Fragment binding prediction using unsupervised learning of ligand substructure binding sites

Grace Tang
Altman Lab
July 20, 2013
Berlin, Germany

STANFORD BIOENGINEERING

Structure Based Virtual Screening

Databases ..... Virtual Screening ..... End Goal

***ZINC*[1]**

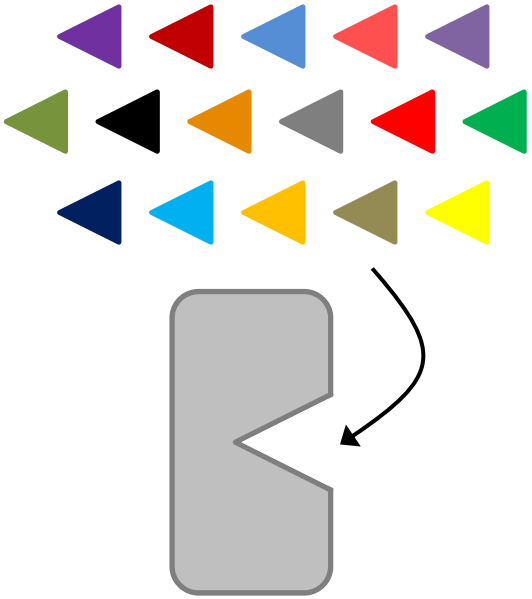
21 million purchasable compounds

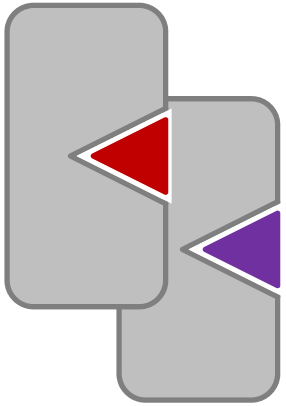
***PubChem*[2]**

48 million compounds

***GDB-13*[3]**

970 million drug-like small molecules





1. Irwin, J.J., et al., *ZINC: A Free Tool to Discover Chemistry for Biology.* J Chem Inf Model, 2012.
2. Bolton E., et al., *PubChem: Integrated Platform of Small Molecules and Biological Activities*. Annual Reports in Computational Chemistry, 2008.

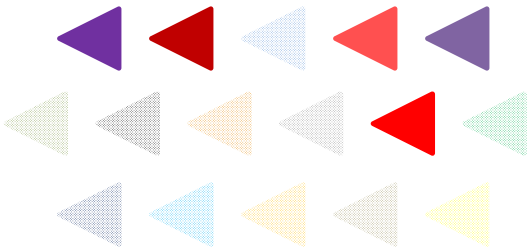
Adding Prior Knowledge

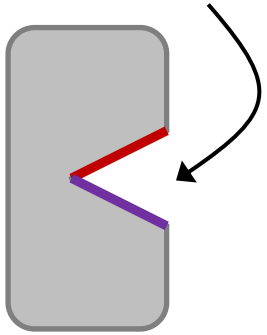
Databases ····· Virtual Screening ····· End Goal

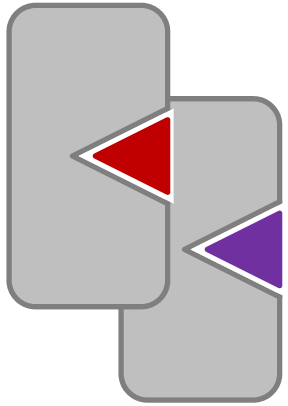
ZINC[1]
21 million purchasable compounds

PubChem[2]
48 million compounds

GDB-13[3]
970 million drug-like small molecules







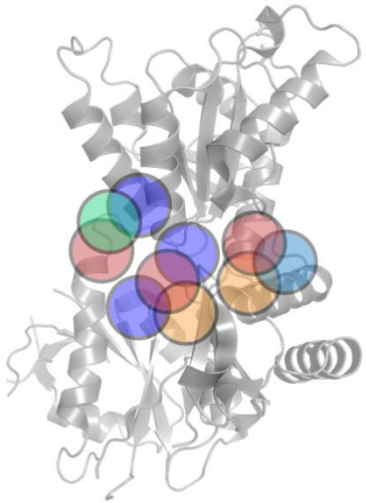
contains fragment: X

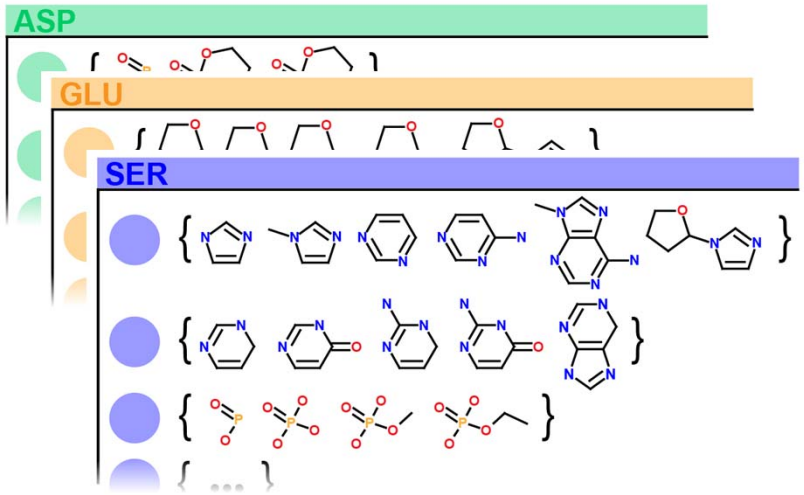
contains fragment: Y

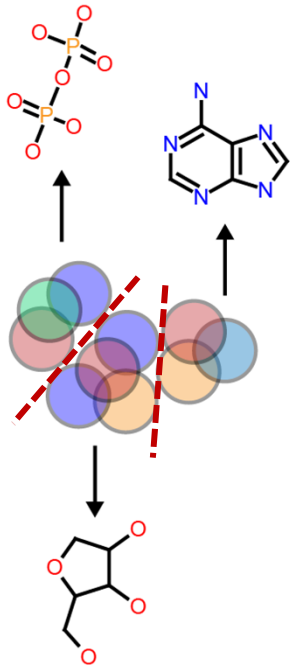
3. Blum, L.C. and J.L. Reymond, *970 million druglike small molecules for virtual screening in the chemical universe database GDB-13.* J Am Chem Soc, 2009. **131**(25): p. 8732-3.

Method Overview

Query Protein ····· Knowledge Base ····· Fragment Predictions



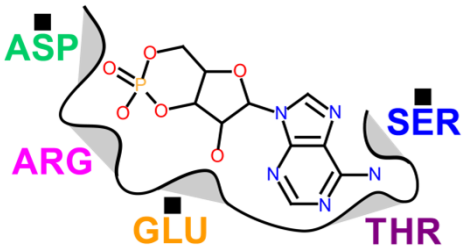


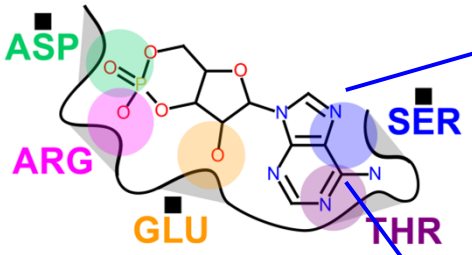


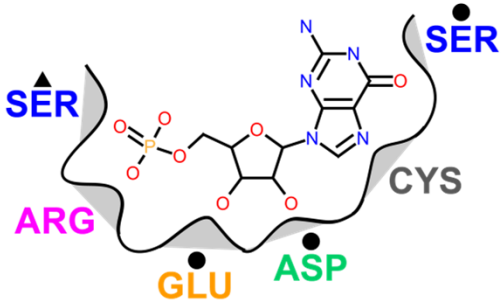
STANFORD BIOENGINEERING

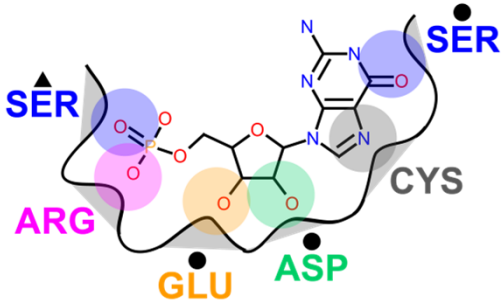
Derivation of Knowledge Base

Protein-Ligand Complexes[1] ⋯⋯▸ Structural Information[2]

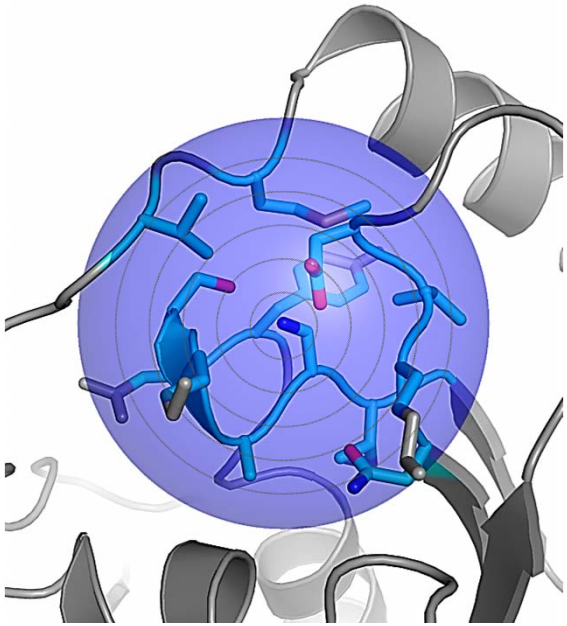








FEATURE



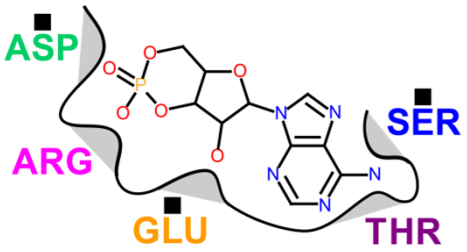
Atom Type	Partial Charge
Atom Element	Hydrophobicity
Residue Name	Aromatic
Residue Class	etc.

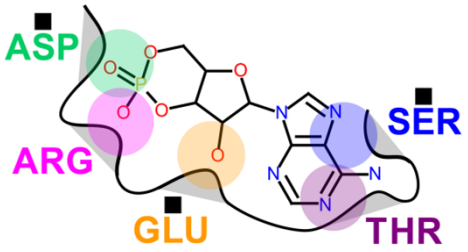
1. Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Res, 2000. **28**(1): p. 235-42.
2. Halperin, I., et al., *The FEATURE framework for protein function annotation.* BMC Genomics, 2008. **9 Suppl 2**: p. S2.

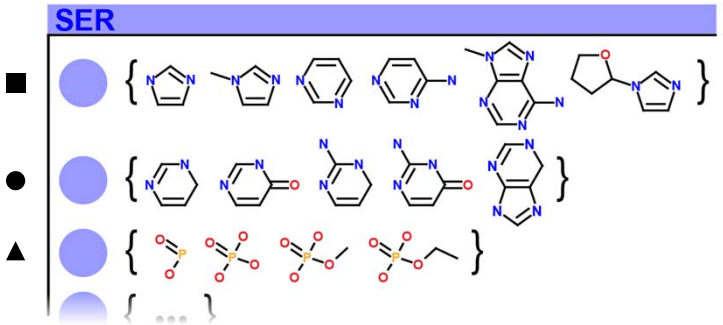
STANFORD BIOENGINEERING

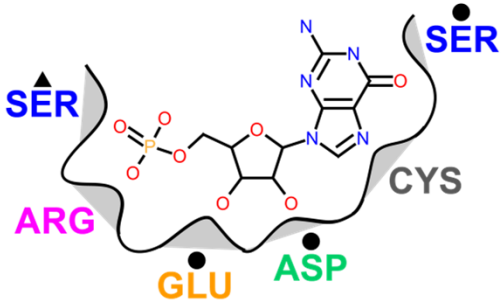
Derivation of Knowledge Base

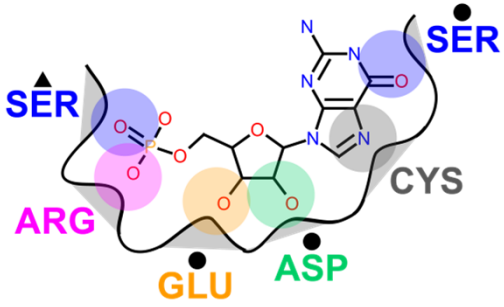
Protein-Ligand Complexes[1] ⋯⋯▸ Fragment Information[2] ⋯⋯▸ Knowledge Base

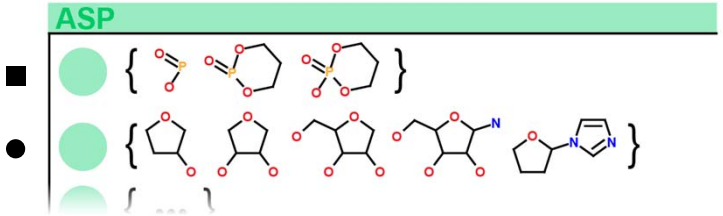


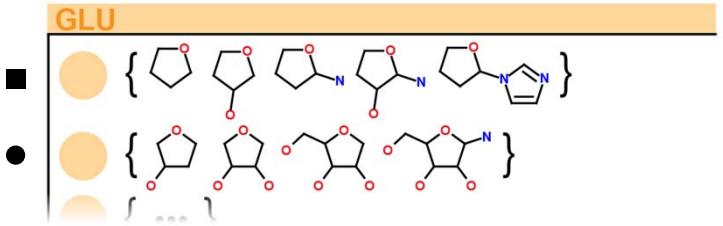








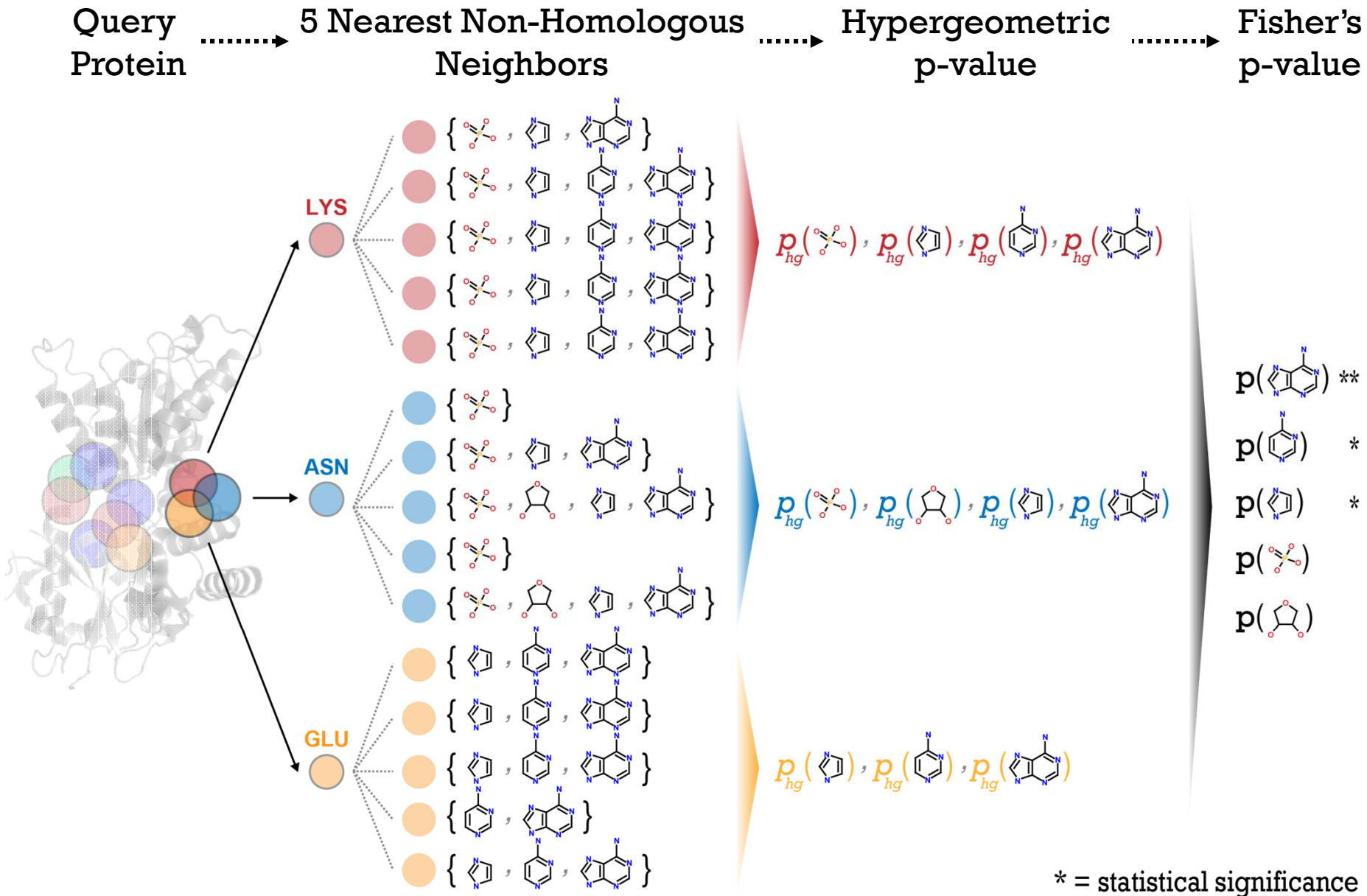




1. Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Res, 2000. **28**(1): p. 235-42.
2. Rahman, S.A., et al., *Small Molecule Subgraph Detector (SMSD) toolkit.* J Cheminform, 2009. **1**(1): p. 12.

STANFORD BIOENGINEERING

Method

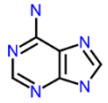


STANFORD BIOENGINEERING

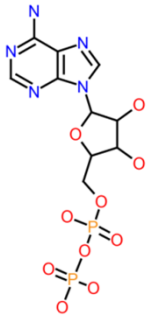
Validation Ligands

ADE	adenine	n = 123
ADP	adenosine-5'-diphosphate	n = 2640
FAD	flavin-adenine dinucleotide	n = 2769
NAD	nicotinamide-adenine dinucleotide	n = 2309
VIB	thiamin, vitamin B1	n = 19
TPP	thiamine diphosphate	n = 217
PLP	pyridoxal-5'-phosphate	n = 1227
TCL	triclosan	n = 88

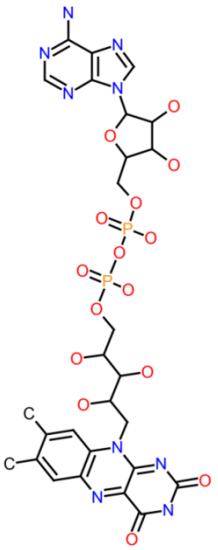
ADE



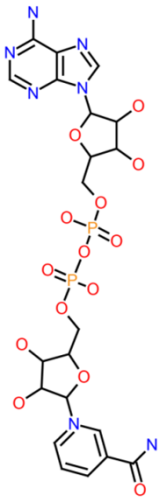
ADP



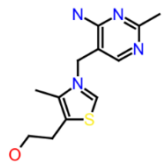
FAD



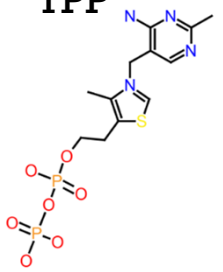
NAD



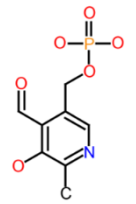
VIB



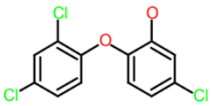
TPP



PLP

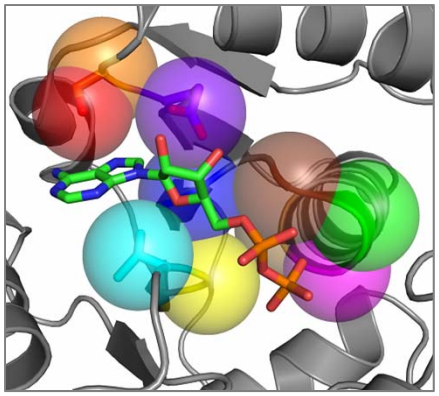


TCL



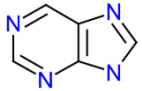
Validation

Protein – ADP Complex

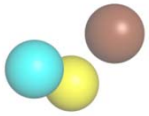


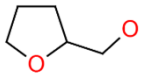
PDB ID: 2X58

Predictions and Moieties

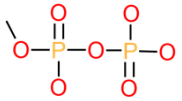


adenine ring 1 & 2





ribose



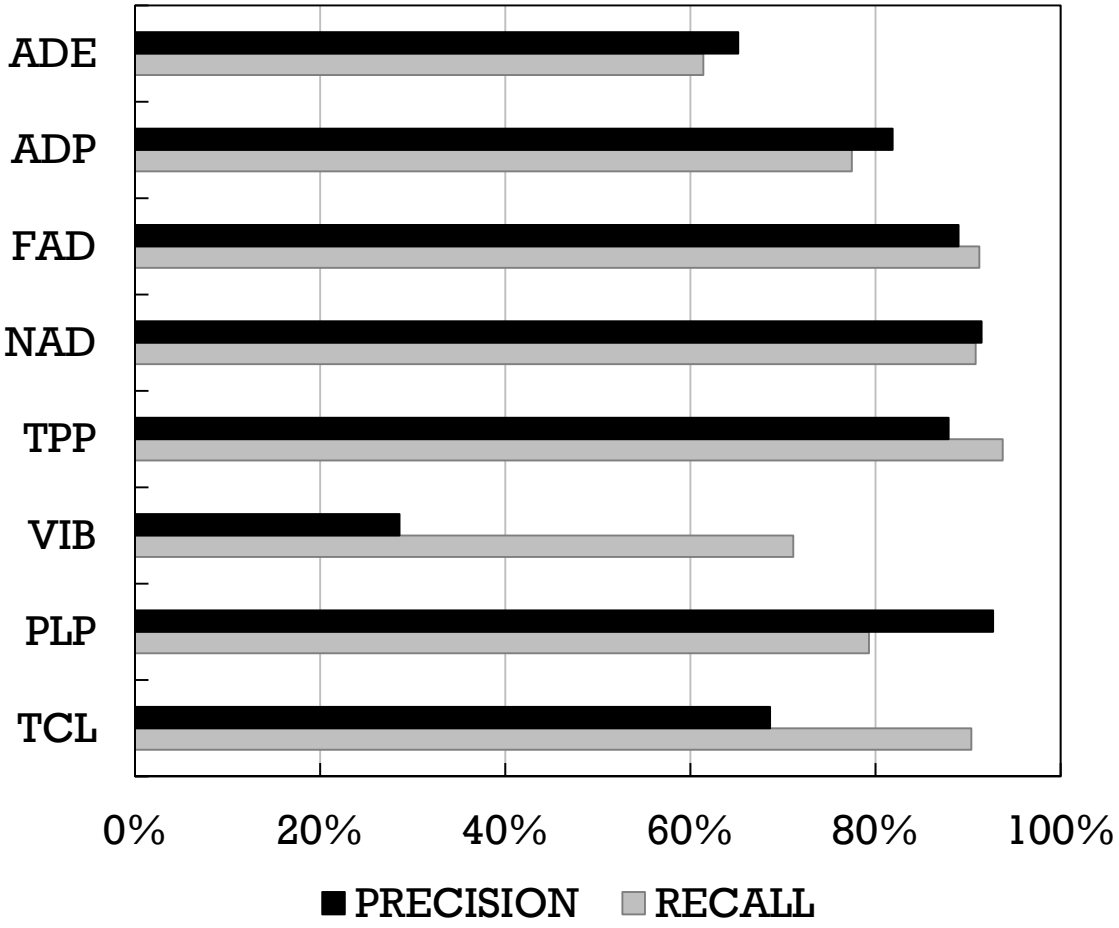
phosphate 1 & 2

Precision:

$$\frac{\text{\# Correct Predictions}}{\text{\# Total Predictions}}$$

Recall:

$$\frac{\text{\# Predicted Moieties}}{\text{\# Bound Moieties}}$$

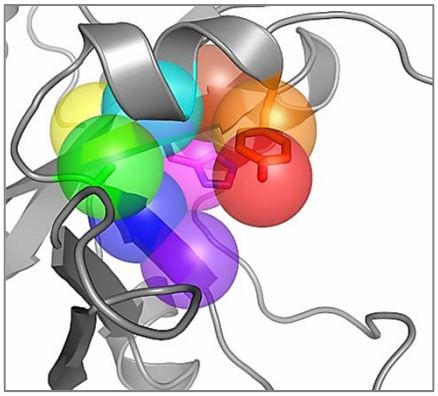


STANFORD BIOENGINEERING

Exotoxin A

Exotoxin A from *Pseudomonas aeruginosa* is an ADP-ribosyltransferase that inactivates eukaryotic ribosomal elongation factor 2, preventing protein synthesis and triggering cell necrosis.

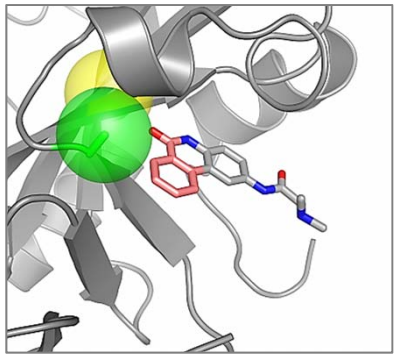
Query



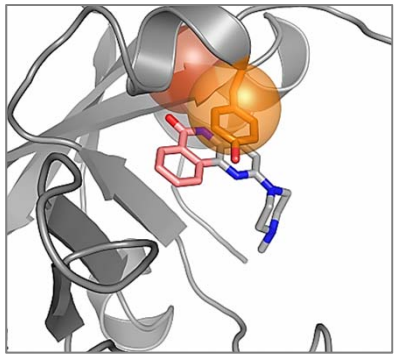
PDB ID: 3B78

Predicted Fragment
CID: 450318
(p-value 1×10^{-28})

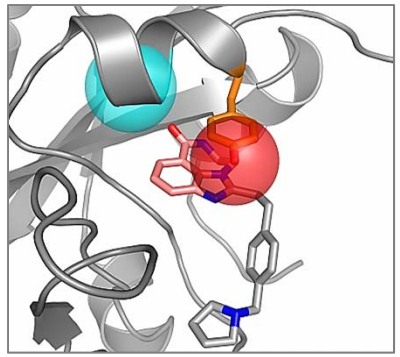
Example Nearest Microenvironment Neighbors



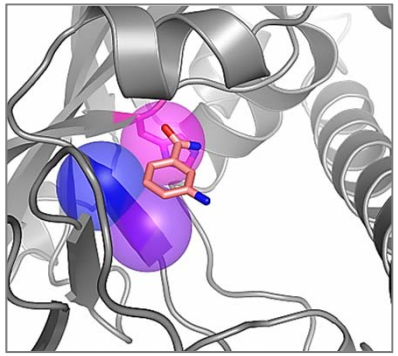
PDB ID: 3GEY



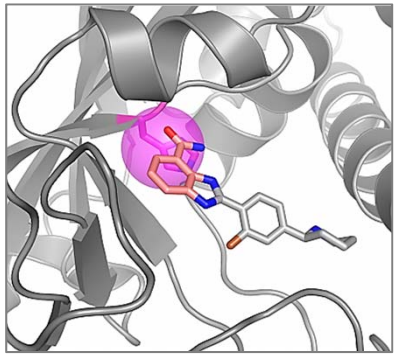
PDB ID: 3KI2



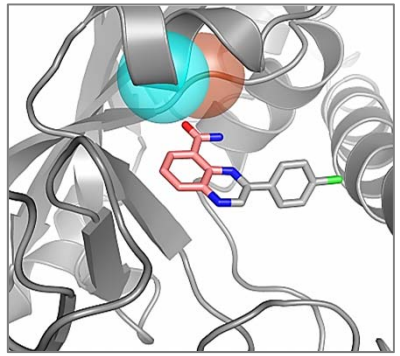
PDB ID: 3KI7



PDB ID: 3KCZ



PDB ID: 3L3M

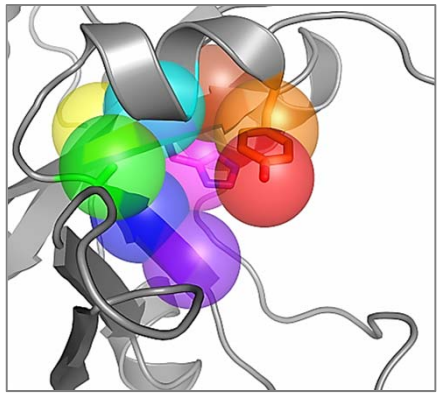


PDB ID: 1WOK

Exotoxin A

Exotoxin A from *Pseudomonas aeruginosa* is an ADP-ribosyltransferase that inactivates eukaryotic ribosomal elongation factor 2, preventing protein synthesis and triggering cell necrosis.

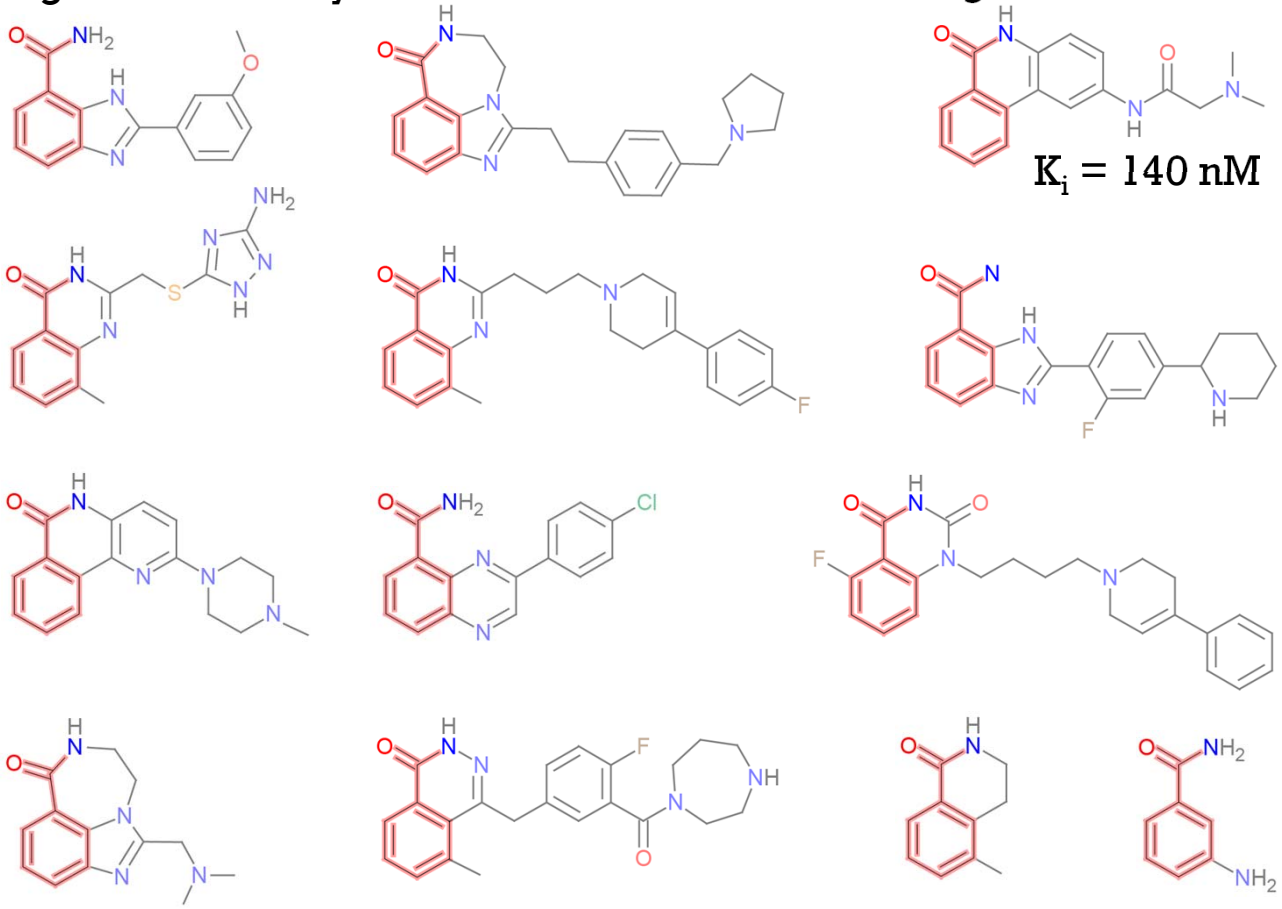
Query



PDB ID: 3B78

Predicted Fragment
CID: 450318
(p-value 1×10^{-28})

Ligands Bound by Nearest Microenvironment Neighbors

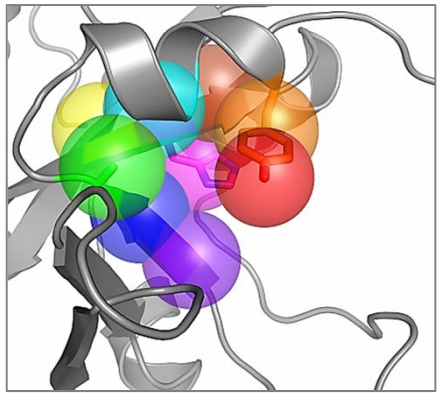


STANFORD BIOENGINEERING

Exotoxin A

Exotoxin A from *Pseudomonas aeruginosa* is an ADP-ribosyltransferase that inactivates eukaryotic ribosomal elongation factor 2, preventing protein synthesis and triggering cell necrosis.

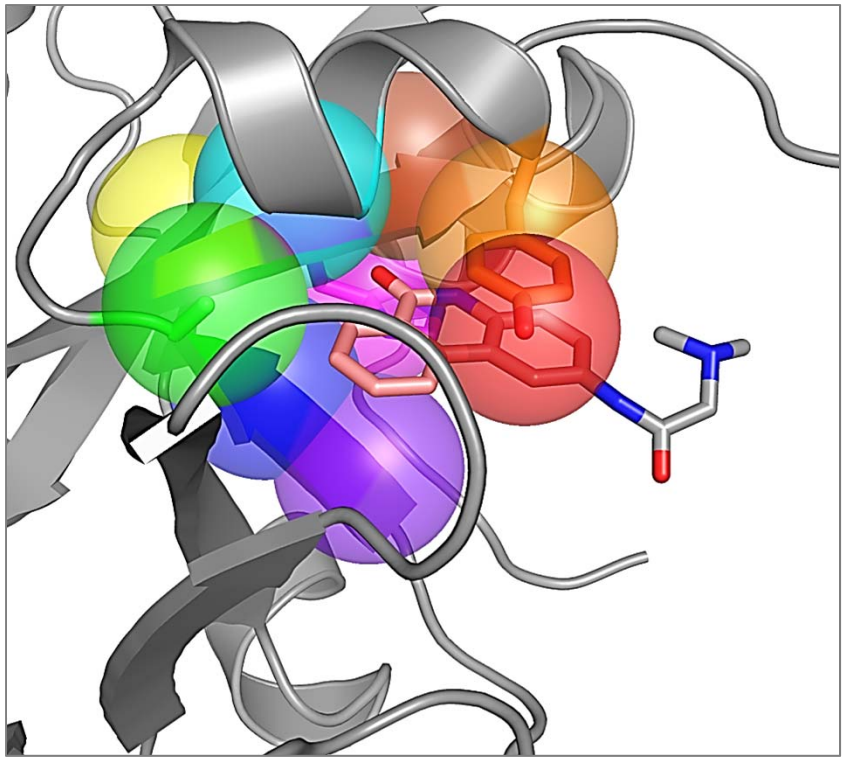
Query



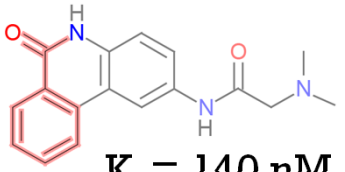
PDB ID: 3B78

Predicted Fragment
CID: 450318
(p-value 1×10^{-28})

Structural validation of predicted fragment



PDB ID: 1XK9



K_i = 140 nM

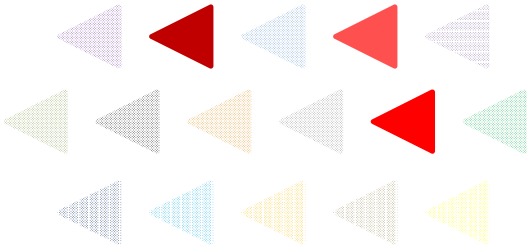
Adding Fragment Knowledge

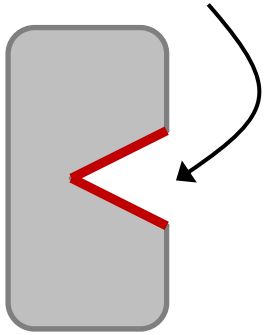
Databases ····· Virtual Screening ····· End Goal

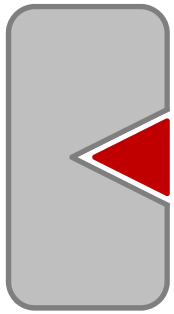
ZINC
21 million purchasable compounds

PubChem
48 million compounds

GDB-13
970 million drug-like small molecules







contains fragment:

STANFORD BIOENGINEERING

Acknowledgements

- Russ Altman
- Altman Lab

- Funding Sources

- Travel funding to ISMB/ECCB 2013 was generously provided by ISCB

Thank You!

gwtang@stanford.edu

3DSIG:

Poster … 046

ISMB/ECCB:

Poster … L081

Fragment binding prediction using unsupervised learning of ligand substructure binding sites

Grace W. Tang[1] and Russ B. Altman[1,2]
Departments of [1]Bioengineering and [2]Genetics, Stanford University

Motivation

In structure-based drug discovery, virtual screening plays an important role in the identification of hit and lead molecules for a protein target of interest. Docking algorithms are popular for this task but have limited throughput in order to account for target and ligand flexibility. These algorithms cannot accommodate the increasing size of small molecule databases and number of target structures without reduction of method complexity and accuracy. Identification of database subsets pre-filtered for molecules with potential bioactivity against a protein target focuses computational resources, enabling sophisticated screening techniques. As bioactive compounds against a target frequently share chemical substructures, we propose an unsupervised machine learning approach to predict small molecule fragments compatible with a target protein structure. We take advantage of the availability of structural data for proteins whose bound ligands share substructures to enhance our understanding of fragment binding to develop a fragment binding predictor. Our method requires as input a 3D structure of the target but requires no prior knowledge of active compounds, enabling broad usage.

Method

For all protein residues involved in ligand binding, we collect their local structural microenvironment using FEATURE[1] and annotate them with the ligand fragments they bind[2]. This serves as the knowledge base of protein-fragment interactions. Comparison to the knowledge base enables retrieval of fragments statistically preferred by the microenvironments of a target structure.

Part 1: Build a knowledge base of protein-fragment associations

Protein-Ligand Complexes (PDB[3]) → Derived Information → Residue-Specific Knowledge Bases

FEATURE microenvironment

6 concentric shells capturing 80 physicochemical features (atom type, residue class, hydrophobicity etc.)

fragments associated to each microenvironment

Part 2: Apply the knowledge base to query binding pockets

Putative Binding Pocket → 5 Nearest Non-Homologous Neighbors → Hypergeometric P-value → Fisher's Combined P-value

spatially proximal microenvironments

p-value ranked fragments

* = statistical significance

Performance

Results on a dataset of proteins binding a variety of ligands show strong ability to rediscover fragments corresponding to the ligand bound, validating the knowledge base and method.

ADE	adenine	n = 123
ADP	adenosine-5'-diphosphate	n = 2640
FAD	flavin-adenine dinucleotide	n = 2769
NAD	nicotinamide-adenine dinucleotide	n = 2309
VIB	thiamin, vitamin B1	n = 19
TPP	thiamine diphosphate	n = 217
PLP	pyridoxal-5'-phosphate	n = 1227
TCL	triclosan	n = 88

ADE FAD NAD VIB PLP ADP TPP TCL

Example protein-ADP complex. Protein in cartoon representation with ADP in stick representation. Microenvironments shown as semi-transparent spheres

Precision = # Correct Predictions / # Total Predictions

Recall = # Predicted Moieties / # Bound Moieties

ADE ADP FAD NAD TPP VIB PLP TCL

0% 20% 40% 60% 80% 100%

■ PRECISION ■ RECALL

Correct Predictions & Moieties

adenine ring 1 & 2 ribose phosphate 1 & 2

References

1. Halperin, I., Glazer, D. S., Wu, S. & Altman, R. B. The FEATURE framework for protein function annotation: modeling new functions, improving performance, and extending to novel applications. *BMC Genomics* **9 Suppl 2**, S2 (2008).
2. Rahman, S. A., Bashton, M., Holliday, G. L., Schrader, R. & Thornton, J. M. Small Molecule Subgraph Detector (SMSD) toolkit. *J Cheminform* **1**, 12 (2009).
3. Berman, H. M. et al. The Protein Data Bank. *Nucleic Acids Res.* **28**, 235-242 (2000).

This work is supported by NIH LM05652 and fimbrio SIEBEL SCHOLARS Travel funding provided by ISCB

Inhibitor Fragment Predictions

For multiple protein targets, we identify high scoring fragments that are substructures of known inhibitors. Our method therefore predicts fragments suitable for preprocessing small molecule databases to enrich for bioactive compounds for a given protein target.

Exotoxin A from *Pseudomonas aeruginosa* is an ADP-ribosyltransferase that inactivates eukaryotic ribosomal elongation factor 2, preventing protein synthesis and triggering cell necrosis. Shown are nearest neighbors for the microenvironments predicting benzamide.

Predicted fragment benzamide

P34 (Exotoxin A Inhibitor) K_i = 140 nM

Microenvironments predicting benzamide

Exotoxin A bound to P34

Protein Kinase A (PKA) phosphorylates a large number of proteins in the cytoplasm and nucleus, regulating processes including lipolysis, ion channel permeability, and cell proliferation. Shown are nearest neighbors for the microenvironments predicting fragments 4989215, 11182733, and 6386.

Predicted fragment 4989215 | Predicted fragment 11182733 | Predicted fragment 6386

Microenvironments predicting fragment 4989215 | Microenvironments predicting fragment 11182733 | Microenvironments predicting fragment 6386

R69 (PKA Inhibitor) *IC50 = 30 nM | CQP (PKA Inhibitor) *IC50 = 400 nM | G98 (PKA Inhibitor) *IC50 = 19 nM

PKA bound to R69 | PKA bound to CQP | PKA bound to G98

* activity against 97% sequence identity homolog

* activity for compound sharing same scaffold

STANFORD BIOENGINEERING