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

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

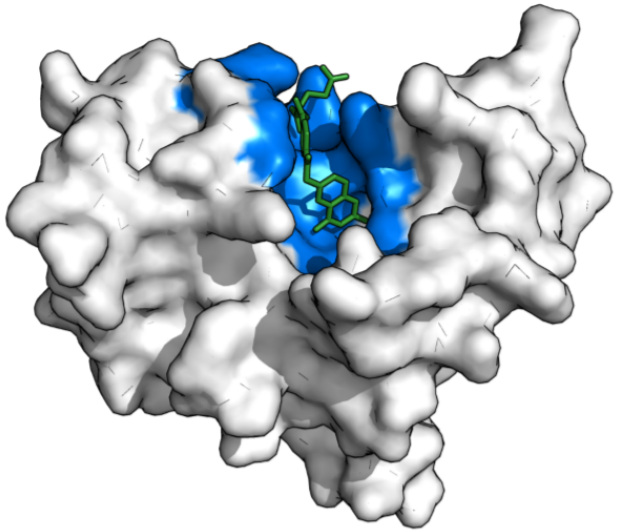
Kateřina Břicháčková
Vedoucí: RNDr. David Hoksza, Ph. D.

Univerzita Karlova
Přírodovědecká fakulta

19. 02. 2021

Proč predikovat vazebná místa?

- Aplikace
  - Vývoj léčiv
  - Predikce vedlejších účinků léčiv
  - Anotace funkce
  - Inverse virtual screening, ...
- Mnoho existujících nástrojů a přístupů
- Otázka: Jak se obecně liší vazebná a nevazebná místa?



Protein-ligand binding. Figure retrieved from commons.wikimedia.org/wiki/File:DHFR_methotrexate_inhibitor.png

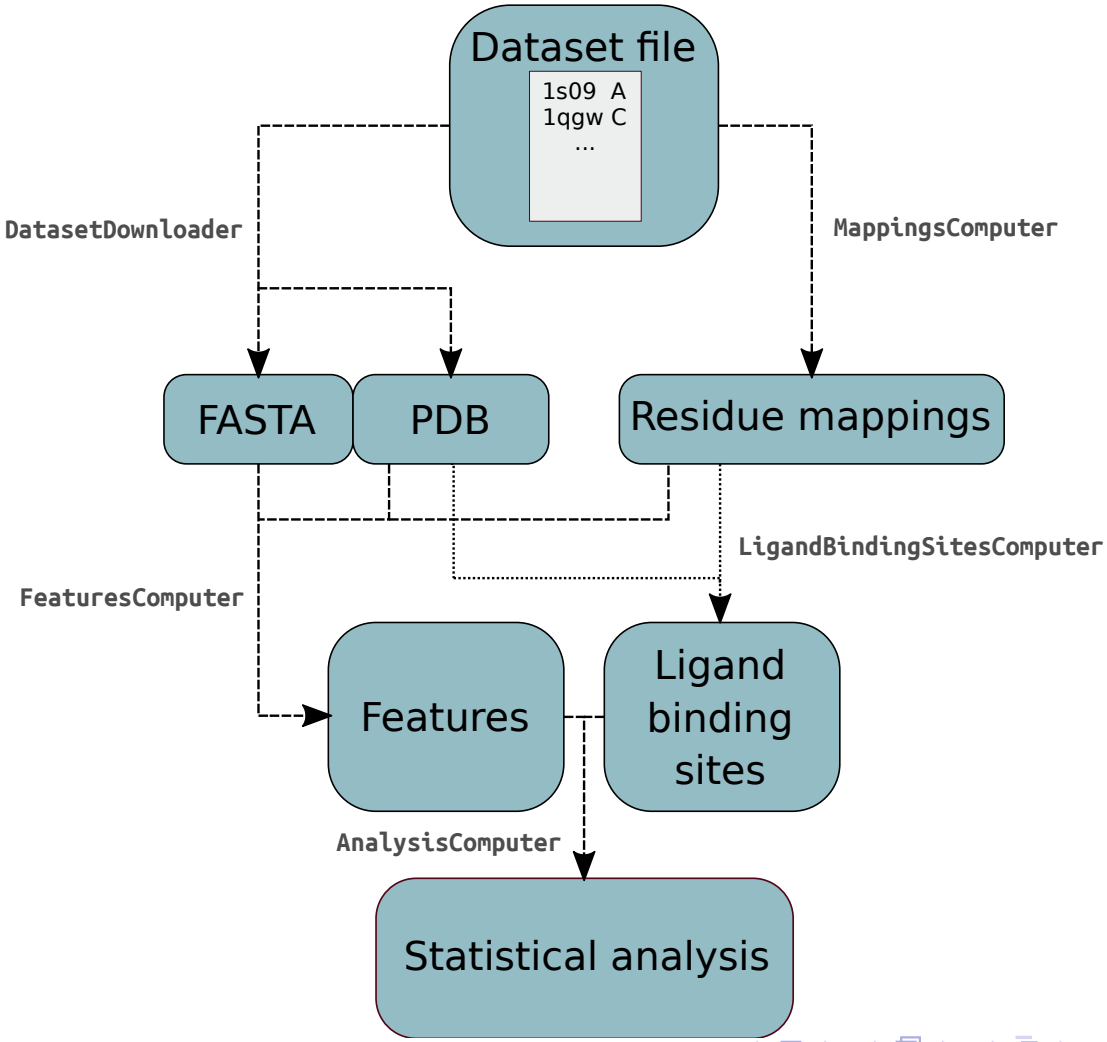
Cíle práce

- Implementace pipeliny pro statistickou analýzu vlastností protein-ligand vazebných míst
- Aplikace pipeliny na existující anotace primární a terciární struktury
- Interpretace výsledků
- Ověření praktické významnosti výsledků metodou P2Rank

Pipeline

- Python
- Testování hypotéz a velikosti účinku
- GitHub repozitář
  `https://github.com/katebrich/LBS_analysis_pipeline`
  - Setup, requirements file
  - Příklady použití
  - Popis vstupu a výstupu
  - Popis parametrů (velikost vzorku, počet iterací, vzdálenost od ligandu při výpočtu vazebných reziduí, počet vláken, ...)

Pipeline - diagram



Pipeline se dá využít různými způsoby

- Základní použití:

```
python3 scripts/source/analysis_pipeline.py \
        -d data/datasets/DATASET_NAME.txt \
        -o output/DATASET_NAME
```

- Definice vlastních features (anotací)
- Spuštění na vlastních datech od určitého kroku
- Random sampling
- Rozšiřující skripty pro trénování P2Rank modelů

Features

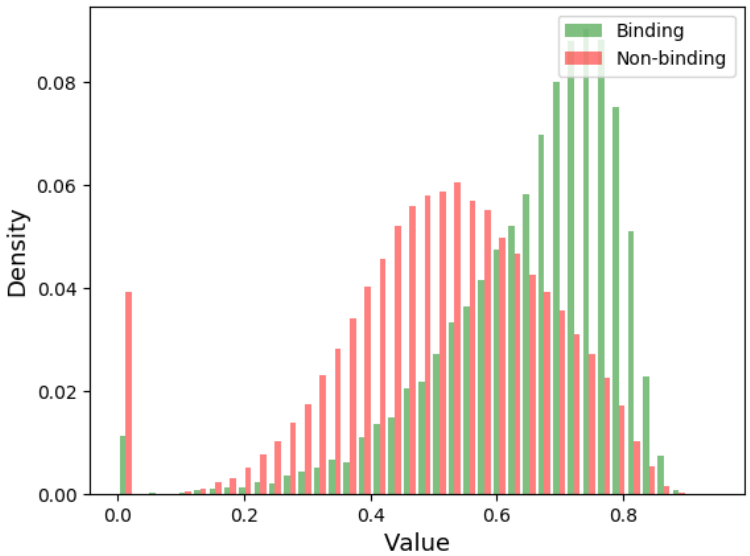
Name	Type	Source
PTM	binary	UniProtKB
lipidation	binary	UniProtKB
glycosylation	binary	UniProtKB
mod_res	binary	UniProtKB
disulfid	binary	UniProtKB
non_standard	binary	UniProtKB
sec_str	categorical	UniProtKB
helix	binary	UniProtKB
turn	binary	UniProtKB
strand	binary	UniProtKB
natural_variant	binary	UniProtKB
variation	binary	UniProtKB
compbias	binary	UniProtKB
pdbekb_conservation	ordinal	PDBe-KB
dynamine	continuous	PDBe-KB
efoldmine	continuous	PDBe-KB
depth	continuous	PDBe-KB

Name	Type	Source
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

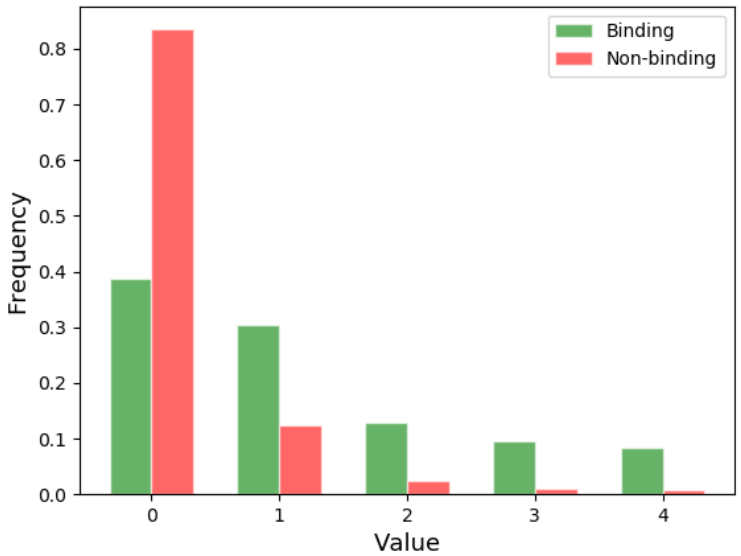
Statistická analýza

- Testování hypotéz
  - Welch's test
    - spojitá data
    - H_0 - Průměr hodnot pro vazebná a nevazebná rezidua je stejný
  - Chi-kvadrát (χ^2) test nezávislosti
    - data v kontingenčních tabulkách
    - H_0 - Hodnota nezávisí na tom, jestli je reziduum vazebné
- Velikost účinku
  - Cohenovo *d*
    - spojitá data
  - Cohenovo *w*
    - binární, ordinální a kategorická data
- 4 datasety, celkem 5047 různých protein-ligand komplexů
- 2 přístupy:
  - Všechna rezidua z datasetu zároveň
  - Random sampling - 500 vazebných a 500 nevazebných reziduí, 1000 iterací

Vazebná rezidua jsou více konzervovaná

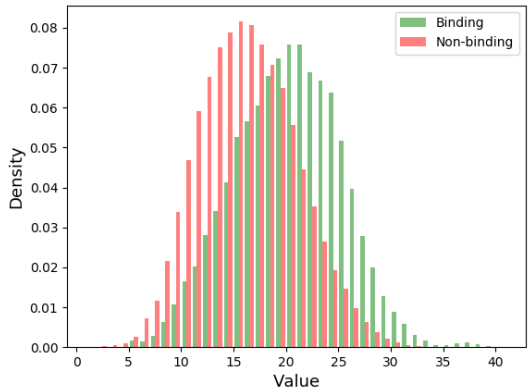


(a) `conservation`

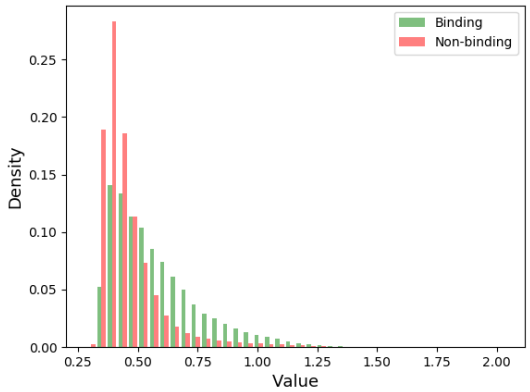


(b) `pdbekb_conservation`

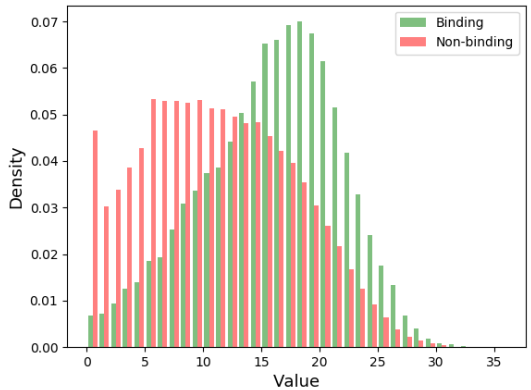
Vazebná místa se nachází v kapsách na povrchu proteinu



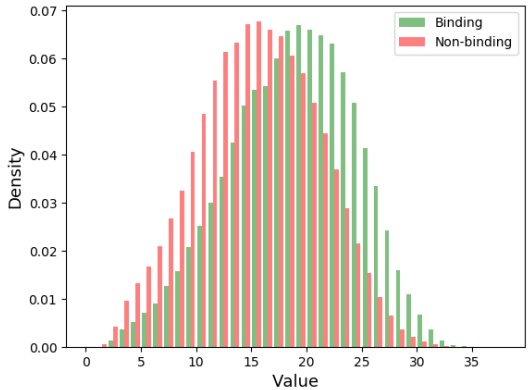
(a) `exposure_CN`



(b) `depth`

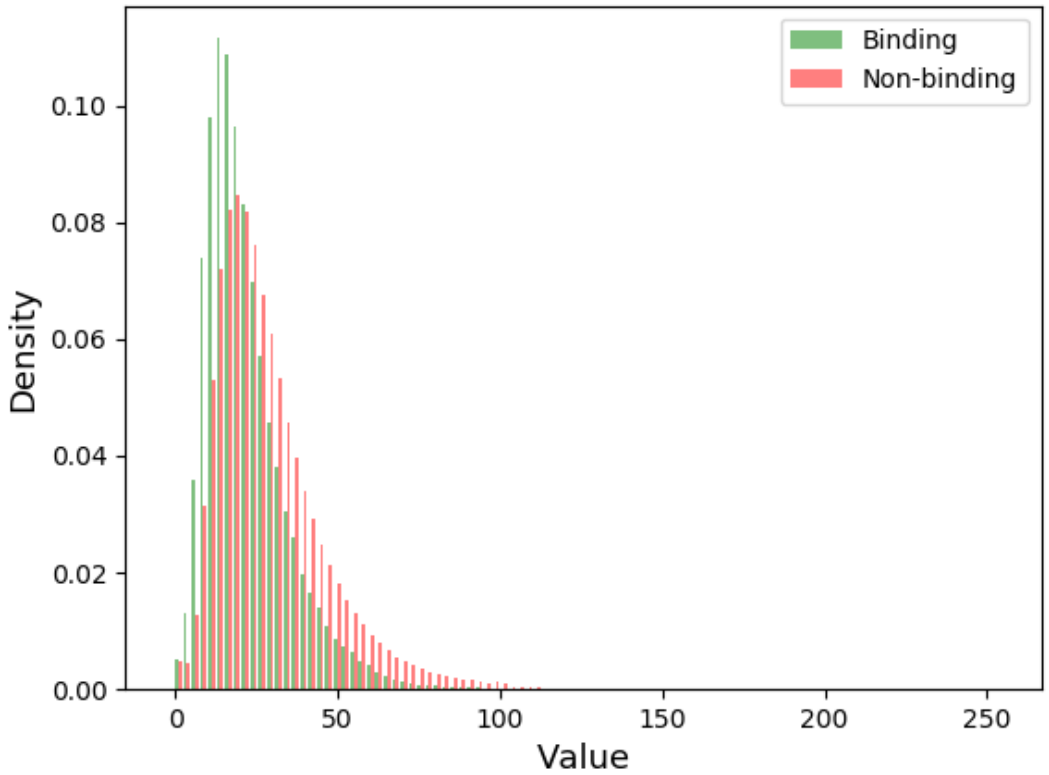


(c) `HSE_up`

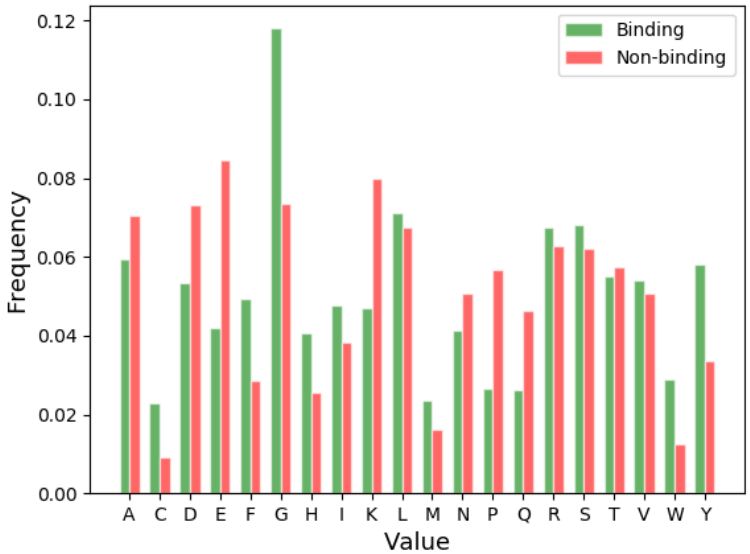


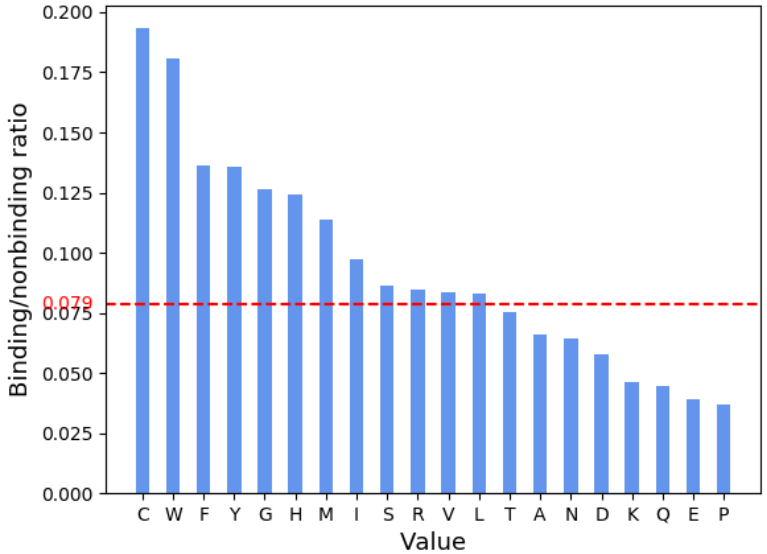
(d) `HSE_down`

Vazebná rezidua mají nižší B faktor



Vazebná a nevazebná místa mají odlišné aminokyselinové složení





Trénování P2Rank modelů

- P2Rank
  - Univerzita Karlova (Radoslav Krivák, David Hoksza)
  - Random forests
  - Rychlost (1s/protein) a úspěšnost predikce
- *chen11* - training dataset
- *coach420* - evaluation dataset
- parametry stejné jako pro defaultní P2Rank model (100 stromů, neomezená hloubka, 6 featur / strom)
- pouze konzervovanost zlepšila úspěšnost

Model	**Úspěšnost**
baseline model	73.6
pdbekb_conservation	78.9
conservation	76.8

Důležitost vlastností koreluje s výsledky analýzy

feature	importance
pdbekb_conservation	0.008592
HSE_up	0.006906
conservation	0.004021
depth	0.003474
HSE_down	0.002467
exposure_CN	0.002011
aromaticity	0.001627
bfactor	0.000891
helix	0.000807
aa_HIS	0.000783
hydropathy	0.000706
efoldmine	0.000675
dynamine	0.000661
strand	0.000645
mobiDB	0.000578
charged	0.000564
mol_weight	0.000533
aa_PHE	0.000516
phi_angle	0.000457
aa_GLY	0.000454

feature	importance
psi_angle	0.000433
H_bond_atoms	0.00042
aa_CYS	0.000389
aa_LEU	0.000363
aa_TYR	0.000305
aa_GLU	0.000221
aa_PRO	0.000221
aa_SER	0.000211
aa_ILE	0.000209
aa_ARG	0.000209
aa_MET	0.000196
aa_LYS	0.000194
aa_ASP	0.000138
aa_VAL	0.000137
aa_TRP	0.000121
aa_THR	0.000116
aa_ALA	0.000107
aa_ASN	0.000102
aa_GLN	9.1E-05
turn	9E-05

Podmnožiny vlastností podle velikosti účinku

- **small** - `pdbekb_conservation,conservation,depth,HSE_up,HSE_down, exposure_CN,bfactor,mol_weight,hydropathy,aromaticity, H_bond_atoms`
- **medium** - `pdbekb_conservation, conservation, depth, HSE_up, HSE_down, exposure_CN, bfactor`
- **large** - `pdbekb_conservation, conservation, depth, HSE_up, HSE_down, exposure_CN`

Model	**Úspěšnost**
small	71.6
medium	67.4
large	66.5
all pipeline features	70.3
small + all P2Rank	77.4
medium + all P2Rank	77.3
large + all P2Rank	77.9
all pipeline + all P2Rank	75.1
baseline model	73.6

Podmnožina P2Rank vlastností podle výsledků z pipeline

- Místo 35 původních featur pouze 9
- **P2Rank subset** - `protrusion`, `bfactor`, `hydrophatyIndex`, `aromatic`, `hBondDonor`, `hBondAcceptor`, `hBondDonorAcceptor`, `hydrophobic`, `hydrophilic`

Model	**Úspěšnost**
all P2Rank + pdbekb_conservation	78.9
P2Rank subset + pdbekb_conservation	77.3

Srovnání konzervovaností

- `conservation`
  - UniProtKB/SwissProt, UniRef90 a TrEMBL
  - používaná v P2Rank
  - pomalá
- `INTAA_conservation`
  - UniProtKB/Swiss-Prot
- `pdbekb_conservation`
  - předpočítaná v PDBe-KB

Model	**Úspěšnost**
baseline model	73.6
conservation	77.1
INTAA_conservation	77.6
pdbekb_conservation	78.4

Závěr

- Analýza vlastností vazebných reziduí vypočítaná z 5047 protein-ligand komplexů a 33 anotací
- Konzervovanost používaná v metodě P2Rank by mohla být nahrazena rychlejší alternativou
- Statistická analýza může pomoci odhalit důležité vlastnosti vazebných reziduí
- Implementace pipeline - jednoduchý způsob, jak odhalit nadějné featury
  - možné využití i pro DNA a RNA vazebná místa

Děkuji za pozornost

Podobnost featur

- Stejné vlastnosti:
  - B faktor
  - vlastnosti aminokyselin (aromaticita, hydrofobicita, donor/akceptor, mol. hmotnost, ...)
- Podobné vlastnosti:
  - zanořenost reziduí
- Nové vlastnosti:
  - posttranslační modifikace
  - konzervovanost
  - phi a psi úhel
  - sekundární struktury
  - jednotlivé aminokyseliny
  - early folding regions
  - backbone dynamics
  - intrinstic disorder regions

Výsledky - Cohenovo *d*

	Průměr	Směrodatná odchylka
conservation	0.8812	0.06919
exposure_CN	0.7829	0.06753
HSE_up	0.7511	0.06866
depth	0.6927	0.06815
HSE_down	0.5649	0.06453
bfactor	0.4744	0.06165
phi_angle	0.07136	0.05012
mobiDB	0.07066	0.04874
psi_angle	0.06029	0.045
efoldmine	0.05961	0.04477
dynamine	0.05236	0.04086
random_cont (test)	0.04968	0.03807

Cohenovo d	Efekt
0.2	Malý
0.5	Střední
0.8	Velký

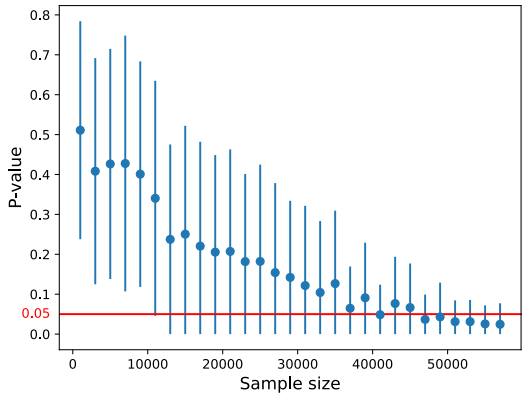
Výsledky - Cohenovo *w*

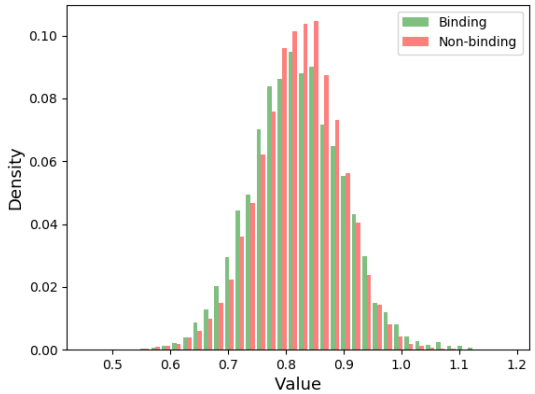
	Průměr	Směrodatná odchylka
lbs (test)	0.998	2.22E-16
pdbekb_conservation	0.4742	0.02518
aa	0.2443	0.02677
mol_weight	0.2414	0.0259
hydropathy	0.2336	0.02724
H_bond_atoms	0.1411	0.02893
aromaticity	0.1067	0.03028
sec_str	0.09798	0.02917
polarity	0.09157	0.0296
charged	0.08189	0.032
strand	0.07289	0.03048
helix	0.07052	0.02998
random_binary (test)	0.02375	0.01976
variation	0.02298	0.01929
turn	0.02023	0.01797

Cohenovo w	Efekt
0.1	Malý
0.3	Střední
0.5	Velký

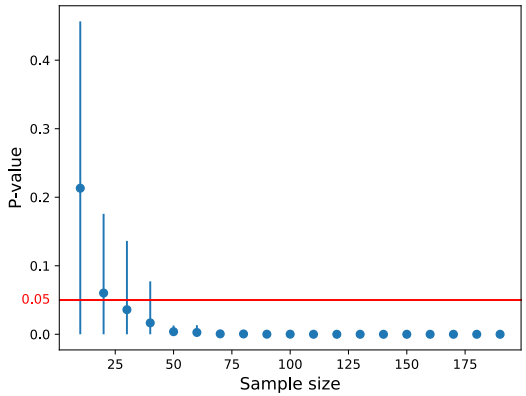
Závislost P hodnoty na velikosti vzorku

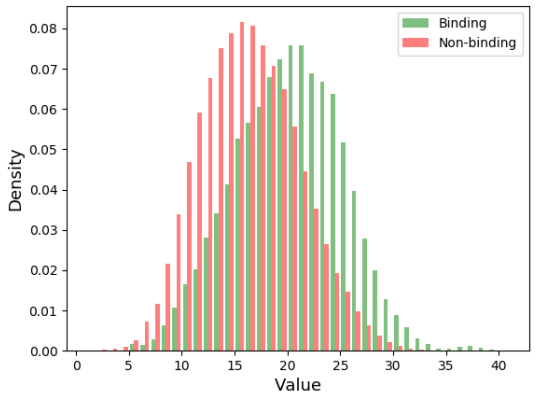
- `dynamine`





- `exposure_CN`





Datasety

- 4 datasety: *chen11*, *coach420*, *joined*, *holo4k*
- Filtrování ligandů: žádné, P2Rank, MOAD

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