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 cile prace

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

...

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 custom features.

The pipeline is able to do all steps needed for the task, from downloading the structures from PDB, to computing the statistical significance of individual features. Moreover, there are two scripts that further extend the analysis pipeline and can be used to train and evaluate P2Rank models with new features. The only needed input is a dataset file with listed proteins.

The pipeline is implemented in Python, making use of several Python packages, such as BioPython [?], NumPy [?] or Matplotlib [?]. BioPython is an open-source collection of Python tools for computational biology and it was very useful for this work. The main script `analysis_pipeline.py` defines the user API, parses and checks arguments, takes care of logging and runs individual parts of the pipeline. See `https://github.com/katebrich/Master-thesis/tree/dev/Scripts` TODO for more details about options, requirements, input, output and examples of usage.

The structure of the pipeline is depicted in Diagram 3.1. The details about individual parts are described in this chapter.

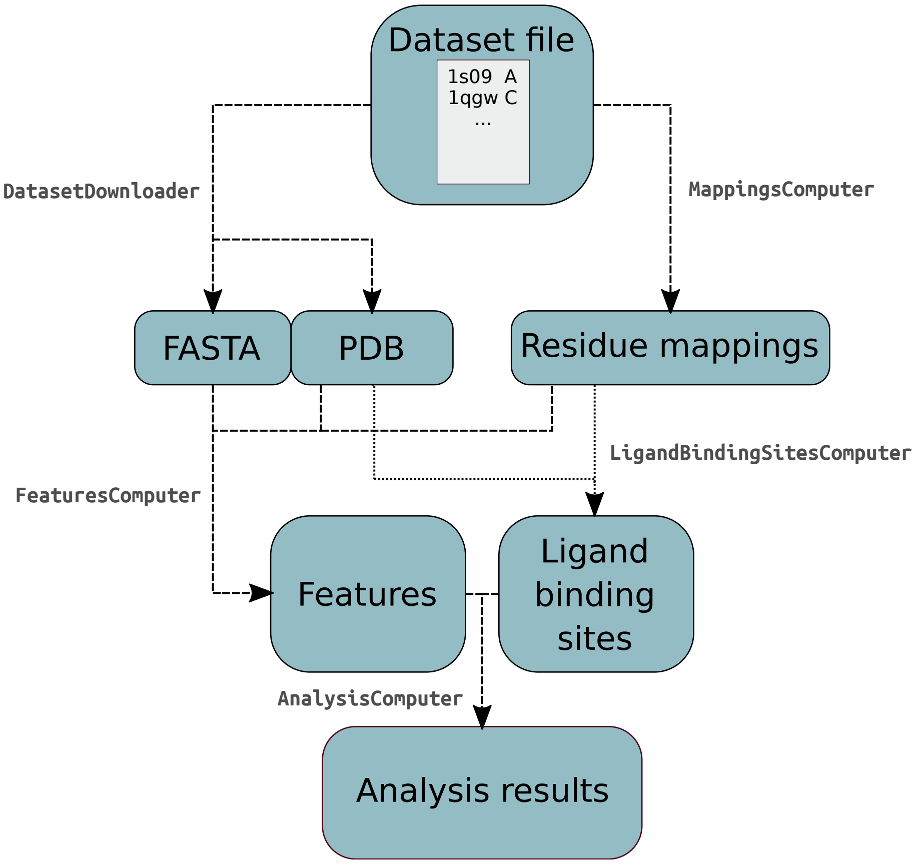


Figure 3.1: Diagram of the pipeline structure.

TODO verze programu a databazi? Kam to dat? TODO options, parameters - tasks, threads, output_dir, dataset TODO default values

3.1 Dataset file

TODO co je REST API

3.2 Data download

3.3 Residue mappings

- zminit chybu s mapovanim segmentu, insertion codes, proc nesly SIFTS

3.4 Ligand binding sites labelling

- options - lbs_distance_threshold - problemy s HETATM - SASA + cutoff

3.5 Features

....TODO - options - features, config_path

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 and it provides detailed, accurate and consistent annotations of the sequences [**?**].

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 more information, such as available evidence, source or related publications. The features are arranged in a well-organized manner on the 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 [**?**]. 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 [?].

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 [?].

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 [?].

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 [?] based on the coordinate data sets extracted from the Protein Data Bank (PDB). They are neither predicted computationally, nor propagated to related species [?].

3.5.1.5 Natural variant

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

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 [?].

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. [**?**]

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. [**?**].

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 was developed by the Bio2Byte group [**?**] 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 [**?**]. 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 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 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

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” [**?**]. 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 [**?**].

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 Half sphere exposure

3.5.3.3 Exposure CN

3.5.3.4 Phi and psi angles

3.5.3.5 Cis peptide

3.5.4 FASTA

3.5.5 Other resources

3.5.5.1 MobiDB

3.5.5.2 Conservation

3.5.6 Custom features

3.6 Statistical analysis

TODO P-value TODO power of test

TODO implementace testu v Pythonu

To find properties mapped on the protein primary structure which are possibly important for prediction of protein-ligand binding sites, statistical analysis will have the crucial role. It is a great way to explore the big amounts of accessible data and it can potentially help to discover underlying patterns and draw inferences from the data.

This chapter describes the method that was used to analyse the *statistical significance* of the properties and to distinguish the ones that stand out in the known protein-ligand binding sites.

Hypothesis testing is a method of statistical inference. Its goal is to infer properties of a *statistical population*, i.e. a set of similar items or events. In this work, two populations will be compared: we take values of a property for all the amino acids across all the proteins in the dataset and then compare the ones in binding sites and outside of binding sites.

A dataset usually contains a subset sampled from a larger population, rather than the whole population. This subset is called a *statistical sample.* It should represent the population well and be unbiased.

A *hypothesis* makes a statement about an unknown population parameter. In a hypothesis testing problem, an experimenter states two complementary hypotheses: the *null hypothesis* and the *alternative hypothesis*, denoted by H_0 and H_1, respectively. The null hypothesis comprises a subset of possible parameters and the alternative hypothesis comprises the supplement, so that all the possible parameters are covered.

In a hypothesis testing problem, an experimenter should come to one of the conclusions: to either accept H_0, or to reject H_0 and accept H_1.

To decide which one of two complementary hypotheses is true, an experimenter employes a suitable *hypothesis test.* A hypothesis test is a rule that specifies for which sample values the H_0 is accepted as true and for which sample values it is rejected, and therefore H_1 is accepted as true. A hypothesis test is usually specified in terms of a test statistic (i.e. a function of the sample) [3].

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 3.1. 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 [3].

		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 3.1: Type I and II Error in hypothesis testing.

To control statistical significance of the result, a study defines a threshold called *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.

One way to report the result of the test would be simply to tell whether the null hypothesis was accepted or rejected at the given significance level. However, most researchers choose to report a certain kind of test statistic (function of a sample X), the so-called *p-value*. Smaller values of $p(X)$ give stronger evidence for rejecting the null hypothesis. The null hypothesis is rejected when $p(X) \leq \alpha$. Hence, we are able to determine the smallest significance level at which the hypothesis would be accepted/rejected. 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 [3, 5, 11].

It is suggested that the significance level for a study is set prior to any data collection [9]. The typical choices in practice are $\alpha = 0.01$, 0.05 or 0.10 [3]. One should be aware that by fixing the significance level of the test, the experimenter 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 [11].

Let's suppose an experimenter has a research hypothesis that he or she hopes to prove, but does not want to risk accepting it without convincing data support. In this case, the test should be set up in such a way that the research hypothesis corresponds to the alternative hypothesis, not the null hypothesis. By specifying a small significance level α, the experimenter thus controls the probability of the Type I error. In other words, the probability of accepting the research hypothesis when in is not true would be α at most [3].

3.6.1 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 [12].

Both Student's and Welch's t-test assume that the two examined populations follow a normal distribution [12]. 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 [5]. 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 [7, 15, 16]. According to Lehmann and Romano [5], 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 [7, 13]. The simulations made by Lumley *et al.* [7] 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 [10].

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 [4].

Some researchers tend to pre-test for variance equality by a preliminary test of variances (such as Levene's [6], 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 [14]. 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 [14], "a higher Type I error rate of the preliminary test actually improves the performance of the compound test". 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 [10] 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 [4]. 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.

3.6.2 Chi-squared test of independence

A different kind of tests will be needed for the analysis of categorical features. An example of a categorical feature is XXX. Moreover, quantitative features can be grouped into categories and analysed in the same way as categorical features. In this section, two widely-used tests of such kind will be presented and discussed.

Both Fisher's exact test and Chi-squared test of independence are well-known hypothesis tests used for the analysis of data in contigency tables. 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 3.2. The sums of frequencies in rows and columns are called *marginal totals.*

	PTM	Without PTM	Total
Binding sites	XX	XX	XX
Non-binding sites	XX	XX	XX
Total	XX	XX	XX

Table 3.2: A 2×2 contingency table. TODO real data

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 Chi-squared 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 [8]. 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 [2].

Chi-squared (χ^2) 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 as $\frac{row\ total \times column\ total}{grand\ total}$. 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 [2]. TODO vysvetlit chi-sq rozdeleni a to jak se pocita ta testova statistika

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. TODO FISHERUV TEST OBECNE NA MENSI VZORKY, ... W. G. Cochran (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:

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

TODO: for small, sparse, or unbalanced data, the exact and asymptotic p-values can be quite different and may lead to opposite conclusions concerning the hypothesis of interest. (wikipedia)

3.7 P2Rank models training and evaluation

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 [?]. 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 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 [?]. It comprises at most one representative chain for every SCOP family [?] 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 [?]; after re-downloading and filtering, the numbers are slightly smaller.

- **Coach420** - a dataset that was originally taken from a benchmark study [?] and used in other studies [? ?]. 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[?], a benchmark set for the validation of protein-ligand docking performance [?], and a dataset with bound and unbound structures used for evaluation of a ligand binding sites predictor [?].

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

TODO single chain

4.1.1 Ligands filtering

4.2 Statistical analysis

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 4.1. 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 4.1 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

	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 4.1: 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.

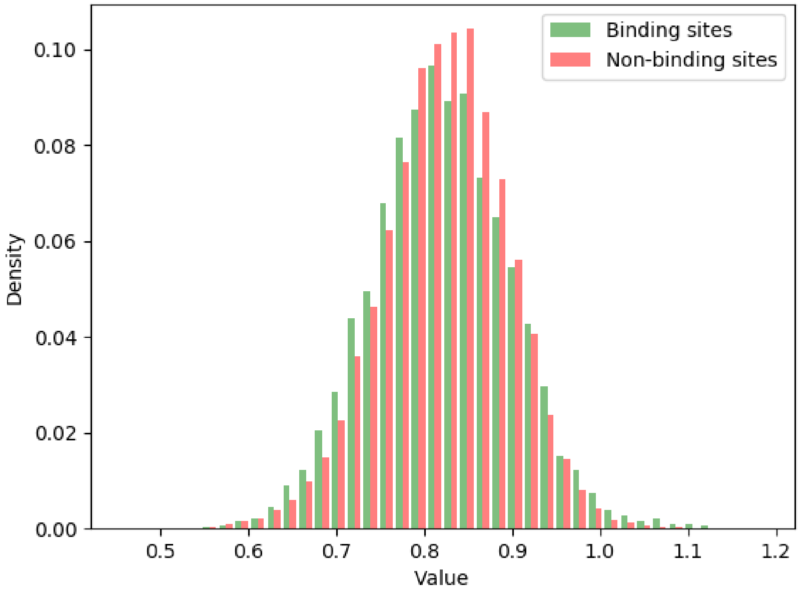


Figure 4.1: Histogram of feature `dynamine` computed on holo4k dataset. Density on the y-axis is computed with respect to the number of binding or non-binding sites. Difference in means: 0.0014; difference in variances: 0.0015.

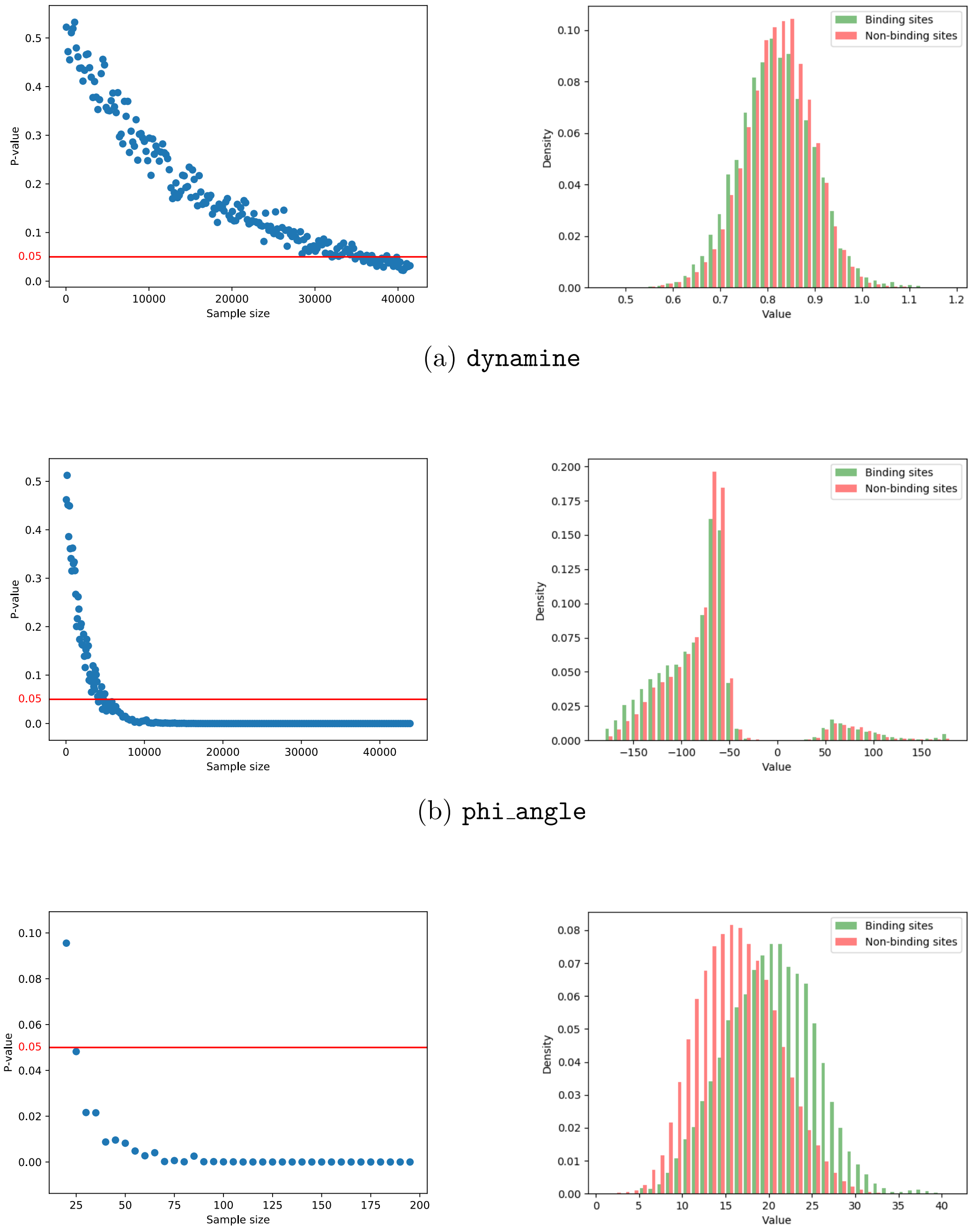
subtler and more complex effects, which can be advantageous in some cases, but also misleading. It all depends on the purpose of the statistical testing. The question we should ask is not whether the results are statistically significant (which there almost always will be for large samples), but whether they are interesting for our research [**?**].

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. [**?**] reviewed articles in two leading Information System (IS) journals and 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 4.1 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 4.1 is that the results for some features vary across datasets. Let's take a look at feature `turn`, for example. The P-value



(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 $\alpha = 0.05$.

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 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 4.2. 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.* [7] 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

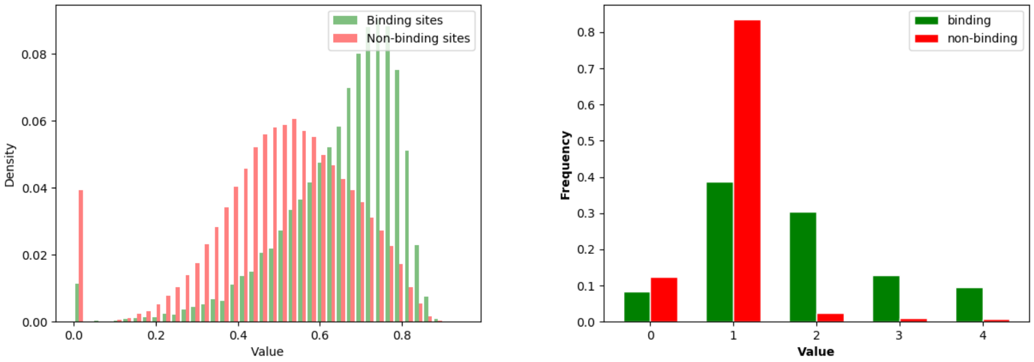
The results for the datasets with different ligands filters does not seem to differ widely and does not reveal any additional informations. For that reason, following features analysis and computed plots will be shown only for the dataset 'mix' with P2Rank ligands filter. The results obtained from the 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 [? ? ? ?]. 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 '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 [?] or PocketPicker [?], as well as other approached, make use of this property. The histograms for these continuous

	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 4.2: 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.

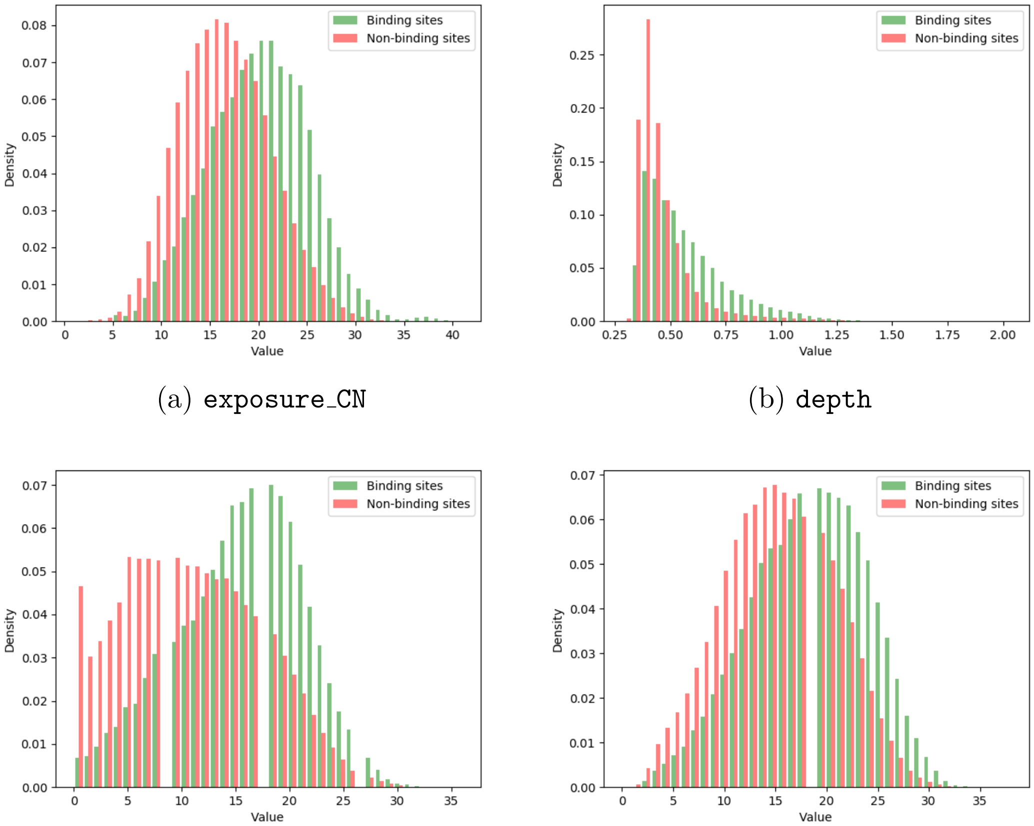


(a) `conservation`

(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.

features are depicted in Figure 4.4



(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.

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.

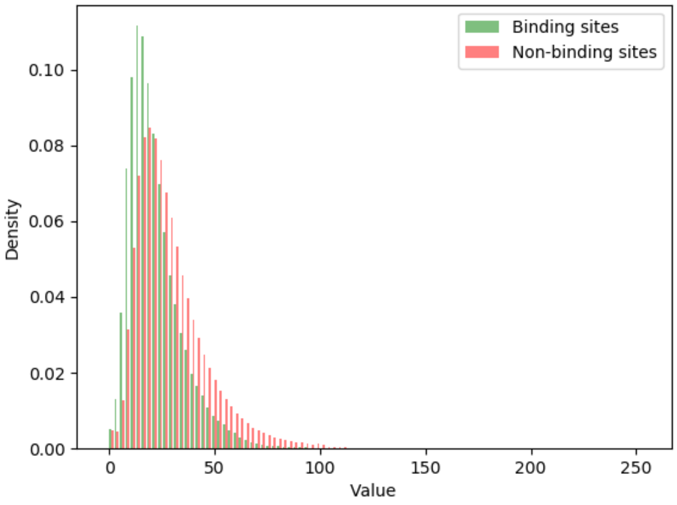


Figure 4.5: `bfactor`: Binding sites have lower B factor values on average.

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 [**?**]. 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?

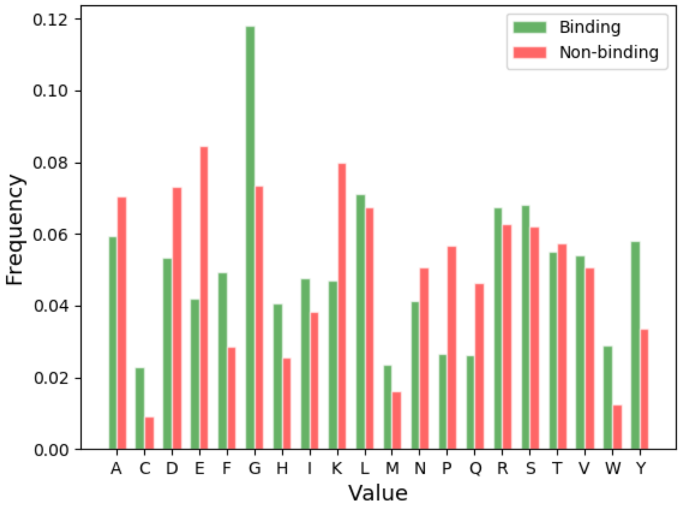
TODO !!!! It was discovered that high B-factor-characterized regions show a higher average flexibility index, more pronounced average hydrophilicity, and higher absolute net charge. - (11) Radivojac, P.; Obradovic, Z.; Smith, D. K.; Zhu, G.; Vucetic, S.; Brown, C. J.; Lawson, J. D.; Dunker, A. K. Protein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

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.

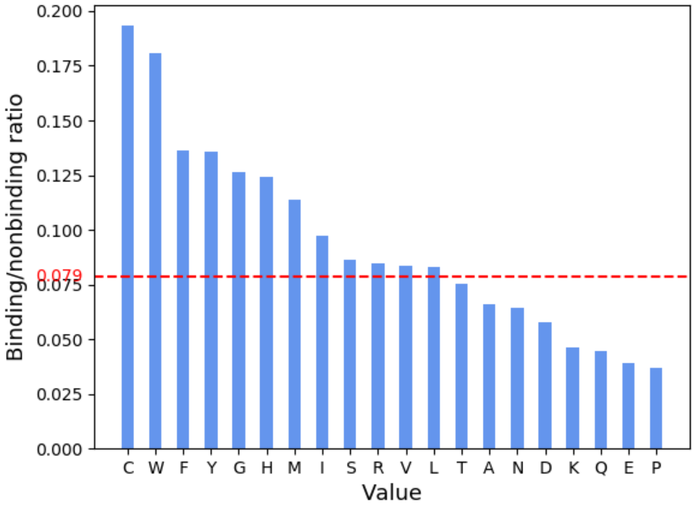
4.3 P2Rank models

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



(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`.

	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 4.3: Comparison of the performance of models with different sets of features.

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.

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 4.3. Our baseline model performs a little better in the Top-n category than the default P2Rank model described in the P2Rank article [**?**], 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.

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 4.4 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 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 4.5

	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 4.4: 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.

	Top-n	Top-(n+2)
lbs	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
efoldmine	7.2	13.5
strand	6.3	10.9
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 4.5

Conclusion

Conclusion.

List of Abbreviations

AA	Amino acid
atd	a tak dale

Bibliography

[1]

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

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

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] Jerzy Neyman and Egon S Pearson. *The testing of statistical hypotheses in relation to probabilities a priori*, volume 29. 1933.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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