356.

[51] The UniProt Consortium. Uniprot: the universal protein knowledgebase. *Nucleic acids research*, 45:D158–D169, January 2017. ISSN 1362-4962. doi: 10.1093/nar/gkw1099.

[52] Eldon L. Ulrich, Hideo Akutsu, Jurgen F. Doreleijers, Yoko Harano, Yannis E. Ioannidis, Jundong Lin, Miron Livny, Steve Mading, Dimitri Maziuk,

Zachary Miller, Eiichi Nakatani, Christopher F. Schulte, David E. Tolmie, R. Kent Wenger, Hongyang Yao, and John L. Markley. Biomagresbank. *Nucleic acids research*, 36:D402–D408, January 2008. ISSN 1362-4962. doi: 10.1093/nar/gkm957.

[53] Mark N. Wass and Michael J. E. Sternberg. Prediction of ligand binding sites using homologous structures and conservation at casp8. *Proteins*, 77 Suppl 9:147–151, 2009. ISSN 1097-0134. doi: 10.1002/prot.22513.

[54] B. L. WELCH. THE GENERALIZATION OF 'STUDENT's' PROBLEM WHEN SEVERAL DIFFERENT POPULATION VARLANCES ARE INVOLVED. *Biometrika*, 34(1-2):28–35, 1947. doi: 10.1093/biomet/34.1-2.28.

[55] Jianyi Yang, Ambrish Roy, and Yang Zhang. Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. *Bioinformatics (Oxford, England)*, 29:2588–2595, October 2013. ISSN 1367-4811. doi: 10.1093/bioinformatics/btt447.

[56] Zengming Zhang, Yu Li, Biaoyang Lin, Michael Schroeder, and Bingding Huang. Identification of cavities on protein surface using multiple computational approaches for drug binding site prediction. *Bioinformatics (Oxford, England)*, 27:2083–2088, August 2011. ISSN 1367-4811. doi: 10.1093/bioinformatics/btr331.

[57] Donald W. Zimmerman. Invalidation of parametric and nonparametric statistical tests by concurrent violation of two assumptions. *The Journal of Experimental Education*, 67(1):55–68, jan 1998. doi: 10.1080/00220979809598344.

[58] Donald W. Zimmerman. A note on preliminary tests of equality of variances. *British Journal of Mathematical and Statistical Psychology*, 57(1):173–181, may 2004. doi: 10.1348/000711004849222.

[59] Donald W. Zimmerman and Bruno D. Zumbo. Rank transformations and the power of the student t test and welch t' test for non-normal populations with unequal variances. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 47(3):523–539, 1993. doi: 10.1037/h0078850.

[60] Bruno D. Zumbo and Daniel Coulombe. Investigation of the robust rank-order test for non-normal populations with unequal variances: The case of

reaction time. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 51(2):139–150, 1997. doi: 10.1037/1196-1961.51.2.139.

A. Attachments

The attached CD contains two attachments:

A.1 attachment1

blabla

A.2 attachment2

blabla