3D-Sensitive Encoding of Pharmacophore Features

François Bérenger (postdoc)
Tsuda Laboratory - The University of Tokyo

21/09/2023

Outline

What is a Pharmacophore?

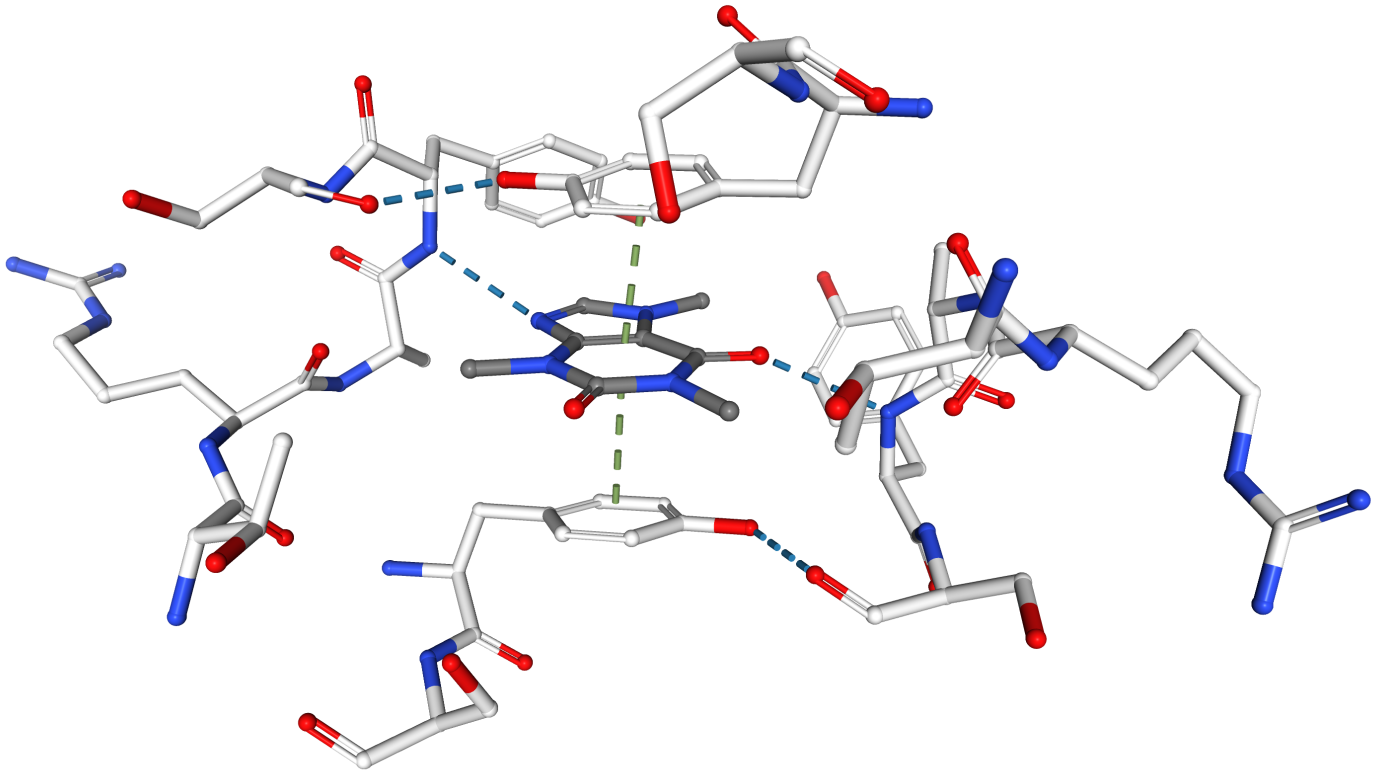


Figure 1: A pharmacophore for caffeine (PDB:6QTL "caffeine-recognizing nanobody"; RCSB NGL Viewer).

What Are Pharmacophore Features/Points?

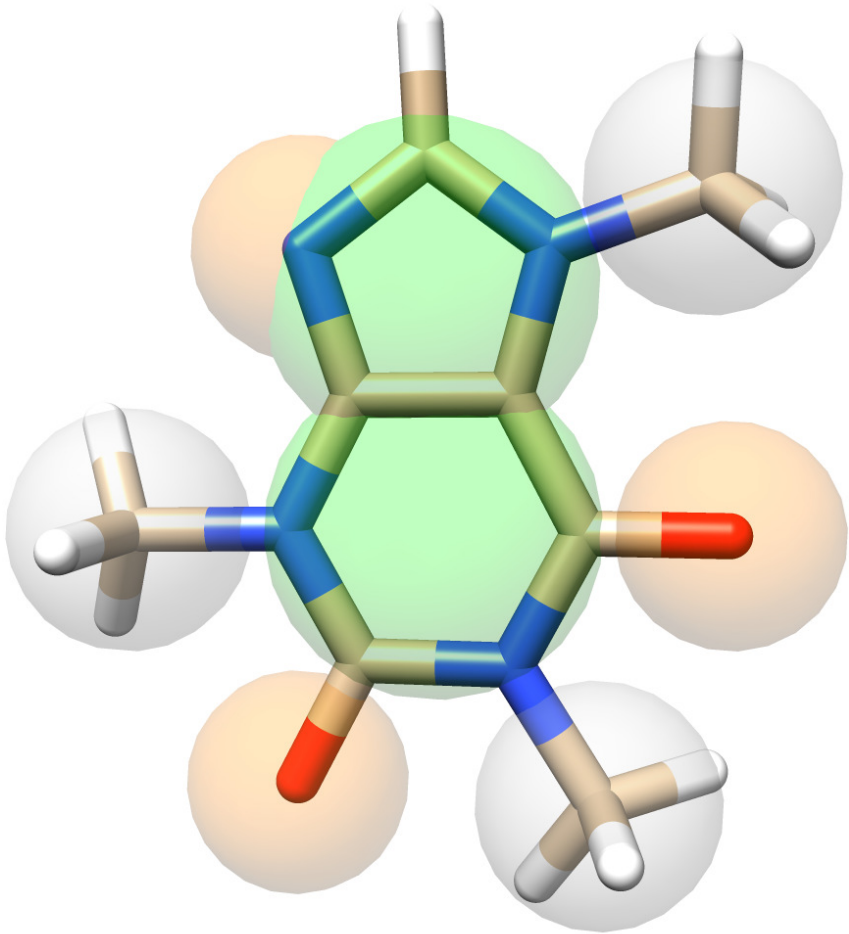


Figure 2: Pharmacophore features of caffeine (extracted using SMARTS strings and RDKit; UCSF Chimera).

Molecular Encoding Overview

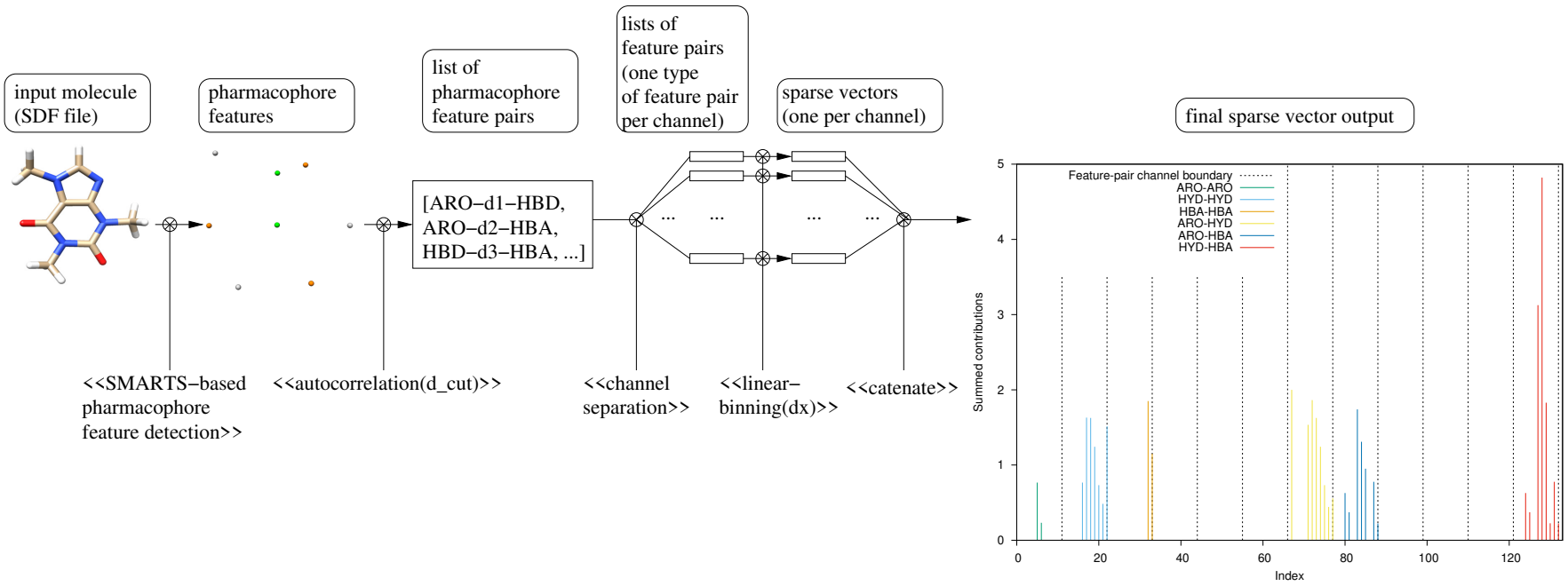


Figure 3: Method overview of the ACP4 encoding: starting from an input molecule (extreme left) and up to the corresponding sparse vector output (extreme right). On top of the figure are data explanations. Under the figure are data transformations.

The LIT-PCBA Dataset

Protein	Actives	Total	Random hit-rate
ADRB2	17	310984	0.00005
ALDH1	7140	144312	0.04948
ESR1+	13	5008	0.00260
ESR1-	99	5548	0.01784
FEN1	368	354135	0.00104
GBA	166	294821	0.00056
IDH1	39	359678	0.00011
KAT2A	193	347104	0.00056
MAPK1	307	62470	0.00491
MTORC1	97	33069	0.00293
OPRK1	24	268815	0.00009
PKM2	545	245112	0.00222
PPARG	27	5191	0.00520
TP53	79	4213	0.01875
VDR	868	354626	0.00245

Figure 4: Number of actives, total and random hit-rates for the 15 protein targets of the LIT-PCBA dataset.

Comparison Vs. Pharao

	ACP4	Pharao	Cohen d
ADRB2	0.50+/-0.09	**0.59+/-0.04**	-1.11
ALDH1	**0.54+/-0.02**	0.50+/-0.01	1.37
ESR1p	**0.66+/-0.07**	0.57+/-0.11	0.86
ESR1m	**0.60+/-0.07**	0.54+/-0.07	0.92
FEN1	0.49+/-0.08	**0.52+/-0.07**	-0.39
GBA	**0.59+/-0.05**	0.50+/-0.05	1.35
IDH1	**0.61+/-0.05**	0.50+/-0.06	1.40
KAT2A	0.46+/-0.04	**0.52+/-0.03**	-1.28
MAPK1	**0.60+/-0.04**	0.55+/-0.03	1.11
MTORC1	0.46+/-0.04	**0.52+/-0.03**	-1.30
OPRK1	0.67+/-0.06	0.67+/-0.07	0.02
PKM2	**0.58+/-0.03**	0.51+/-0.05	1.18
PPARG	**0.77+/-0.12**	0.56+/-0.08	1.42
TP53	**0.55+/-0.05**	0.52+/-0.04	0.58
VDR	0.46+/-0.05	**0.54+/-0.04**	-1.45

Table 1: AUROC values on the LIT-PCBA dataset for ACP4 Vs Pharao; up to 50 randomly drawn queries per protein target. A bold font target name indicates the difference between distributions is statistically significant according to a two-sample Kolmogorov-Smirnov test with p-value $\leq 5\%$. A bold $\mu \pm \sigma$ value means the difference between means is not small ($|d| \geq 0.2$).

Comparison Vs. RDKit's 3D Pharmacophore Fingerprints

	ACP4	RdPh4	Cohen d
ADRB2	0.50+/-0.09	0.50+/-0.07	-0.02
ALDH1	**0.53+/-0.02**	0.52+/-0.03	0.41
ESR1p	**0.58+/-0.07**	0.55+/-0.08	0.42
ESR1m	**0.61+/-0.06**	0.55+/-0.06	0.81
FEN1	0.51+/-0.08	**0.56+/-0.07**	-0.63
GBA	0.57+/-0.06	0.57+/-0.05	0.08
IDH1	**0.61+/-0.05**	0.52+/-0.05	1.26
KAT2A	0.46+/-0.04	**0.48+/-0.04**	-0.50
MAPK1	**0.59+/-0.05**	0.57+/-0.05	0.43
MTORC1	0.46+/-0.03	0.47+/-0.05	-0.14
OPRK1	**0.67+/-0.06**	0.58+/-0.18	0.62
PKM2	0.58+/-0.03	**0.61+/-0.03**	-0.85
PPARG	**0.77+/-0.12**	0.71+/-0.10	0.53
TP53	0.54+/-0.05	0.55+/-0.05	-0.19
VDR	0.47+/-0.05	**0.52+/-0.08**	-0.72

Table 2: AUROC values on the LIT-PCBA dataset for ACP4 Vs RdPh4. Same legend as previous table but independent experiment.

Bioactive Conformer Vs Calculated Lowest Energy Conformer

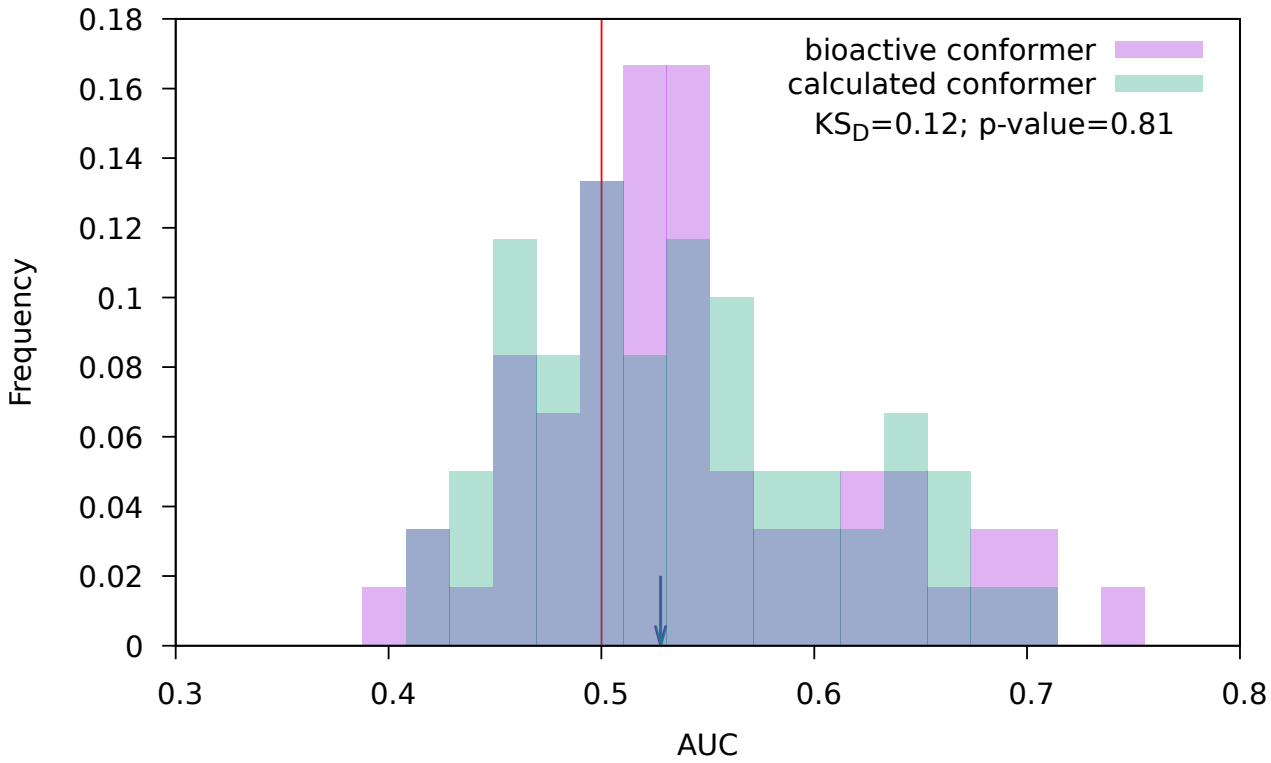


Figure 5: Similarity searches using a ligand's bioactive conformer or its calculated lowest-energy conformer (OpenEye OMEGA-4.1.2). A Kolmogorov-Smirnov statistical test does not discriminate the two distributions. The two arrows pointing at the median AUROC values on the X axis overlap.

ACP4 Comes in Two Flavors

- Ligand comparison mode: $\delta x = 0.5$Å, $d_{cut} = 5.0$Å.
- Binding-site comparison mode: $\delta x = 0.9$Å, $d_{cut} = 40$Å, $ligand_{HA_{cut}} = 5.0$Å.
- Low number of fittable parameters: 2 (+ 1 for binding-site extraction from bound ligand).
- Sparse encoding (sparsity 15% to 18%).
- Not very high dimensional (ligand:232; binding-sites:946 dimensions).

Comparing Protein Ligand-binding Sites: Same Business!

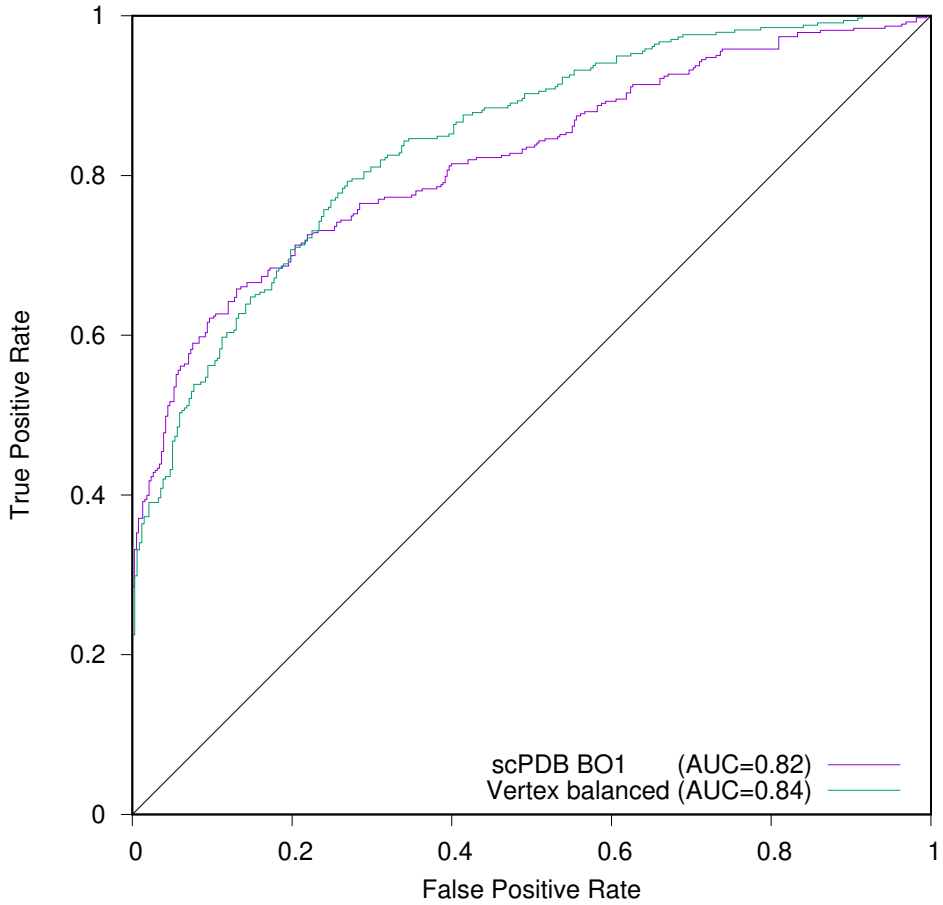


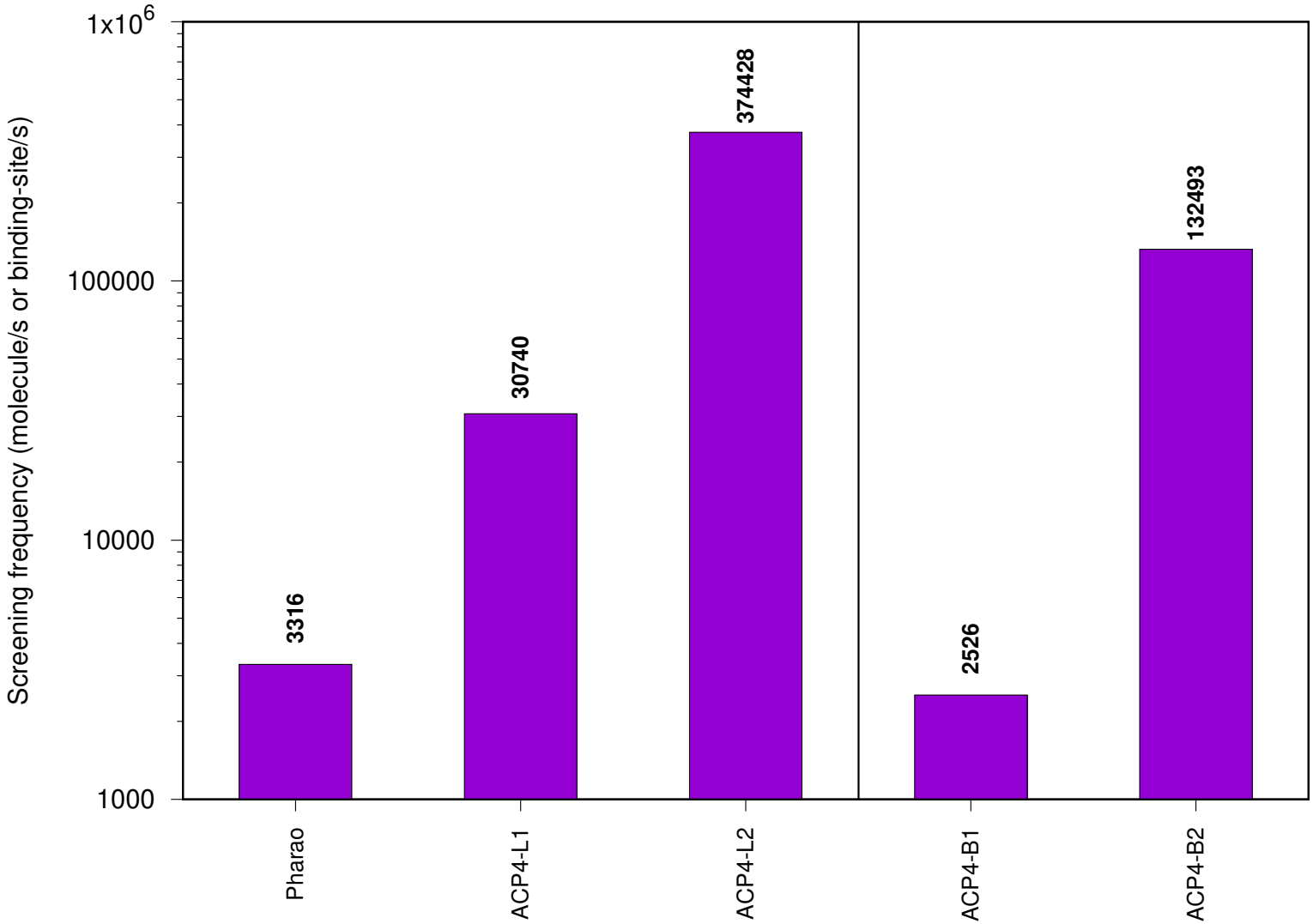
Figure 6: Performance at discriminating similar from dissimilar binding-sites on the sc-PDB BO1 subset (purple ROC curve; training set) or the Vertex balanced dataset (green curve; test set).

Comparison Vs Other Binding-site Comparison Methods

Method	Completeness (%)	AUC
ProBiS	64.2	0.896
PocketMatch	99.4	0.895
Kripo	95.2	0.862
SiteAlign	100.0	0.859
ACP4	100.0	0.842
FuzCav	100.0	0.831
ProCare	99.7	0.811
Shaper	99.7	0.774

Table 3: Performance comparison with binding-site comparison methods on the Vertex balanced dataset. All results, except for ACP4, were taken from the literature. Completeness indicates how much of the dataset a method can process.

Screening Speed



Tanimoto Distance and Activity Probability (in ph4 space)

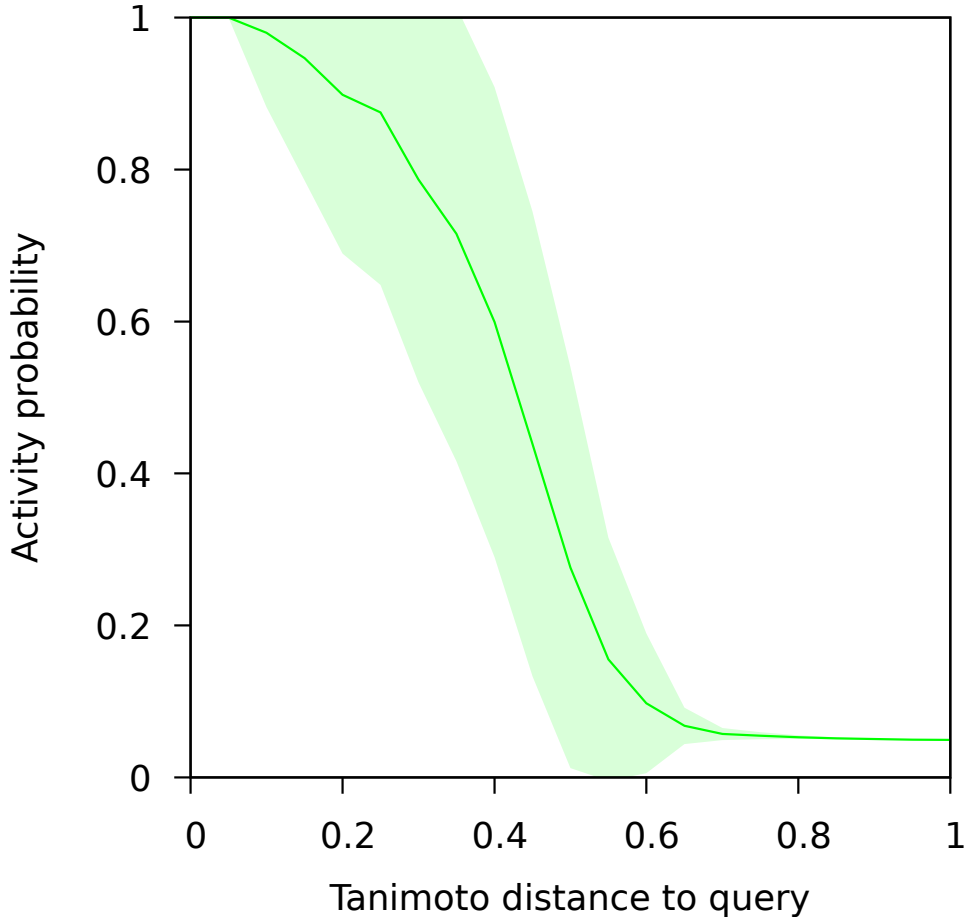


Figure 7: Your query is an active molecule. What is the probability that another molecule at distance d from your query is also active?

Binding-site Similarity Searches

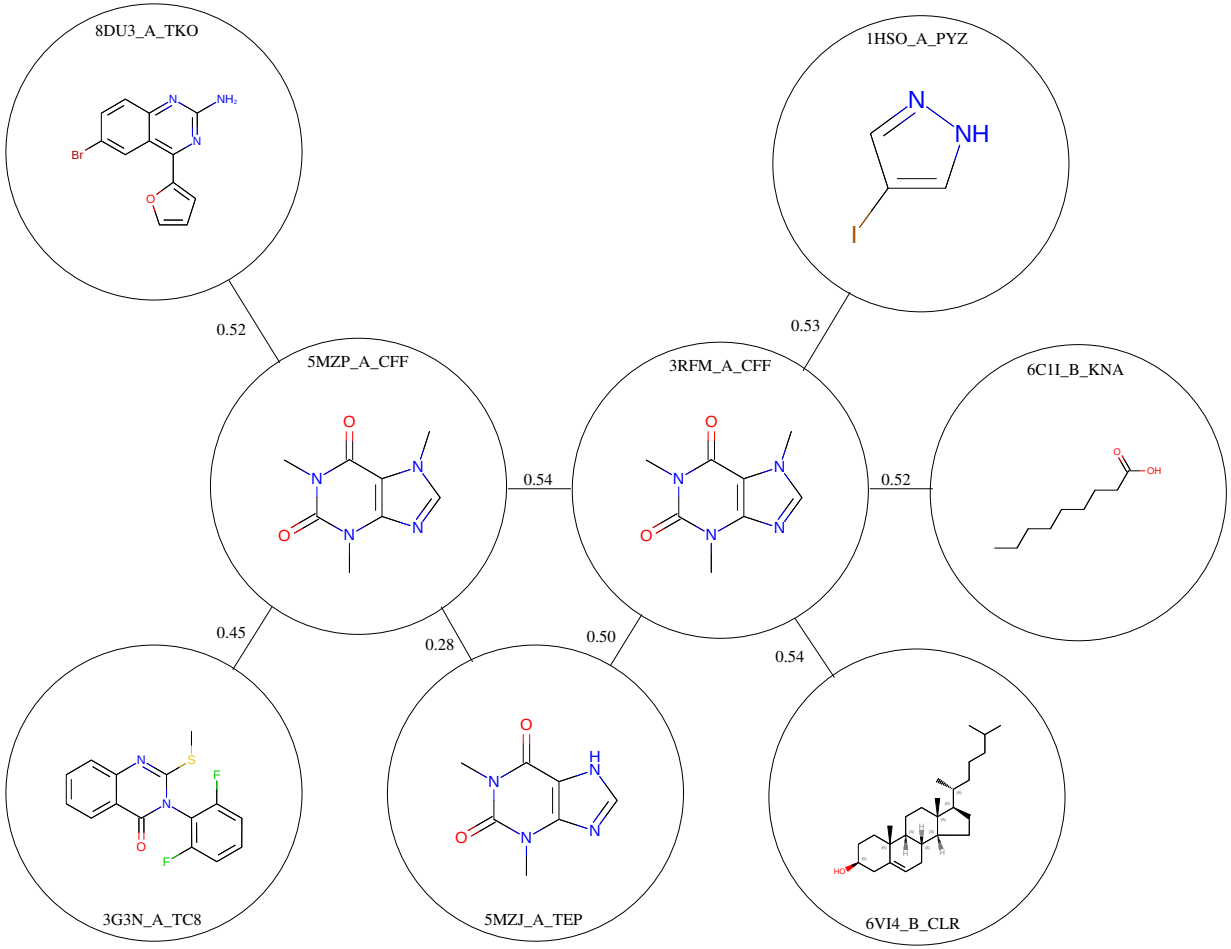


Figure 8: Ligands found among the most similar ligand-binding sites retrieved among 667 human protein structures using two caffeine binding sites as queries (5MZP_A_CFF and 3RFM_A_CFF).

All questions and comments are welcome!

Uses RDKit and is open-source:
`https://github.com/tsudalab/ACP4/`