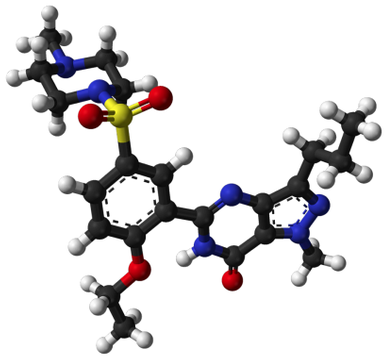
# nAnnolyze : ligand-target prediction by structural network biology

Francisco Martínez-Jiménez

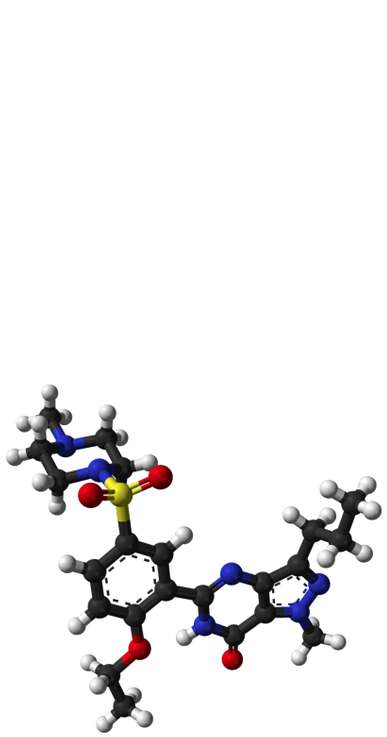
*Drug Development Workshop, ECCB, Strasbourg*

# Finding out the mode of action..



Phenotype

# Finding out the mode of action..



?

Phenotype

# Existing computational methods

Prediction details & accuracy

Computational time

# Existing computational methods

Prediction details & accuracy

**free structure methods**

★Based on previous knowledge.
★Many different methods.
★Good performance.
★Poor information about the interaction.

Computational time

# Existing computational methods

Prediction details & accuracy

**free structure methods**

★Based on previous knowledge.
★Many different methods.
★Good performance.
★Poor information about the interaction.

**structure based methods**

Virtual Docking

★Very precise. Ligand and receptor orientation.
★Needs the binding-site.
★Needs the structure or a reliable 3D-model.
★Not applicable at wide scale.

Computational time

# Existing computational methods

Prediction details & accuracy

## free structure methods

★Based on previous knowledge.
★Many different methods.
★Good performance.
★Poor information about the interaction.

## structure based methods

### Comparative Docking

★Outputs binding-site localization.
★Based on structural comparisons.
★Applicable at wide scale.
★Needs the structure or a reliable 3D-model.

### Virtual Docking

★Very precise. Ligand and receptor orientation.
★Needs the binding-site.
★Needs the structure or a reliable 3D-model.
★Not applicable at wide scale.

Computational time

# Comparative Docking

**BMC Bioinformatics**

BioMed Central

Proceedings

Open Access

## The AnnoLite and AnnoLyze programs for comparative annotation of protein structures

Marc A Marti-Renom*[1], Andrea Rossi[2], Fátima Al-Shahrour[3], Fred P Davis[2], Ursula Pieper[2], Joaquín Dopazo[3] and Andrej Sali[2]

Address: [1]Structural Genomics Unit, Bioinformatics Department, Centro de Investigación Príncipe Felipe (CIPF), Valencia, Spain, [2]Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry, and California Institute for Quantitative Biomedical Research, University of California at San Francisco, San Francisco, CA 94143, USA and [3]Functional Genomics Unit, Bioinformatics Department, Centro de Investigación Príncipe Felipe (CIPF), Valencia, Spain

Email: Marc A Marti-Renom* - mmarti@cipf.es; Andrea Rossi - andrea@salilab.org; Fátima Al-Shahrour - falshahrour@cipf.es; Fred P Davis - fred@salilab.org; Ursula Pieper - Ursula@salilab.org; Joaquín Dopazo - jdopazo@cipf.es; Andrej Sali - sali@salilab.org

* Corresponding author

from The Second Automated Function Prediction Meeting
La Jolla, CA, USA. 30 August – 1 September 2006

Published: 22 May 2007

*BMC Bioinformatics* 2007, 8(Suppl 4):S4 doi:10.1186/1471-2105-8-S4-S4

This article is available from: http://www.biomedcentral.com/1471-2105/8/S4/S4

© 2007 Marti-Renom et al; licensee BioMed Central Ltd.
This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**Abstract**

**Background:** Advances in structural biology, including structural genomics, have resulted in a rapid increase in the number of experimentally determined protein structures. However, about half of the structures deposited by the structural genomics consortia have little or no information about their biological function. Therefore, there is a need for tools for automatically and comprehensively annotating the function of protein structures. We aim to provide such tools by applying comparative protein structure annotation that relies on detectable relationships between protein structures to transfer functional annotations. Here we introduce two programs, AnnoLite and AnnoLyze, which use the structural alignments deposited in the DBAli database.

**Description:** AnnoLite predicts the SCOP, CATH, EC, InterPro, PfamA, and GO terms with an average sensitivity of ~90% and average precision of ~80%. AnnoLyze predicts ligand binding site and domain interaction patches with an average sensitivity of ~70% and average precision of ~30%, correctly localizing binding sites for small molecules in ~95% of its predictions.

**Conclusion:** The AnnoLite and AnnoLyze programs for comparative annotation of protein structures can reliably and automatically annotate new protein structures. The programs are fully accessible via the Internet as part of the DBAli suite of tools at http://salilab.org/DBAli/.

### Background

Genomic efforts are providing us with complete genetic blueprints for hundreds of organisms, including humans. We are now faced with assigning, understanding, and modifying the functions of proteins encoded by these genomes. This task is generally facilitated by protein 3D



*Published online 21 November 2011* *Nucleic Acids Research, 2012, Vol. 40, Database issue* **D549–D553**
*doi:10.1093/nar/gkr1049*

## ProtChemSI: a network of protein–chemical structural interactions

**Olga V. Kalinina[1,2], Oliver Wichmann[1], Gordana Apic[1,3] and Robert B. Russell[1,*]**

[1]Cell Networks, BioQuant, University of Heidelberg, Im Neuenheimer Feld 267, 69120 Heidelberg, Germany, [2]Institute for Information Transmission Problems, RAS, Bolshoi Karetny pereulok 19, Moscow, 127994, Russia and [3]Cambridge Cell Networks Ltd, St John's Innovation Centre, Cowley Road, Cambridge CB4 OWS, UK

Received August 15, 2011; Revised October 4, 2011; Accepted October 25, 2011

**ABSTRACT**

**Progress in structure determination methods means that the set of experimentally determined 3D structures of proteins in complex with small molecules is growing exponentially. ProtChemSI exploits and extends this useful set of structures by both collecting and annotating the existing data as well as providing models of potential complexes inferred by protein or chemical structure similarity. The database currently includes 7704 proteins from 1803 organisms, 11 324 chemical compounds and 202 289 complexes including 178 974 predicted. It is publicly available at http://pcidb.russelllab.org.**

### INTRODUCTION

Protein–chemical interactions are most often not considered in the context of three-dimensional (3D) structures. Most databases, such as DrugBank (1) or STITCH (2) will refer to 3D structures but do not exploit them beyond reporting that a structure for a drug–protein interaction is known. Other databases, such as Binding MOAD (3), PDBbind (4) and BindingDB (5), focus on collecting protein–ligand complexes, but report only those that are experimentally resolved. However, the current network of protein–chemical interactions derived from 3D structures is a rich source of information and provides many possibilities to suggest new protein–chemical interactions.

Recently, we published a method to predict novel protein–chemical interactions using superimposition of known 3D structures (6). The underlying principle is that if two proteins share a common ligand, and the first protein is known to bind a second ligand, the 3D structures of protein–ligand complexes can be superimposed to build a model that can be used to evaluate a complex of the second protein with that second ligand (Figure 1, lower). Here we present ProtChemSI, a database providing these computed complexes. The database also contains known structures of protein–chemical complexes, and several other predicted complexes. Specifically, we also construct models for all interactions with molecules similar to known interaction partners of a protein or a chemical of interest (Figure 1, explained in detail below), and provide a method to traverse the network of interactions to identify possibilities for building a structural model of any protein chemical pair of interest (Figure 2).

Being primarily based on structural interactions, ProtChemSI has little overlap with other databases for protein–chemical interactions, such as DrugBank (1), STITCH (2) and ChEMBL (7) (Table 1). Theoretically, protein–chemical interactions viewed as a network provide a possibility to construct a model of a complex of any given protein and chemical, superimposing molecules along the path that connects them. ProtChemSI implements a routine to construct and evaluate these models on user demand, so the total number of theoretically possible models in ProtChemSI is very large and impossible to quantify. However, including first-order models (i.e. where we consider interactions no more than two-steps away in the network), we have a total of 23 315 known complexes, and predictions, where 65 502 are modeled by obvious homology, 18 917 are modeled by obvious chemical similarity and 94 555 are modeled by superimpositions as detailed in our original study (6).

### FUNCTIONS OF THE DATABASE

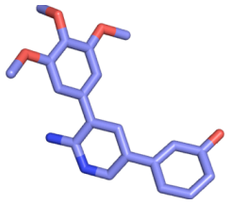
ProtChemSI is intended for those interested in structural details of interactions between proteins and small molecules. It provides details at two levels of certainty: first, it lists all experimentally resolved 3D structures involving the query protein or chemical; second, it constructs a number of models as detailed below.

The workflow of the model construction is schematically represented in Figure 1. For a query protein, models of the following complexes are constructed: (i) with

*To whom correspondence should be addressed. Tel: +49 6221 54 51 362; Fax: +49 6221 54 51 486; Email: robert.russell@bioquant.uni-heidelberg.de
Present address:
Olga V. Kalinina, Max-Planck-Institut für Informatik, Campus E1 4, 66123 Saarbrücken, Germany.

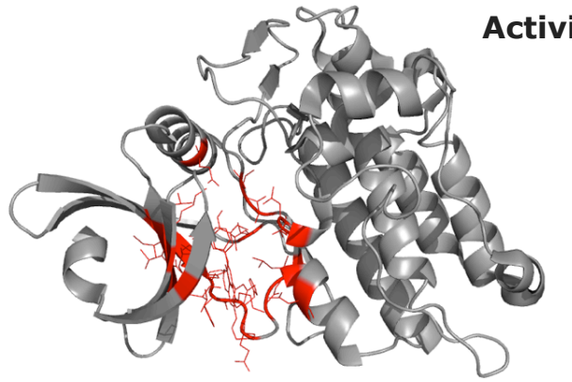
© The Author(s) 2011. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

# Similar binding-sites tend to bind similar ligands



A3F

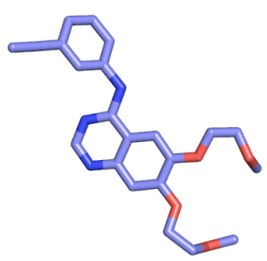
co-crystallized



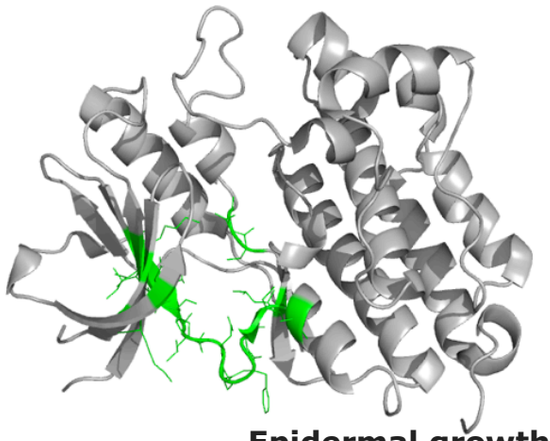
**Activin receptor type-1**

Similar binding-sites

AQ4

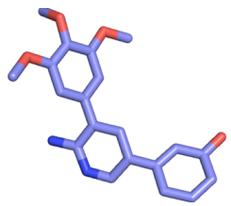


co-crystallized



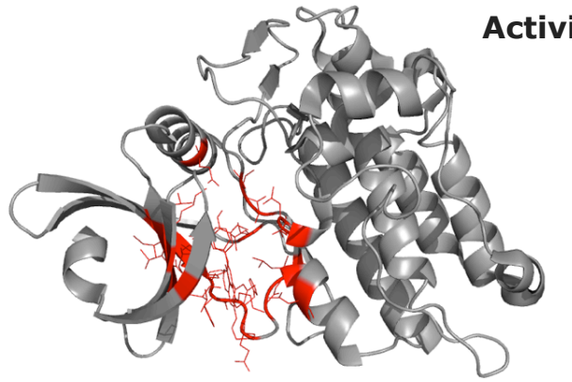
**Epidermal growth factor receptor**

# Similar binding-sites tend to bind similar ligands

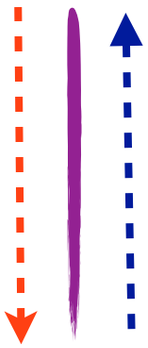


A3F

co-crystallized

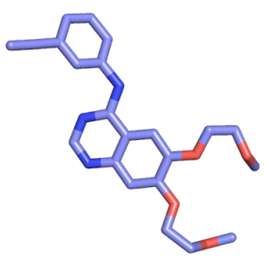


Activin receptor type-1

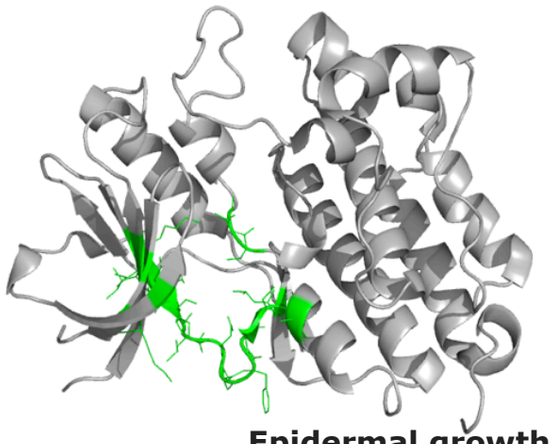


Similar binding-sites

AQ4

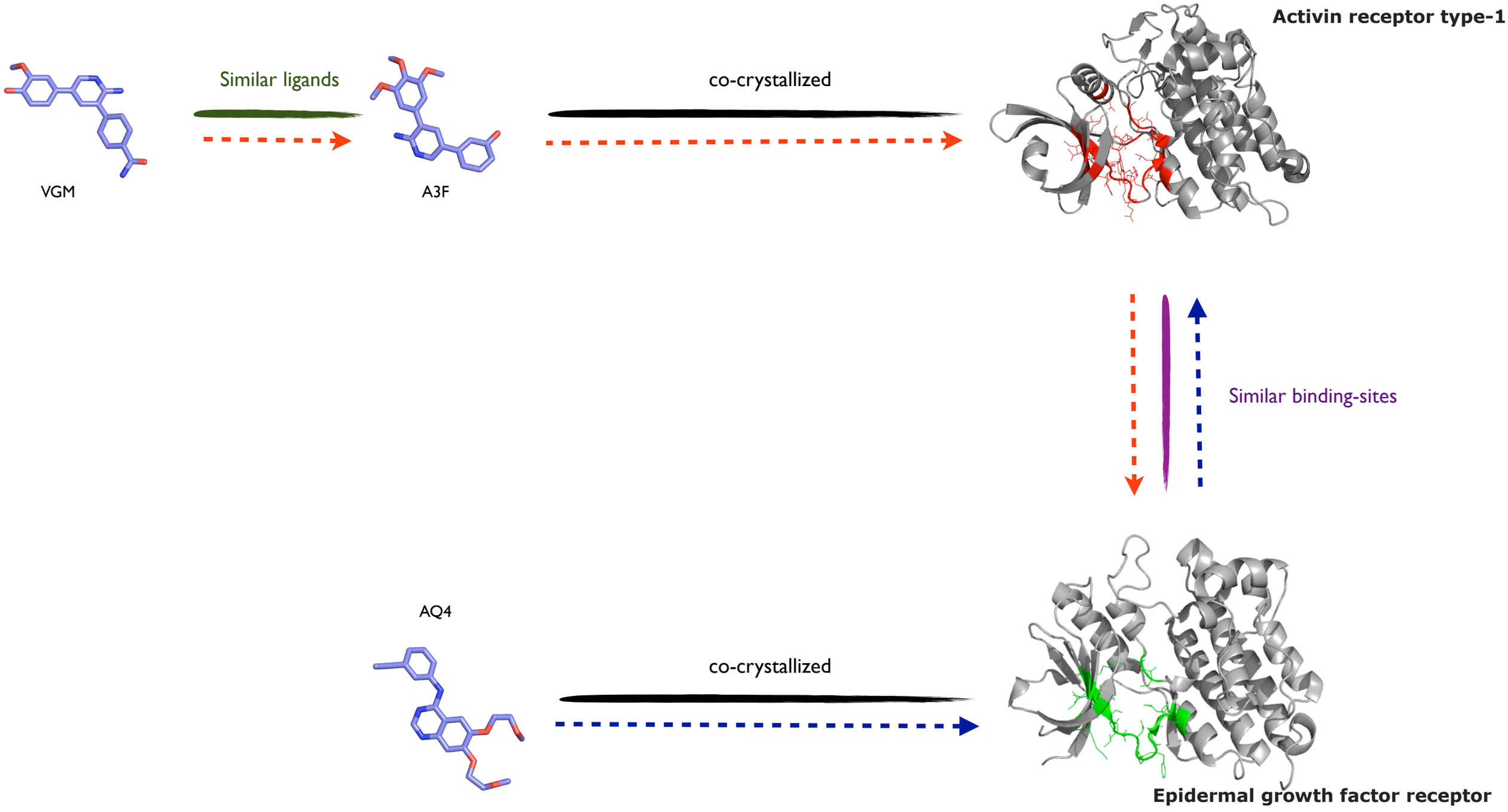


co-crystallized

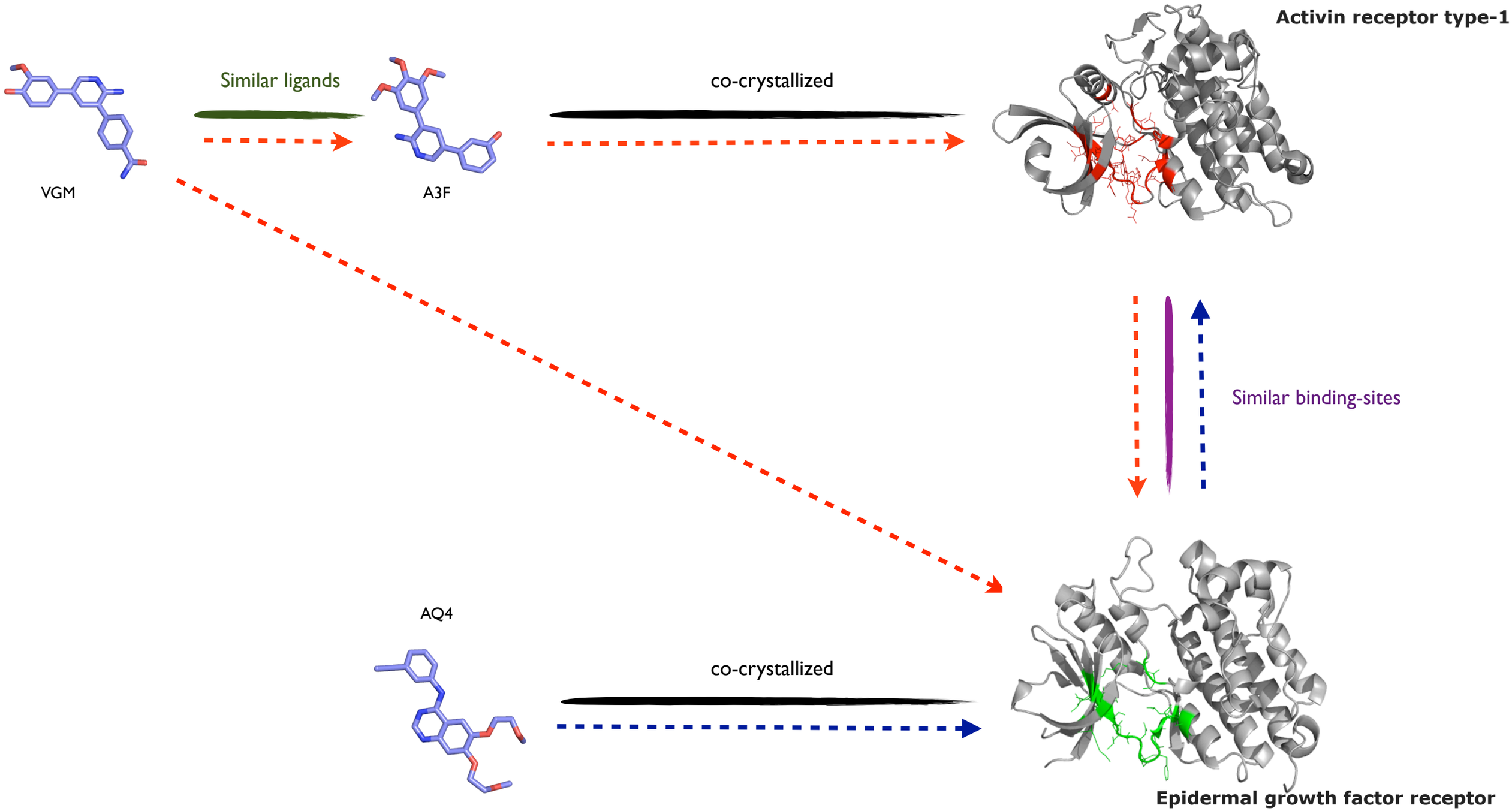


Epidermal growth factor receptor

# Similar binding-sites tend to bind similar ligands

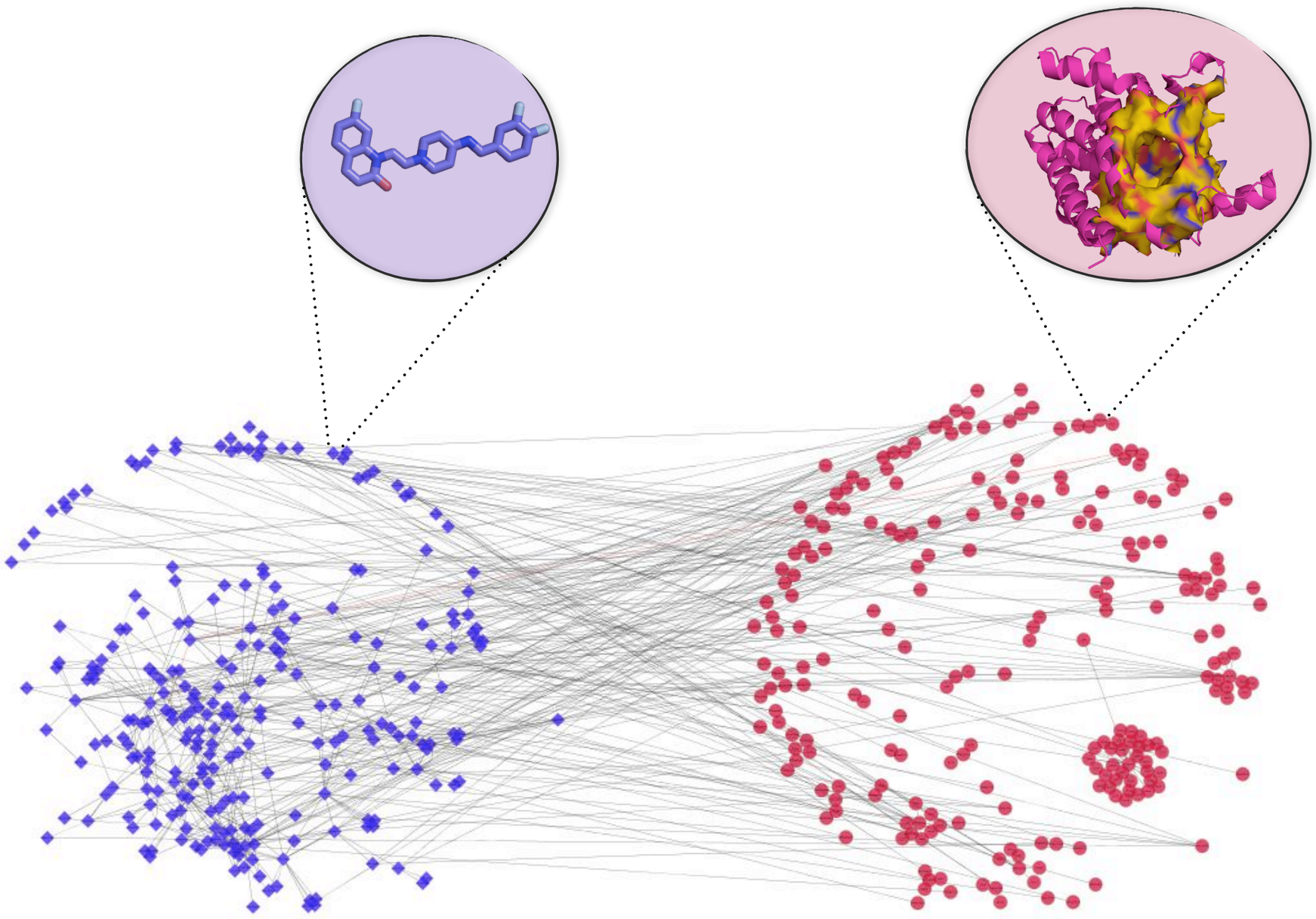


# Similar binding-sites tend to bind similar ligands



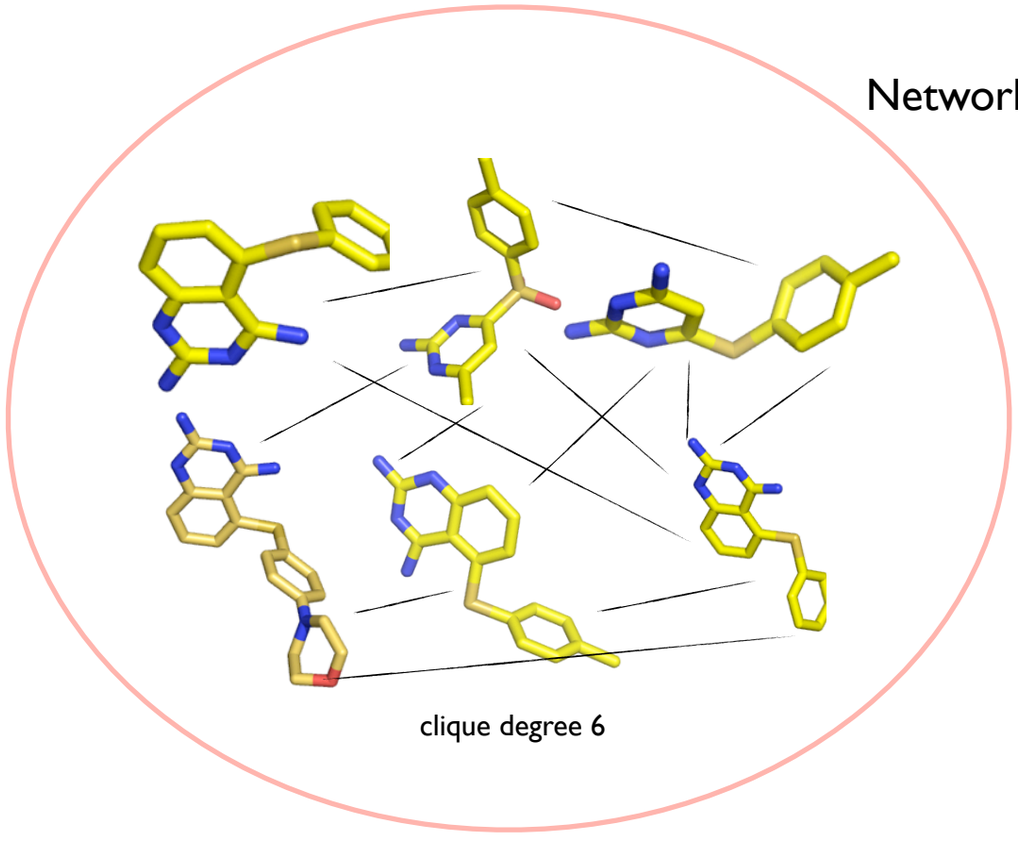
# Network-based Annolyze

## *nAnnolyze*



# Ligand subnetwork

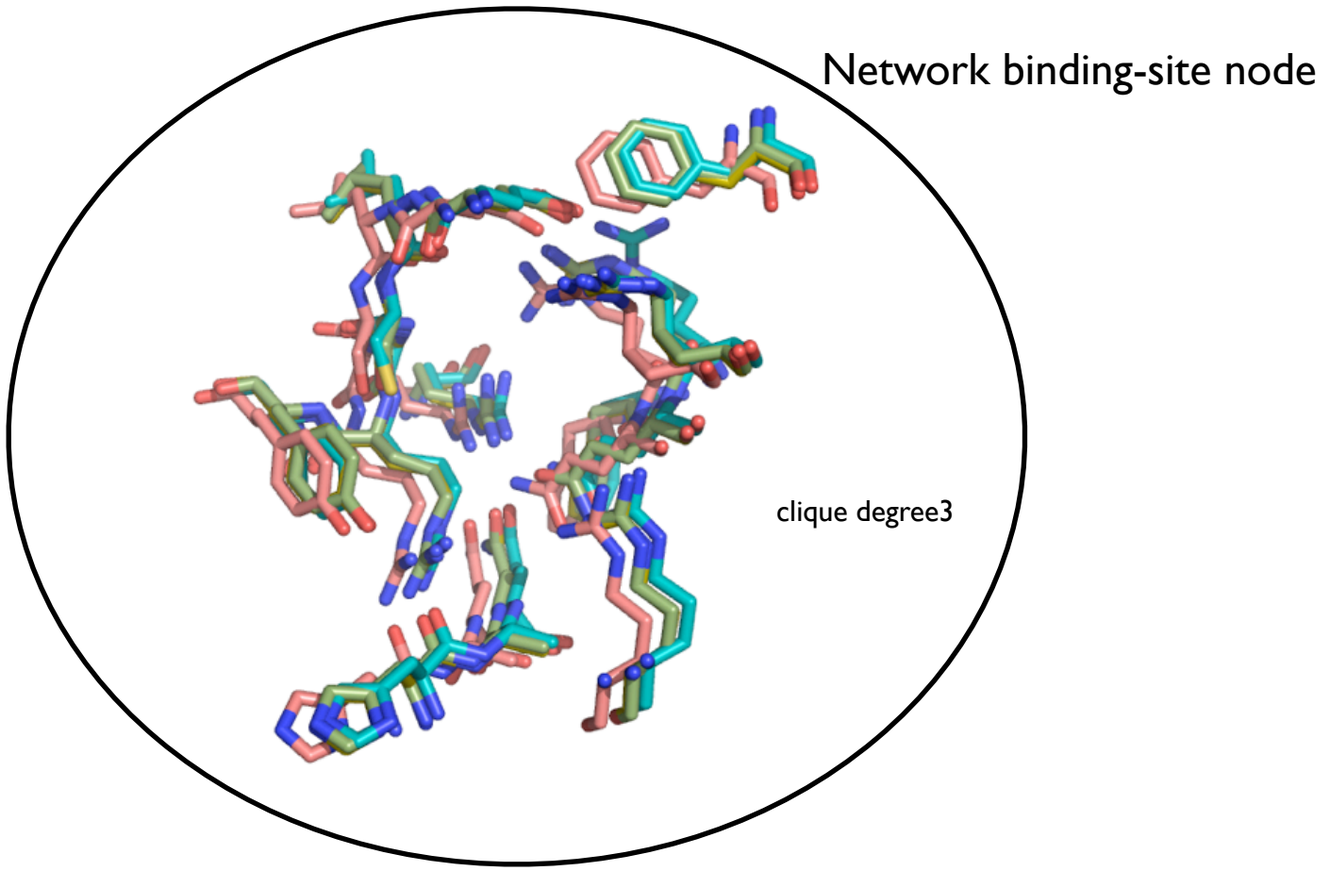
- Retrieved 7,609 high drug-likeness* compounds from PDB.
- Nodes of highly similar compounds: cliques of similarities.
- 4,101 **nodes** of ligand clusters and 24,856 **edges**.
- Edges weight = normalized similarity score.




* Bickerton, G. R., Paolini, G. V, Besnard, J., Muresan, S., & Hopkins, A. L. (2012). Quantifying the chemical beauty of drugs. *Nature chemistry*, *4*(2), 90–8.

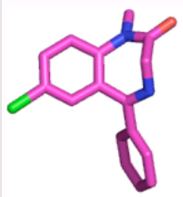
# Protein binding-site network

- Retrieved binding-sites for the 7,609 compounds: 28,299 binding-sites.
- Similarities between proteins by structural comparisons of the binding-site.
- Cluster highly similar groups of binding-sites: cliques of binding-sites.
- 19,483 **nodes** of binding-sites and 29,811 **edges**.
- Edges weight = normalized binding-site similarity score.

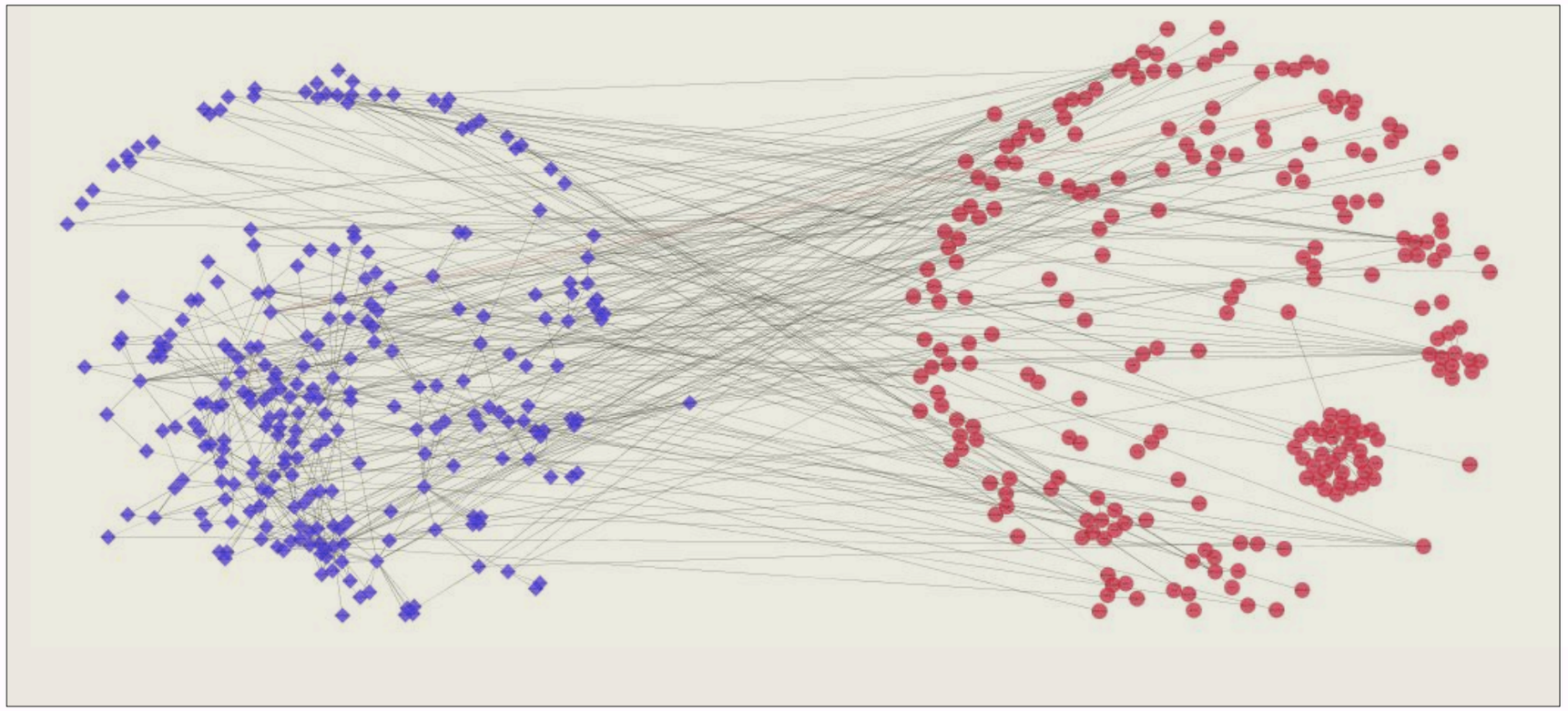


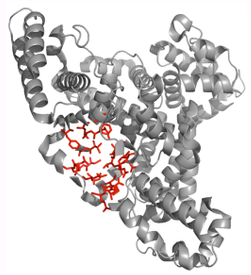

Link the two subnetworks by edges between protein structures and their co-crystallized ligands.

# Looking for targets...

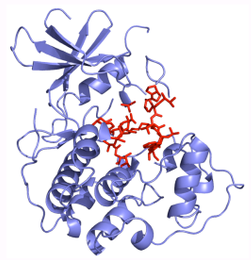


Query
DZP





t1

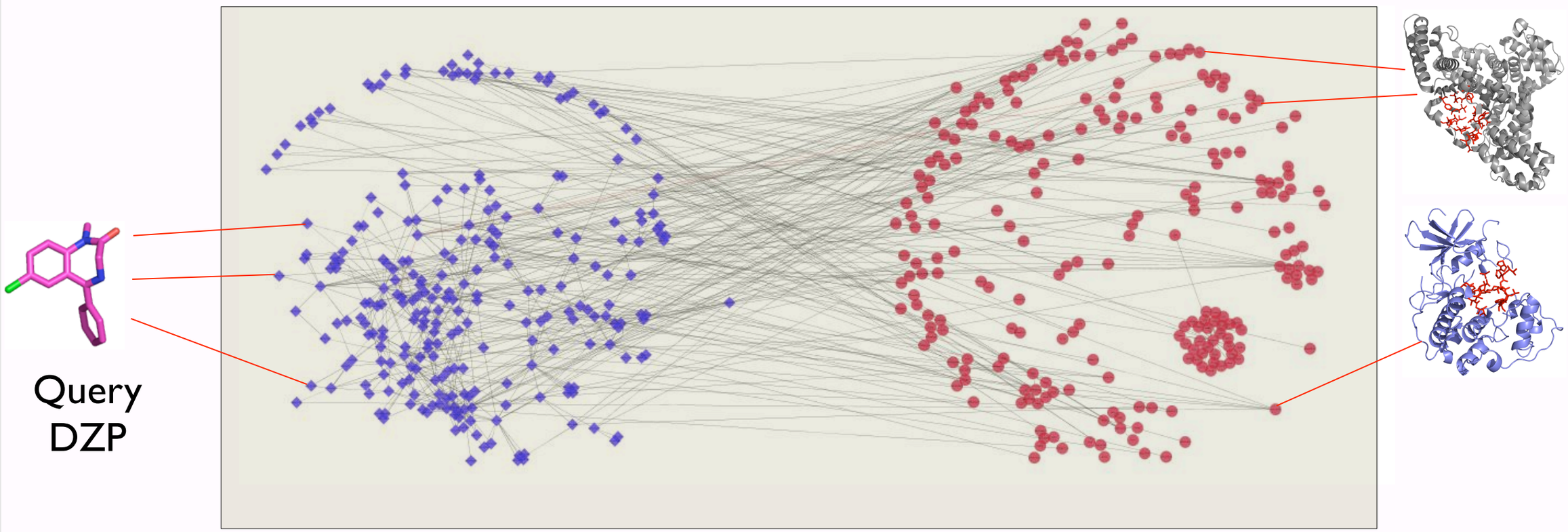


t2

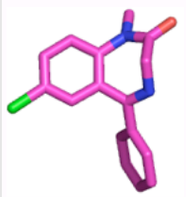
.
.
.

tN

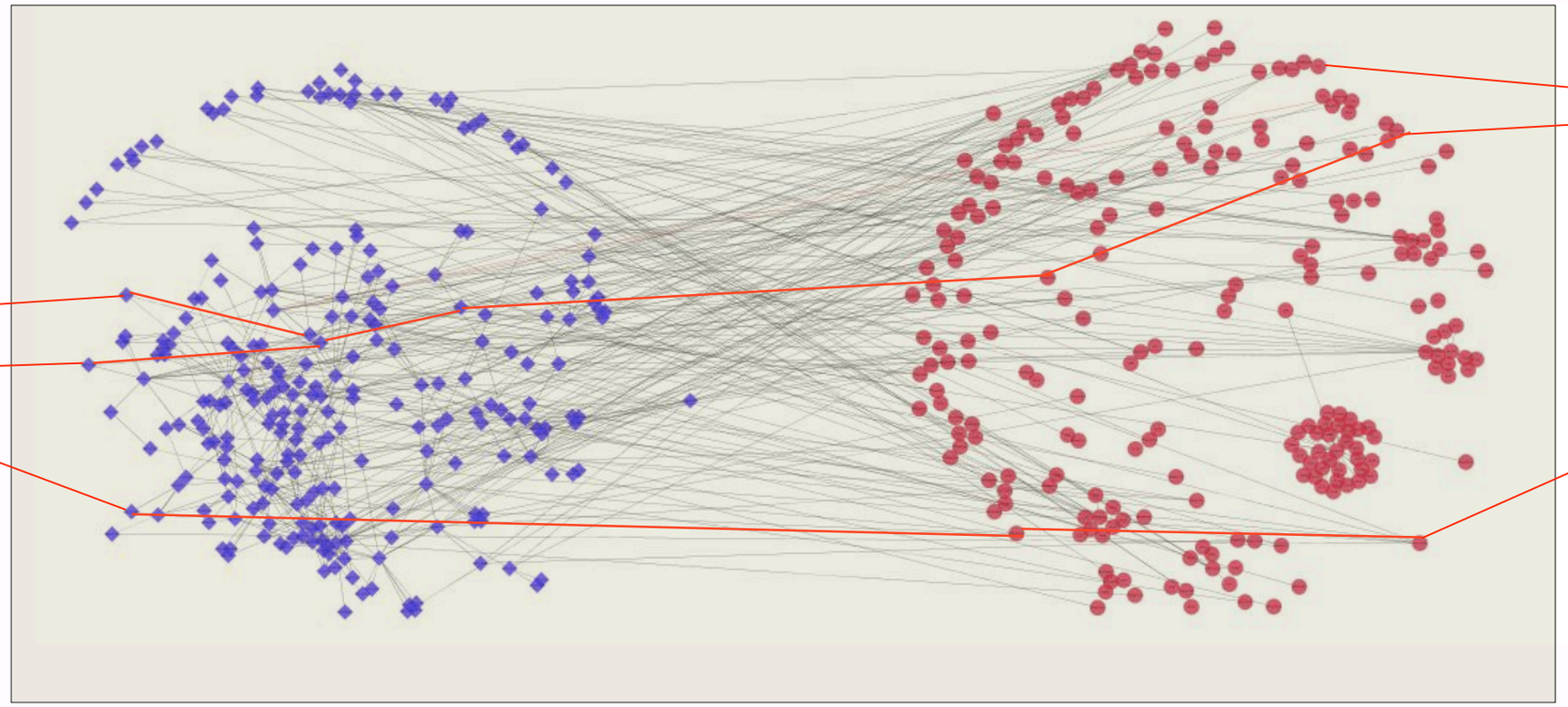
# Looking for targets...

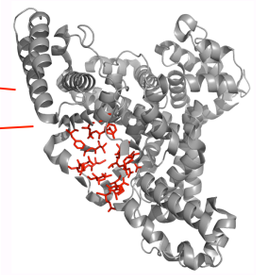


# Looking for targets...

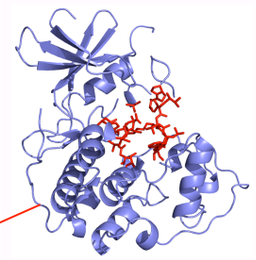


Query
DZP





t1

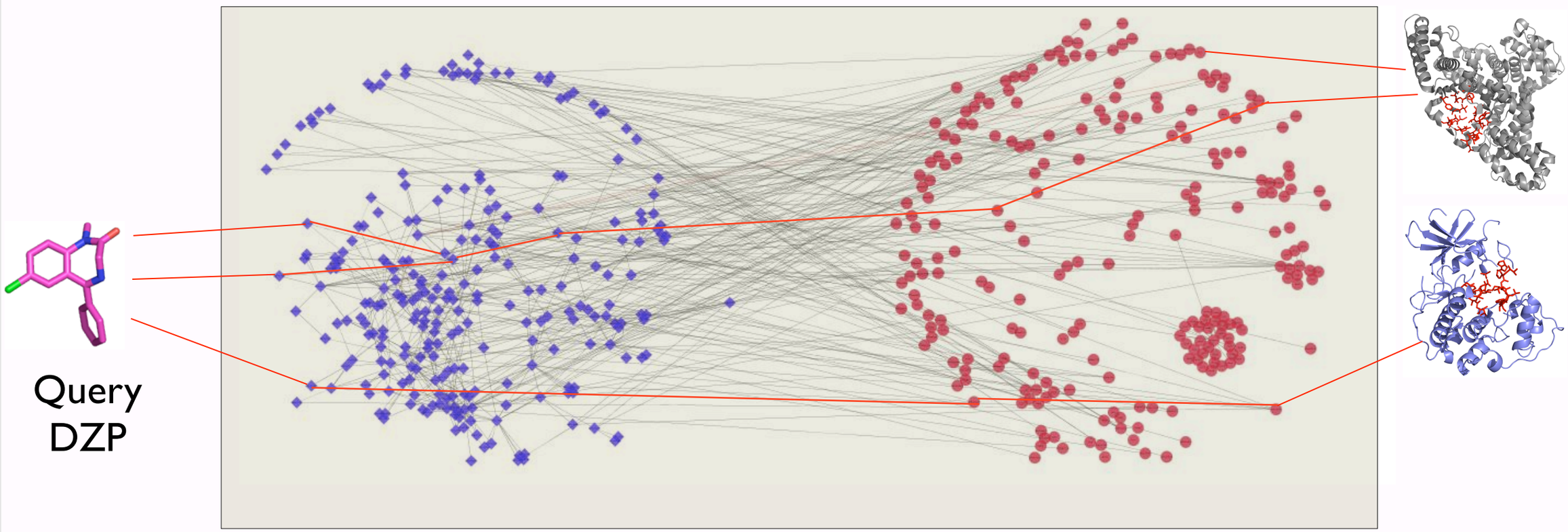


t2

.
.
.

tN

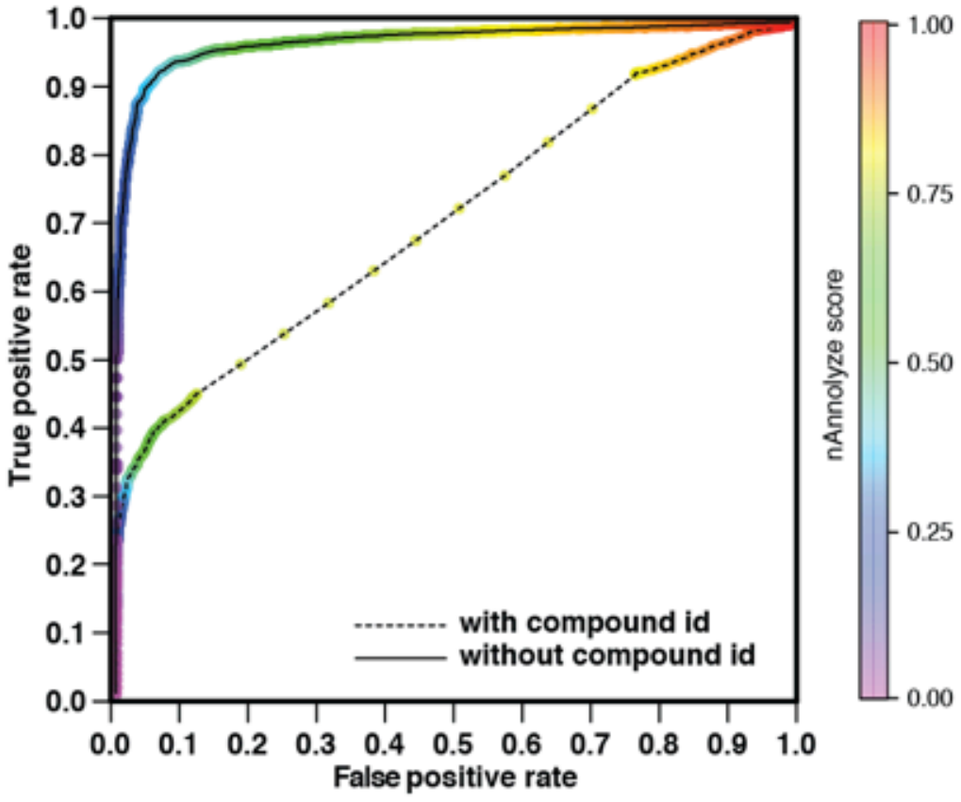
# Looking for targets...

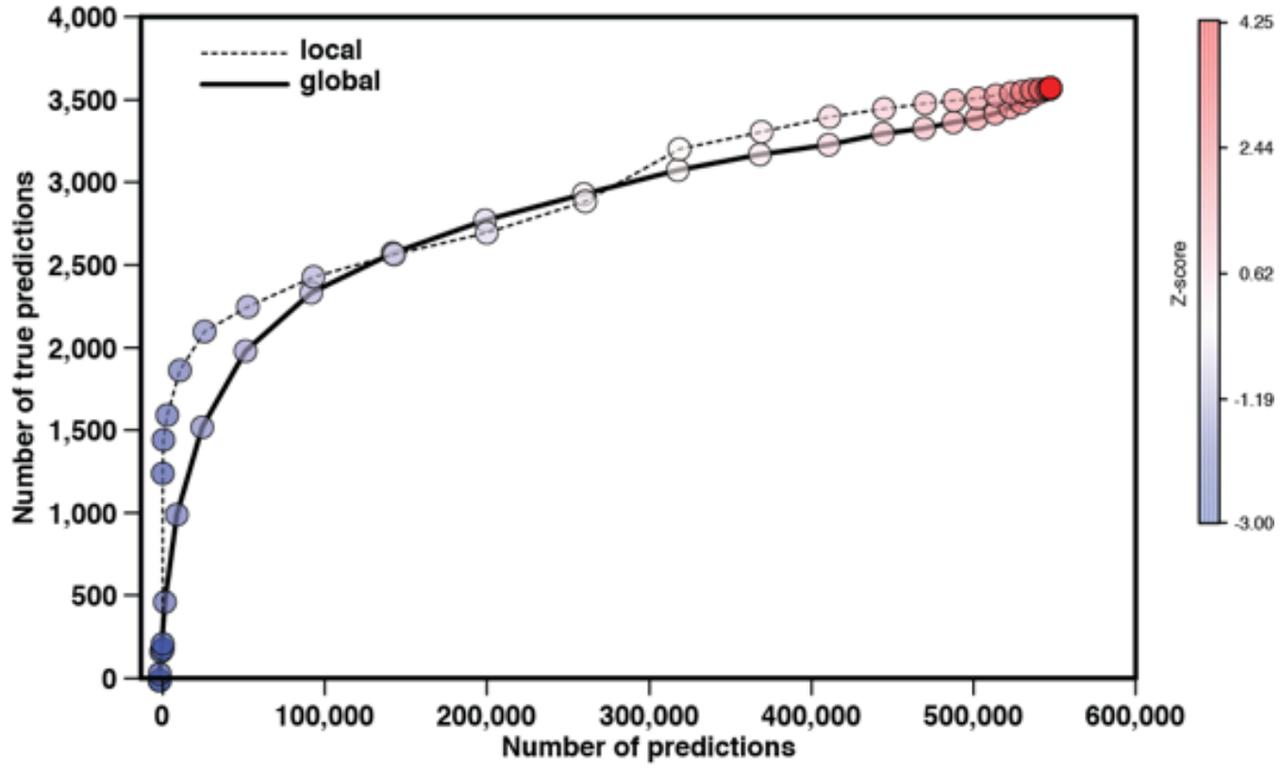



| Ligand | Target | Distance | Global Z-score | Local Z-score |
|---|---|---|---|---|
| DZP | t1 | 1.3 | -1.6 | -2.5 |
| DZP | t2 | 2.5 | 2.3 | 1.02 |
| DZP | tM | 1.9 | -1.6 | -3.16 |
| DZP | tN | 2.6 | 2.42 | 2.97 |

# Benchmarking

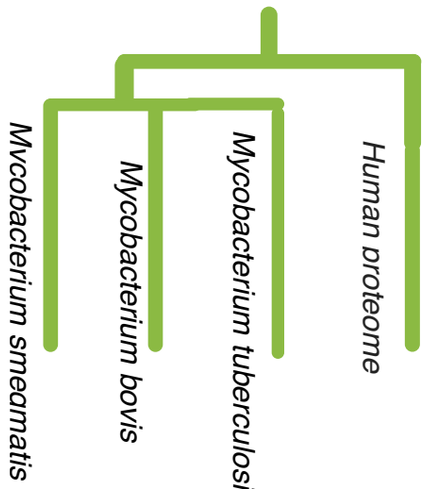
- 232 approved FDA drugs co-crystallized with a protein.
- Test-set  = 6,282 true drug-protein pairs and 5,981 negative pairs.
- Drug ID = 0.97 AUC
- Anonymous compounds = 0.73 AUC



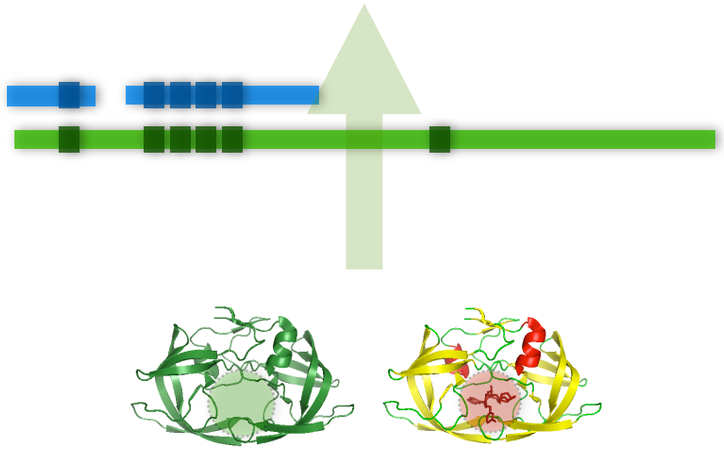
# Applying the method, modeling genomes...

## 1. Modeling




## 2. Binding-site inheritance

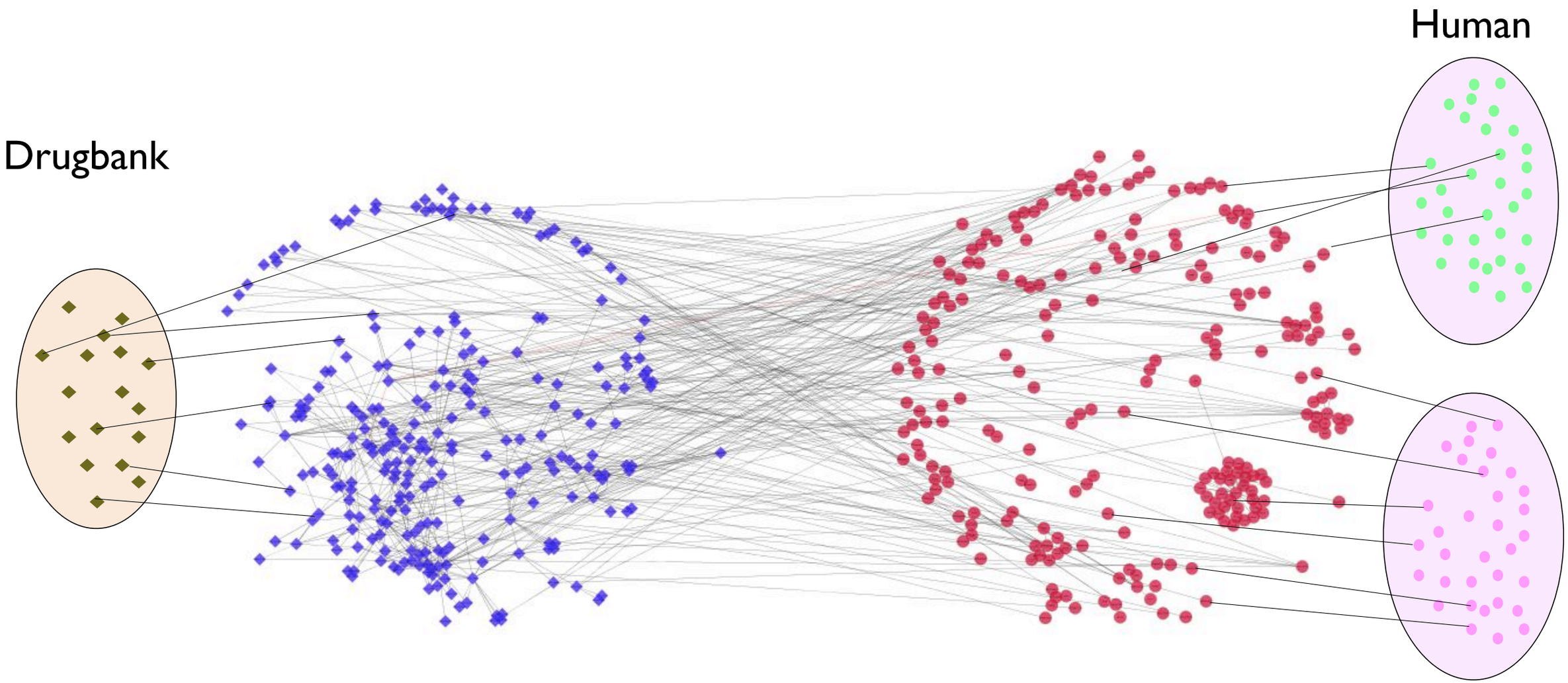
3D model



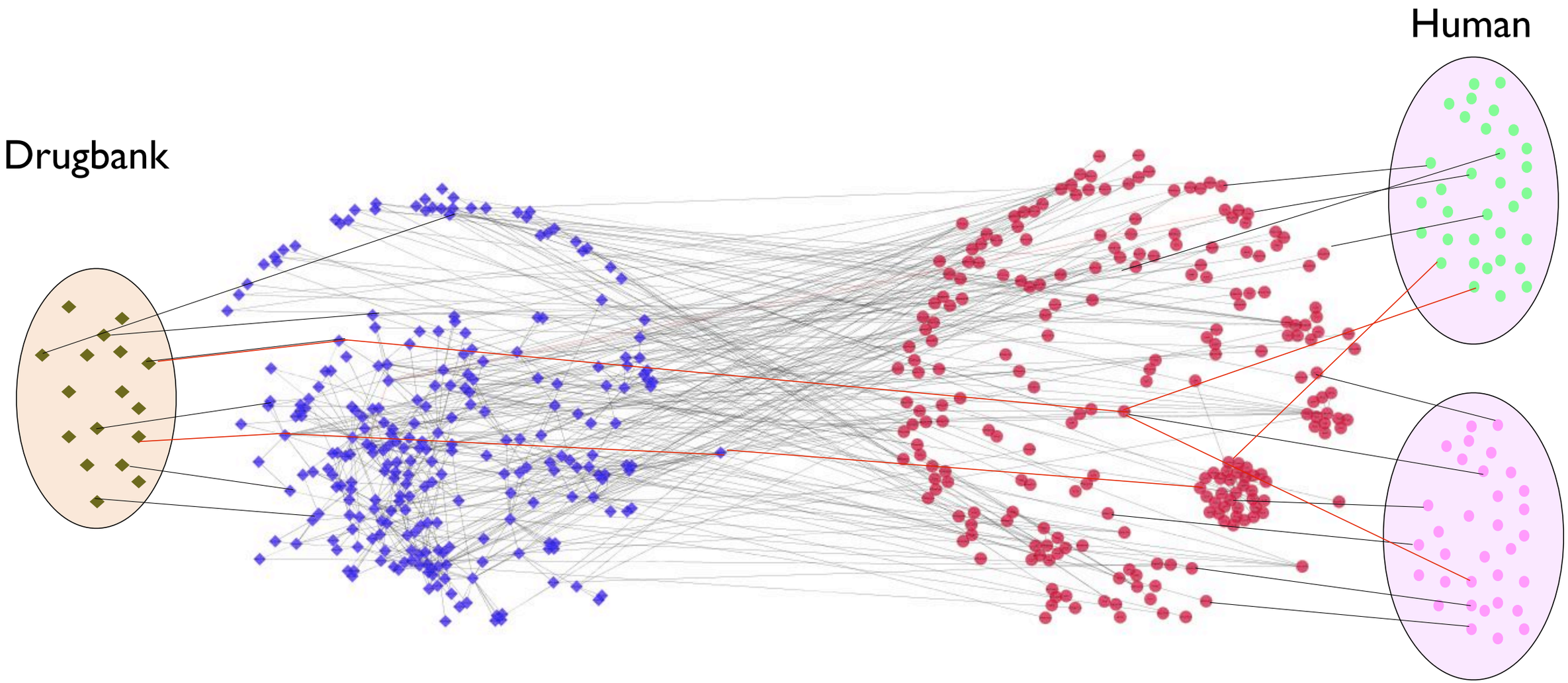
PDB templates

| | Human | Bacterial proteomes |
|---|---|---|
| **3D reliable models** | 31,734 with overlapping | 5,008 no overlapping |
| **Different Proteins** | 14,000 | 5,008 different proteins |
| **Inherited binding-sites** | 64,000 | 30,000 |

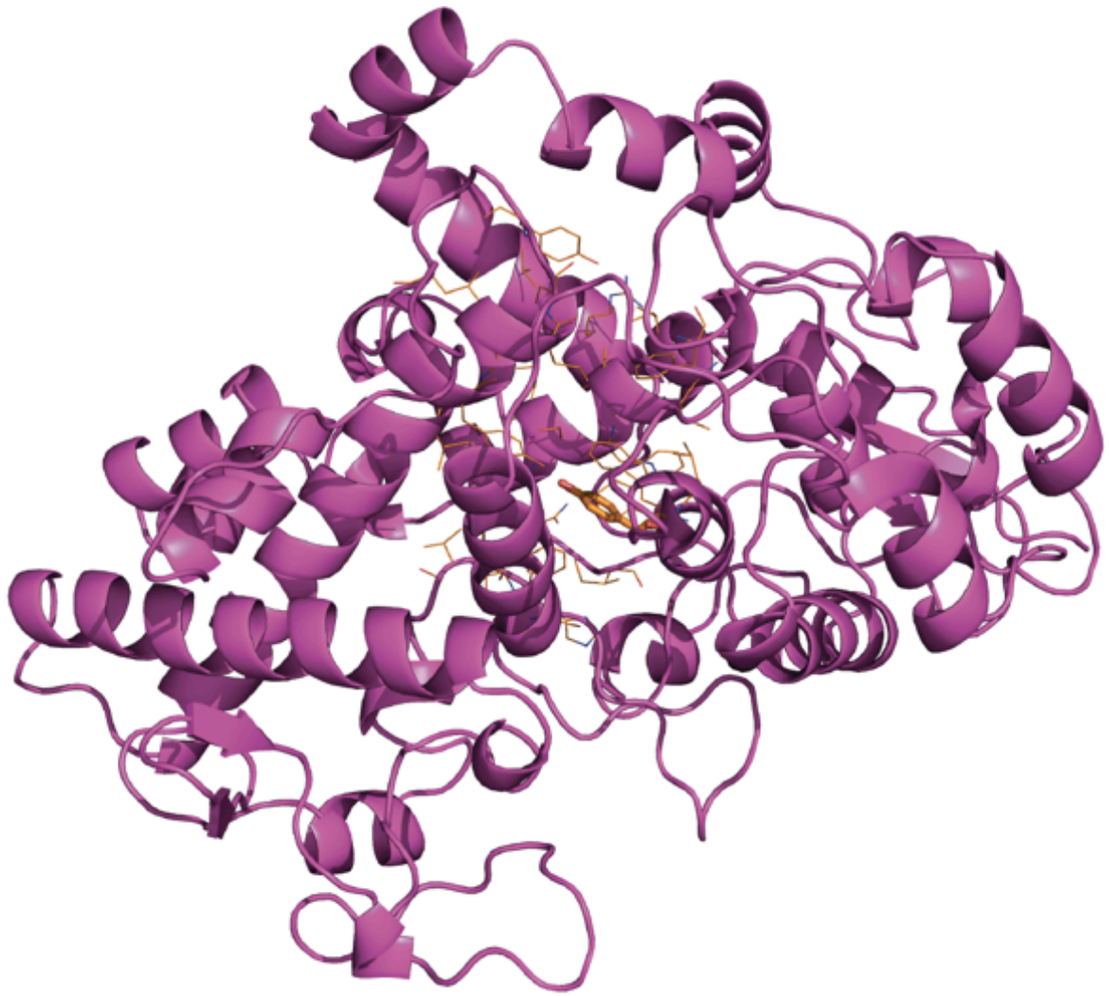
# Searching for Drugbank drugs interactions...



# Searching for Drugbank drugs interactions...



# Human Cyclooxygenase-1 targeted by NSAID drugs



- 21 out of the 44 approved FDA drugs against COX-1 ( score > 0.85 ).
- Human structure model from the sheep COX-1.
- Predicted binding site includes Tyrosine 385.

| Drug ID | Drug name | nAnnoLyze score |
|---|---|---|
| DB00712 | Flurbiprofen | 0.97 |
| DB00328 | Indomethacin | 0.97 |
| DB01600 | Tiaprofenicacid | 0.96 |
| DB00870 | Suprofen | 0.96 |
| DB00821 | Carprofen | 0.96 |
| DB00788 | Naproxen | 0.96 |
| DB00500 | Tolmetin | 0.94 |
| DB00465 | Ketorolac | 0.94 |
| DB00963 | Bromfenac | 0.92 |
| DB00586 | Diclofenac | 0.91 |
| DB06802 | Nepafenac | 0.90 |
| DB01283 | Lumiracoxib | 0.90 |
| DB00784 | Mefenamicacid | 0.89 |
| DB00861 | Diflunisal | 0.88 |
| DB04552 | NiflumicAcid | 0.88 |
| DB00991 | Oxaprozin | 0.88 |
| DB01050 | Ibuprofen | 0.87 |
| DB00939 | Meclofenamicacid | 0.86 |
| DB01399 | Salsalate | 0.86 |
| DB01009 | Ketoprofen | 0.86 |
| DB00605 | Sulindac | 0.85 |

# *Sorafenib* pathway targeting through binding of several protein

# *Sorafenib* pathway targeting through binding of several protein

| Target | Score | Structure | KEGG Pathway |
|---|---|---|---|
| **MAPK 14** | **0.99** | **Yes** | MAPK signaling<br>Fox0 signaling<br>VEGF signaling<br>Rap1 signaling<br>RIG-I-like receptor signaling<br>Acute myeloid leukemia |
| **CDK19** | **0.97** | **No** | - |
| **FLT1** | **0.90** | **Yes** | Ras signaling pathway |
| **RAF 1** | **0.89** | **Yes** | MAPK signaling<br>Ras signaling<br>Rap1 signaling<br>VEGF signaling<br>Fox0 signaling pathway<br>Acute myeloid leukemia |
| **ARAF** | **0.88** | **Yes** | Fox0 signaling<br>Acute myeloid leukemia |
| **CDK10** | **0.88** | **No** | - |
| **BRAF** | **0.88** | **Yes** | MAPK signaling<br>Rap1 signaling<br>Fox0 signaling<br>Acute myeloid leukemia |
| **CDK8** | **0.87** | **Yes** | - |
| **FLT3** | **0.86** | **Yes** | Acute myeloid leukemia |
| **MAPK 15** | **0.86** | **No** | - |

Annotated ( Chembl, PubChem, Drugbank, PDB )

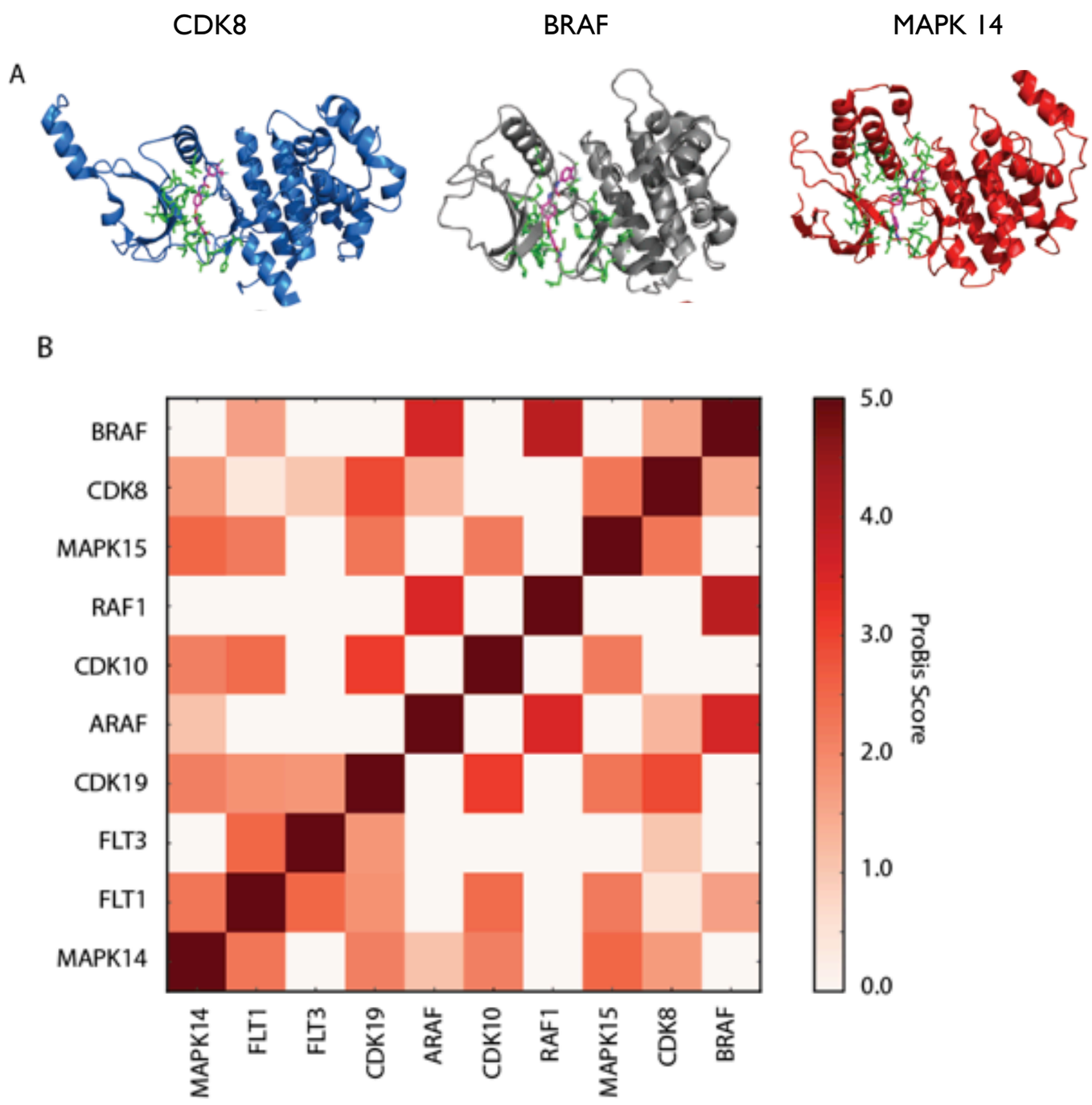
Not Annotated

# *Sorafenib* pathway targeting through binding of several protein

| Target | Score | Structure | KEGG Pathway |
|---|---|---|---|
| **MAPK 14** | **0.99** | **Yes** | MAPK signaling<br>Fox0 signaling<br>VEGF signaling<br>Rap1 signaling<br>RIG-I-like receptor signaling<br>Acute myeloid leukemia |
| **CDK19** | **0.97** | **No** | - |
| **FLT1** | **0.90** | **Yes** | Ras signaling pathway |
| **RAF 1** | **0.89** | **Yes** | MAPK signaling<br>Ras signaling<br>Rap1 signaling<br>VEGF signaling<br>Fox0 signaling pathway<br>Acute myeloid leukemia |
| **ARAF** | **0.88** | **Yes** | Fox0 signaling<br>Acute myeloid leukemia |
| **CDK10** | **0.88** | **No** | - |
| **BRAF** | **0.88** | **Yes** | MAPK signaling<br>Rap1 signaling<br>Fox0 signaling<br>Acute myeloid leukemia |
| **CDK8** | **0.87** | **Yes** | - |
| **FLT3** | **0.86** | **Yes** | Acute myeloid leukemia |
| **MAPK 15** | **0.86** | **No** | - |

Annotated ( Chembl, PubChem, Drugbank, PDB )

Not Annotated



# Antimicrobial drugs against Mycobacterium tuberculosis

OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

## Target Prediction for an Open Access Set of Compounds Active against *Mycobacterium tuberculosis*

**Francisco Martínez-Jiménez[1,2], George Papadatos[3], Lun Yang[4], Iain M. Wallace[3], Vinod Kumar[4], Ursula Pieper[5], Andrej Sali[5], James R. Brown[4]*, John P. Overington[3]*, Marc A. Marti-Renom[1,2]***

**1** Genome Biology Group, Centre Nacional d'Anàlisi Genòmica (CNAG), Barcelona, Spain, **2** Gene Regulation Stem Cells and Cancer Program, Centre for Genomic Regulation (CRG), Barcelona, Spain, **3** European Molecular Biology Laboratory – European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom, **4** Computational Biology, Quantitative Sciences, GlaxoSmithKline, Collegeville, Pennsylvania, United States of America, **5** Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, San Francisco, California, United States of America

nAnnolyze

Home Help About Download

**Introduce your query molecule**

ID of the molecule: DB00398

e.g: DB00712

SMILE: CNC(=O)C1=NC=CC(OC2=CC

e.g: CC(C(O)=O)C1=CC(F)=C(C=C1)C1=CC=CC=C1

Molecular Weight: 464.82

e.g: 244

Input an example!

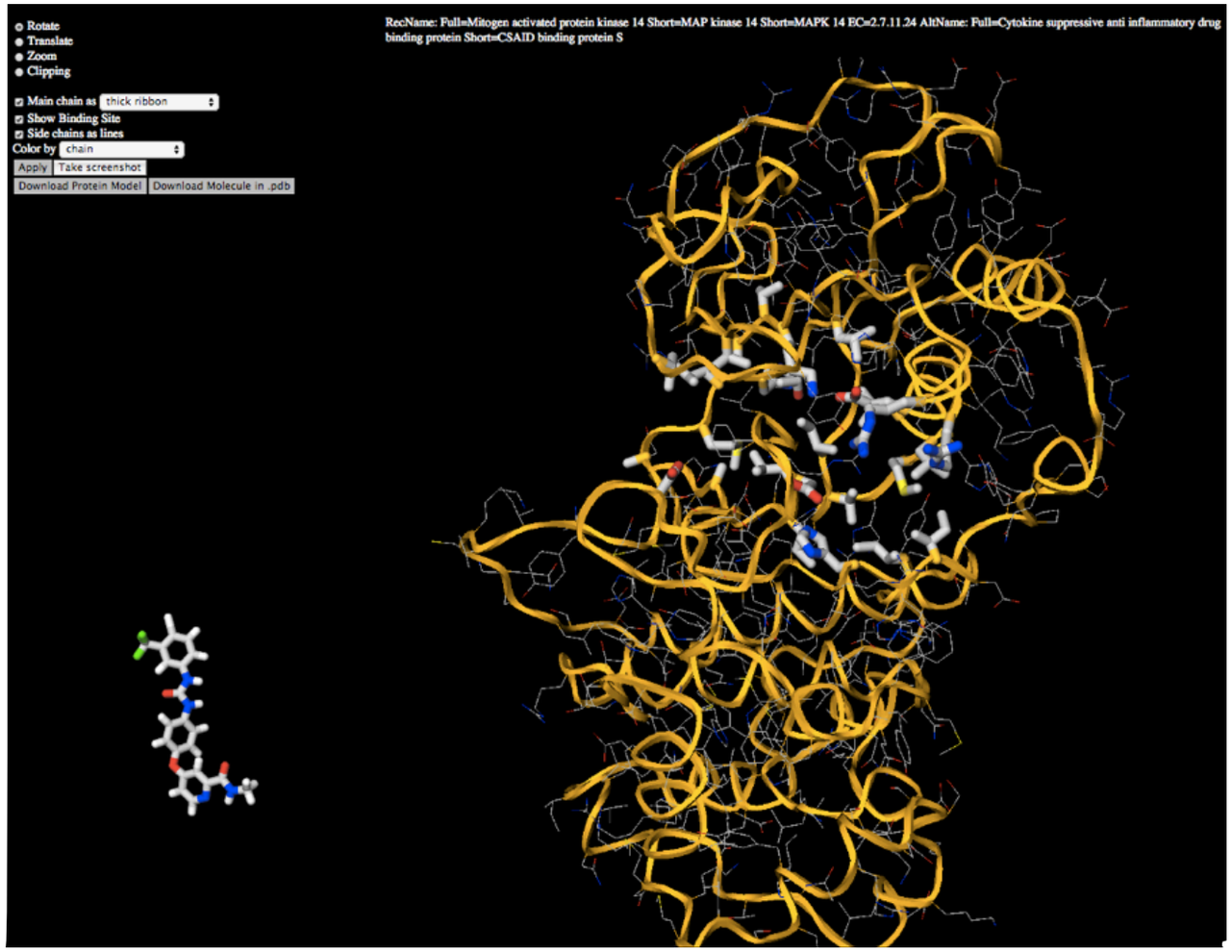
Upload your File with the SMILES: Upload a File — Choose Files — No file chosen — ?

Maximum 10 compounds per search

**Select the organism**

- ○ *Mycobacterium tuberculosis, Mycobacterium bovis, Mycobacterium smegmatis*
- ◉ *Homo Sapiens*

Submit




nAnnolyze

Home Help About Download

+ Toggle Statistics

Graphics | Table

Visualize:

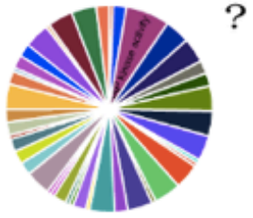
- ◉ *Overrepresented GO terms and Pathways*
- ○ *All Go terms and Pathways*

Filter by global z-score: < -2

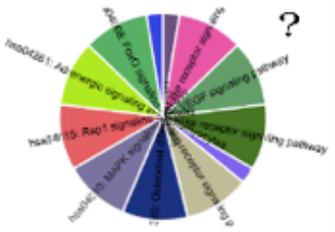
Statistics for Compound: DB00398

Download .csv

**GO terms** ?



**Pathways** ?



Best — Worst

| Molecule ID | External ID | Target UP | Score | Z-Score-local | Z-Score global | Description | Binding-Site residues | Visualize... |
|---|---|---|---|---|---|---|---|---|
| DB00398 | NP_6205( | Q13083 | 0.99 | -3.23 | -3.30 | RecName: Full=Mitogen-activated protein kinase 14;Short=MAP kinase 14;Short=MAPK 14;EC=2.7.11.24;AltName: Full=Cytokine suppressive anti-inflammatory drug-binding protein;Short=CSAID-binding protein;S | [:A and (30;38;51;52;53;67;70;71;74; | |
| DB00398 | NP_0558( | Q5JQZ7 | 0.98 | -4.14 | -3.24 | RecName: Full=Cyclin-dependent kinase 19;EC=2.7.11.22;AltName: Full=CDC2-related protein kinase 6;AltName: Full=Cell division cycle 2-like protein kinase 6;AltName: Full=Cell division protein kinase 1 | [:A and (35;50;51;52;66;69;70;73;78; | |
| DB00398 | NP_0012( | G1T9S5 | 0.97 | -4.39 | -3.14 | SubName: Full=Uncharacterized protein; | [:A and (27;28;32;35;50;51;52;66;69; | |
| DB00398 | GENSCAN | | 0.96 | -3.21 | -3.06 | | [:A and (924;925;932;950;952;969;97 | |
| DB00398 | NP_0013( | L7RSM2 | 0.94 | -2.91 | -2.95 | SubName: Full=Mitogen-activated protein kinase 14;SubName: Full=Mitogen-activated protein kinase 14 isoform CRA_e; | [:A and (30;35;38;51;70;74;83;84;85; | |

http://nannolyze.cnag.cnat

# Acknowledgments


Davide Baù
Gireesh K. Bogu
François le Dily
Marc A. Marti-Renom
David Dufour
François Serra
Michael Goodstadt
Yasmina Cuartero

**COLLABORATORS**

Jim Brown (GSK)
LLuís Ballell (GSK)
John Overington (EBI-EMBL)
Andrej Sali (UCSF)
Anna Tramontano (Sapienza University)

http://marciuslab.org
http://integrativemodeling.org
http://cnag.cat · http://crg.cat

ERA-NET PathoGenoMics
HFSP
GOBIERNO DE ESPAÑA MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD
cnag
CRG Centre for Genomic Regulation