# BioGPS: The Music for the Chemo- and Bioinformatics Walzer

Gabriele Cruciani, Laura Goracci, **University of Perugia, Italy**
Lydia Siragusa, Francesca Spyrakis, Simon Cross, **Molecular Discovery, UK**

**Is a drug repurposable for another target?**

**Given a drug, are we able to find biological targets?**

**Drug: protomerism, tautomerism, flexibility, phys chem properties**

**What is the molecular mechanism of a drug side effects?**

**Can we predict binding kinetics?**

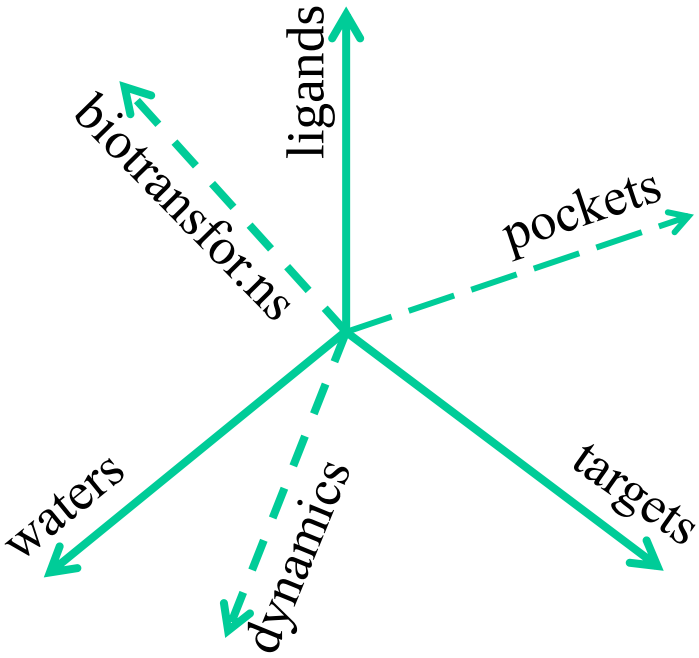
**Biotransformations … ?**

**How can we improve the ligand selectivity?**

**Can we model water molecules interactions?**

**Target flexibility, water network**

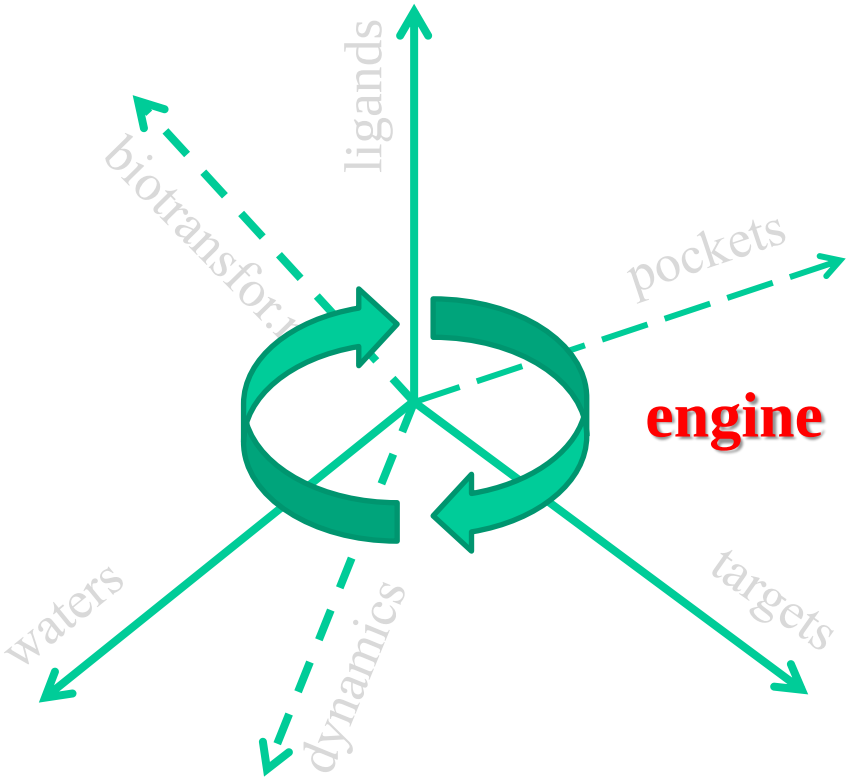
**Is a drug repurposable for another target?**
**What is the molecular mechanism of a drug side effects?**
**How can we improve the ligand selectivity?**




# Holistic approach

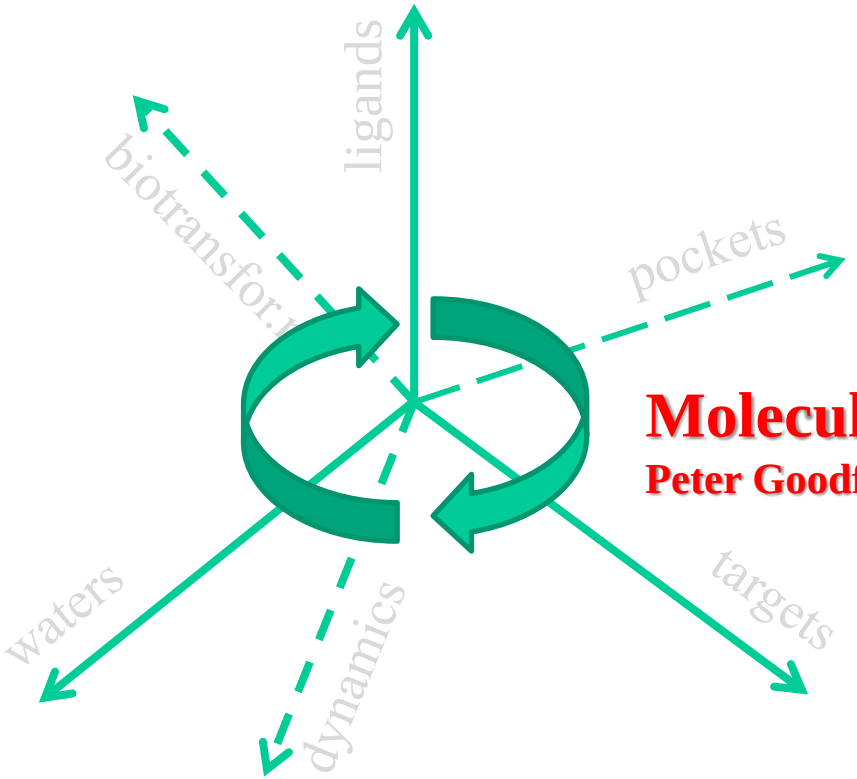
Not 6-dimensional … but still dimensionally demanding

**Is a drug repurposable for another target?**
**What is the molecular mechanism of a drug side effects?**
**How can we improve the ligand selectivity?**



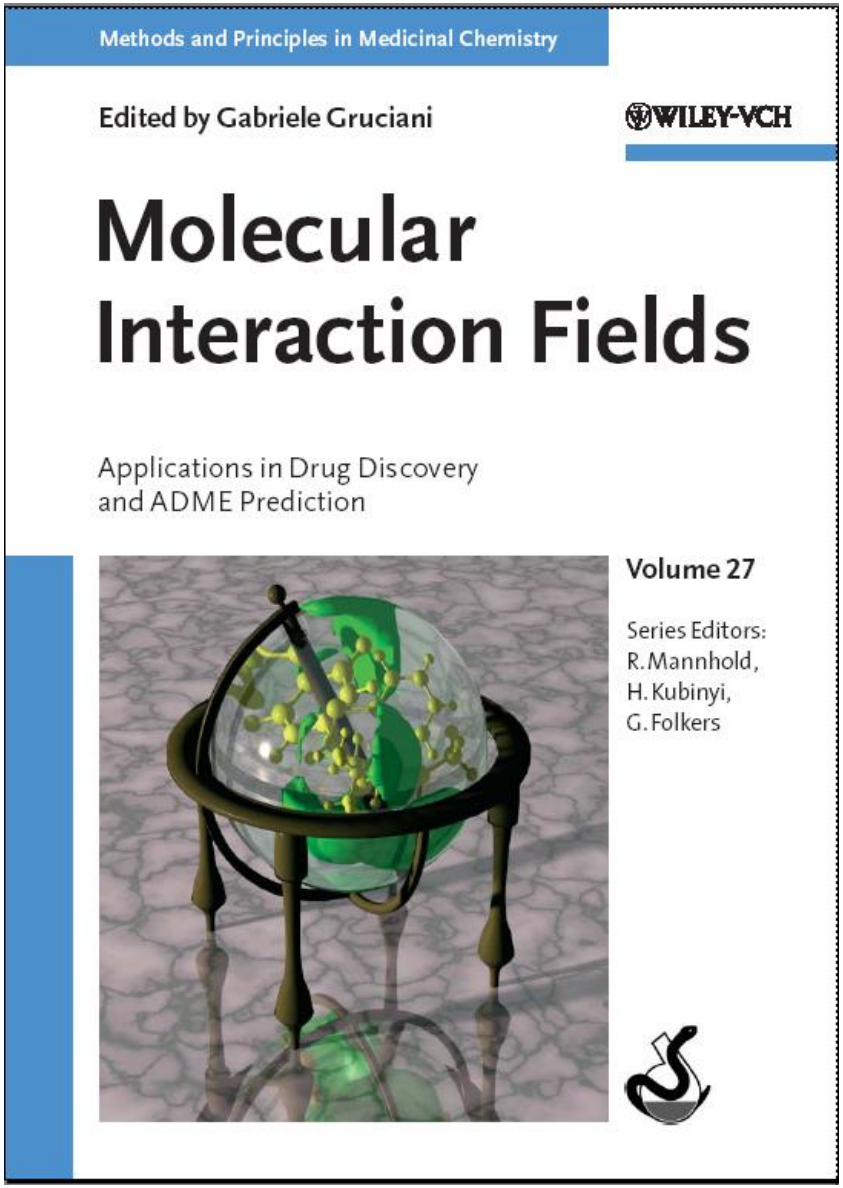

**Holistic approach**

**Is a drug repurposable for another target?**
**What is the molecular mechanism of a drug side effects?**
**How can we improve the ligand selectivity?**




**Molecular Interaction Fields**
**Peter Goodford 1984**

**Holistic approach**

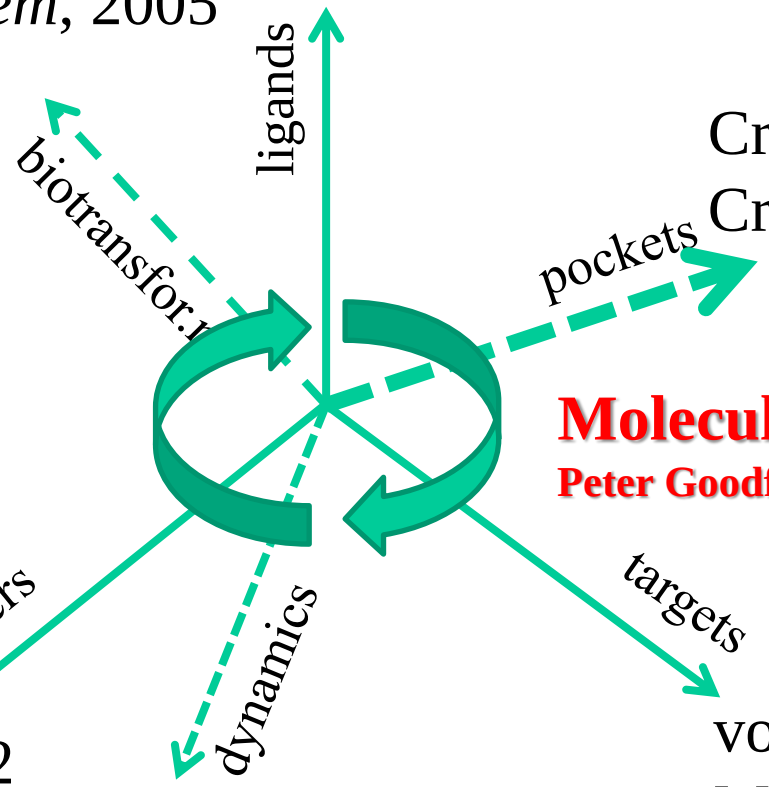



**100 non profit research orgs**
**50 profit research orgs**

**Is a drug repurposable for another target?**
**What is the molecular mechanism of a drug side effects?**
**How can we improve the ligand selectivity?**

Milletti, *JCIM*, 2006

Cruciani, *JMedChem*, 2005




Cruciani, *UK QSAR*, 2005
Cruciani, *JCIM*, 2007

**BioGPS**

**Molecular Interaction Fields**
**Peter Goodford 1984**

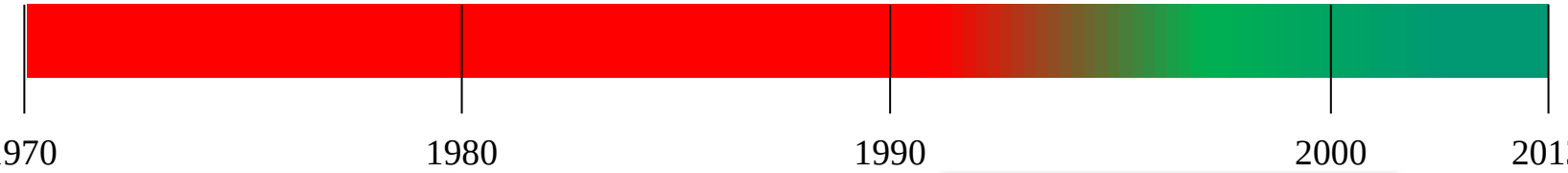
Mason, *TIPS*, 2012

von Itzstein, *Nature*, 1993
Muratore, *PNAS*, 2012

GRID manual 1995
Carosati, *JMedChem*, 2004

**Holistic approach**

- Extracting relevant information from protein structures gives the opportunity to use the biological space for many purposes
- **'Similar entities show similar function' → several methods to compare proteins**




J. Mol. Biol. (1970) 48, 443–453

A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins

Saul B. Needleman and Christian D. Wunsch

JMB

FUGUE: Sequence-structure Homology Recognition Using Environment-specific Substitution Tables and Structure-dependent Gap Penalties

Jiye Shi, Tom L. Blundell and Kenji Mizuguchi*

J. Mol. Biol. (1990) 215, 403–410

BLAST

Basic Local Alignment Search Tool

Stephen F. Altschul[1], Warren Gish[1], Webb Miller[2] Eugene W. Myers[3] and David J. Lipman[1]

Science 22 March 1985: Vol. 227 no. 4693 pp. 1435-1441 DOI: 10.1126/science.2983426

ARTICLES

Rapid and sensitive protein similarity searches

DJ Lipman, WR Pearson

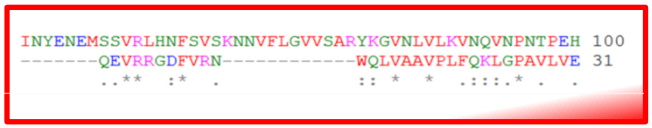
FASTA

J. Chem. Inf. Model. 2010, 50, 123–135

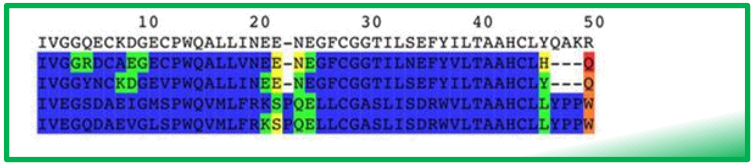
JMB

A New Method to Detect Related Function Among Proteins Independent of Sequence and Fold Homology

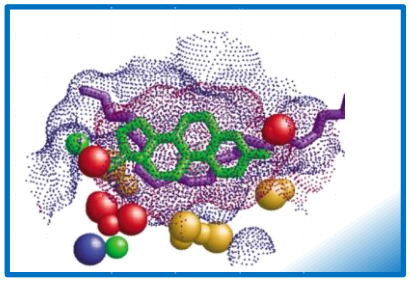
Stefan Schmitt[1,2], Daniel Kuhn[1] and Gerhard Klebe[1]*




*Sequence comparison*

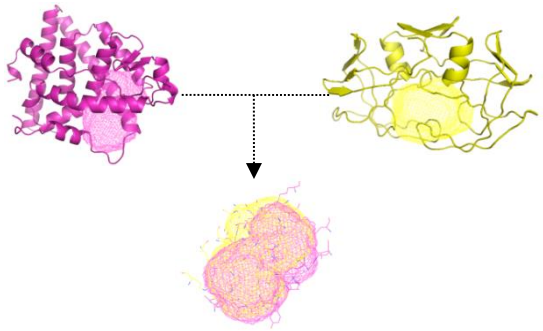



*Annotated sequence comparison*



*3D structure comparison*

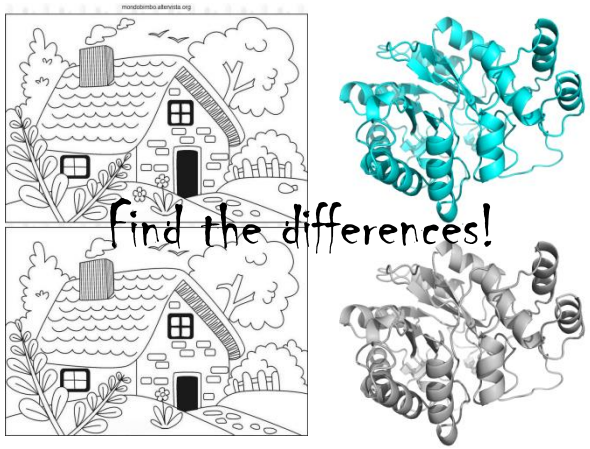
A new computational algorithm for protein binding sites characterization and comparison in terms of their *three-dimensional structure*



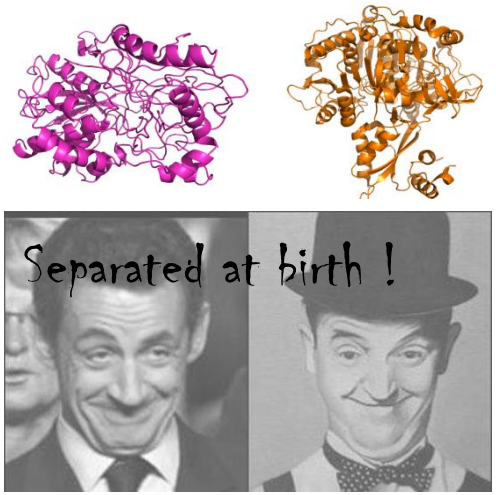
*"the function of a protein does not necessarily depend by the folding or the sequence"*

*J. Struct. Biol. 134, 145-165*

Something like…



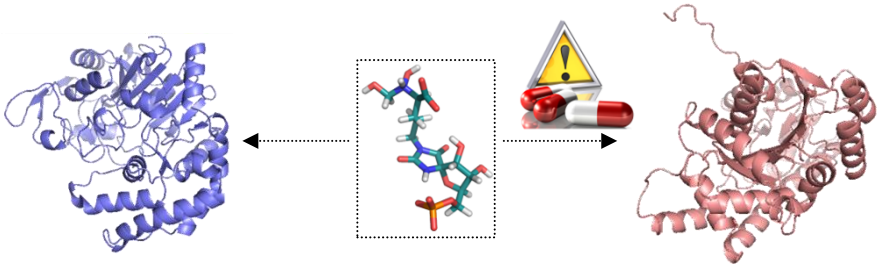

**Slight differences**



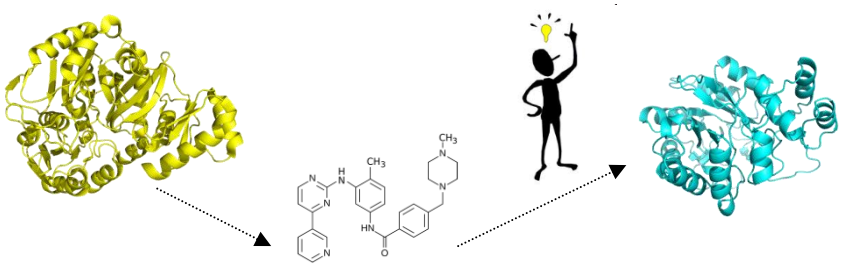

**Unexpected similaritites**

# MOTIVATION

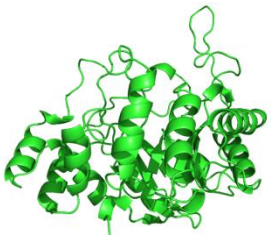
*DRUG SIDE EFFECTS*



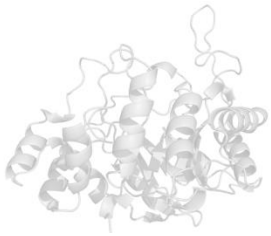
*DRUG REPURPOSING*



*CATALYSIS SPECIFICITY*

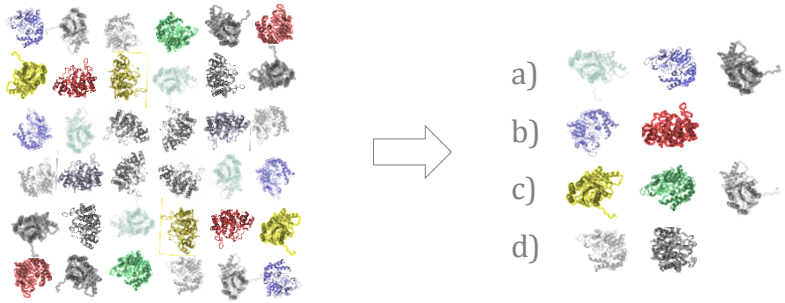


ROOR' + $H_2O$ -> ROOH + R'OH



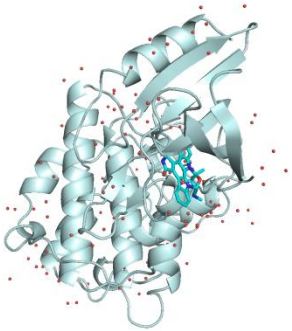
RONHR' + $H_2$ -> ROOH + R'$NH_2$

*LARGE SCALE ANALYSIS*

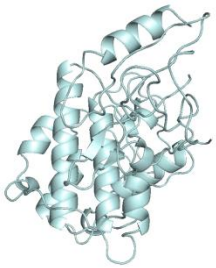


# Methodology: How?

# (1) **Protein refinement**: automatic pre-treatment for protein structures in PDB data format

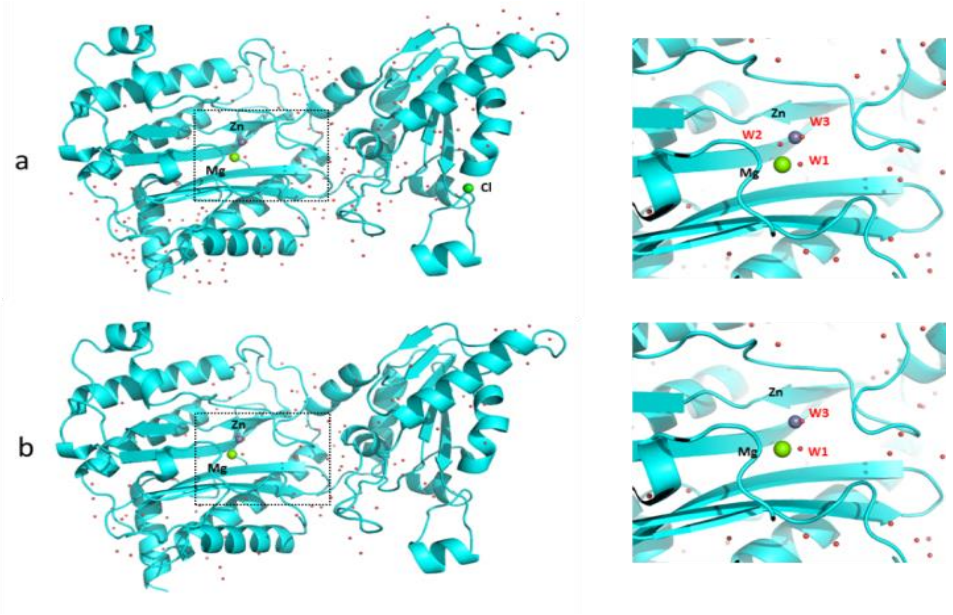


- H2O molecules
- Ions
- Ligands
- Cofactors
- Solvent
- …

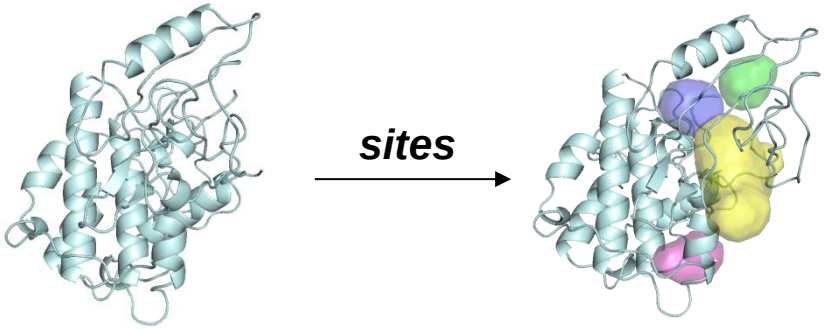


Protein entries are classified according to a web dictionary into nucleic acid, protein, sugar, drug, solvent, ion, inhibitor, coenzyme, ion complex.

Energy-based filters can be used to retain other entries apart from protein residues.

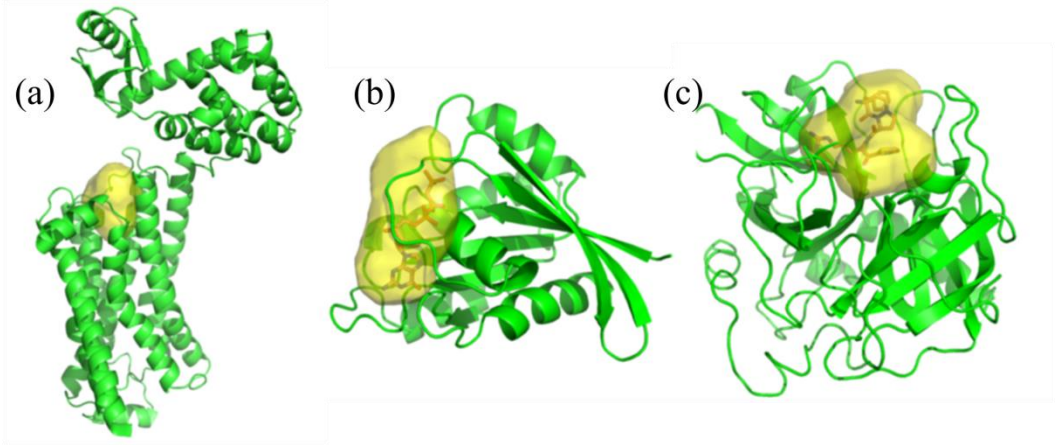


**(2) Cavity detection**: a specialized algorithm is used for the identification of cavities in three-dimensional protein structures




**strength**

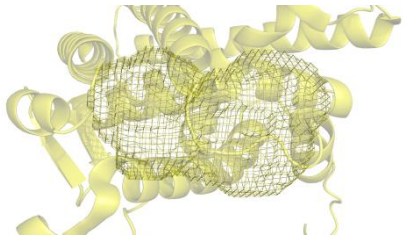
Buriedness index
Erosion and dilation
Hydrophobic probe DRY



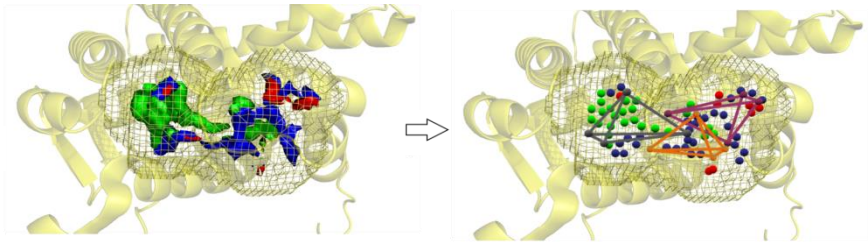
**(3) Cavity characterization**: evaluation of the type, strength and direction of the interactions that a cavity is capable of making

(a) The program GRID is used to calculate the energies of interaction between a chemical group (the "Probe") and another molecule (the "Target")

(b) The resulting MIFs (Molecular Interaction Fields) are then reduced in complexity by selecting a number of representative points using a weighted energy-based and space-coverage function.



*EDRY(xyz)* = ELJ + S
*EDON(xyz)* = ELJ + Ehb + Eel
*EACC(xyz)*= ELJ + Ehb + Eel
*Shape(xyz)* = ELJ



(c) For each quadruplet the four points together with the six distances are stored along with the volume of the quadruplet which retains information about chirality.

(d) All quadruplets generated for a cavity are represented as a bitstring that constitutes the "Common Reference Framework".

| d1 | d2 | d3 | d4 | d5 | d6 | f1 | f2 | f3 | f4 | V |
|---|---|---|---|---|---|---|---|---|---|---|
| d1 | d2 | d3 | d4 | d5 | d6 | f1 | f2 | f3 | f4 | V |
| ... | | | | | | | | | | |
| d1 | d2 | d3 | d4 | d5 | d6 | f1 | f2 | f3 | f4 | V |

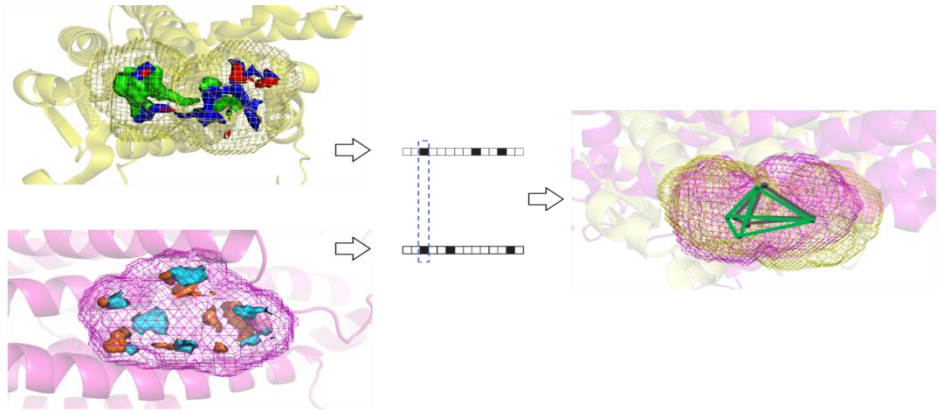
**Structure of quadruplets**

**Common Reference Framework**

**(4) Cavity comparison**: the algorithm compares binding sites via three-dimensional superposition of the "Common Reference Framework"

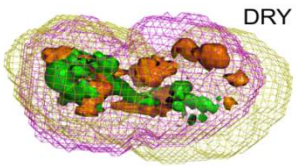
(a) BioGPS performs superpositions by comparing the common reference framework.

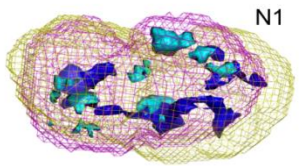
(b) A favorable superposition is said to be found when a pair of quadruplets have all six of their distances coupled in a pair-wise manner (including the type of probe) within a certain distance (1 Å) from each other.

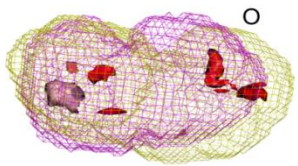


(c) From the quadruplet overlapping, BioGPS overlaps all the region of the MIFs and then 3D structures.

(d) The algorithm calculates for each solution a set of Tanimoto similarity scores.



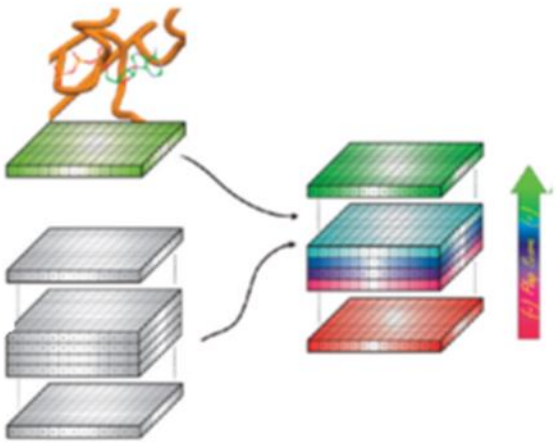




| | H | N1 | O | DRY | Global |
|---|---|---|---|---|---|
| cavity 1 | 0.65 | 0.56 | 0.63 | 0.75 | 0.70 |

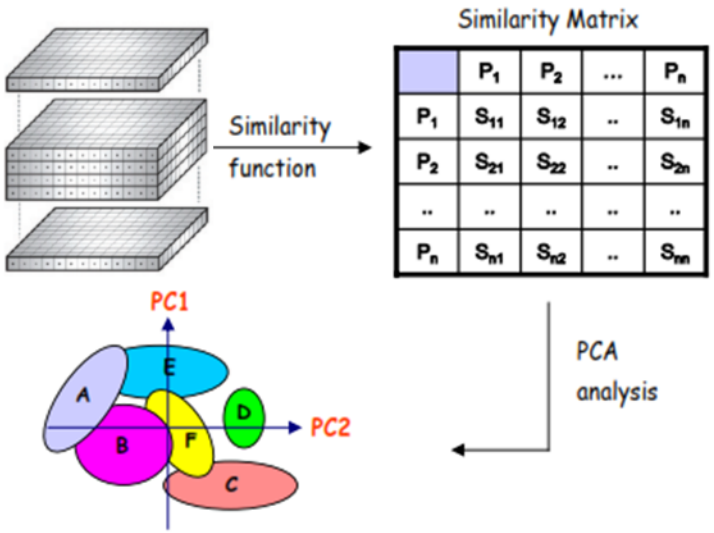
# (5) **Data analysis**: interpretation of similarity scores

## Querying the database



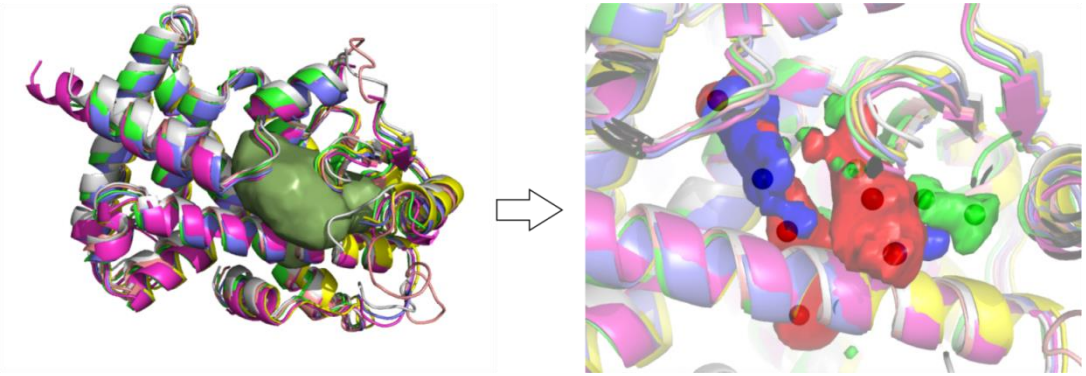
Virtual screening where cavities in the database are ranked accordingly with their degree of similarity against a template (query cavity).

## All against all

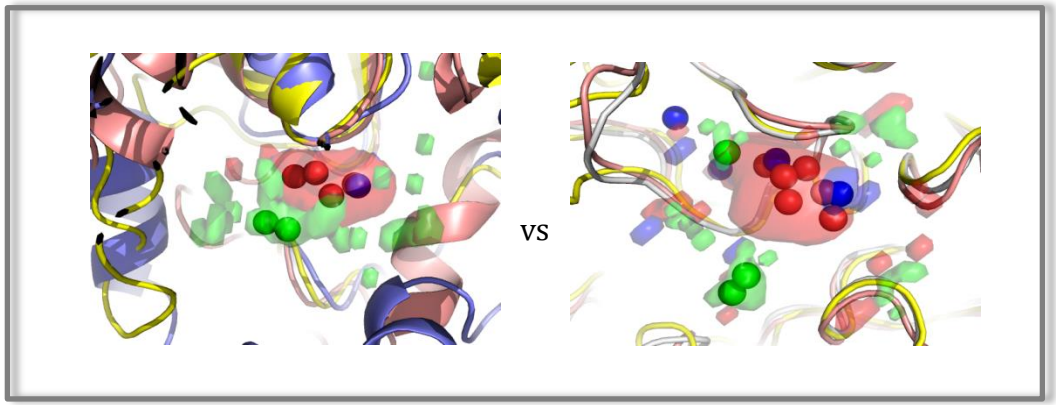



Similarity scores can be used to perform a Principal Component Analysis (PCA).

# (6) *Protein-based pharmacophore*: analysing common features shared by a set of sub-family protein active sites



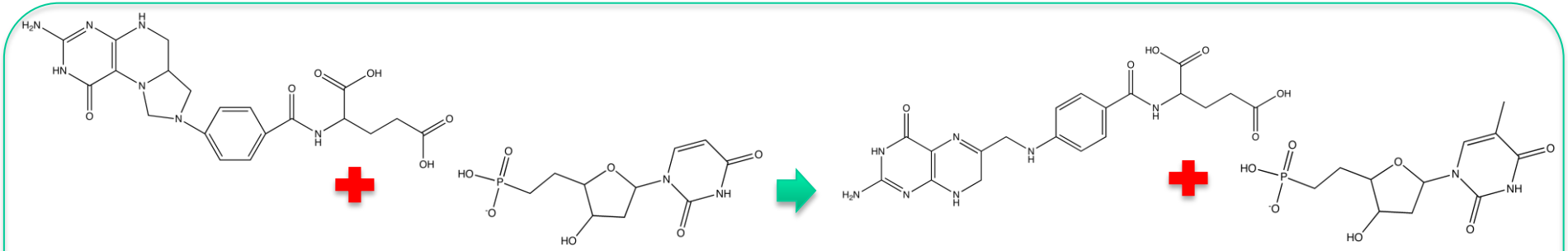
- Three-dimensional arrangement of **common features (PIFs) shared by a set of active sites** of interest (*pseudo-site structure).*
- The minima points of the PIFs are then used to represent *pharmacophoric points*, representing a region where a ligand would favourably interact with all the cavities in the analysis.




- The pharmacophores comparison makes the analysis of similarities and differences very easy and understandable
- The pharmacophore is able to capture and to quantify differences between protein classes

# Applications: What?

# DRUG REPURPOSING: THYMIDILATE SYNTHASE (TS)




## TS ligands

*human TS inhibitors*
➔ *potential anticancer agents*

*bacterial TS inhibitors*
➔ *potential antimicrobic agents*

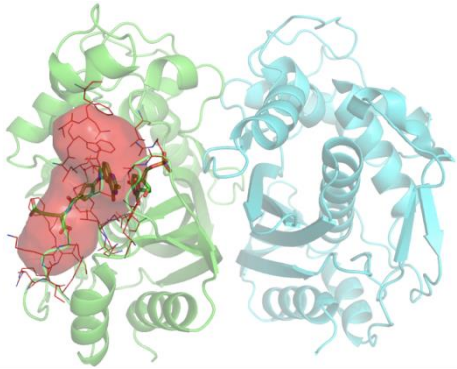
## Research AIM

*Searching for new TS inhibitors candidates*

# STRATEGY

**THE TEMPLATE: human TS complexed with dUMP**

**THE POCKET**



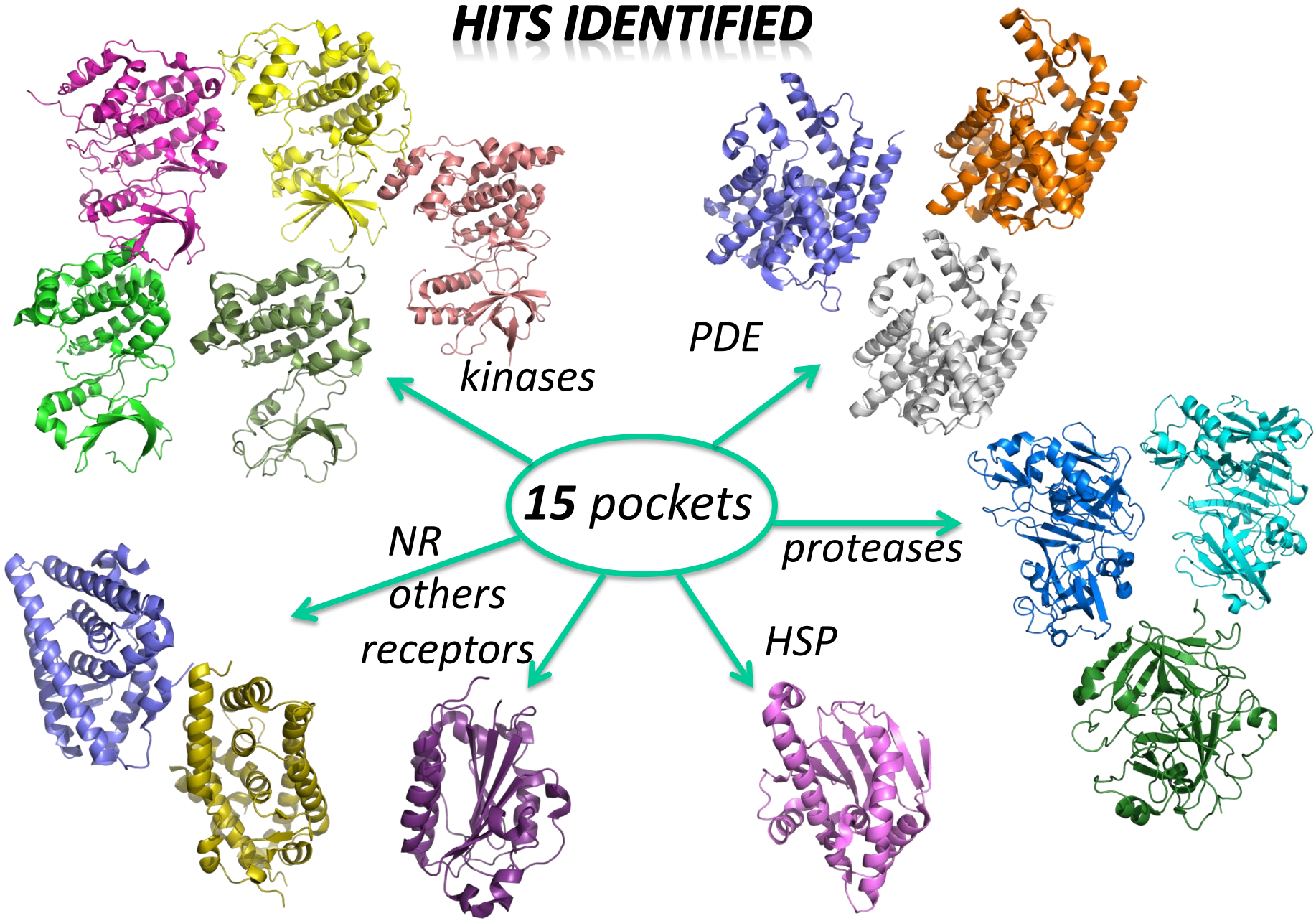
Use TS cavity as template for a virtual screening against all PDB cavities containing a ligand (~ 70.803)

Verify if in the top-ranked solutions we found cavities known to be similar to the TS ones

Select cavities similar to the TS cavity (top-ranked solutions). Select ligands contained in the new cavities as potential TS inhibitors.

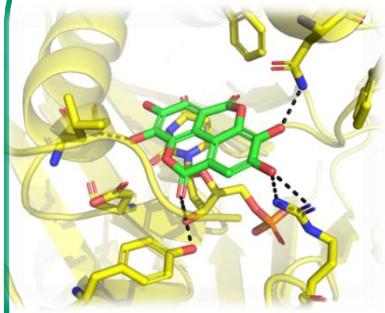
Docking of the potential candidates into TS cavity with FLAPdock

*Virtual Screening PP comparison* → *Validation* → *Candidates selection* → *Docking*

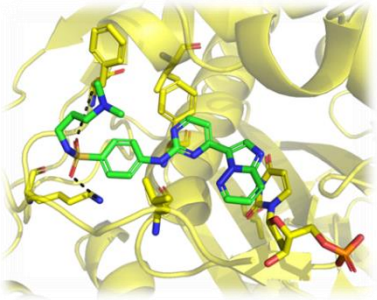

HITS IDENTIFIED
15 pockets
kinases
PDE
proteases
HSP
NR
others
receptors

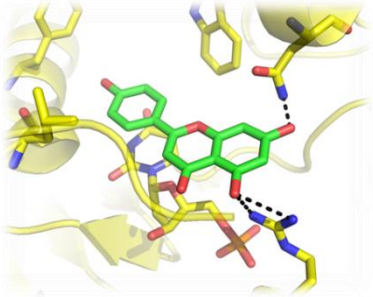
## LIGANDS SELECTED FROM SIMILAR BINDING POCKETS AND DOCKED INTO TS

from CASEIN KINASE

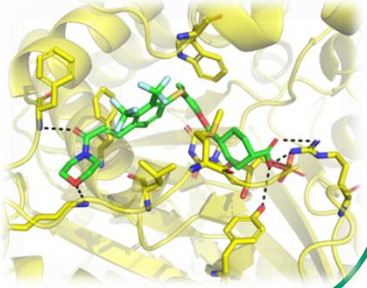


from INTEGRIN





from CYCLIN-DEPENDENT KINASE 2



**30 HITS**

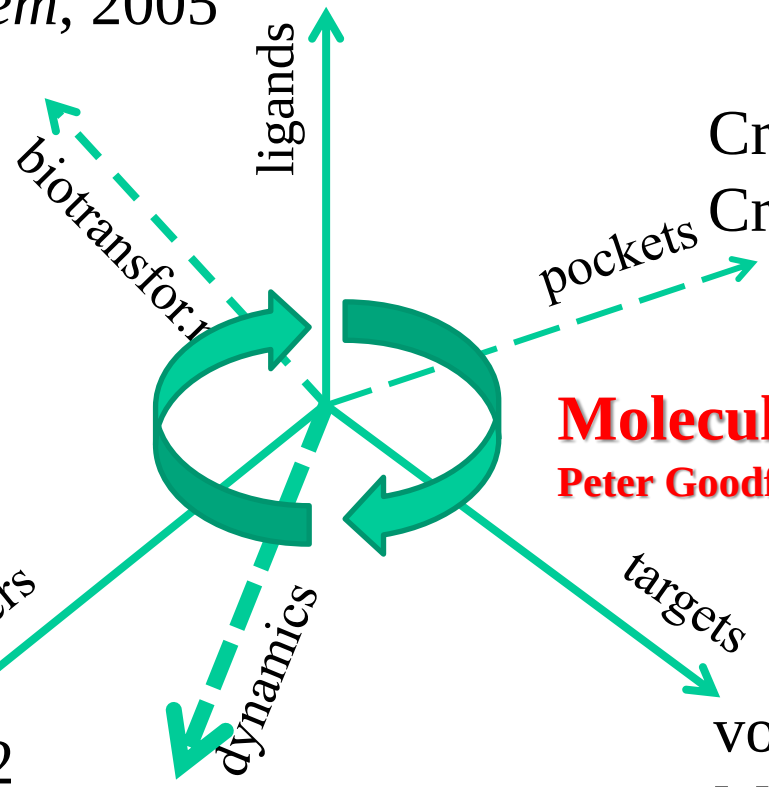
3 MOLECULES

**1 ACTIVE MOLECULE**
**($k_i$ ~ 1 µm)**

**Is a drug repurposable for another target?**
**What is the molecular mechanism of a drug side effects?**
**How can we improve the ligand selectivity?**

Milletti, *JCIM*, 2006

Cruciani, *JMedChem*, 2005




Cruciani, *UK QSAR*, 2005
Cruciani, *JCIM*, 2007

**BioGPS**

**Molecular Interaction Fields**
**Peter Goodford 1984**

Mason, *TIPS*, 2012

von Itzstein, *Nature*, 1993
Muratore, *PNAS*, 2012

GRID manual 1995
Carosati, *JMedChem*, 2004

**Holistic approach**

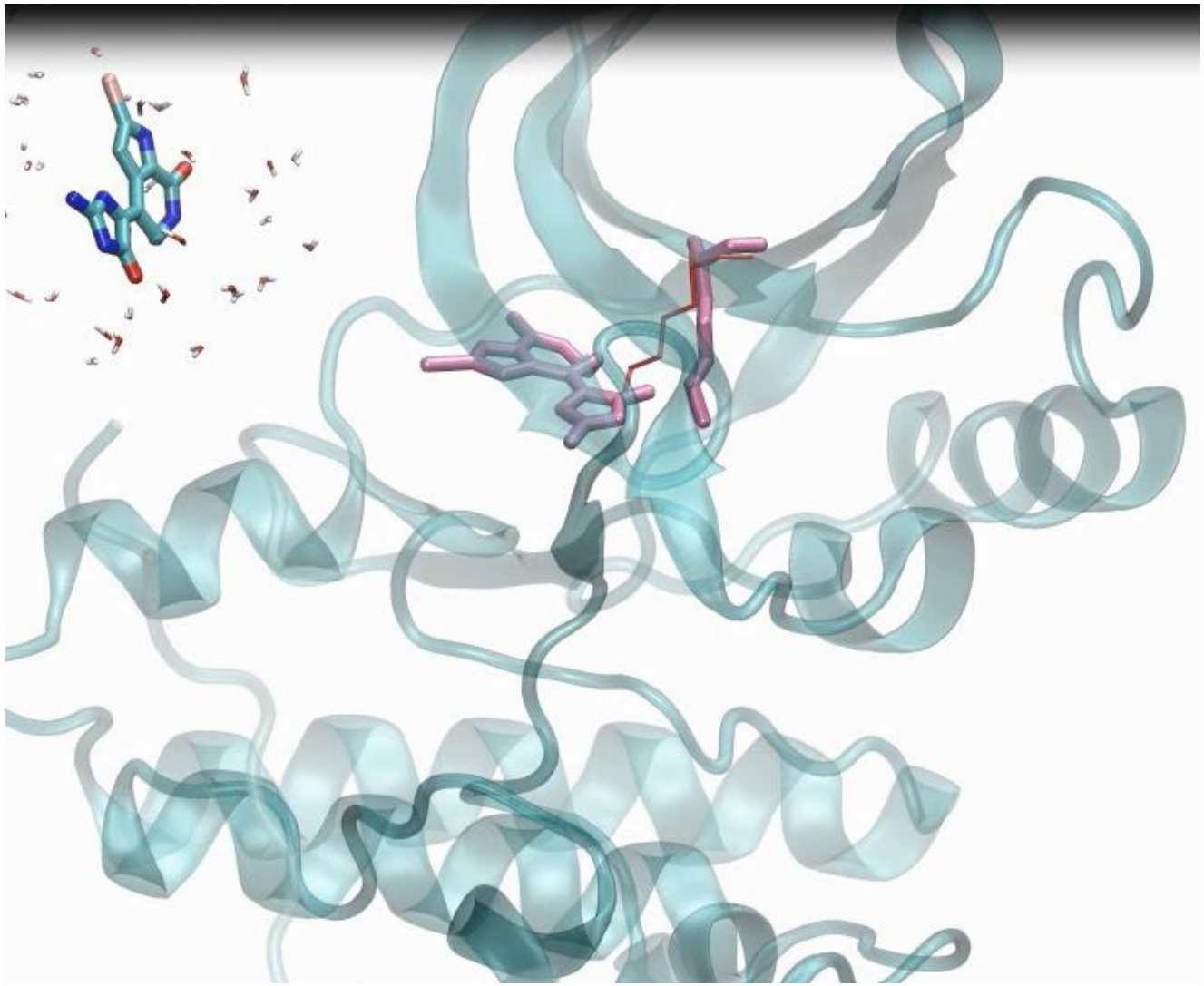
# MD & BioGPS:

finding transient pockets &
using flexibility to search for
off-targets

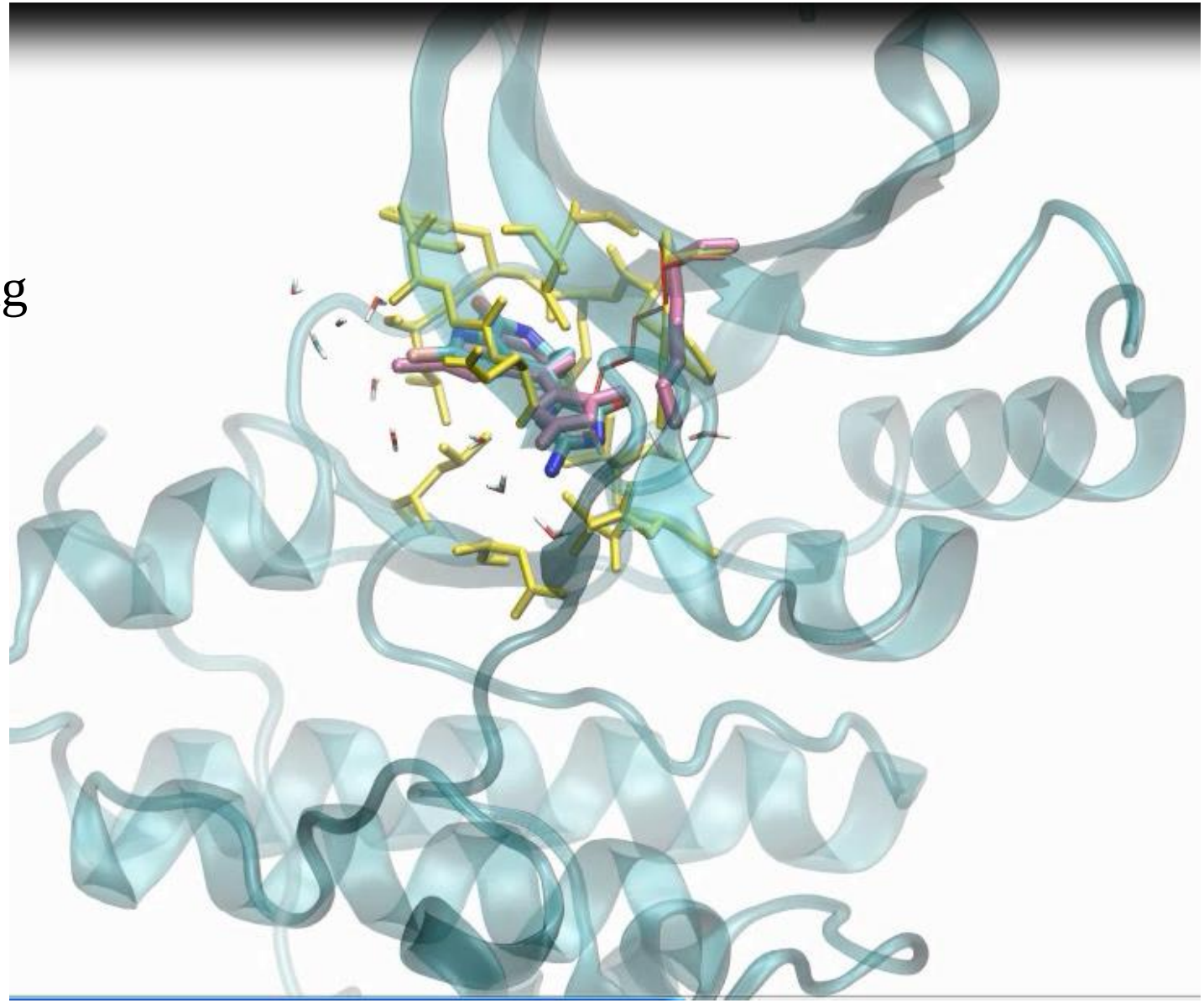



Gabriele Cruciani, Perugia University

Hymenialdisine docked into '*apo*' unbound protein

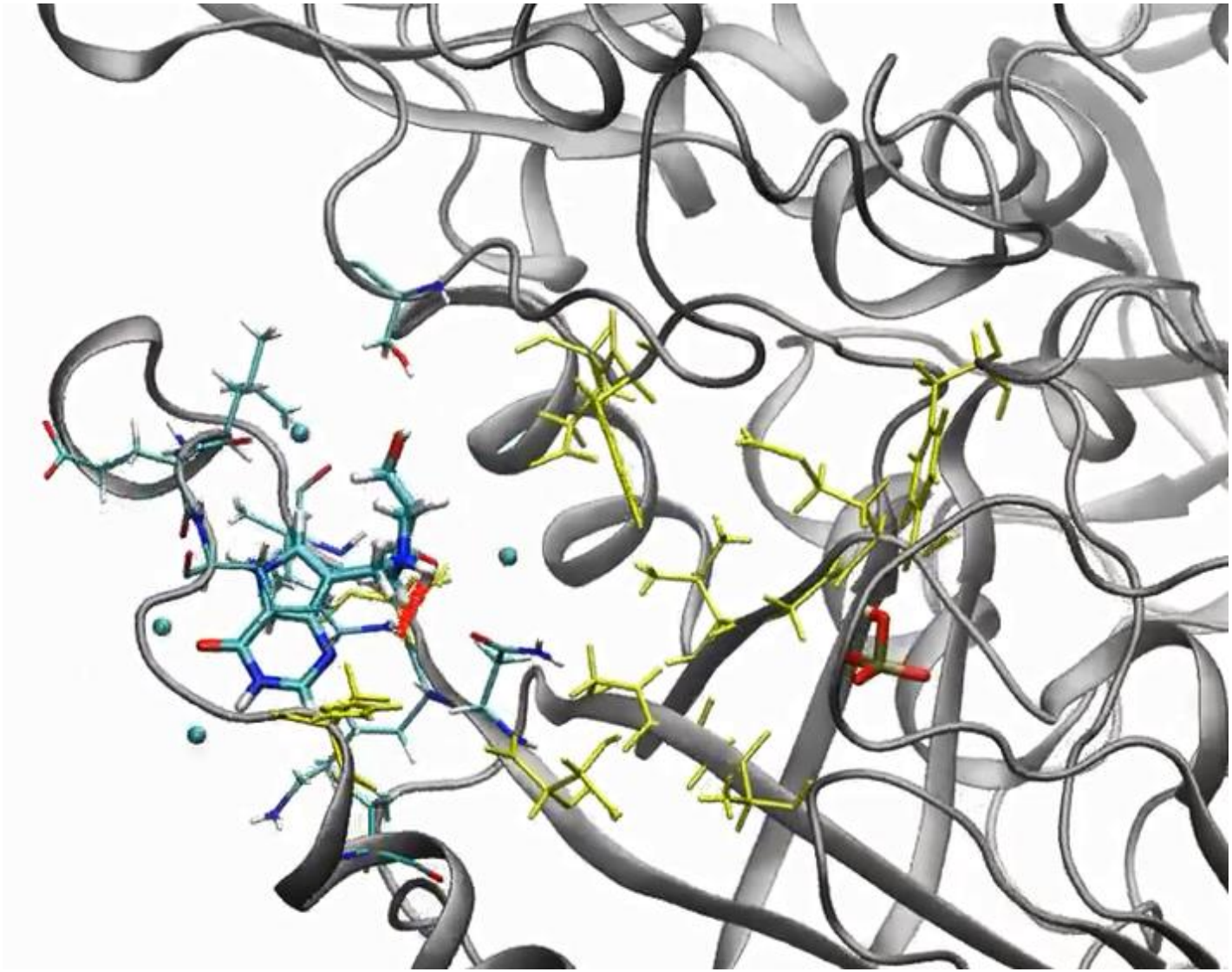


Lys moves
4A away
upon docking

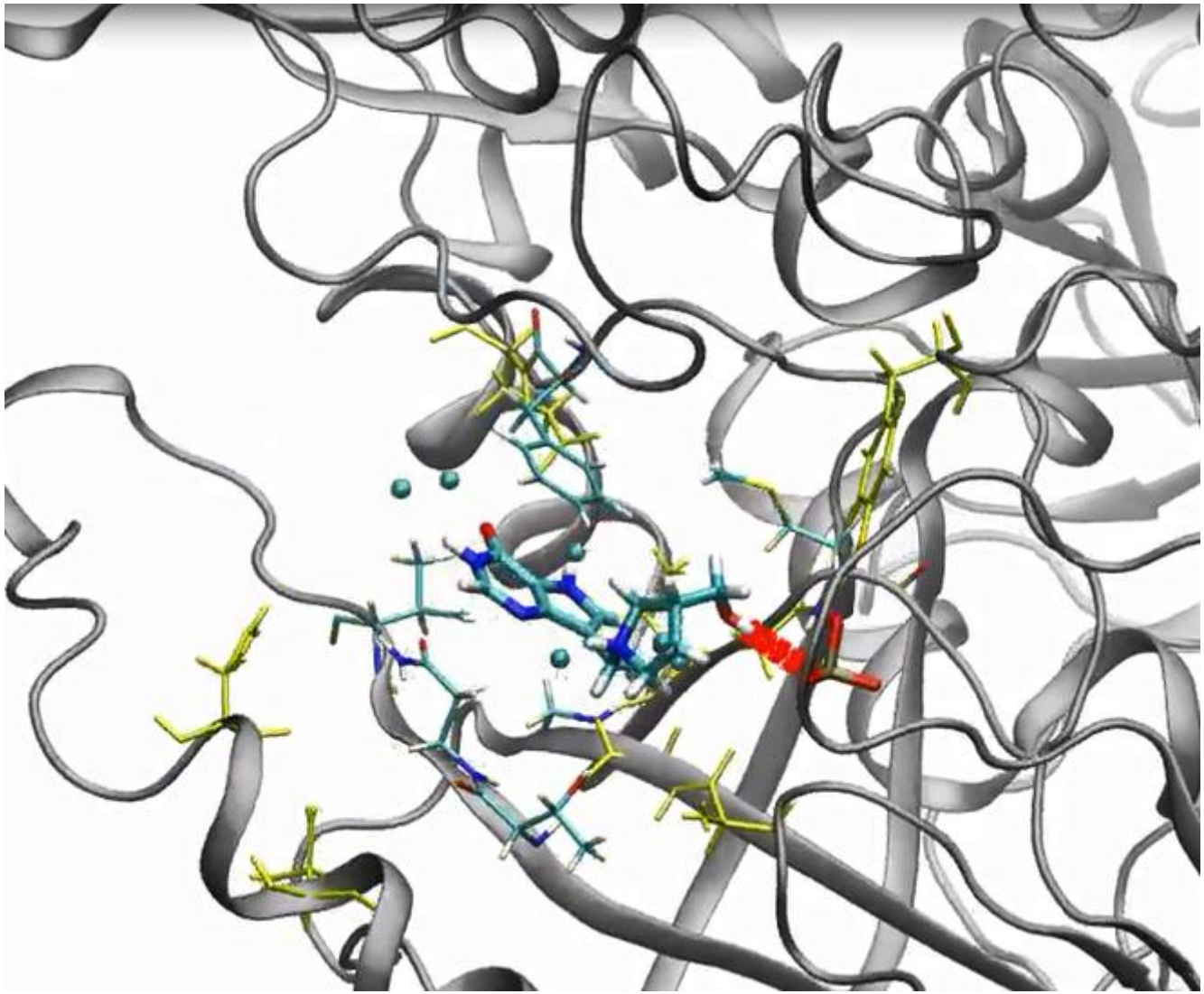


DATMe-ImmH docked into '*apo*' unbound protein

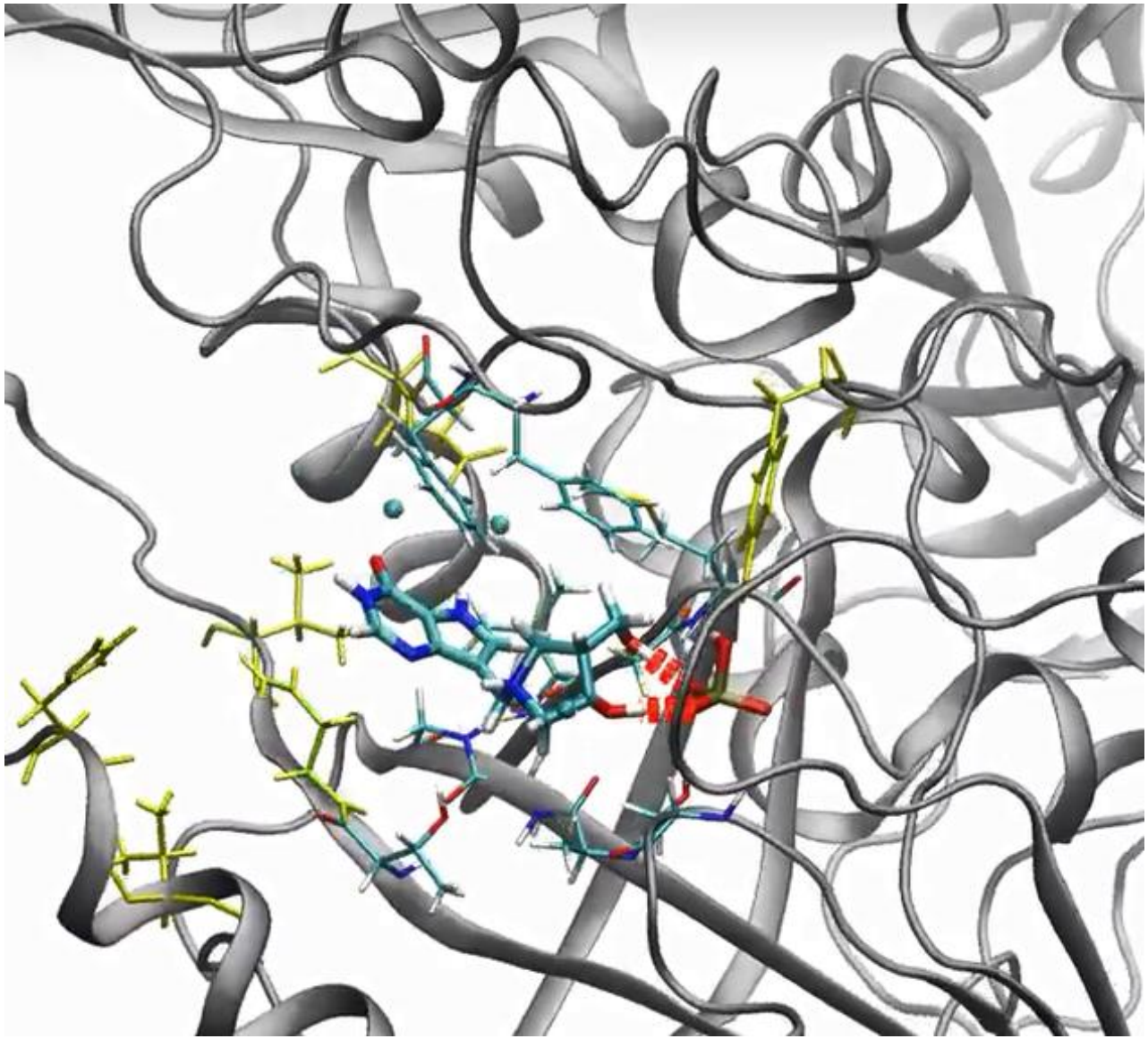
The role of a transient pocket



***Purine Nucleoside Phosphorilase (PNP)***

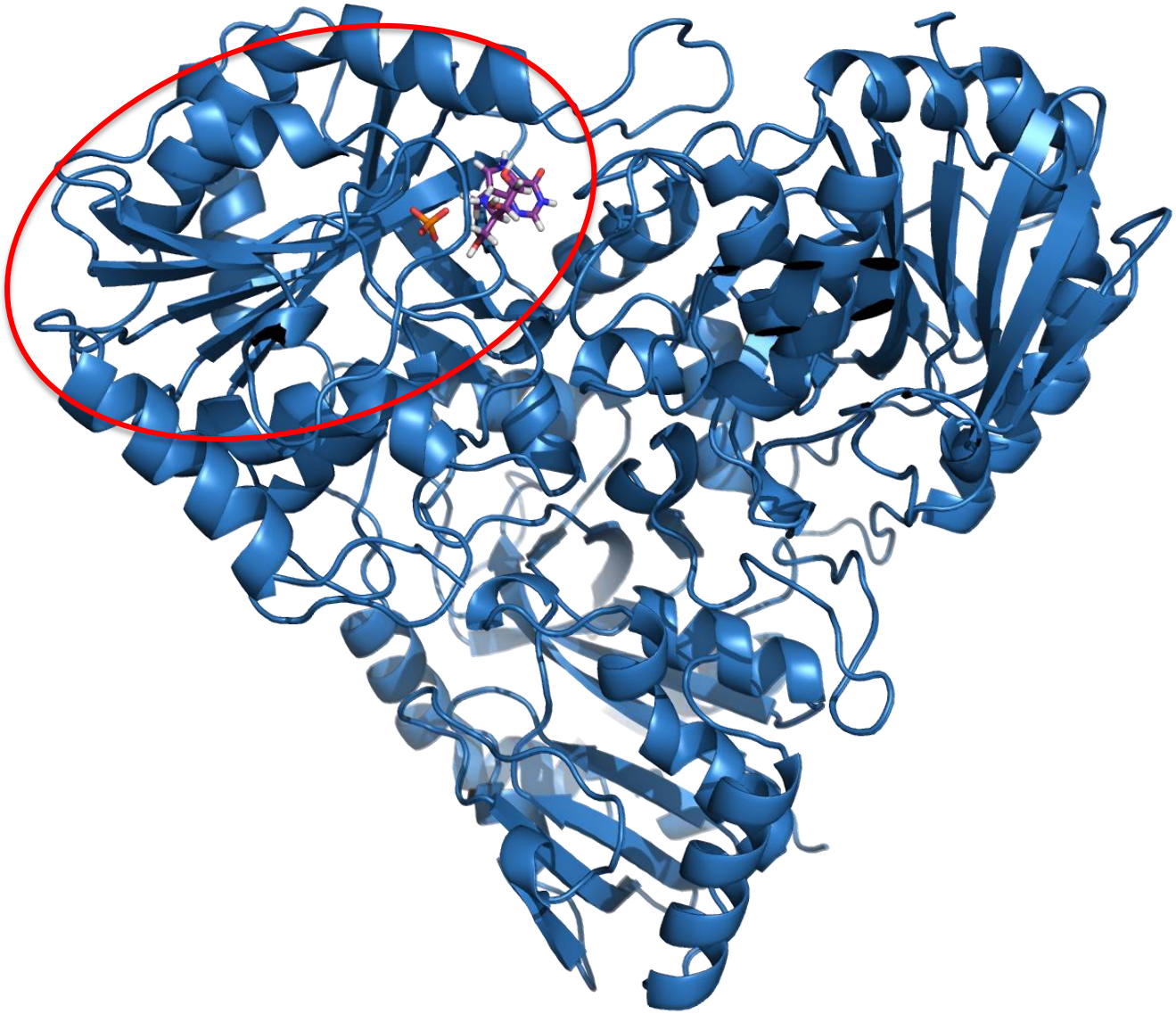


# BioGPS: The Music for the Chemo- and Bioinformatics Walzer

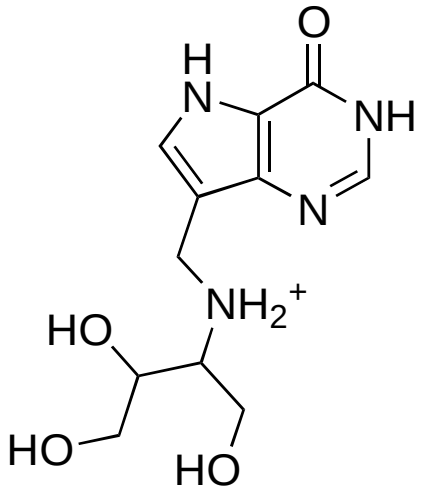


# The test case

## *Purine Nucleoside Phosphorilase (PNP)*



PDBcode 3k8o
(2.40 Å)



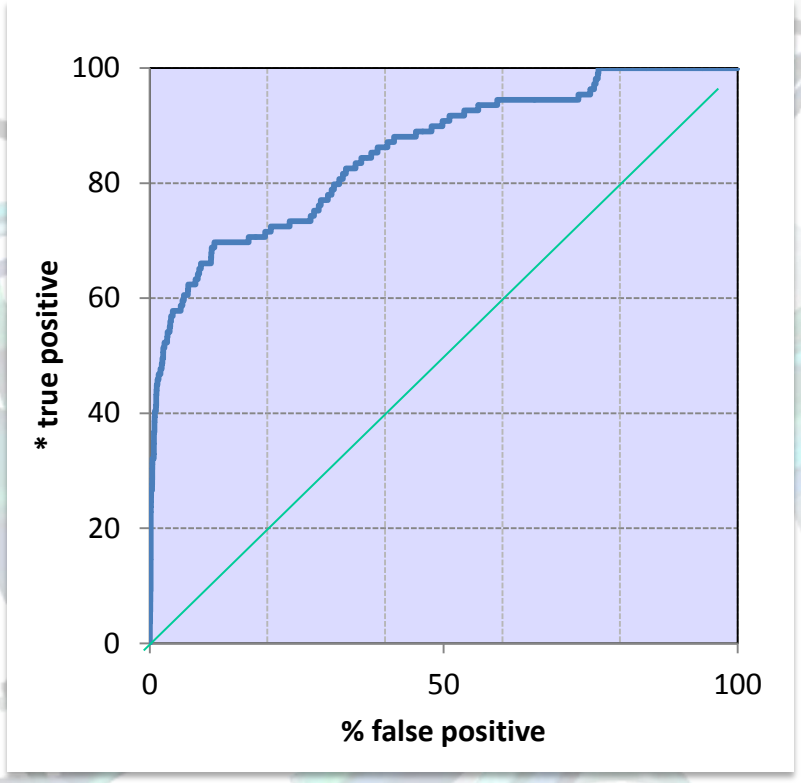

DATMe-ImmH
$K_d$ = 8.6 pM

# The ligands dataset


It contains 102 targets, including 38 of the original 40 DUD targets

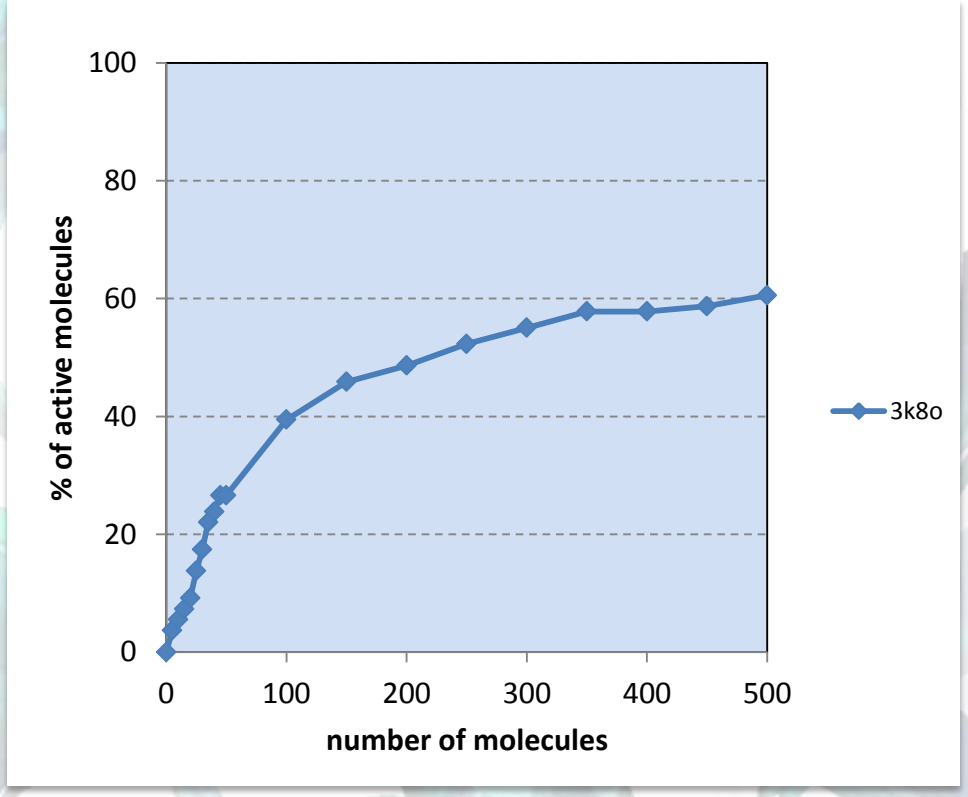
**PNP dataset** {
- 109 actives (229 tautomers, protomers, stereoisomers)
- 7000 decoys

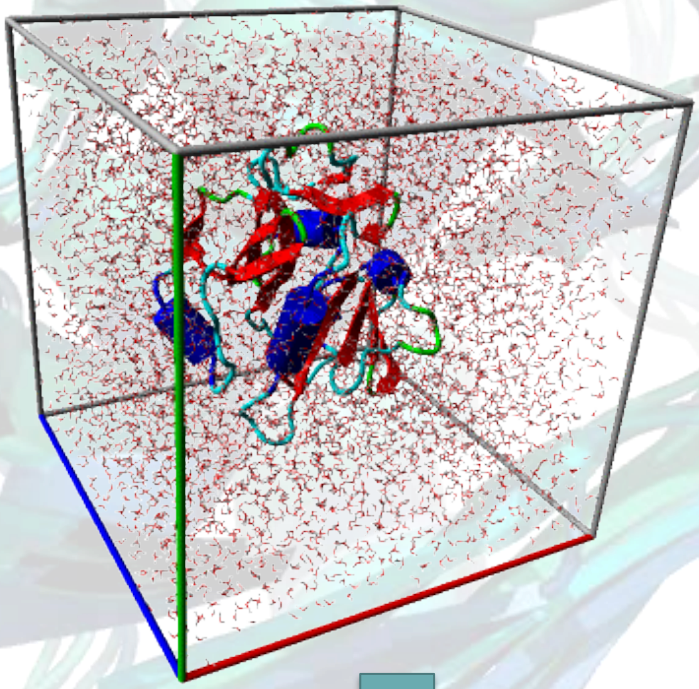
# SBVS against 3k8o



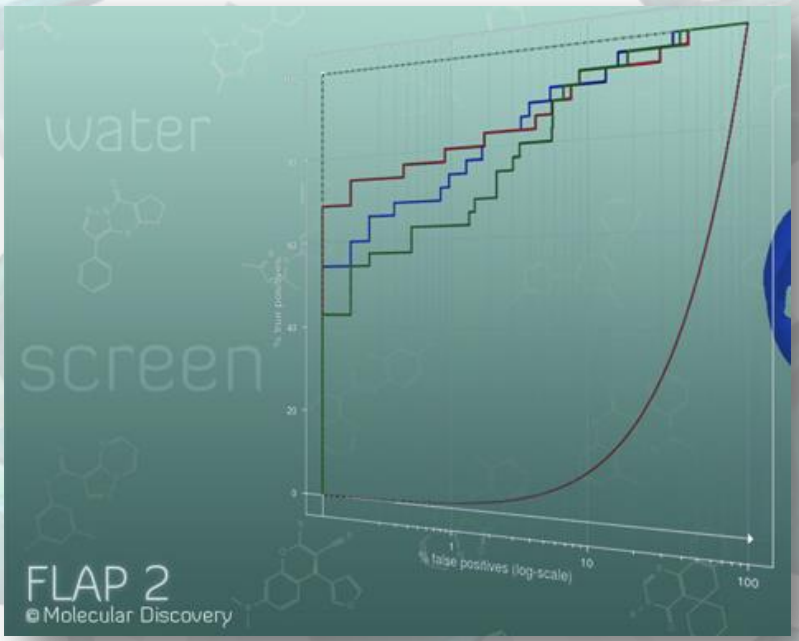

## AUC 0.85

| 0.5% | 1% | 2% | 5% |
|---|---|---|---|
| 0.32 | 0.40 | 0.48 | 0.58 |





# From Molecular Dynamics to Virtual Screening



# Which are the advantages of combining MD and VS?

We allow the structure to relax …

➔ we are free from the structural ligand bias

➔ we allow larger and different ligands to fit the binding site

➔ we can find new or different hits

**most important VS success**

# Which are the problems of combining MD and VS?

From our trajectory ←

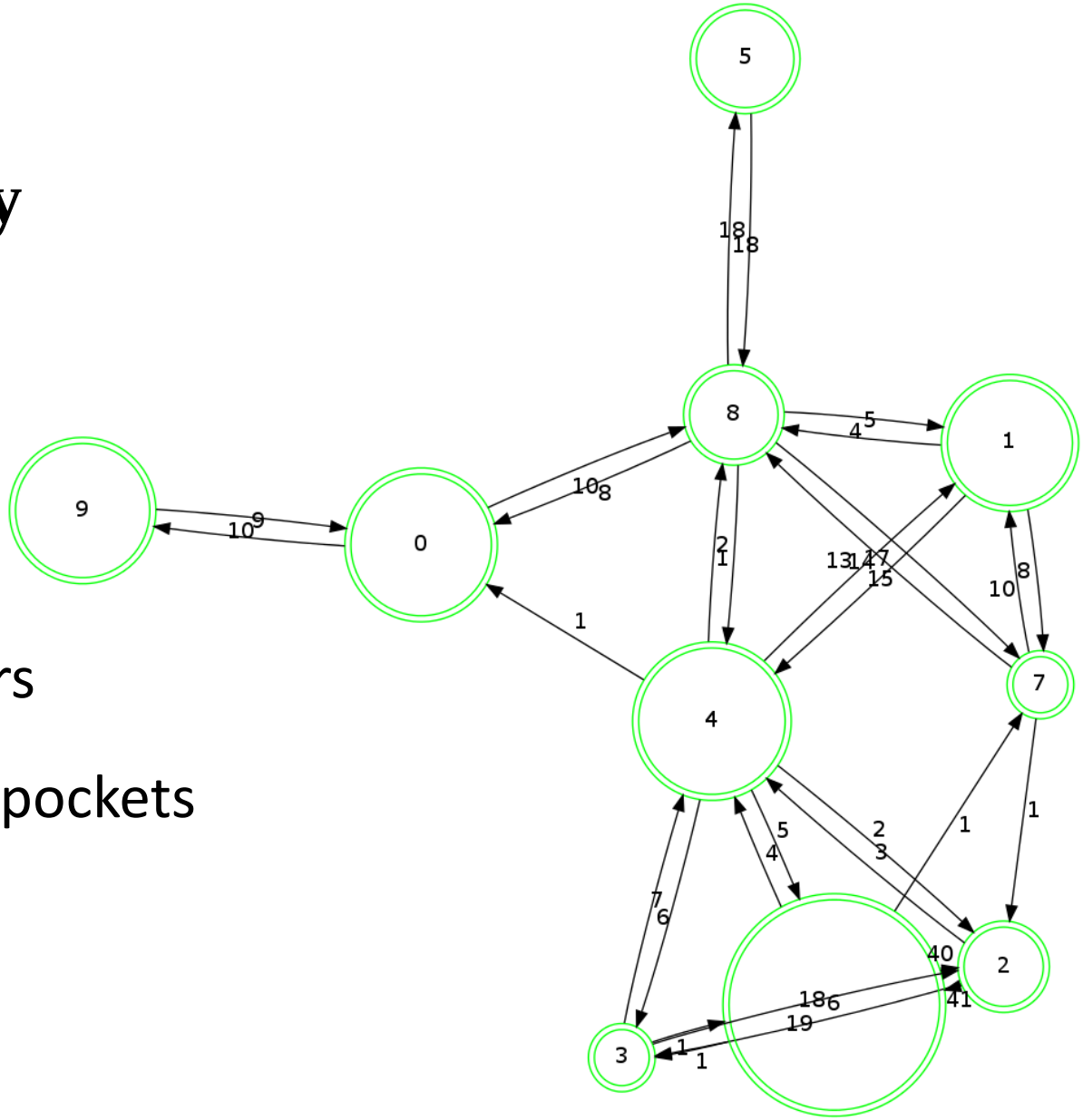
➔ we have to select the right structures among 50.000 snapshots

➔ we might add noise rather than information

*How can we select the MD structures for the screening?*

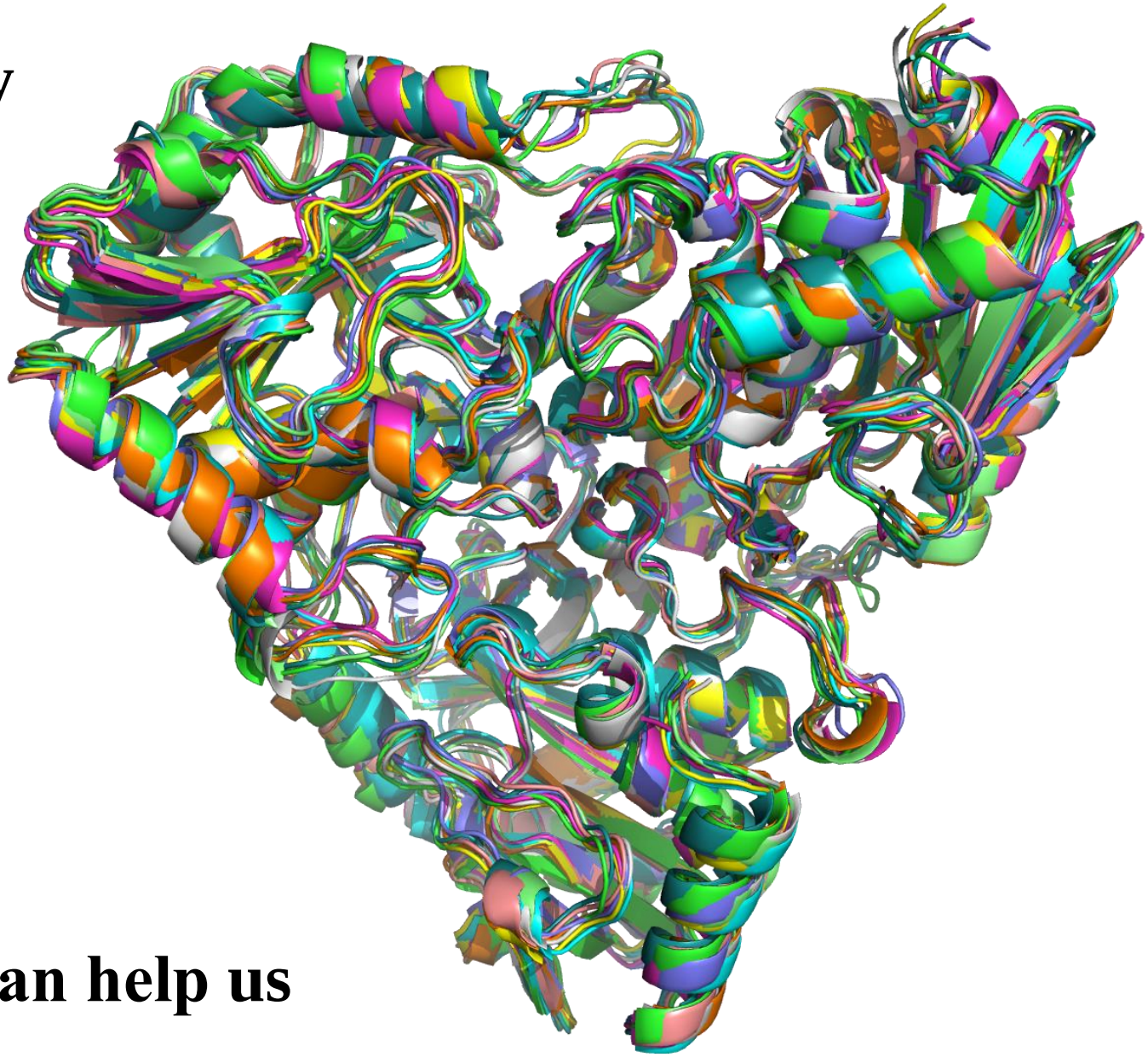
# The clustering technique

# The clustering for the PNP trajectory




- ➔ Most distant clusters
- ➔ PCA analysis of the pockets
- ➔ RMSD calculation
- ➔ AUC calculation

**We are lucky but…**



**…the LDA can help us**

# The Linear Discriminant Analysis

x-ray
md0
md1
md2
md3
md4
md5
…..

11 possible candidates

3 templates

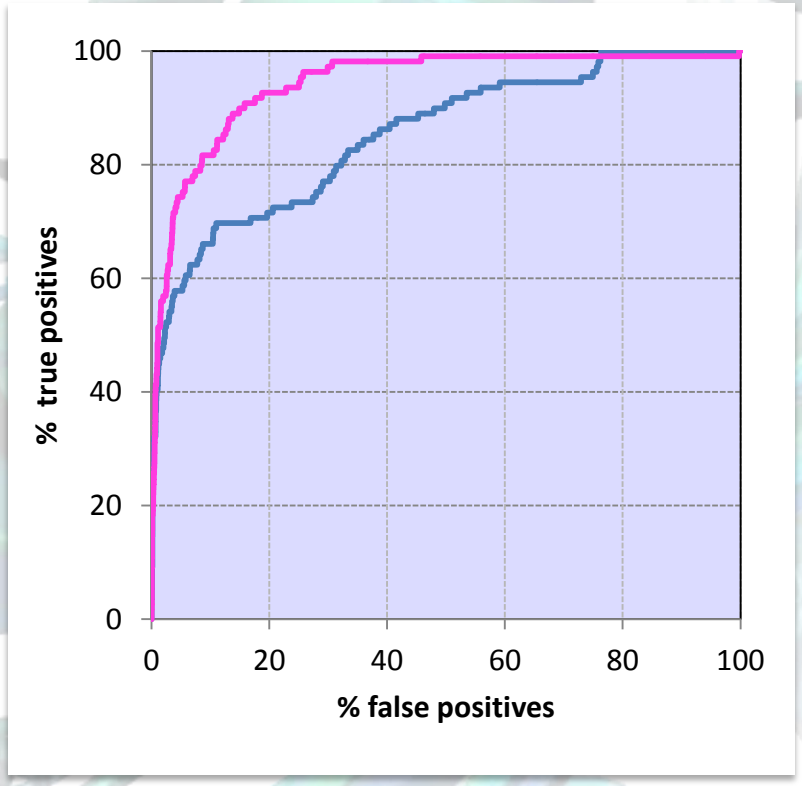
3 scores

17 possible scores

H
N1
O
DRY
DRY*O
H*O*H
H*DRY
H*O*N1
…..

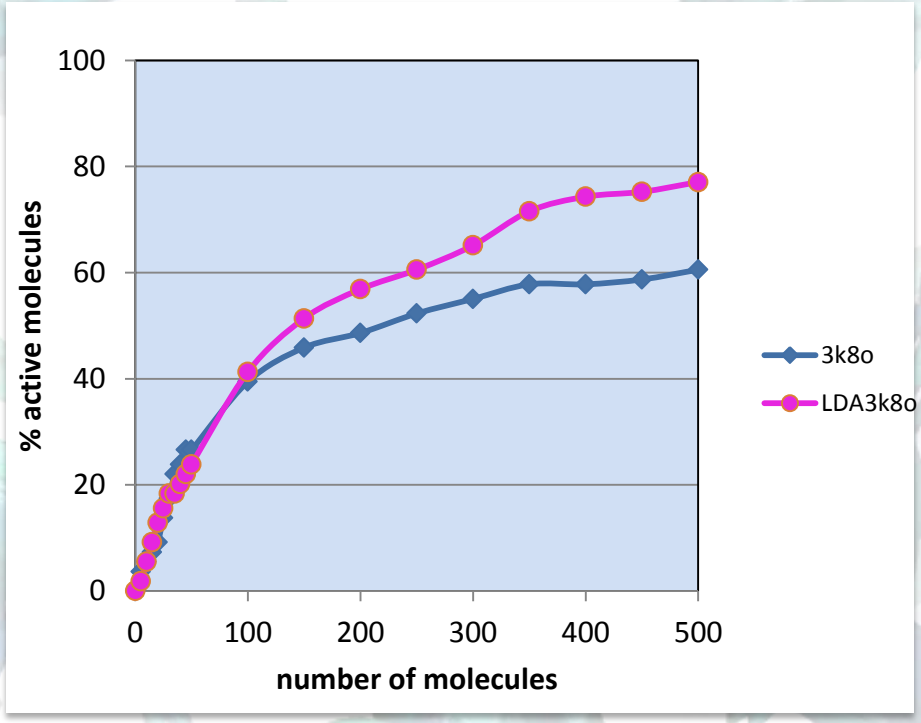
*Best combination to separate actives from decoys*

# SBVS against LDA (3k8o + MD medoids)

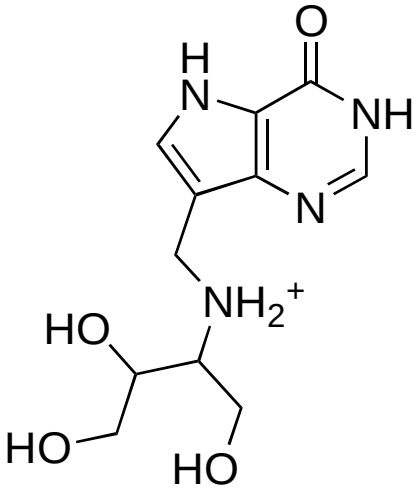



## AUC 0.85 vs 0.94

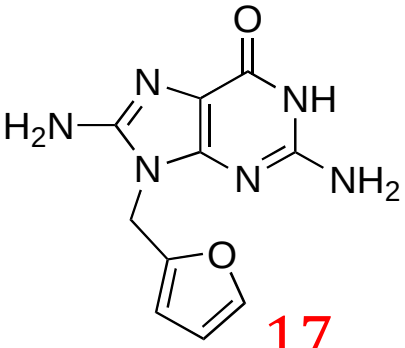
| | 0.5% | 1% | 2% | 5% |
|---|---|---|---|---|
| 3k8o | 0.32 | 0.40 | 0.48 | 0.58 |
| LDA | 0.28 | 0.45 | 0.57 | 0.74 |



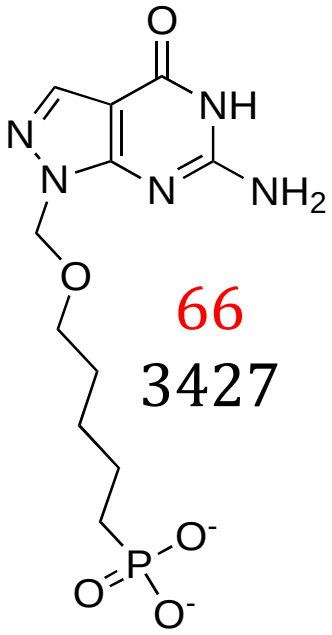
# The different ligands and the different ranking of the LDA-based SBVS



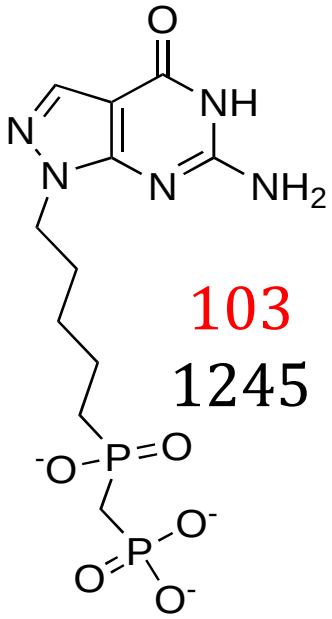
DATMe-ImmH



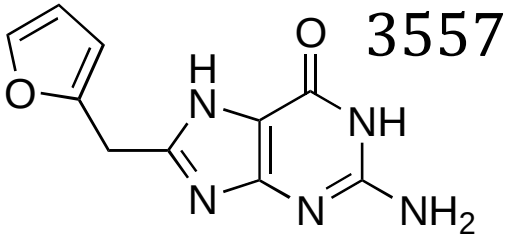
17
215



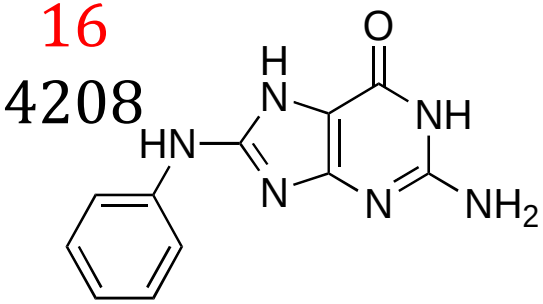
66
3427



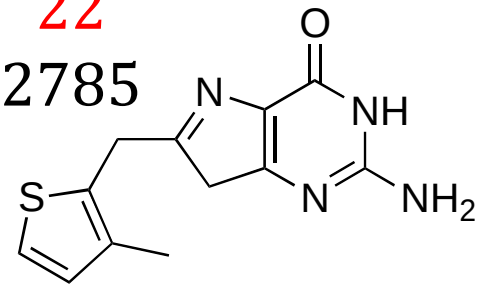
103
1245



131
3557

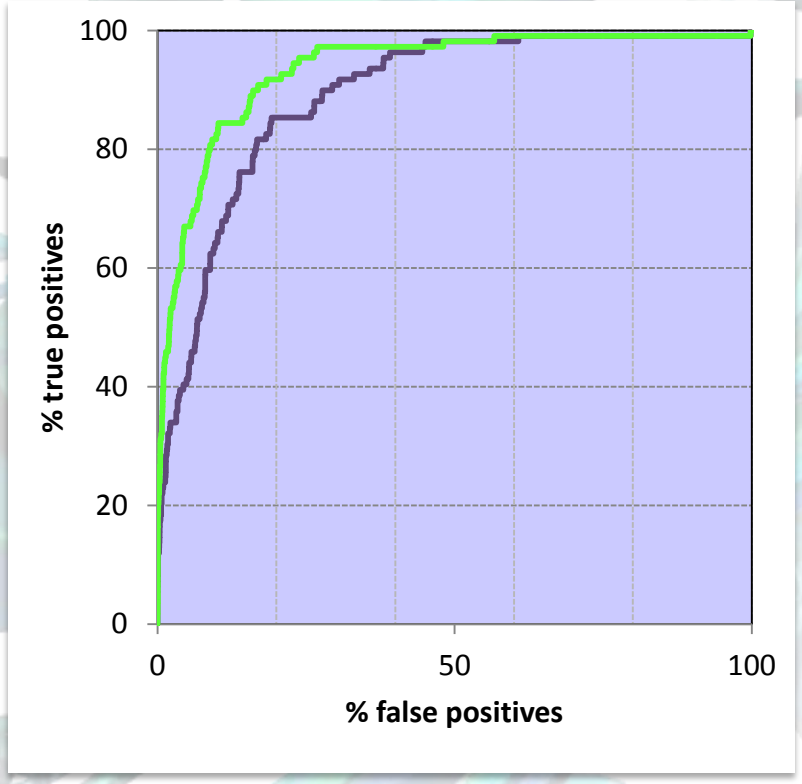


16
4208



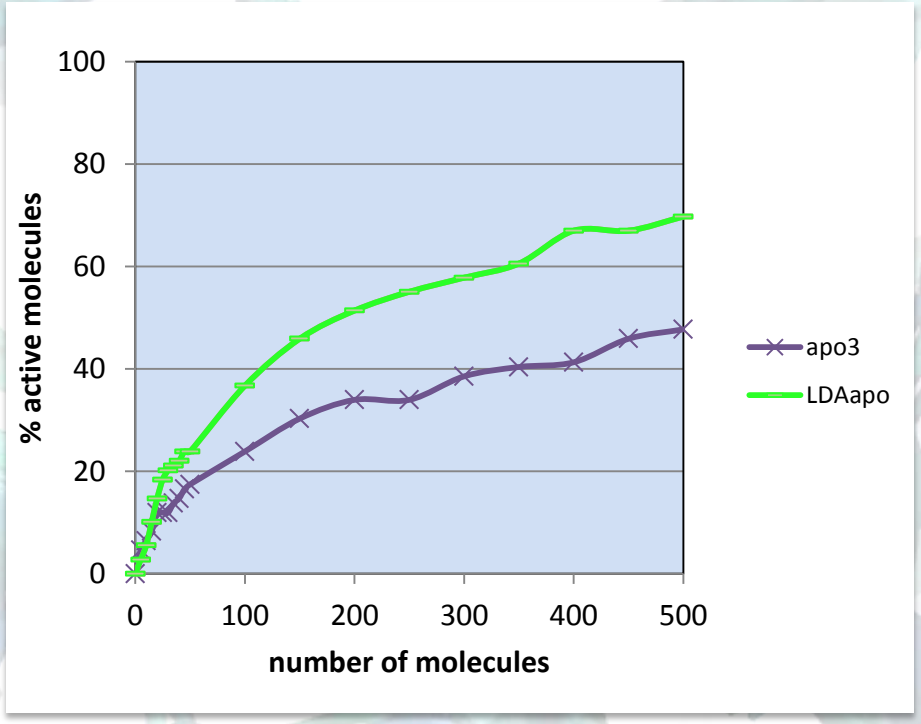
22
2785

# SBVS against LDA (apo + MD medoids)

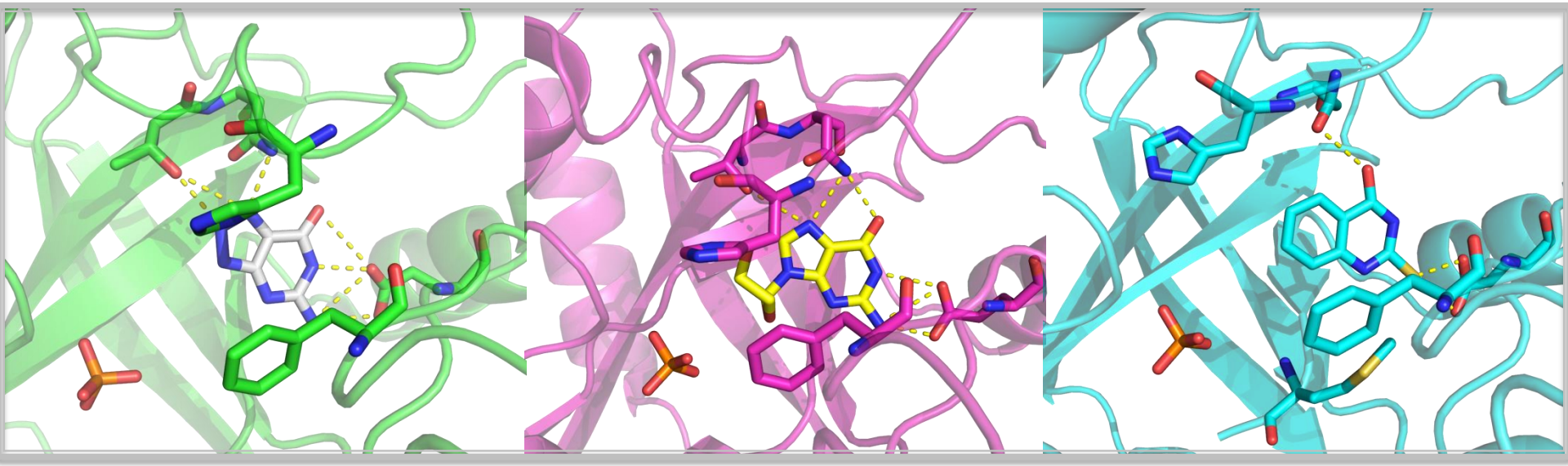



## AUC 0.89 vs 0.93

| | 0.5% | 1% | 2% | 5% |
|---|---|---|---|---|
| apo | 0.17 | 0.24 | 0.32 | 0.41 |
| LDA | 0.31 | 0.41 | 0.50 | 0.67 |

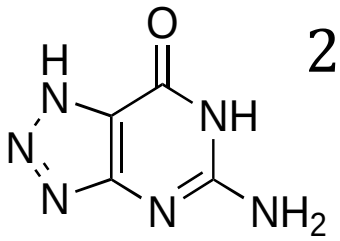


# and if we analyze different X-ray???



**1v41**

2.85 Å

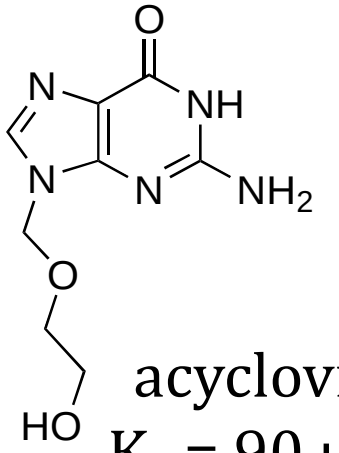



8-azaguanine

$K_d$ = 20 µM

**1pwy**

2.80 Å

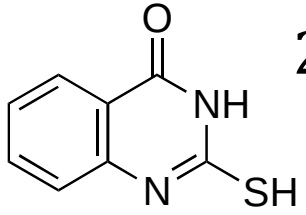



acyclovir

$K_d$ = 90 µM

**3d1v**

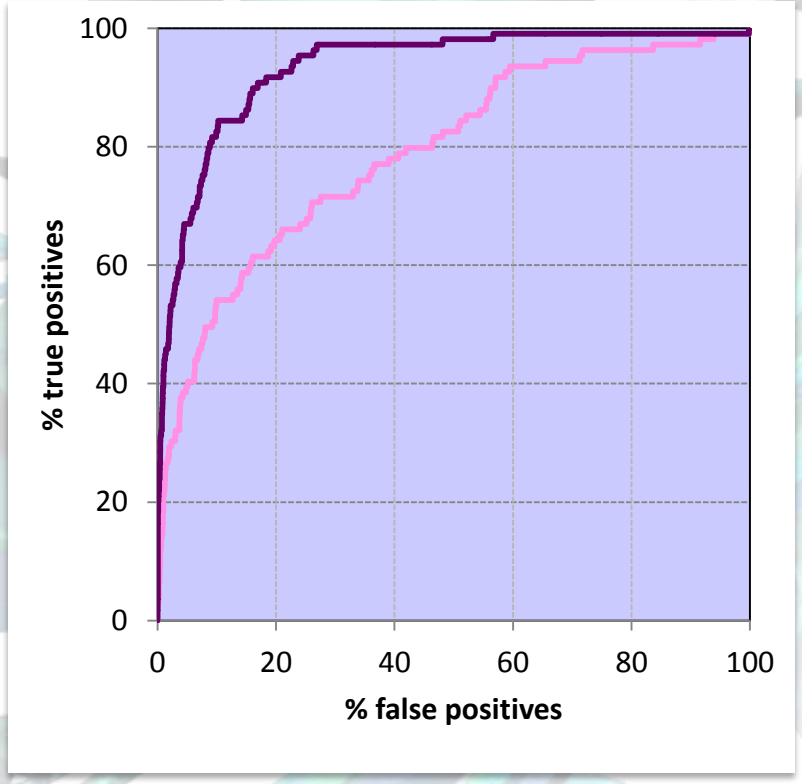
2.70 Å




2-mercapto (3H) quinazolinone

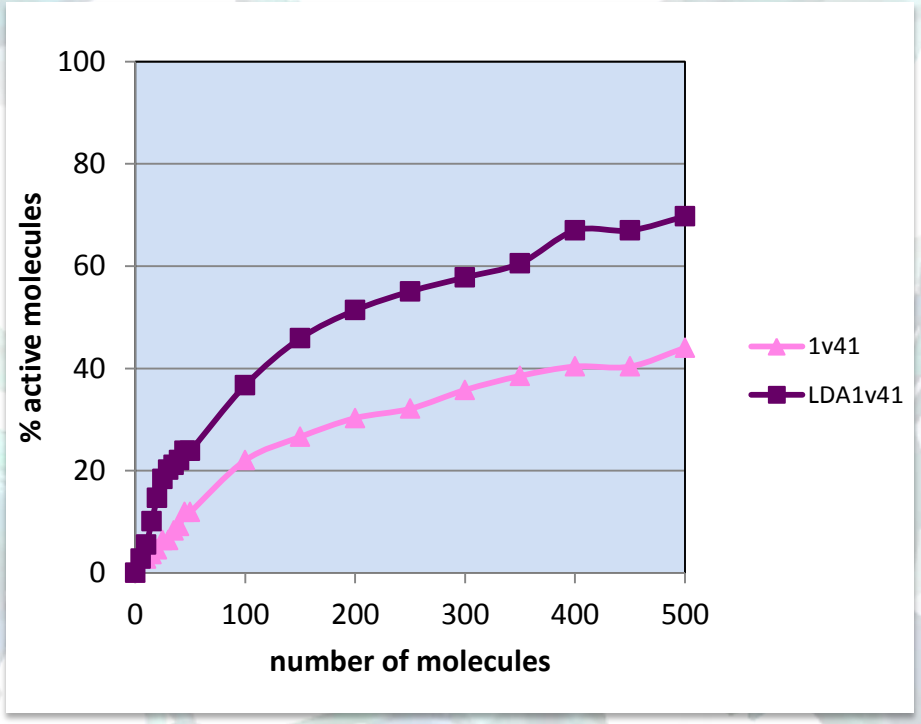
$K_d$ = 324 µM

# SBVS against LDA (1v41 + MD medoids)

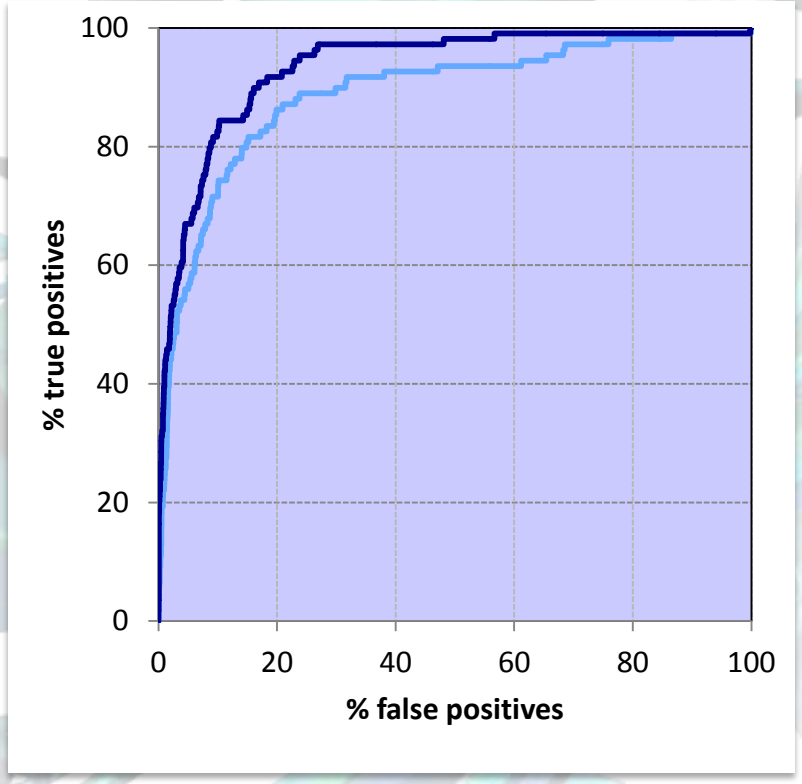



## AUC 0.79 vs 0.93

| | 0.5% | 1% | 2% | 5% |
|---|---|---|---|---|
| 1v41 | 0.12 | 0.21 | 0.29 | 0.39 |
| LDA | 0.31 | 0.41 | 0.50 | 0.67 |

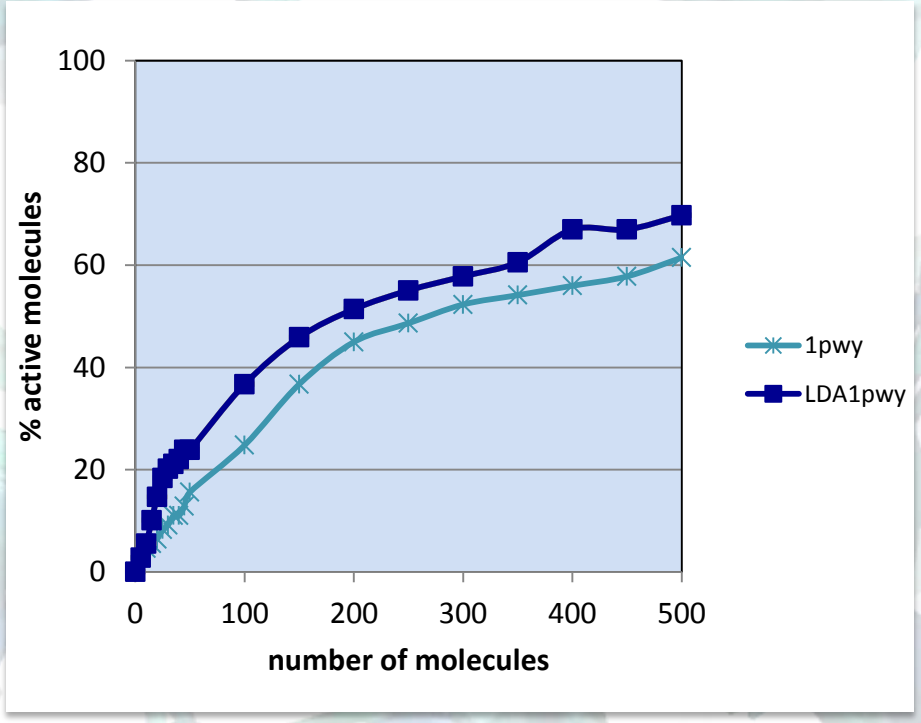


# SBVS against LDA (1pwy + MD medoids)

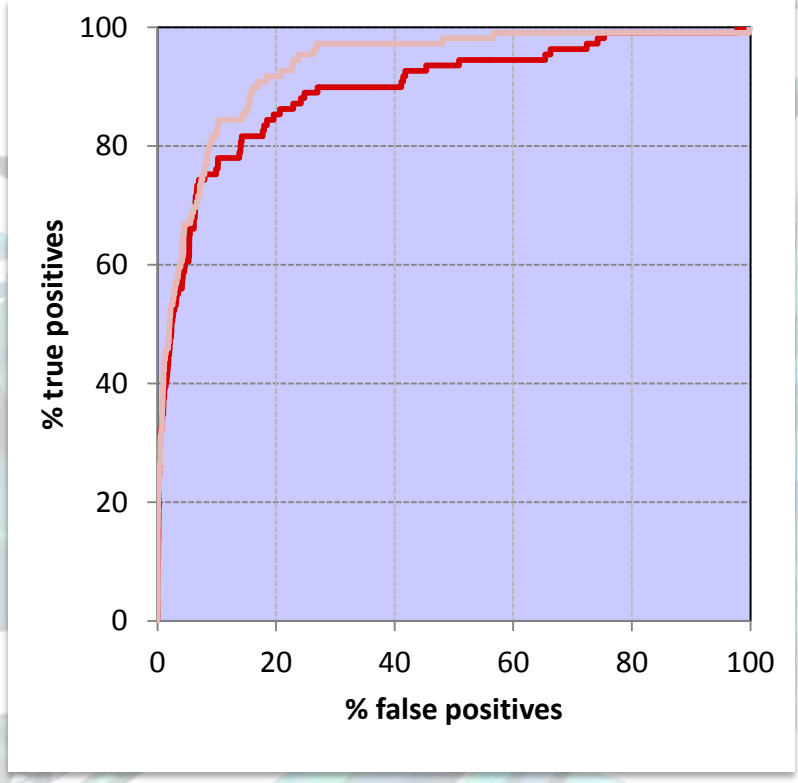



## AUC 0.89 vs 0.93

| | 0.5% | 1% | 2% | 5% |
|---|---|---|---|---|
| 1pwy | 0.16 | 0.24 | 0.44 | 0.56 |
| LDA | 0.31 | 0.41 | 0.50 | 0.67 |

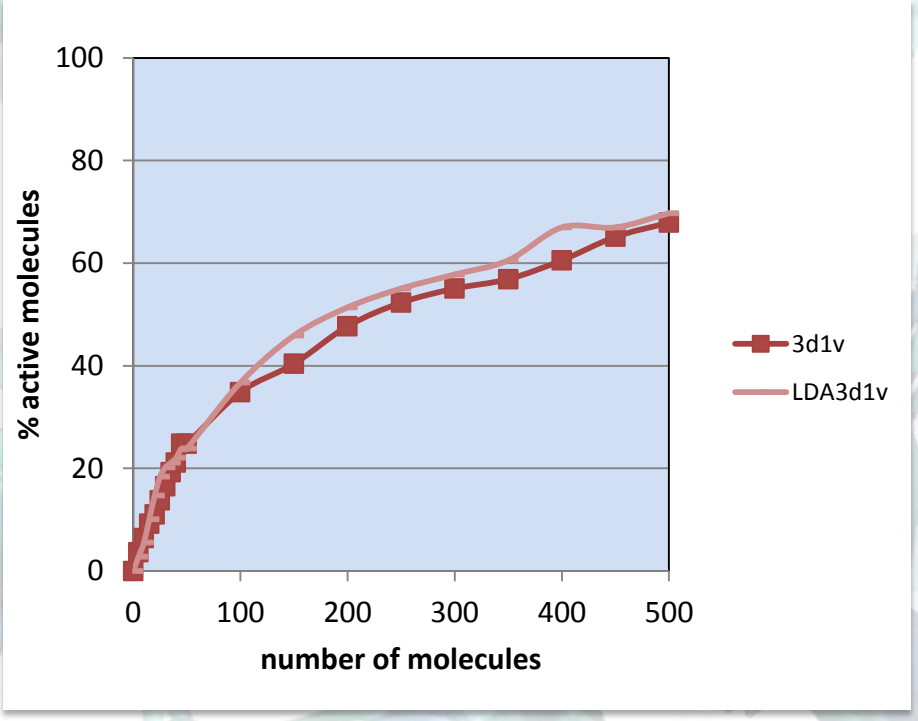


# SBVS against LDA (3d1v + MD medoids)




## AUC 0.90 vs 0.93

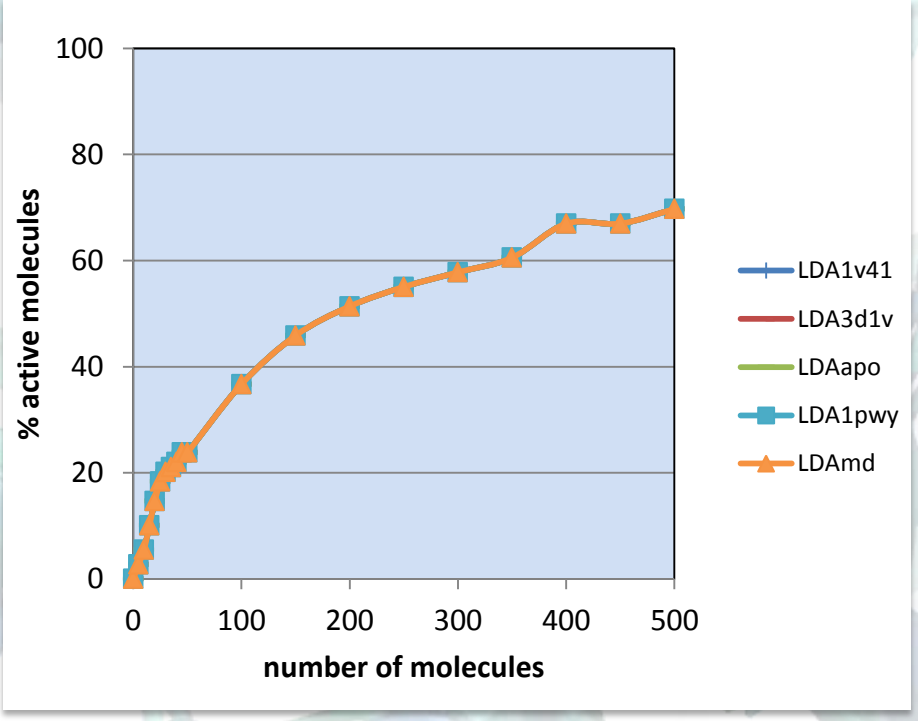
| | 0.5% | 1% | 2% | 5% |
|---|---|---|---|---|
| 3d1v | 0.32 | 0.37 | 0.46 | 0.61 |
| LDA | 0.31 | 0.41 | 0.50 | 0.67 |



| | AUC | 0.5 % | 1% | 2% | 5% |
|---|---|---|---|---|---|
| apo | 0.89 | 0.17 | 0.24 | 0.32 | 0.41 |
| 1v41 | 0.79 | 0.12 | 0.21 | 0.29 | 0.39 |
| 1pwy | 0.89 | 0.16 | 0.24 | 0.44 | 0.56 |
| 3d1v | 0.90 | 0.32 | 0.37 | 0.46 | 0.61 |
| LDA | 0.93 | 0.31 | 0.41 | 0.50 | 0.67 |

# LDA comparison

Same performances regardless by the original x-ray structure



# *conclusions*

- ✔ *The combination of Molecular Dynamics and Virtual Screening can improve the quality of our predictions!*
- ✔ *The inclusion of the flexibility can remove the structural bias of the original ligand and the induced fit memory.*
- ✔ *The combination of the clustering and of the LDA allows to choose the most representative structures/medoids and to add essential information rather than noise.*


Trasimeno lake
Thanks
gabri@moldiscovery.com