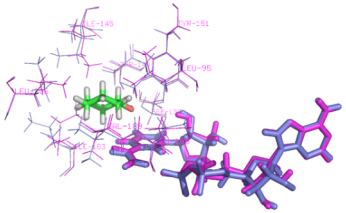
KING'S College LONDON
University of London

Randall Division
of Cell and
Molecular Biophysics

Using molecular docking for substrate identification: a hopeless task?

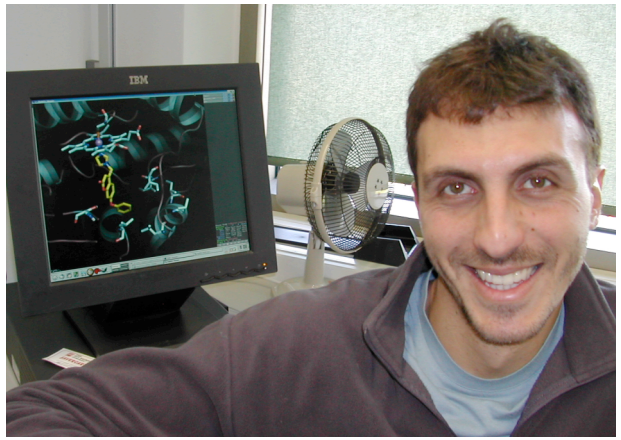
Irilenia Nobeli
Randall Division
King's College London

4th Joint Sheffield Conference on Chemoinformatics
18/06/2007



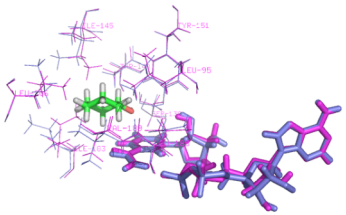
Acknowledgements

Janet Thornton



Angelo Favia

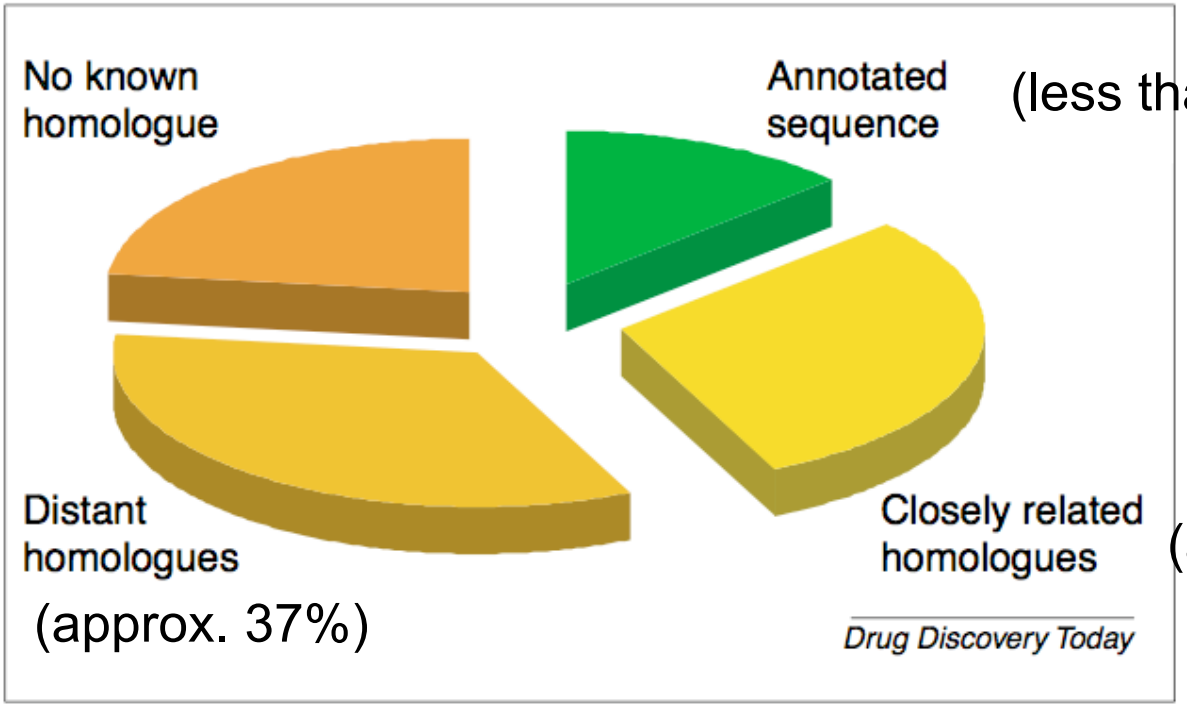
Fabian Glaser



Why we started this work

Distribution of known sequences (~2 million)

(approx. 25%)

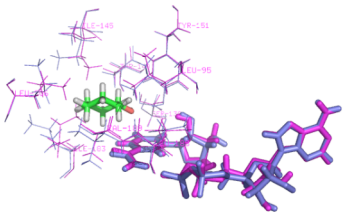


(less than 15%)

(approx. 37%)

(approx. 23%)

From: Ofran et al. (2005). Drug Discovery Today 10, 1475.

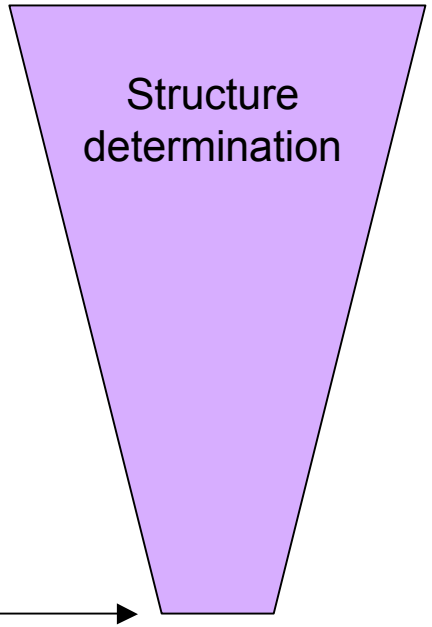


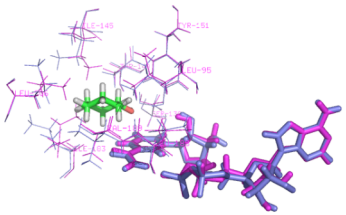
Why we started this work

Structural genomics is an attempt to help but…

e.g. The Structural Genomics Consortium reports:
433 solved structures (June 2007)

but only
53 papers
(incl. analyses)





The plan for this work

Concentrate on a large family of related proteins with known substrate promiscuity within the family

» Short-Chain Dehydrogenase/Reductases (SDRs)
(a target of the Structural Genomics Consortium)

Use molecular docking

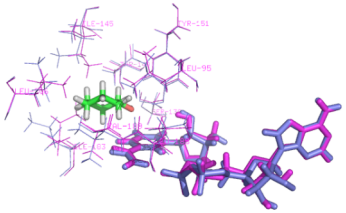
» established technique in virtual screening and inhibitor studies

» evidence that it has been successful for substrate identification in previous studies

Kalyanaraman et al. (2005). Biochemistry 44, 2059.

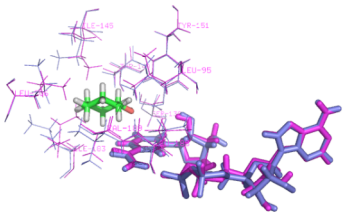
Tyagi & Pleiss (2006). J. Biotechnol. 124, 108..

Hermann et al. (2006). J. Am. Chem. Soc. 128, 15882.



Why docking may be useful in function identification

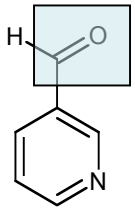
- Sequence similarity is helpful in predicting function/substrates but it is not, on its own, a reliable predictor of function!



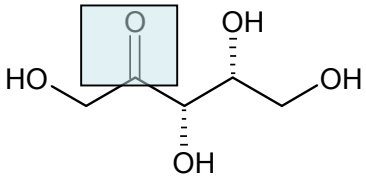
How is sequence similarity related to substrate similarity?

1cyd (mouse carbonyl reductase)
vs.
1pr9 (human L-xylulose reductase)

67.2% sequence identity

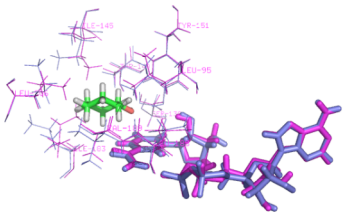


vs.



0.09 similarity score

Carbonyl reductase is known to accept many different substrates

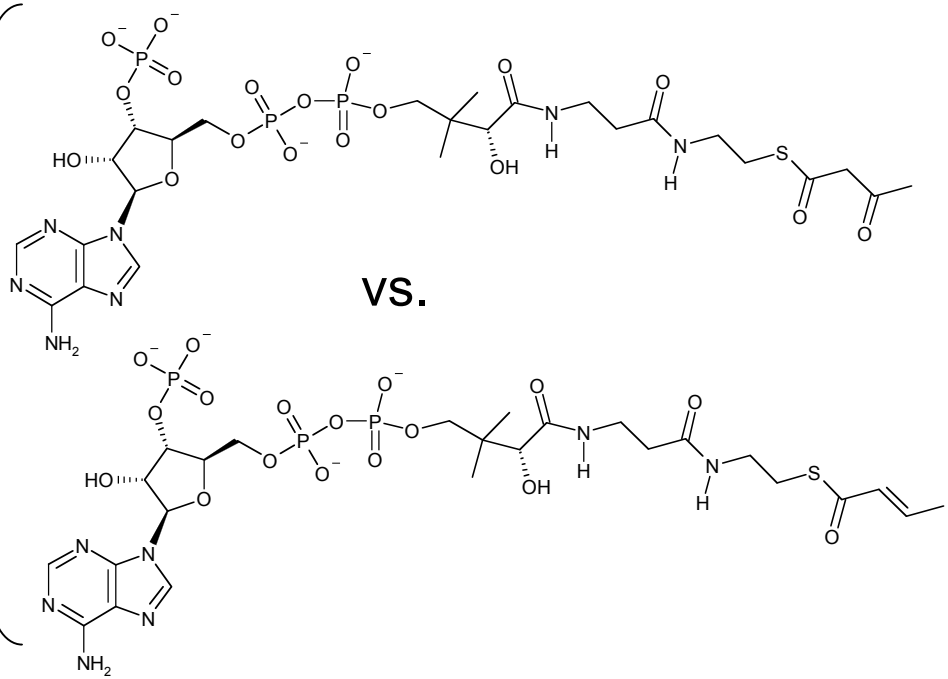


How is sequence similarity related to substrate similarity?

1edo (rapeseed beta-keto ACP reductase)
vs.
1c14 (E.coli enoyl reductase)

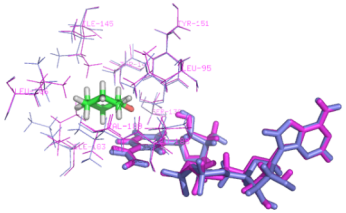
20.2% sequence identity

models of true substrates



vs.

0.95 similarity score



Why docking may be useful in function identification

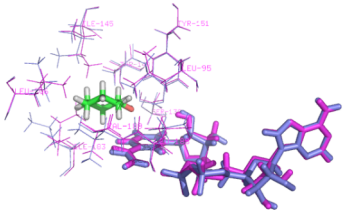
- Many structure-based methods take into account only part of the properties of the binding site (e.g. shape, volume etc)

- Proteins related by evolution may still bind each other's substrates, even if binding is not **productive**.

  can function be uniquely defined?
  is function context-independent?
  ….???

The best we can hope for is that:

The number of experiments required to characterise a protein could be reduced, if they could be guided by *in silico* experiments



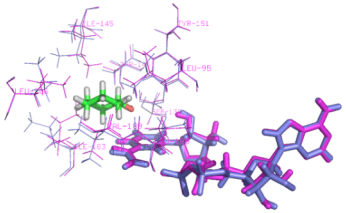
The logic behind our “experiments”

We assume that the substrates of a reaction catalysed by an enzyme must show reasonable binding affinity for that enzyme

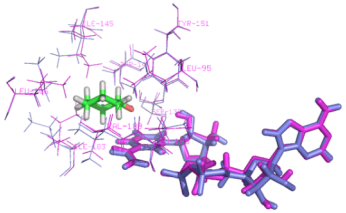
We look for good binders with reasonable binding modes and hence possible substrates using molecular docking

We assume that there is no need to search the whole of the chemical space! Only small molecules found as endogenous metabolites need to be checked

The number of molecules to be checked can, in theory, be further reduced using clustering

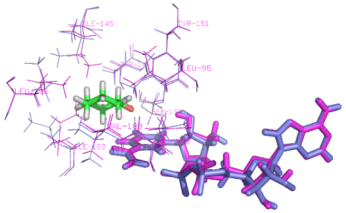


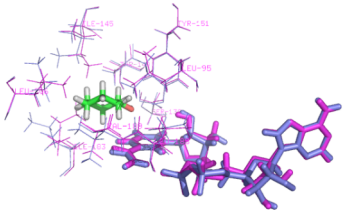
The family of
Short-chain Dehydrogenase/Reductases (SDRs)



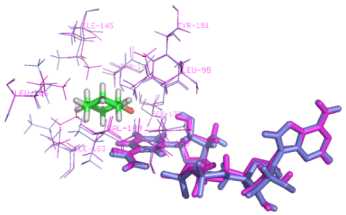
What do we know about SDRs?

- Present in all three domains of life.
- Over 2000 sequences have been deposited and ~200 crystal structures are in the PDB.
- 63 genes identified in the human genome in 2002 (same range as CYP P450).
- Typically one-domain NAD(P)(H)-dependent enzymes of ~250 aa.
- Mostly oxidoreductases, but lyases and isomerases are also known.
- Highly divergent family - recognised by typical sequence motifs
- Wide substrate spectrum including steroids, alcohols, sugars, aromatic compounds, and xenobiotics.

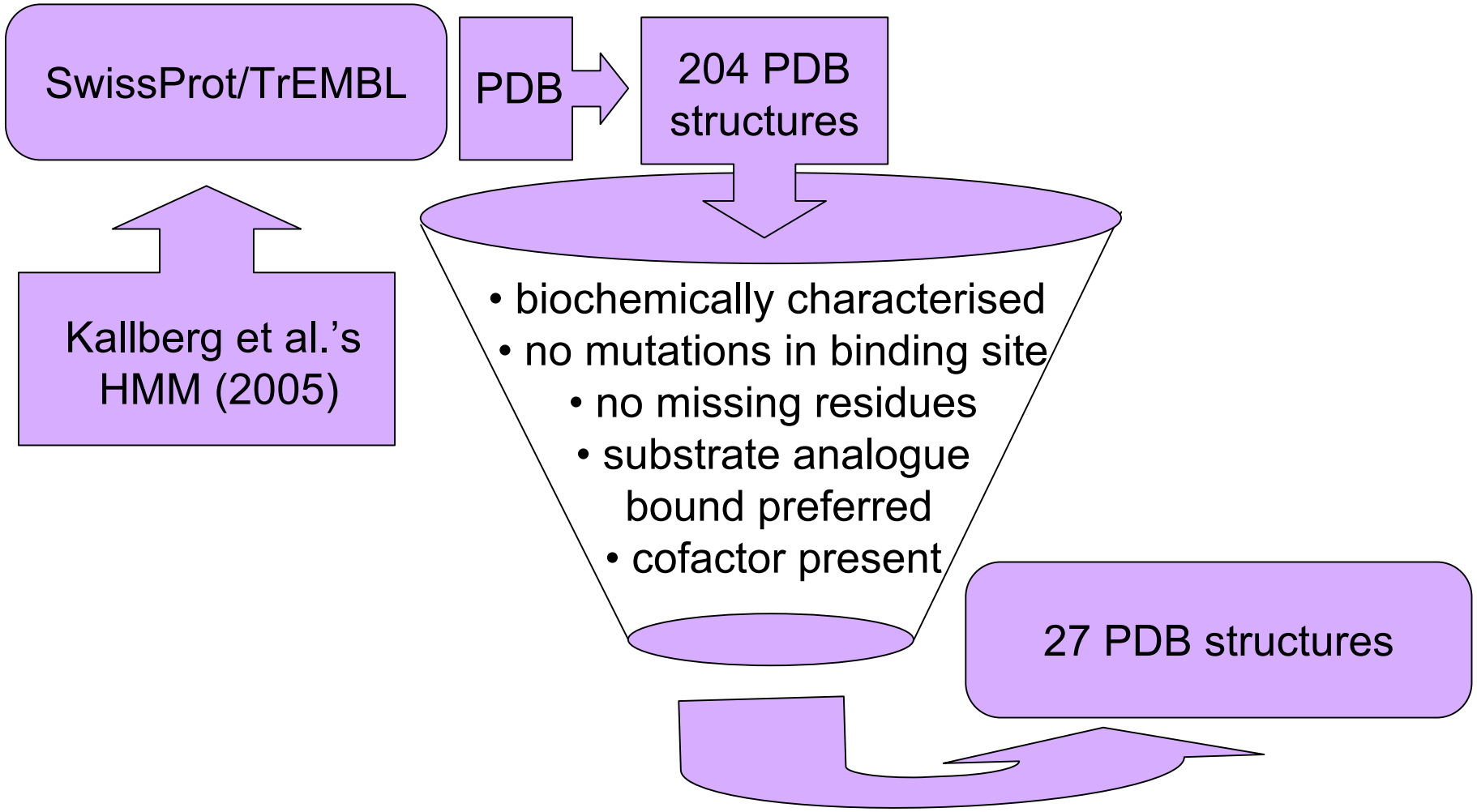


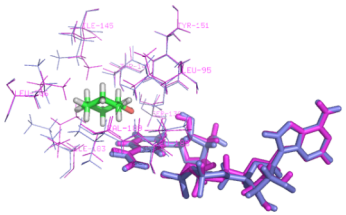


Our dataset of SDRs



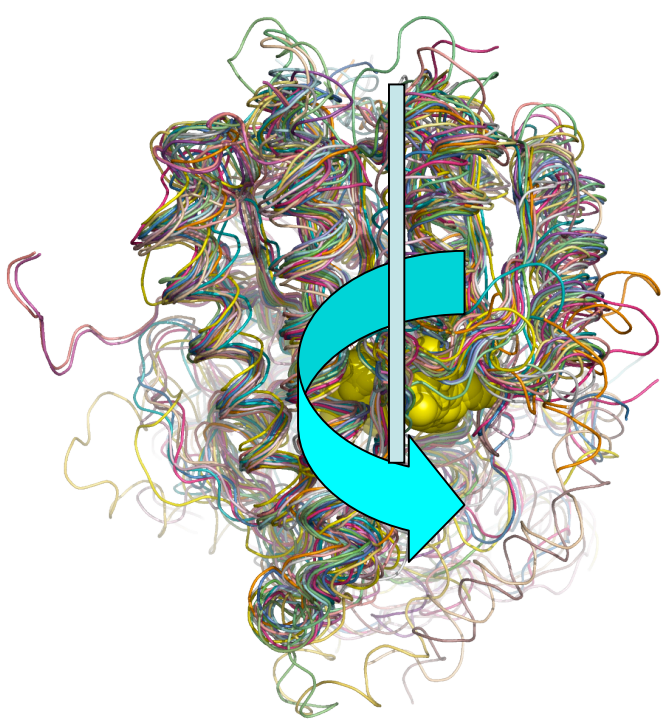
The dataset

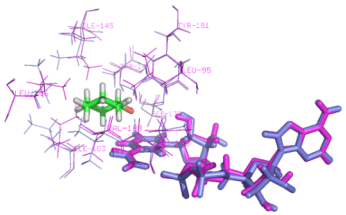

SwissProt/TrEMBL
PDB
204 PDB structures
Kallberg et al.'s HMM (2005)
• biochemically characterised
• no mutations in binding site
• no missing residues
• substrate analogue bound preferred
• cofactor present
27 PDB structures



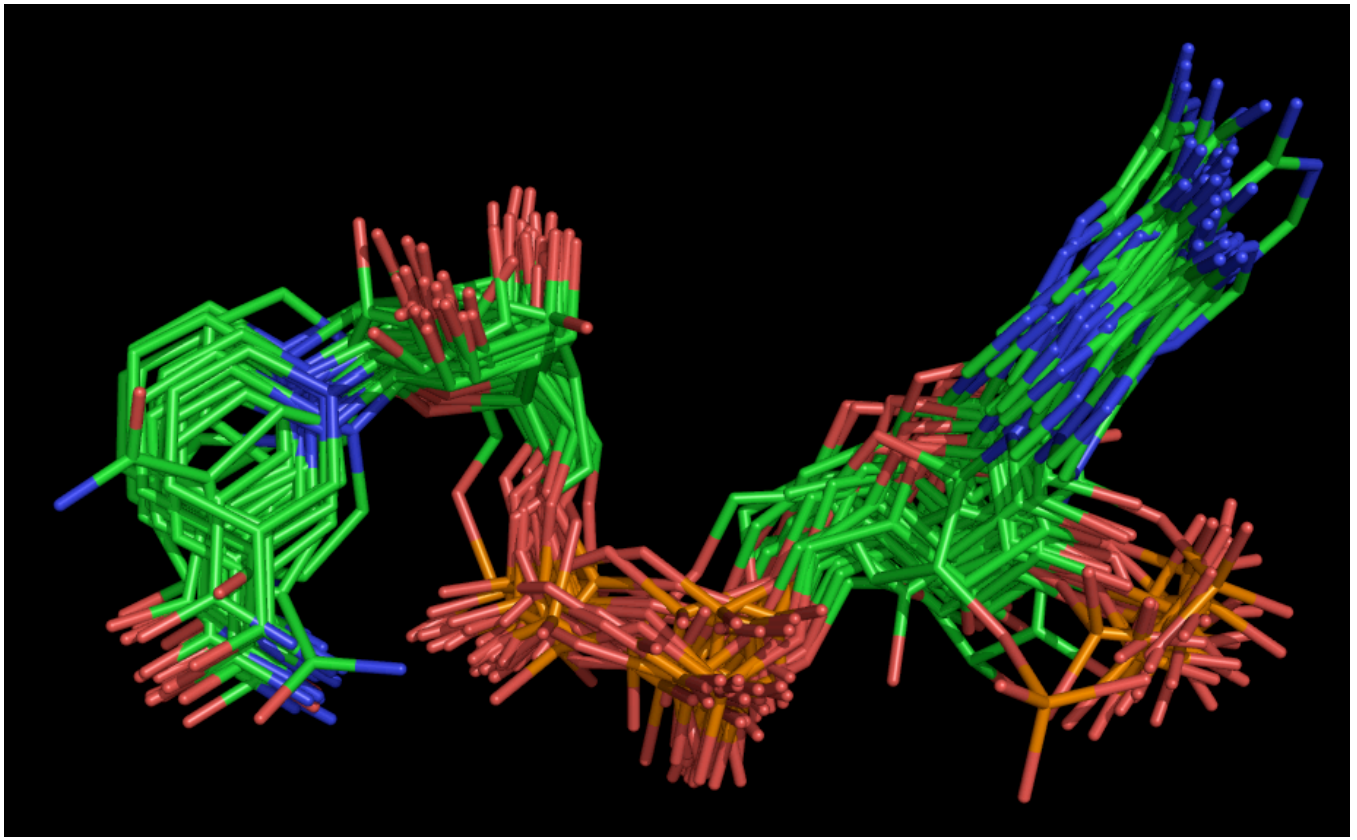
Variation in SDR structures

Significant variation in the SDR structure is observed primarily in the substrate-binding C-terminal domain

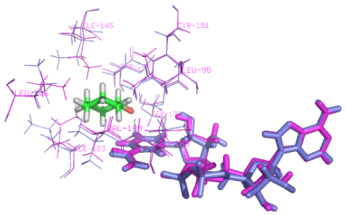




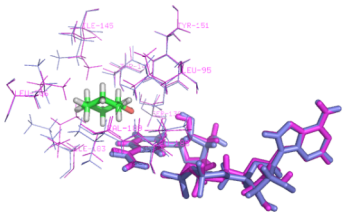
Variation in SDR structures



Superposition of the NAD(P)(H) cofactors in our 27 SDR structures

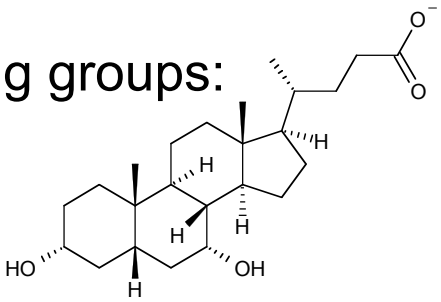


The substrates

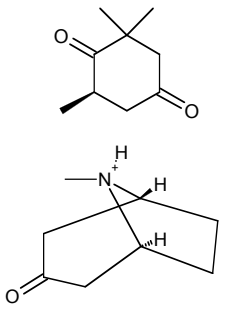


The substrates

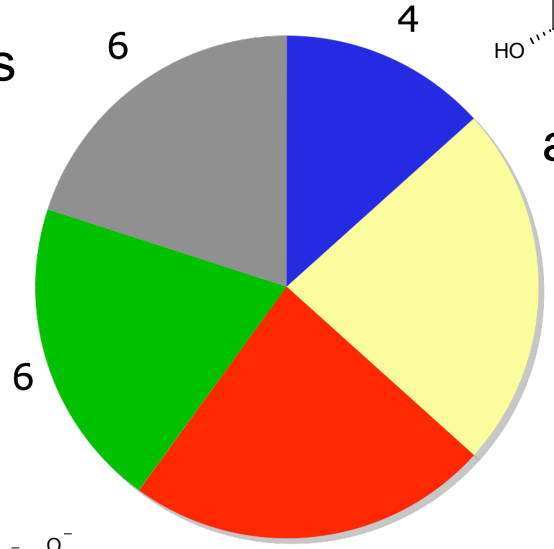
A visual inspection reveals the following groups:

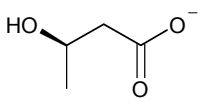


a) steroids & steroid-like

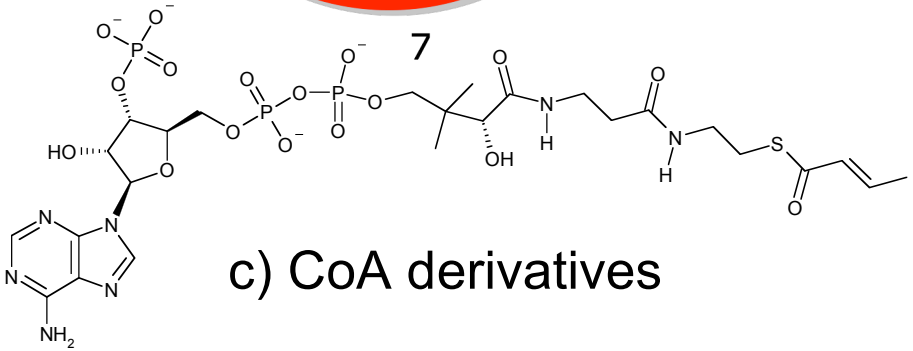


e) others



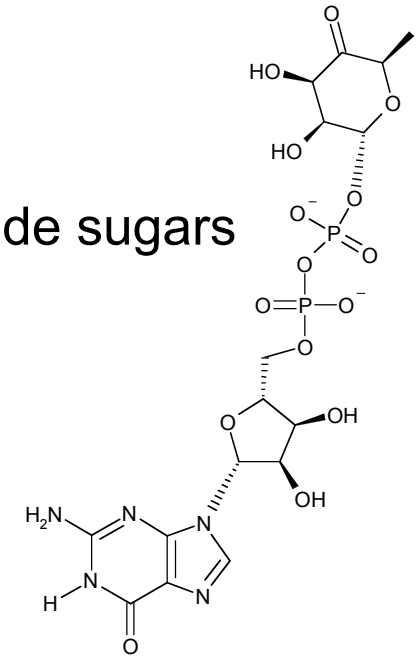


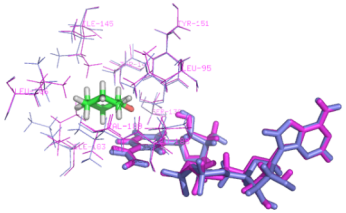
d) polar & small



c) CoA derivatives

b) nucleotide sugars



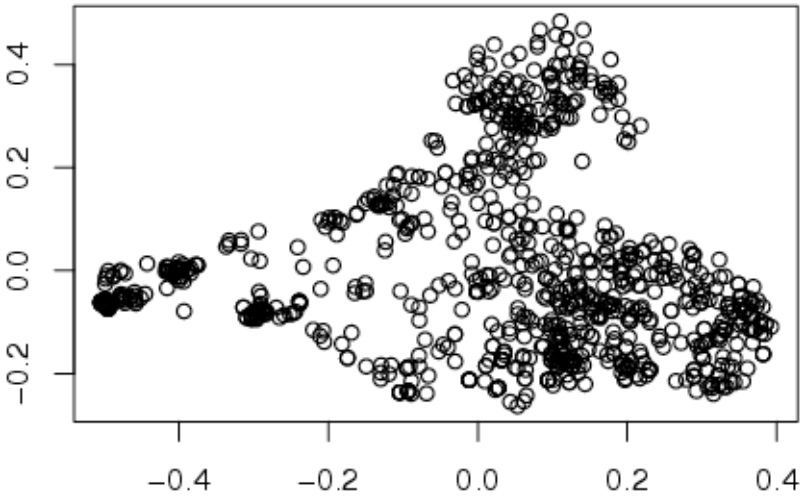


The docking dataset: the metabolite representatives

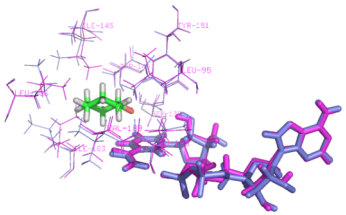
We use the human metabolome for virtual screening

The human metabolome forms a continuum in structural space

Similarity based on hashed fingerprints from CDK

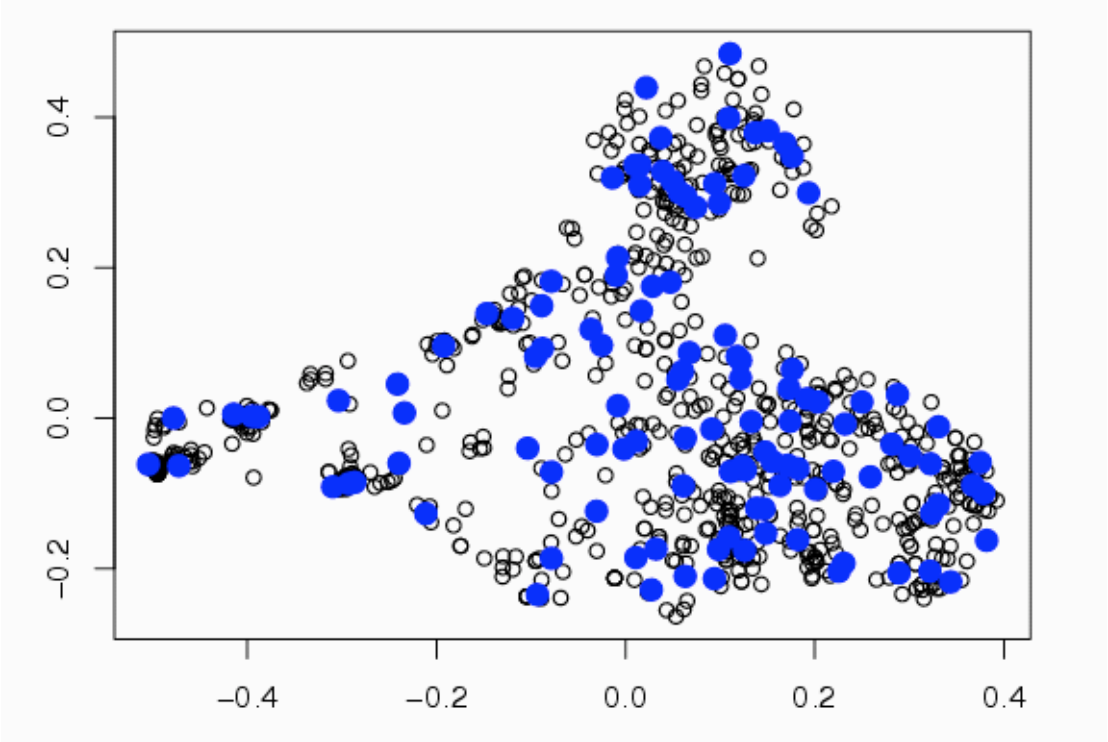


Multidimensional scaling plot (2D) of 931 human metabolites

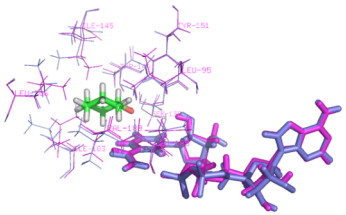


The docking dataset: the metabolite representatives

Clustering reduces significantly the substrate space



Medoids of 115 clusters shown on top of all 931 human metabolites
(using R's partitioning around medoids method)



The docking experiments

27 SDR proteins from the PDB X

61 substrates/products

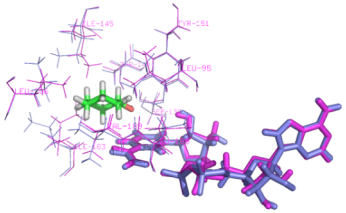
922 human metabolites

115 human metabolite reps

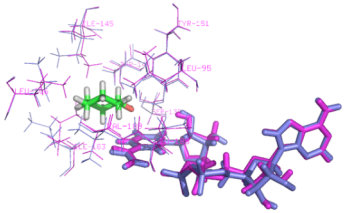
All KEGG ligands (~18K)

We used:

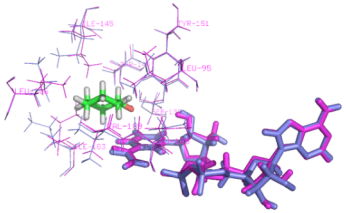
Glide (with two different ways of scoring) and Autodock
Rigid protein - flexible ligand docking
Constraints and filtering to dock the ligands near the NAD C4 atom



Results



How well do we predict the expected binding mode?

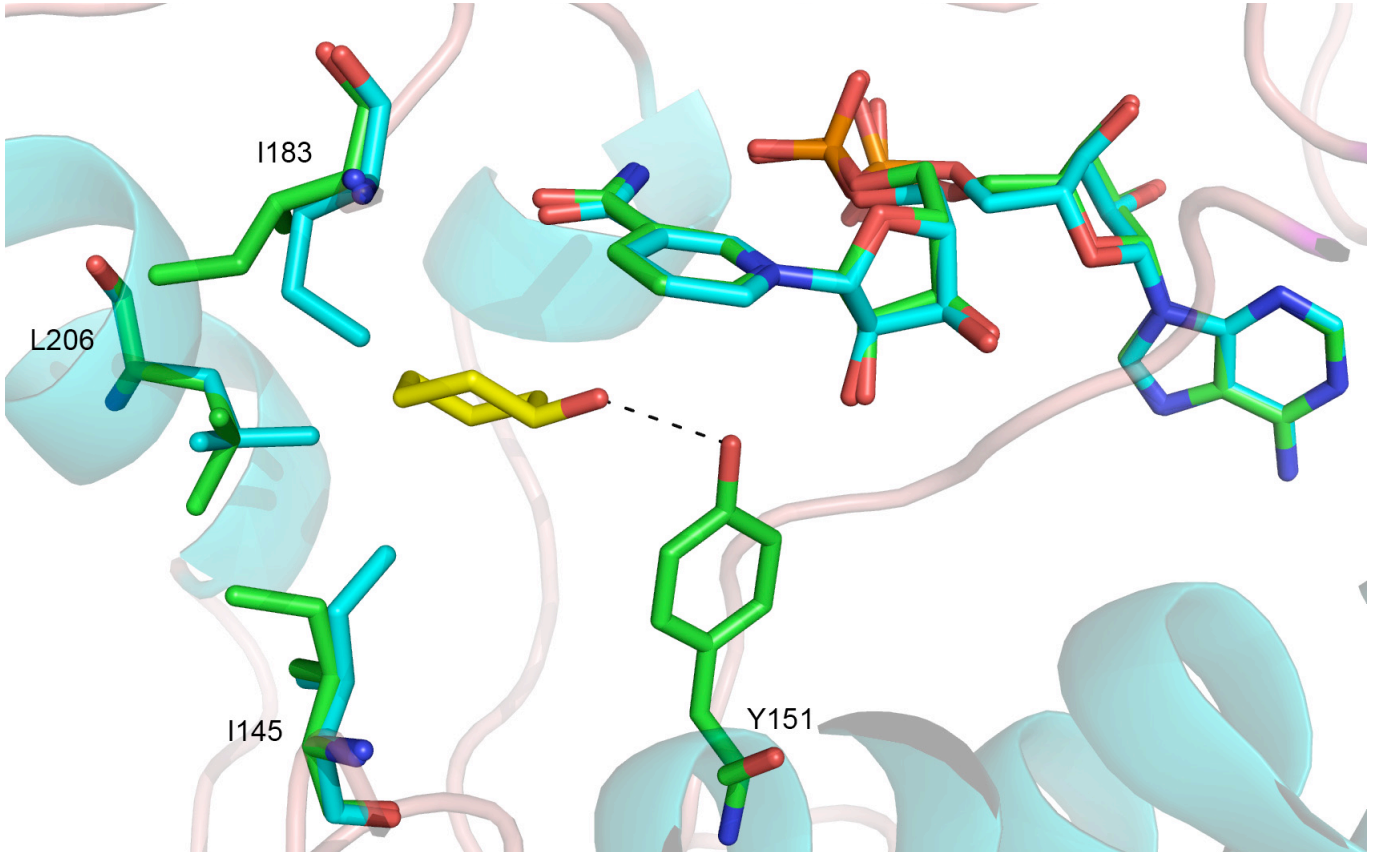


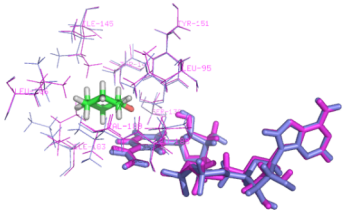
How well do we predict the expected binding mode?

In all cases but one we find at least one binding mode that passes our distance constraint to the NAD C4 atom.

Cyclohexanol docked in Drosophila alcohol dehydrogenase (1b14)

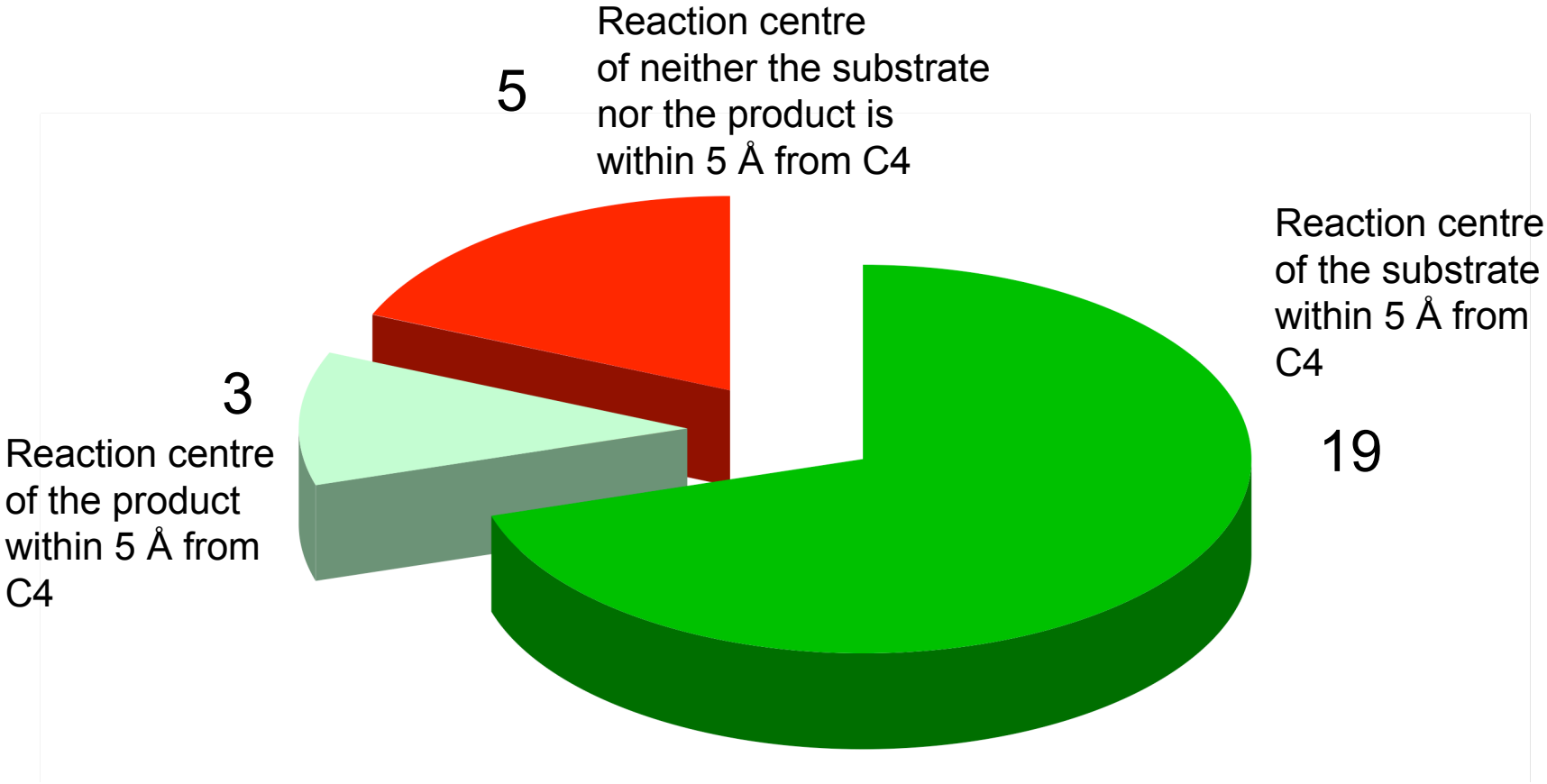
Binding site before (blue) and after (green) application of the Induced Fit Docking protocol

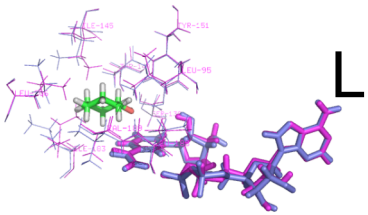




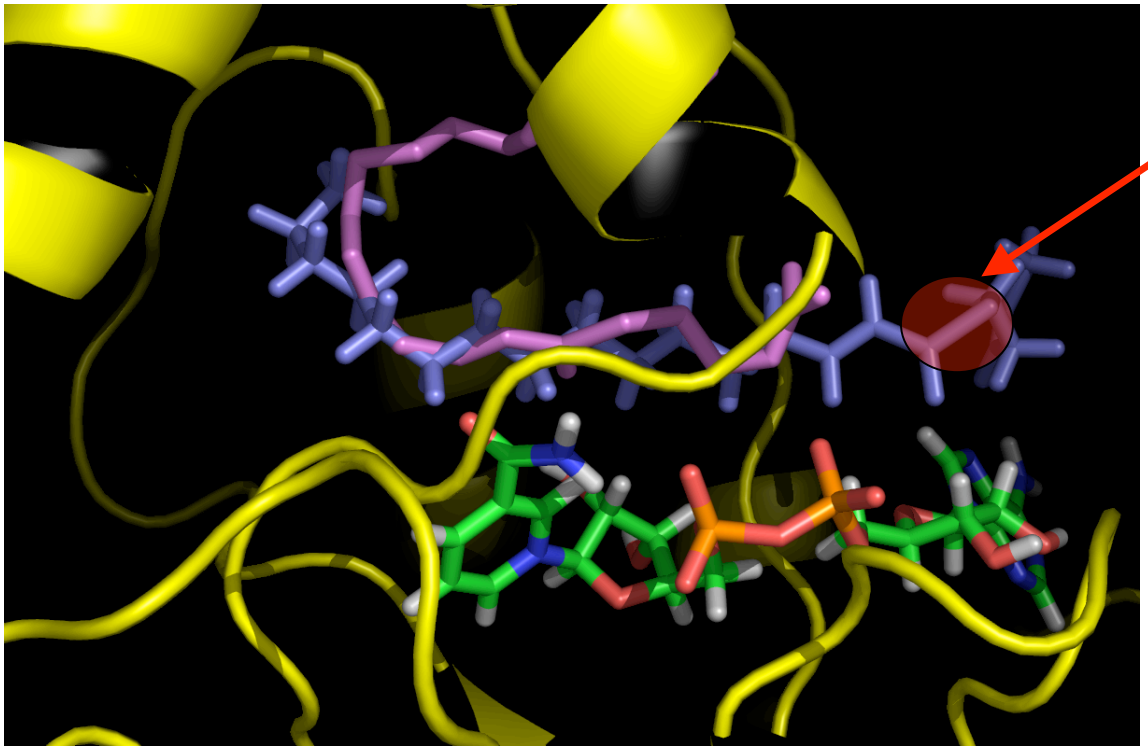
How well do we predict the expected binding mode?

In most cases we can also predict a reasonable orientation for the top-ranking pose of the substrate/product.

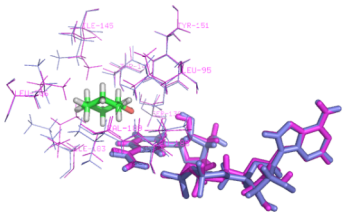




Large and flexible substrates are not usually docking-friendly…

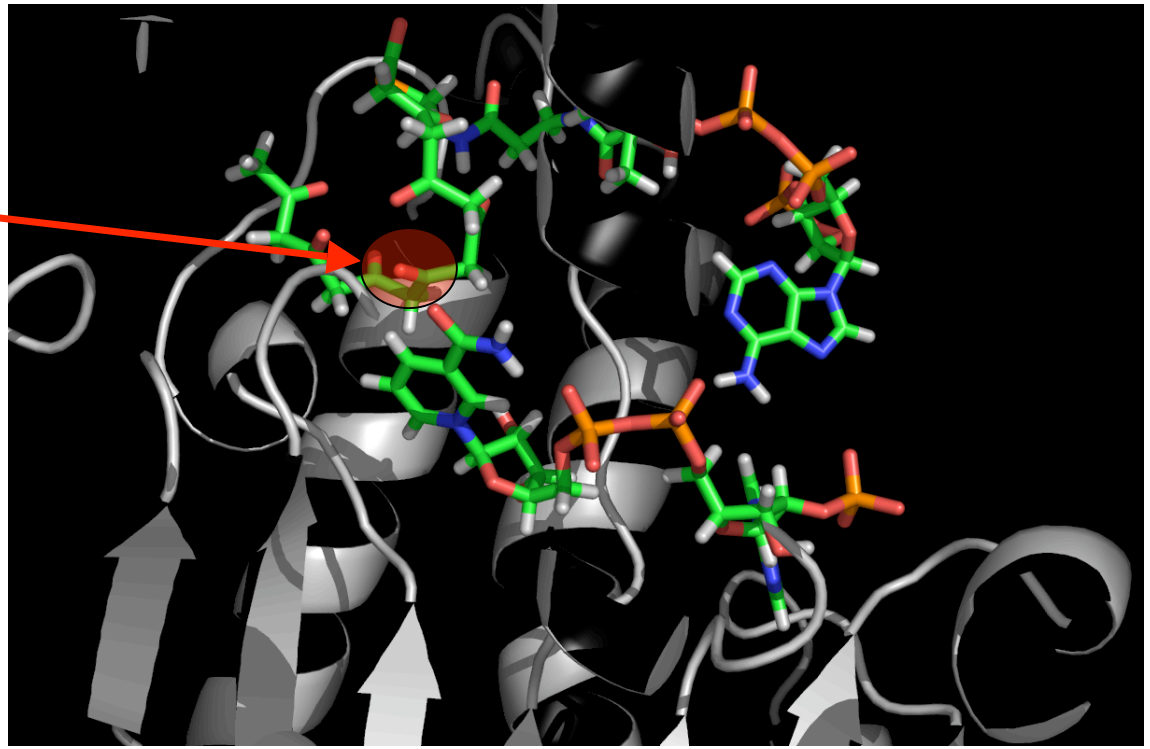


Enoyl-ACP reductase from *M. tuberculosis* in complex with a fatty acyl substrate
Slate blue: protein substrate
Pink: PDB ligand from 1bvr

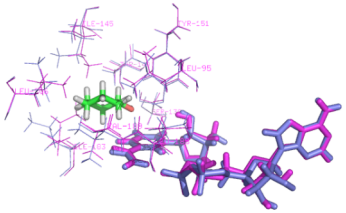


…unless you get very lucky!

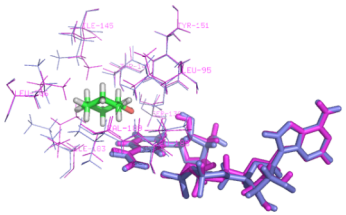
reaction centre



Actinorhodin polyketide reductase from *S. coelicolor* (1w4z)
Substrate docked in expected position
Only PDB ligand present was formic acid

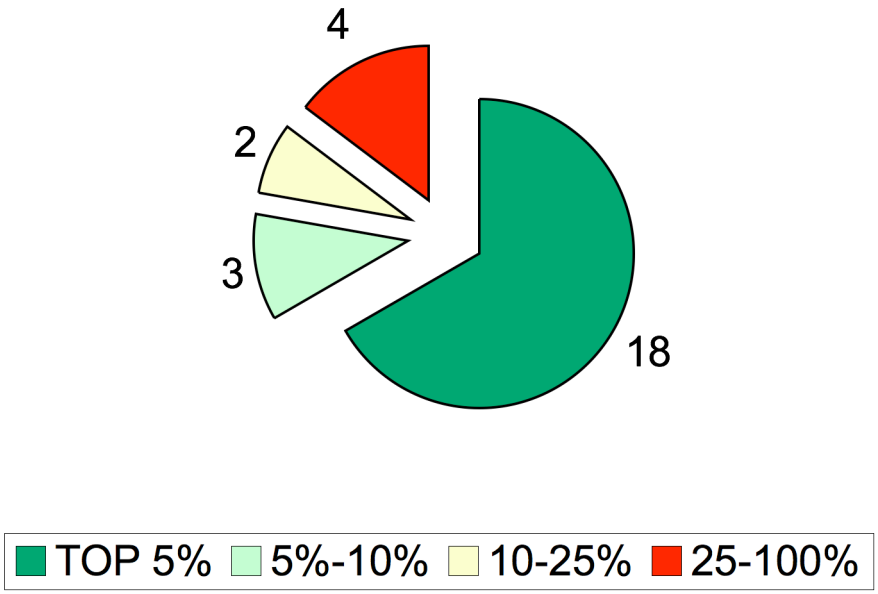


How do the substrates score in a pool of metabolites?

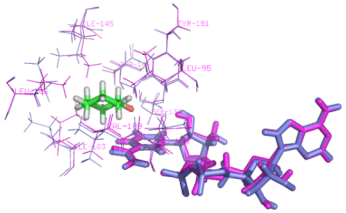


How do the substrates score in a pool of metabolites?

Rank of substrate/product among 922 ligands:

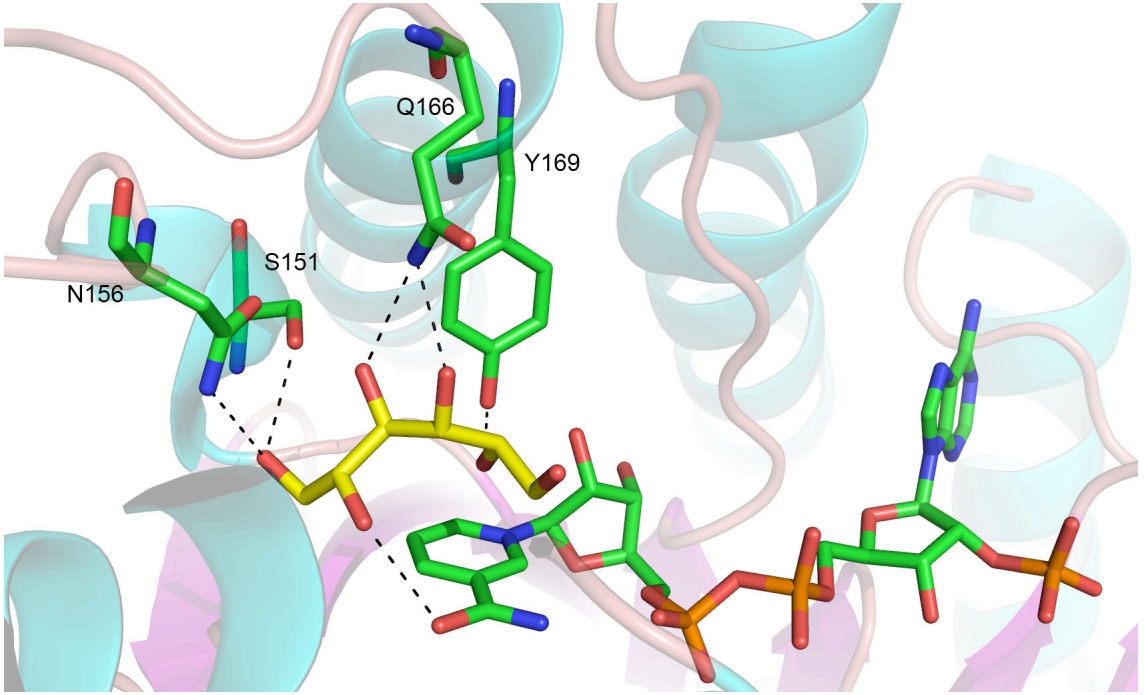


3/4 of the time we find the substrate in the top 10% of energy scores

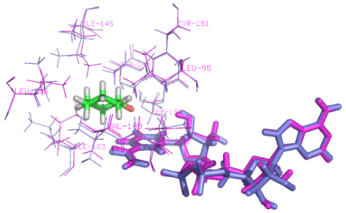


How do the substrates score in a pool of metabolites?

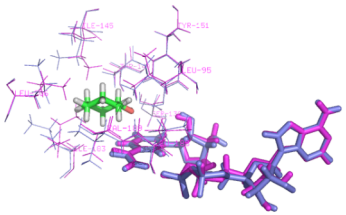
Some of the “failures” can be easily explained, e.g. mannitol docked to mannitol dehydrogenase from common mushroom:



K_M for mannitol is very high (~ 29 mM) - crystallographers have failed to obtain a structure of the complex



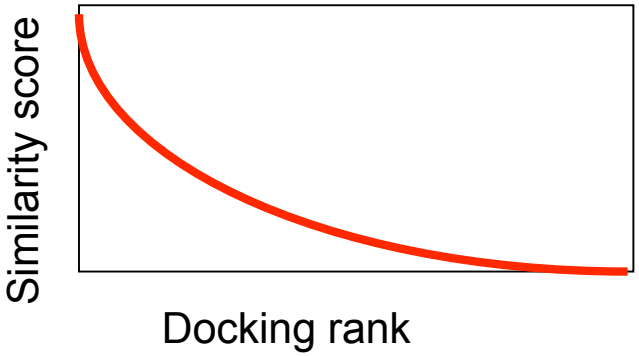
Do top-ranking ligands resemble the substrate?

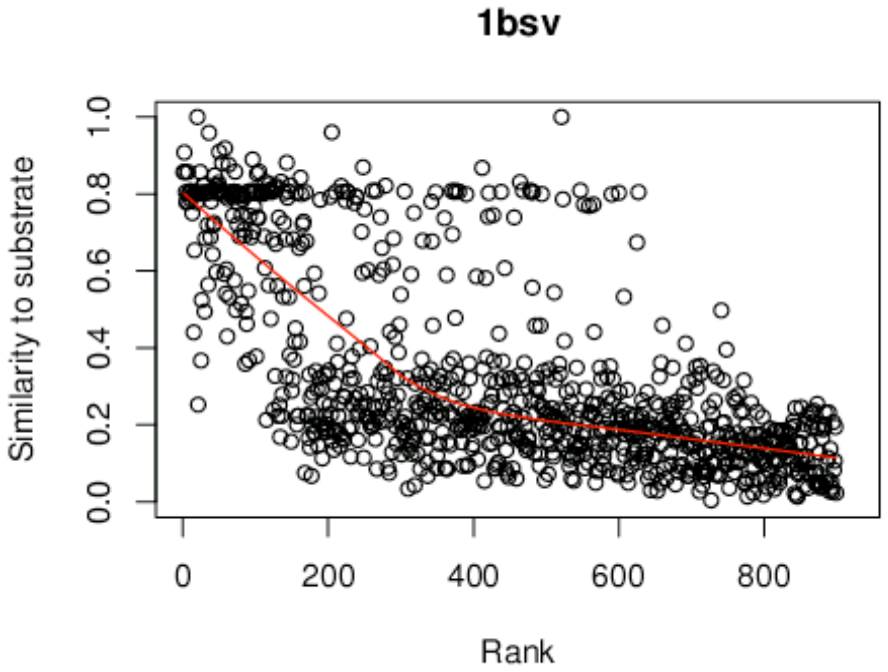


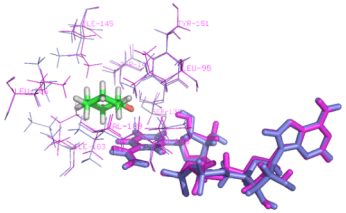
Do top-ranking ligands resemble the substrate?

A strong similarity of top-ranking ligands to substrates was observed in other studies

We see a trend but not a strong correlation - results are very much protein-dependent

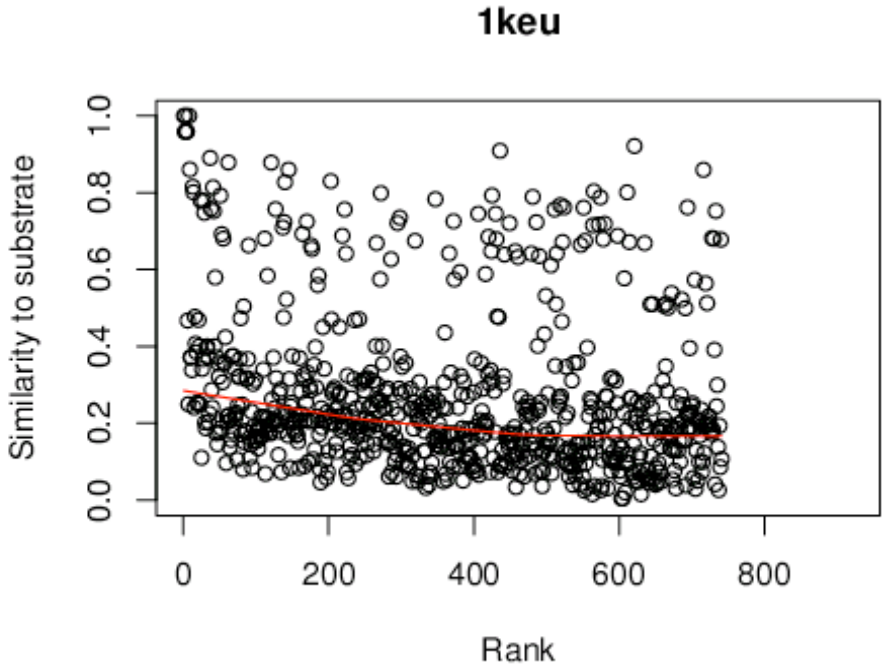




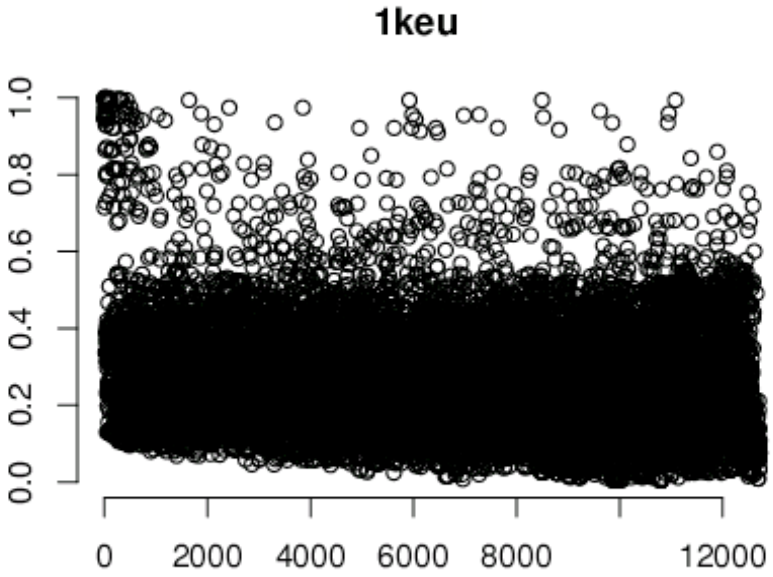


Do top-ranking ligands resemble the substrate?

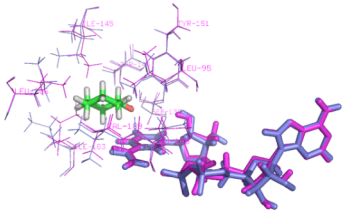
No big difference observed between docking
only representatives, only human metabolites or the whole of KEGG



Results from docking 922 human metabolites

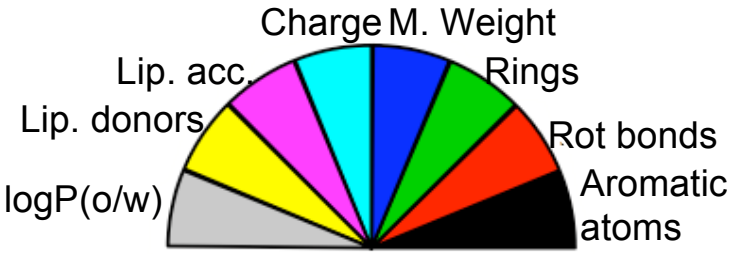


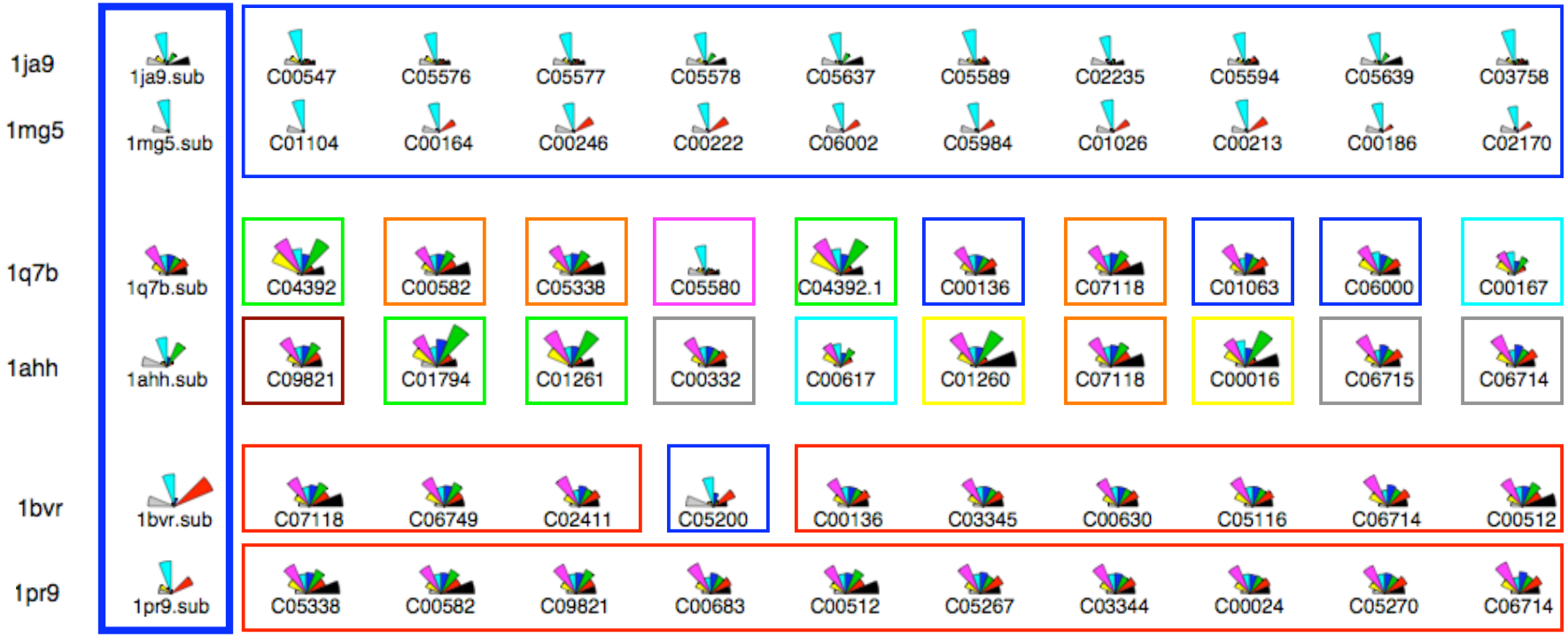
Results from docking the whole of KEGG

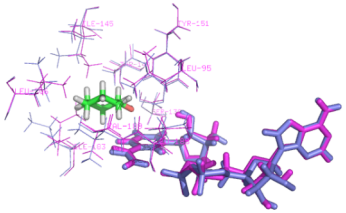


Do top-ranking ligands resemble the substrate?

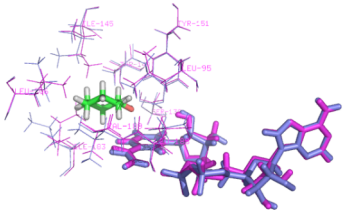
Comparison of physicochemical descriptors for top-ranking ligands and the substrate





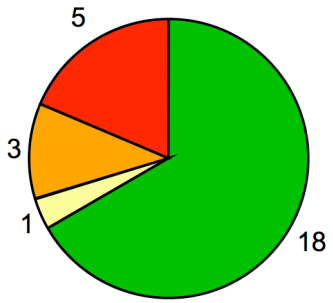


How do different docking/scoring approaches compare?

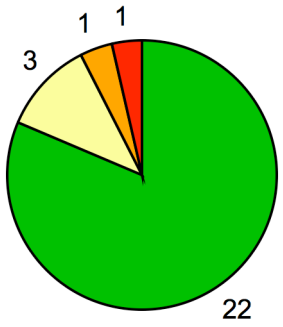


How do different docking/scoring approaches compare?

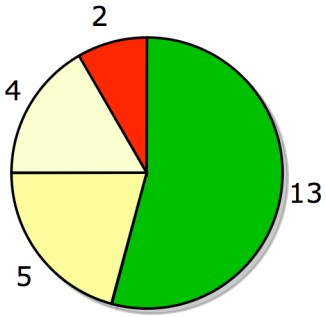
Ranking of substrate/product in a pool of 176 ligands:



Autodock
(docking + scoring)

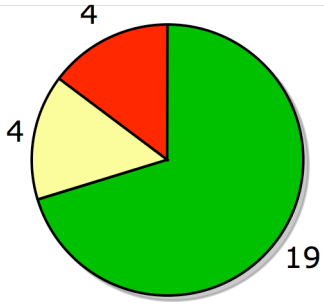


Glide
(docking + SP scoring)



Glide + Prime
(docking and
rescoring -
no protein min)

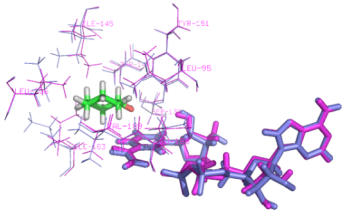
$$E_{binding} = E_{complex} - E_{protein} - E_{ligand}$$



Glide + Prime
(docking and
rescoring -
with protein min)

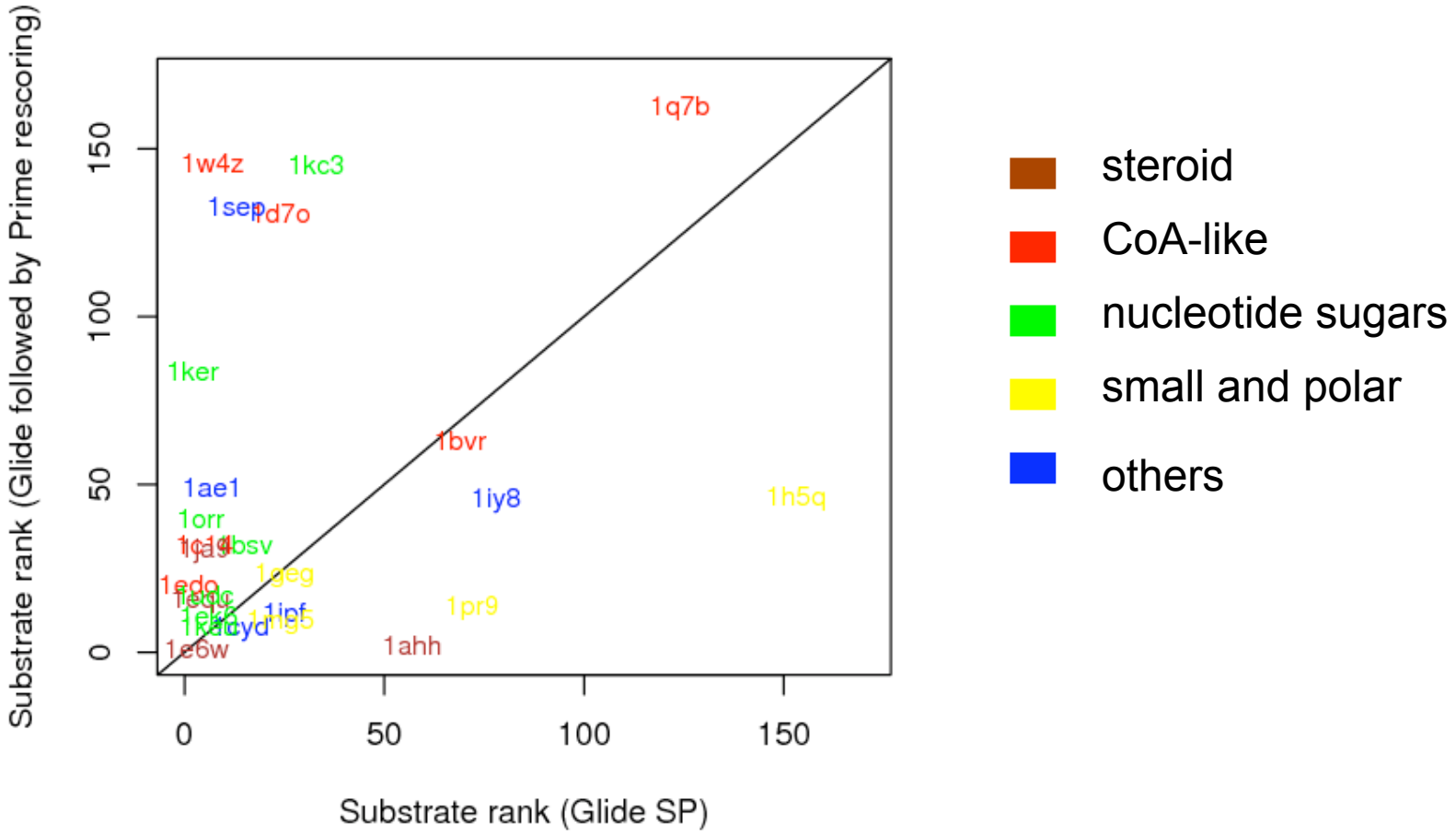
$$E_{binding} = E_{complex} - E_{protein} - E_{ligand}$$

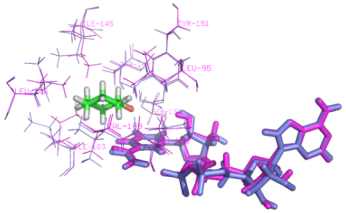
Top 25% 25 -50% 50-75% 75-100%



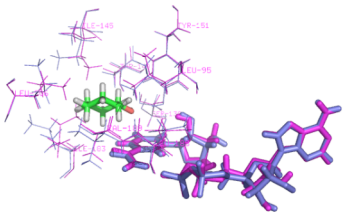
How do different docking/scoring approaches compare?

Different scoring functions perform differently for each complex:





Can we benefit from using cluster representatives?



Can we benefit from using cluster reps?

The basic idea:

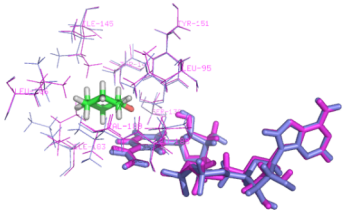
If the function of the protein is unknown, there are too many compounds to try out!

In vitro binding experiments can be expensive

Accurate computational studies are time-consuming

In both cases, we want to exclude as much of the chemical space as possible

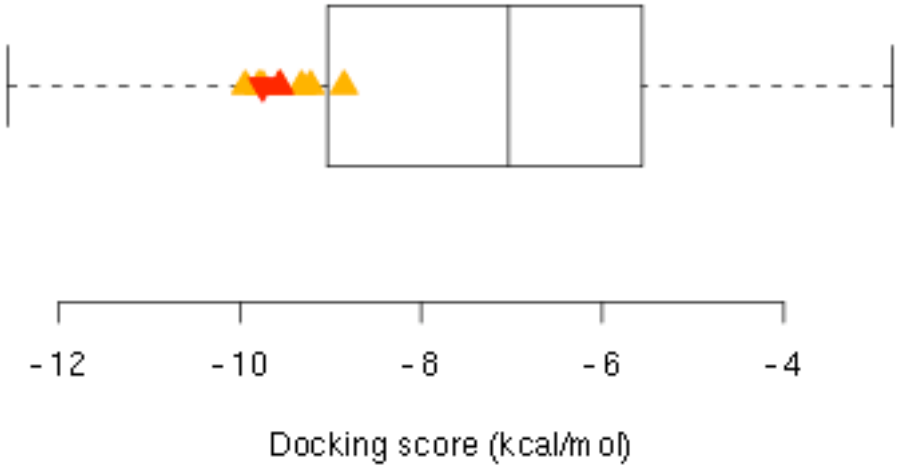
If we can identify classes of compounds that can be excluded, then we can save both time and money



The assumption: Members of the same family should score similarly

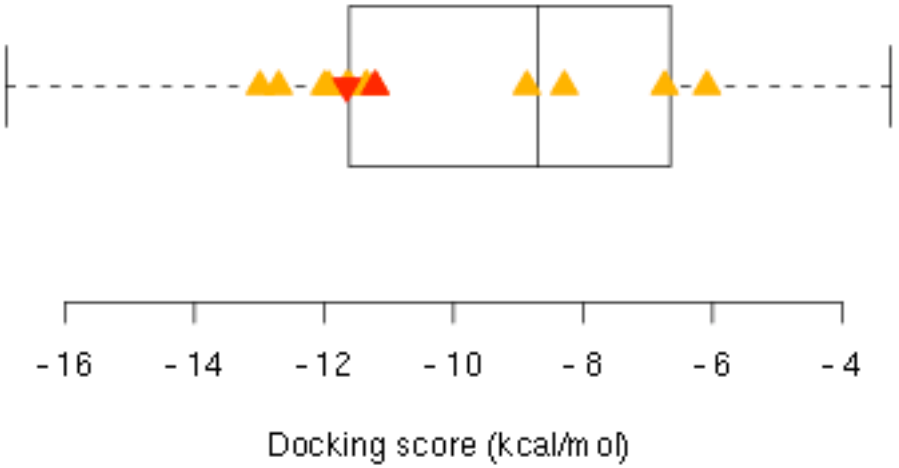
1equ

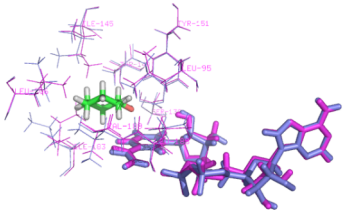
Mean dissimilarity in cluster = 0.18



1cyd

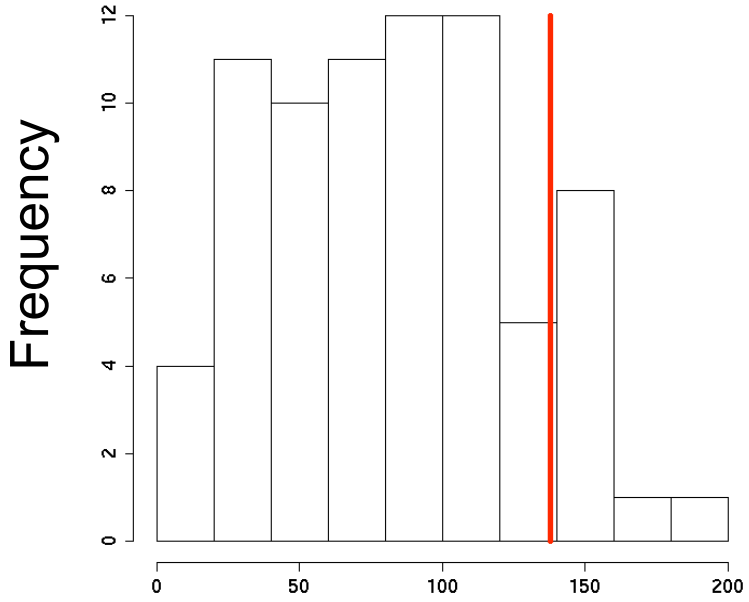
Mean dissimilarity in cluster = 0.34



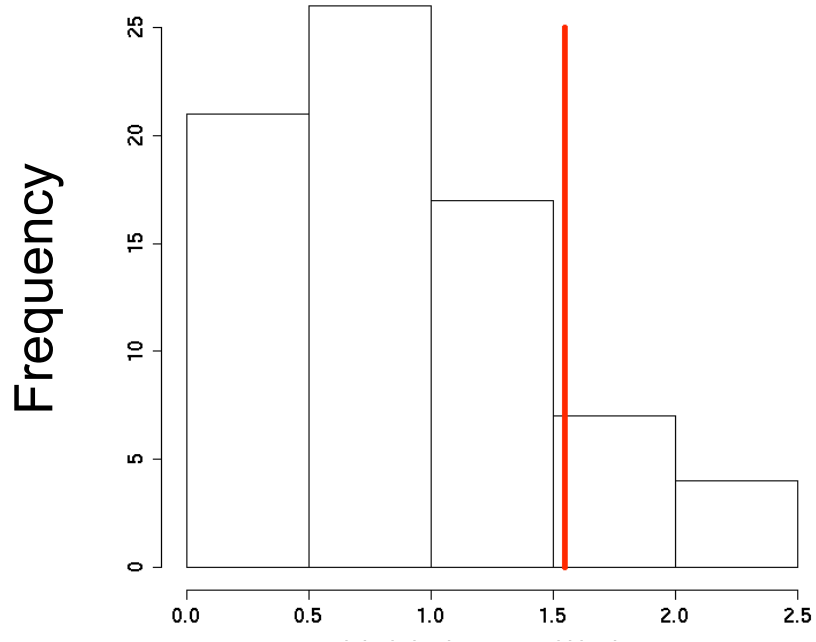


The reality: deviation of energy and ranks within a cluster

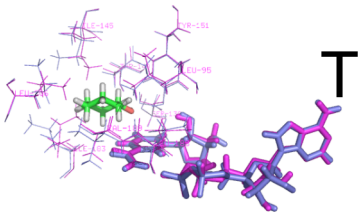
1ae1



Std dev in rank within each cluster

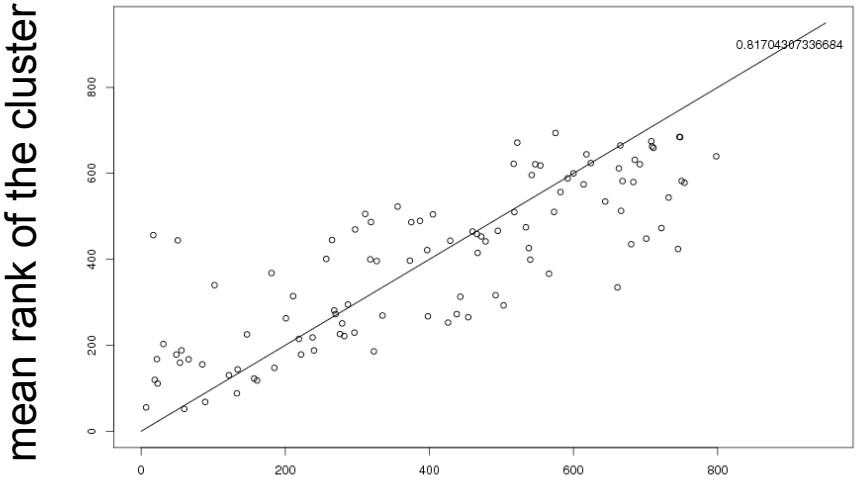


Std dev in energy within each cluster

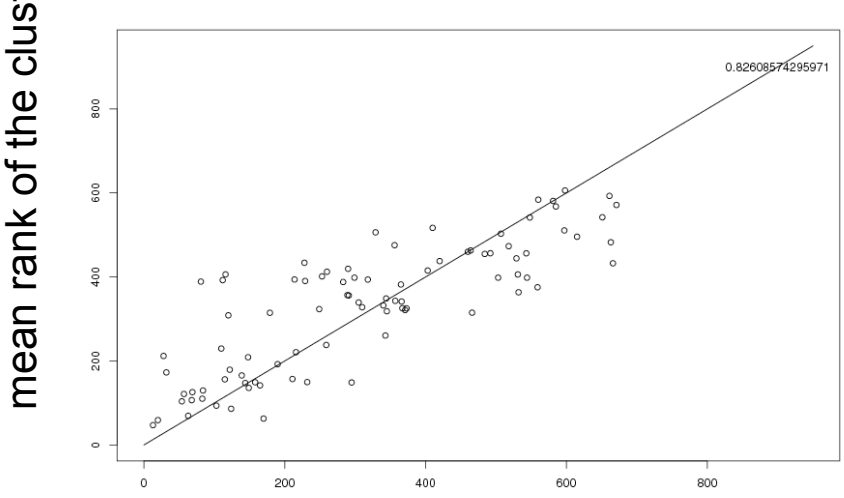


The medoids are good representatives of the average performance of a cluster

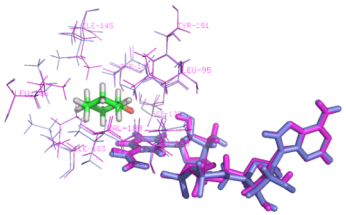
1equ



1ipf



The correlation between the cluster medoid rank and the cluster mean rank is high, regardless of the protein



Conclusions

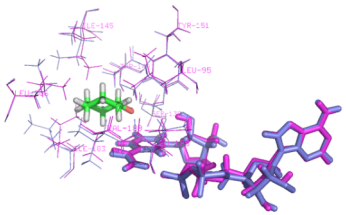
Docking can be a useful tool in elucidating protein function

In the case of SDRs, docking reproduced in most cases realistic binding modes for the natural substrates

…and, most of the time, the substrate was ranked near the top 10% of the distribution of scores

More accurate simulations are eventually needed - use of cluster representatives might make these feasible!

Metabolite docking profiles carry information about the binding site of a protein, and they should be explored



People & Funding sources

- Angelo Favia & Janet Thornton (EBI)
- Fabian Glaser
- Rafael Najmanovich, Eric Blanc, Tim Massingham & Gareth Stockwell (EBI)

and for funding:

- The Medical Research Council and the Royal Society (I.N.)
- The EU for a Marie Curie studentship (A.F.)
- The Structural Genomics Consortium (F.G. & A.F.)