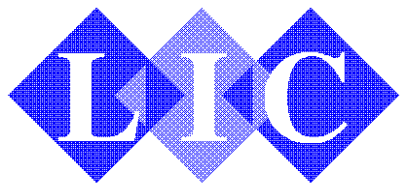
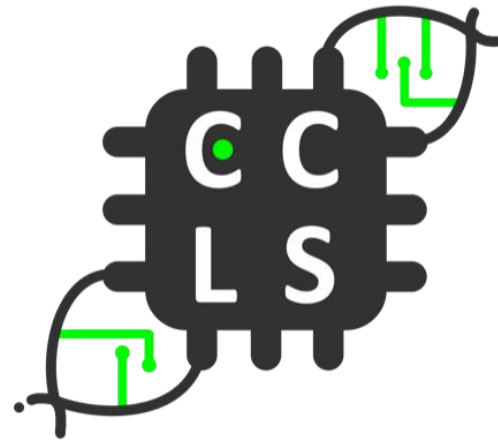


- *CCLS Matchmaking Event*



Leiden Institute of Chemistry



INSTITUTE OF BIOLOGY LEIDEN



Leiden Institute of
Advanced
Computer
Science

About the CCLS

- We started out as the Data-Drive Drug Discovery Network (D4N)
 - Active between 2016 – 2019
 - Focussed on Drug Discovery and Informatics
- Mostly collaborating in student supervision and ad hoc projects
- We realized that there is more potential...

- There is a lot of (scattered) expertise in the BioScience park within LU/LUMC
- Multiple initiatives in parallel
 - Data-Driven Drug Discovery Network
 - Computational Hub
- ...And also the Leiden Center for Data Science (LCDS)...

- In June 2018 we founded the CCLS
 - Institute of Biology Leiden (IBL)
 - Leiden Academic Centre for Drug Research (LACDR),
 - Leiden Institute of Advanced Computer Sciences (LIACS)
 - Leiden Institute for Chemistry (LIC)
 - Leiden University Medical Centre (LUMC).
 - Mathematical Institute (MI)

CCLS so far

- About 8 events a year...
 - Tuesday Seminars
 - Summer events
 - Matchmaking events



Projects

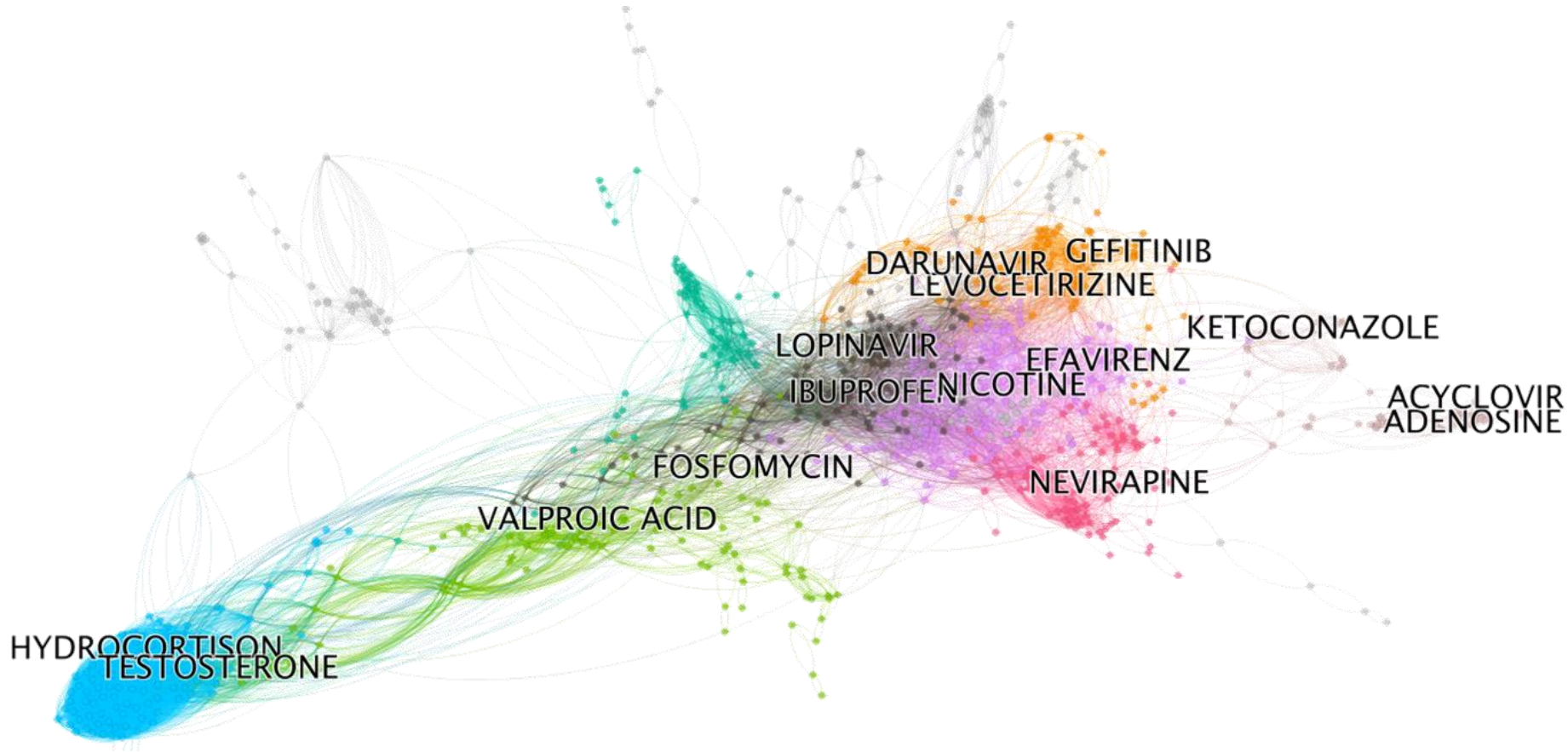
- Several flagship projects
 - Machine learning and drug discovery (LACDR / LIC / LIACS)
 - Machine learning based retrosynthesis (LACDR / LIC / LIACS)
 - Image based machine learning (LIACS / LACDR)
- One public private partnership
 - EXPLORE (NOW funded LIFT) collaboration with Galapagos

Goal for today...

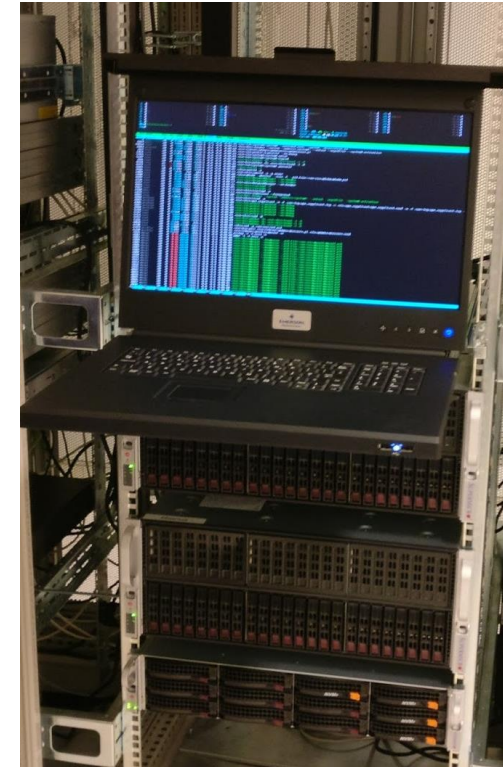
- Meet, interact, get to know each other

Computational Drug Discovery

CCLS Matchmaking Event



Gerard JP van Westen



LACDR

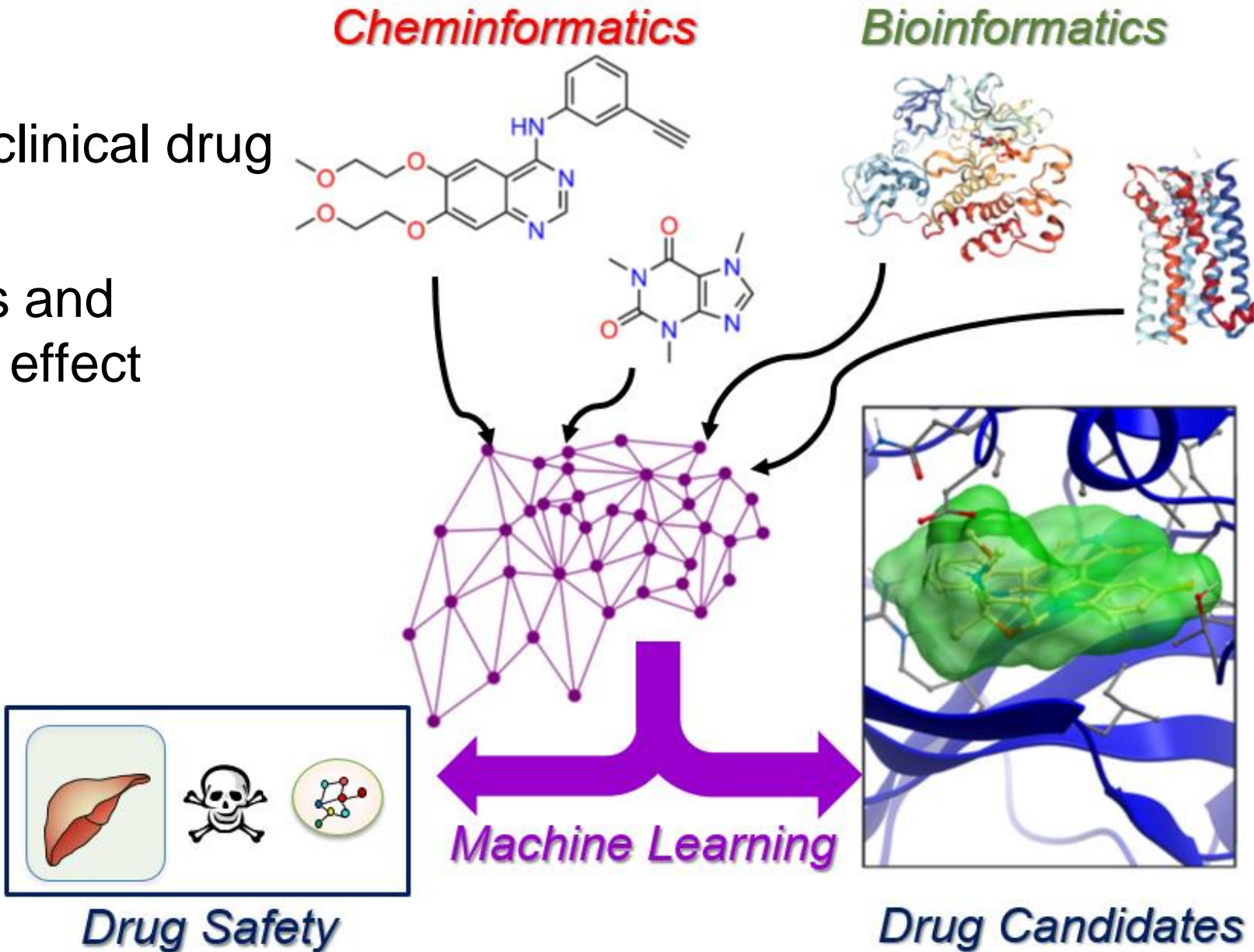


Applied and Engineering Sciences



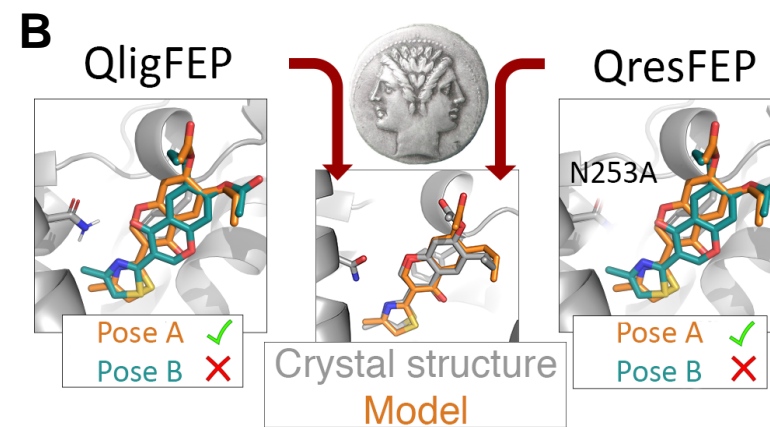
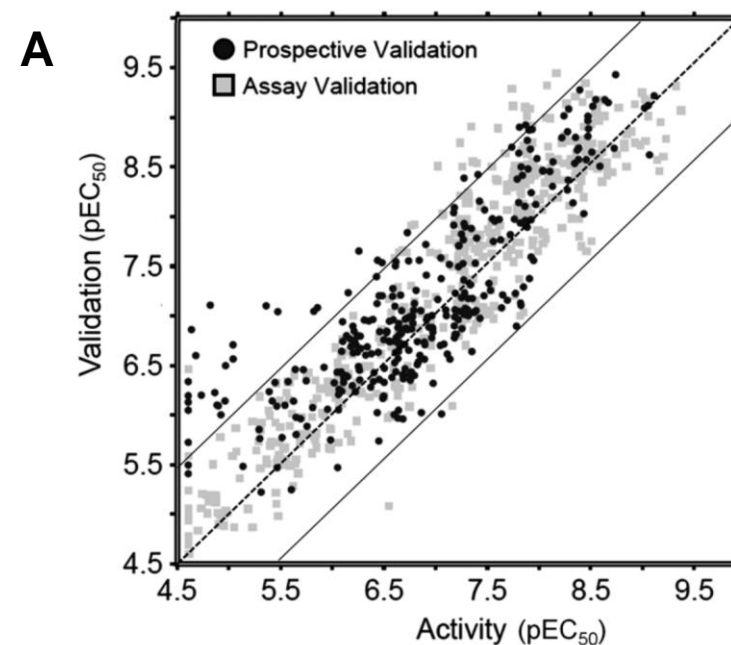
Computational Drug Design

- Artificial Intelligence in pre-clinical drug discovery
- Combining cheminformatics and bioinformatics for biological effect prediction



In general two flavors of computational drug design

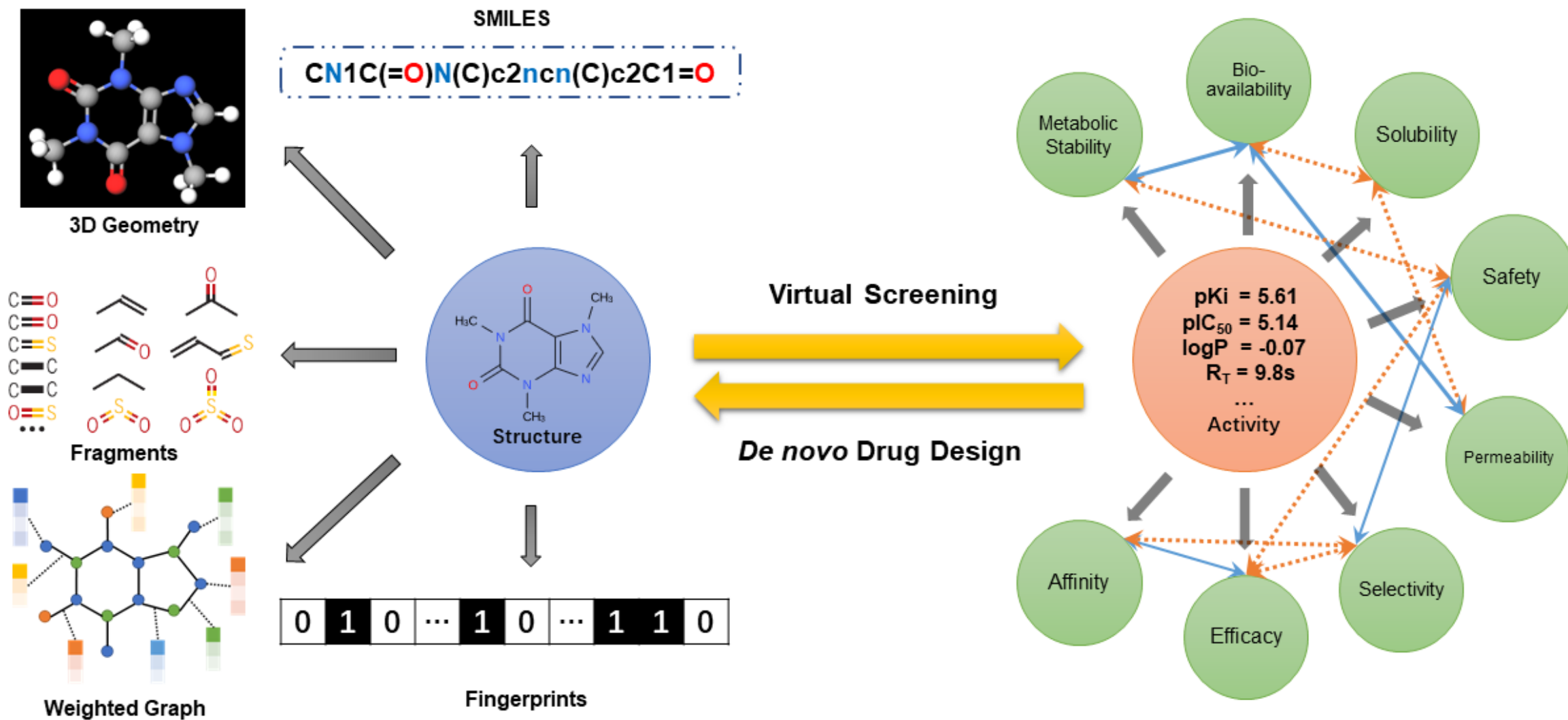
- Ligand based methods
 - Quantitative Structure-Activity Relationship (QSAR)
 - Artificial Intelligence
 - Property prediction (2d chemical structures)
 - *de novo* chemical structure generation
- Structure-based methods
 - Docking and scoring
 - Artificial Intelligence
 - 3D protein structure generation
 - Trajectory analysis



[A] van Westen, G.J.P., et al. (2011), PLoS ONE, 6 (11), e27518

[B] Jespers, W et al. (2020). Ang. Chem. Int. Ed. 59:16536-16543

AI approaches in a ligand based world..



What can I bring to CCLS?

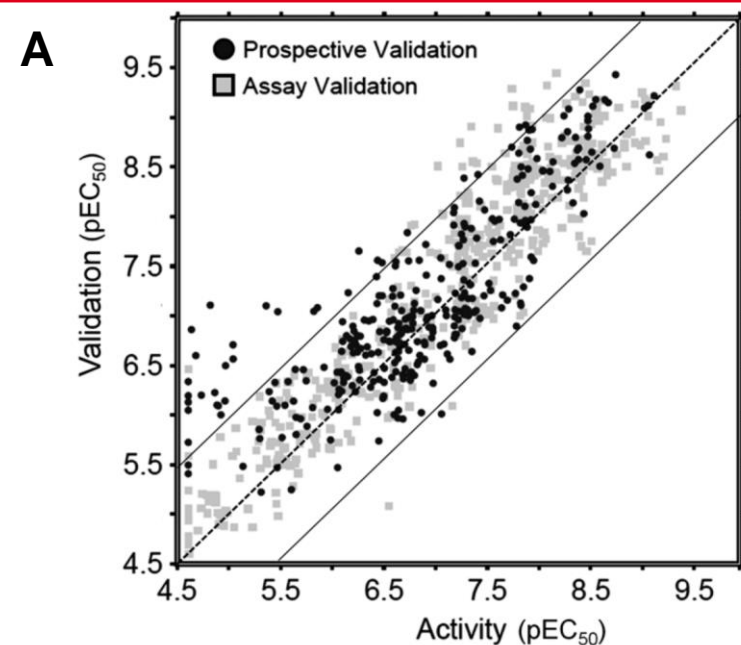
- Expertise in machine learning applied to chemical structures
- Biological data to learn from
- Experimental validation of novel algorithms applied to chemical data

What would I like from CCLS

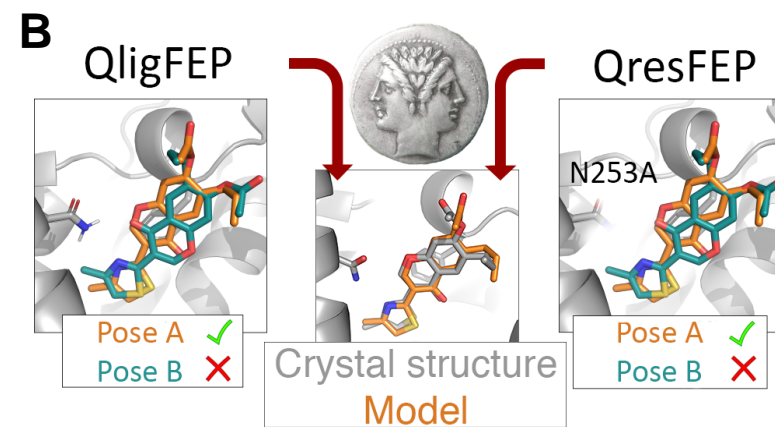
- Novel algorithms to apply to my data
- Critical feedback on computational design of experiment
- Expertise in scaling up or scaling our calculations

In general two flavors of computational drug design

- Ligand based methods
 - Quantitative Structure-Activity Relationship (QSAR)
 - Artificial Intelligence
 - Property prediction (2d chemical structures)
 - *de novo* chemical structure generation

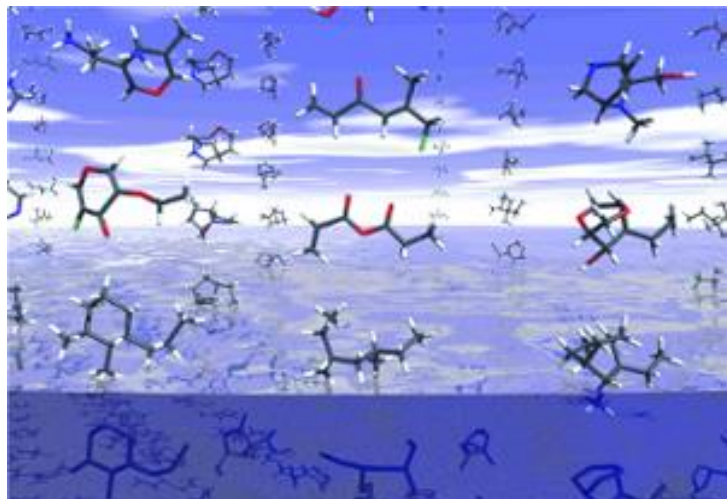


- Structure-based methods
 - Docking and scoring
 - Artificial Intelligence
 - 3D protein structure generation
 - Trajectory analysis



Chemical Space

- Typically $10^1 - 10^2$ molecules are made in a drug discovery project
 - $\sim 10^8$ molecules have been synthesized (CAS, Sept 2020)
 - $\sim 10^{33} - 10^{60}$ Lipinski drug like molecules estimated [1-3]
 - For molecules up to 36 heavy atoms...



1,000,000,000,000,000,000,000,000,000,000,000,000,000,
000,000,000,000,000,000,000,000,000,000,000,000,000

[1] Reymond, J.-L. (2015) *Acc. Chem. Res.*, 48, 722.

[2] Berman, H.M., et al. (2012) *Structure*, 20, 391.

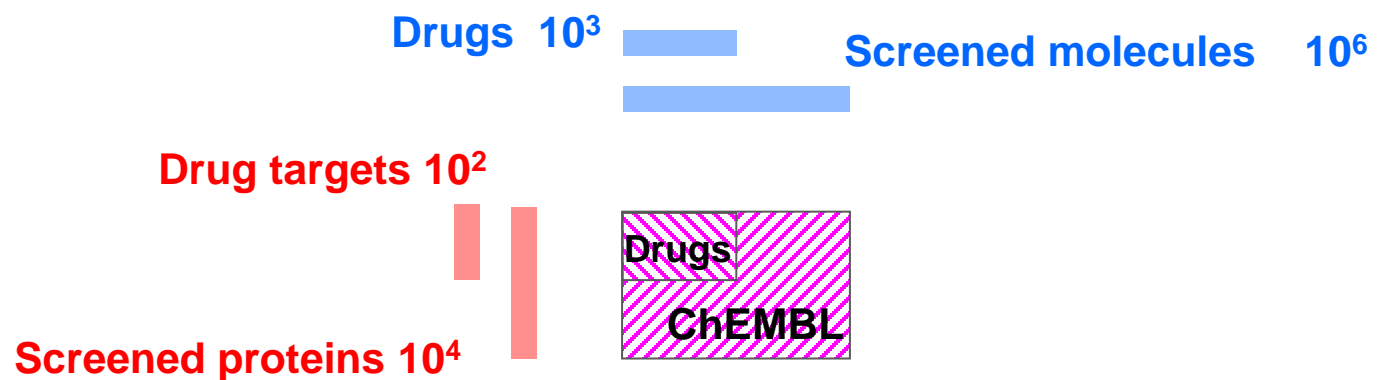
[3] Polishchuk, P.G. (2013) *J. Comput. Aided. Mol. Des.*, 27, 675

So how many drugs are out there...?

Grand challenge... charting and navigating the intersect of chemical and bioactivity space



Grand challenge... charting and navigating the intersect of chemical and bioactivity space



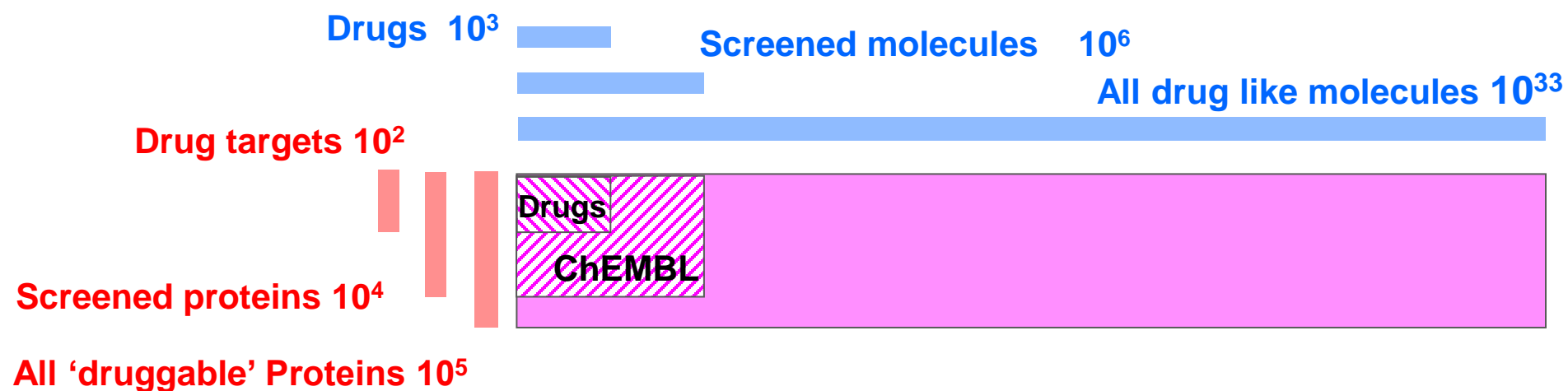
[1] Reymond, J.-L. (2015) *Acc. Chem. Res.*, 48, 722.

[2] Berman, H.M., et al. (2012) *Structure*, 20, 391.

[3] Polishchuk, P.G. (2013) *J. Comput. Aided. Mol. Des.*, 27, 675

Adapted from J.P. Overington

Grand challenge... charting and navigating the intersect of chemical and bioactivity space



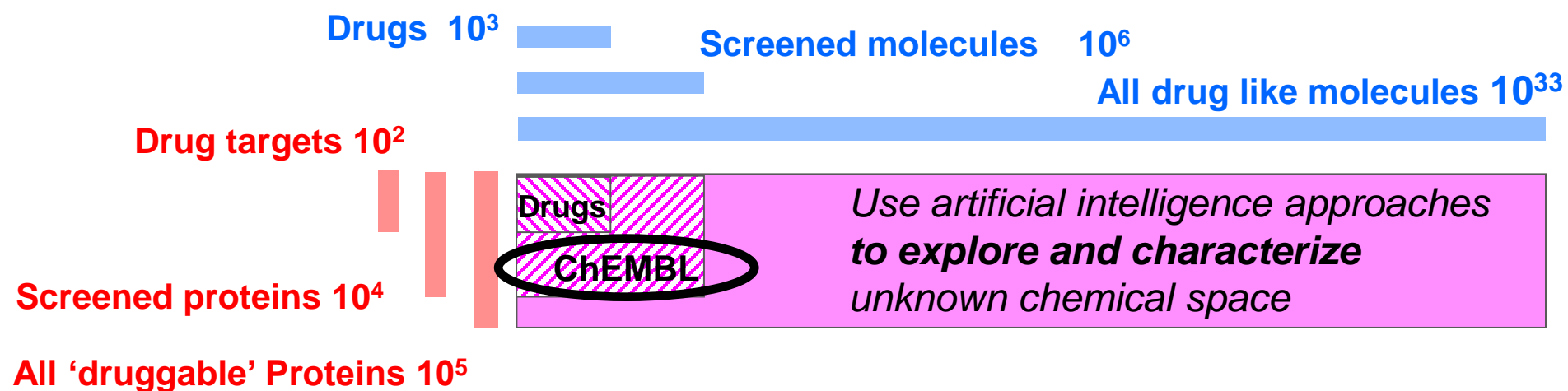
[1] Reymond, J.-L. (2015) *Acc. Chem. Res.*, 48, 722.

[2] Berman, H.M., et al. (2012) *Structure*, 20, 391.

[3] Polishchuk, P.G. (2013) *J. Comput. Aided. Mol. Des.*, 27, 675

Adapted from J.P. Overington

Grand challenge... charting and navigating the intersect of chemical and bioactivity space



[1] Reymond, J.-L. (2015) *Acc. Chem. Res.*, 48, 722.

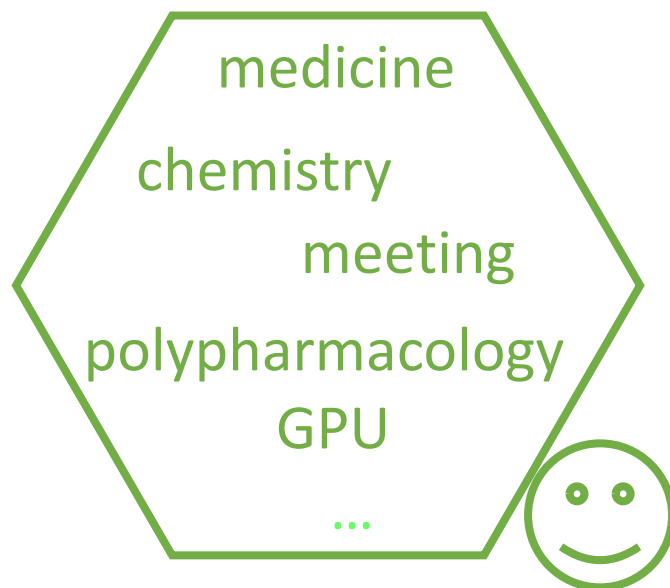
[2] Berman, H.M., et al. (2012) *Structure*, 20, 391.

[3] Polishchuk, P.G. (2013) *J. Comput. Aided. Mol. Des.*, 27, 675

Adapted from J.P. Overington

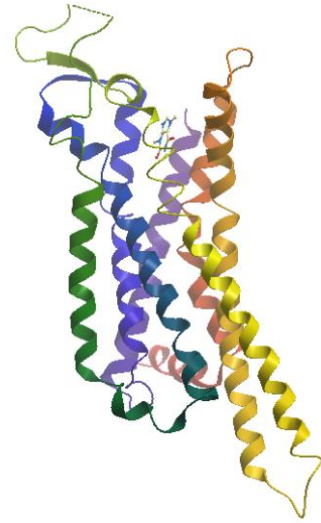
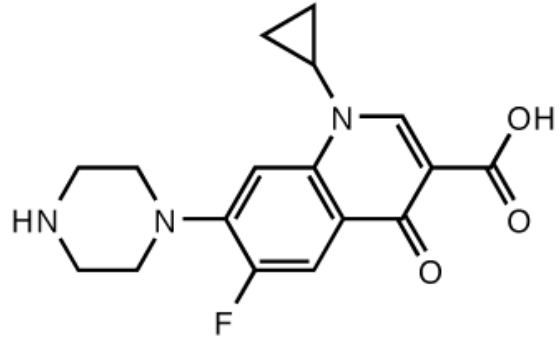
AI – Property Prediction

- Pattern recognition based on chemical structures and (predicted) biological activity
 - Using the input data we can distinguish which features are predictive and then predict the activity of the query (unknown molecule)

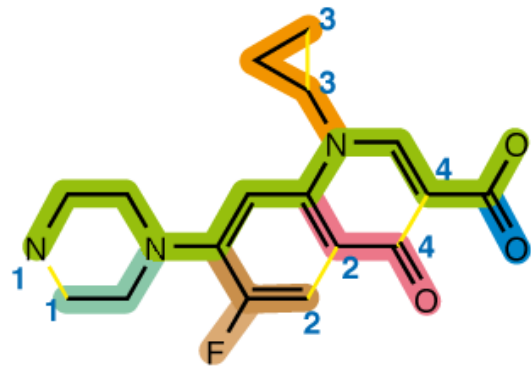
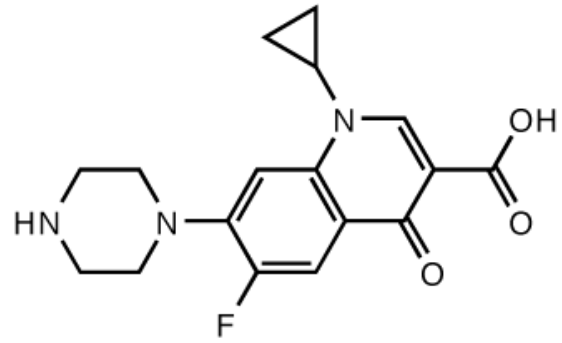


Adapted from A. Bender

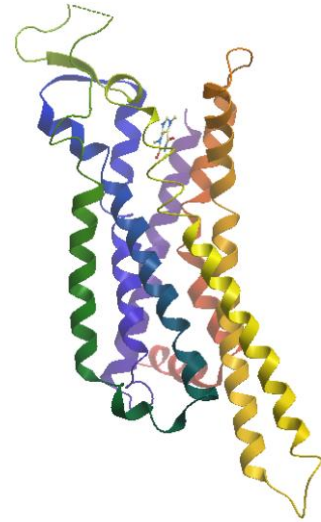
Molecular and protein grammar



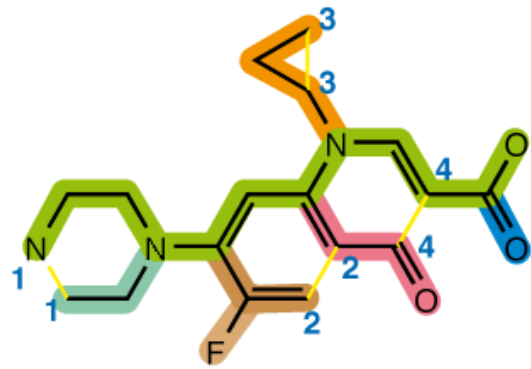
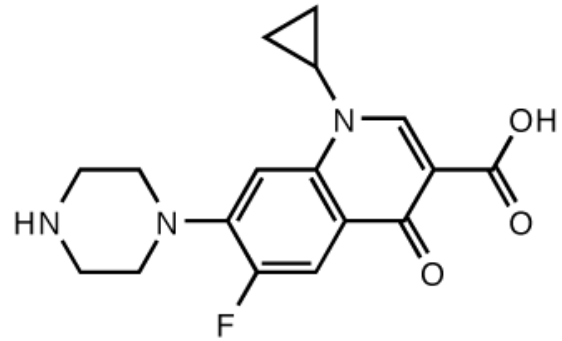
Molecular and protein grammar



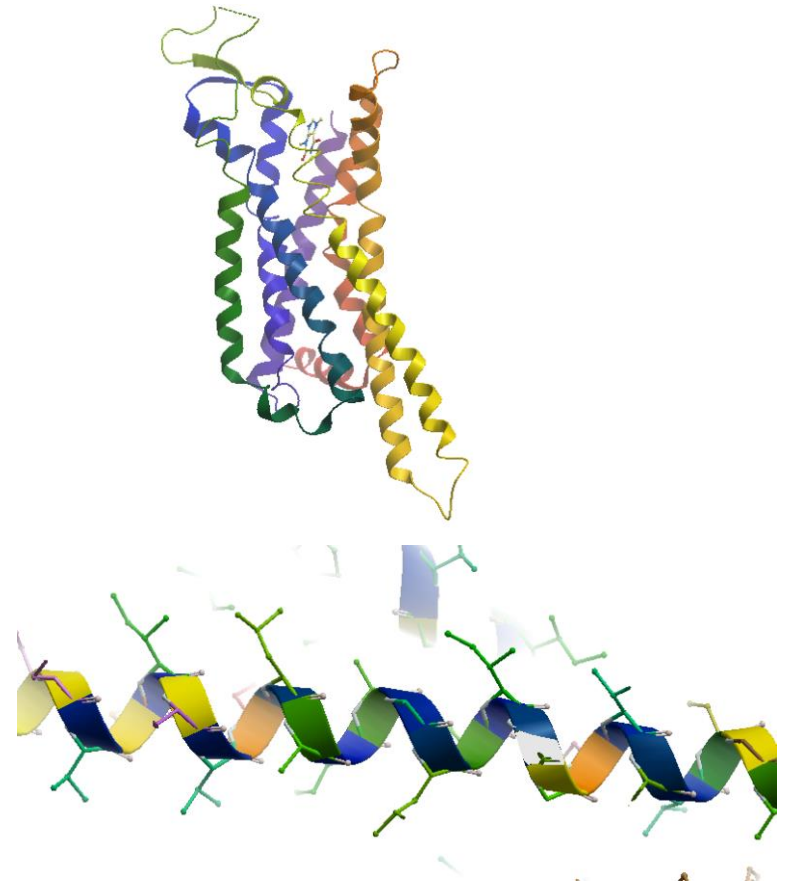
N1CCN(CC1)C(C(F)=C2)=CC(=C2C4=O)N(C3CC3)C=C4C(=O)O



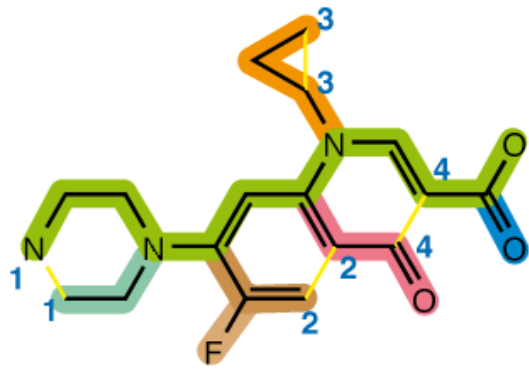
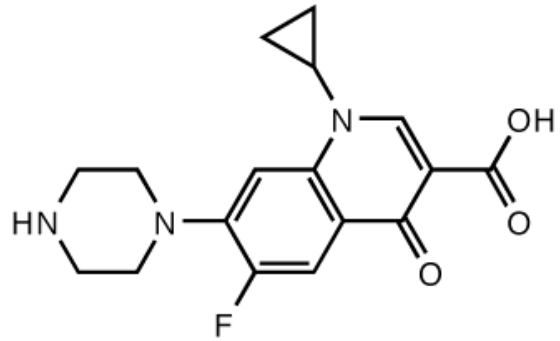
Molecular and protein grammar



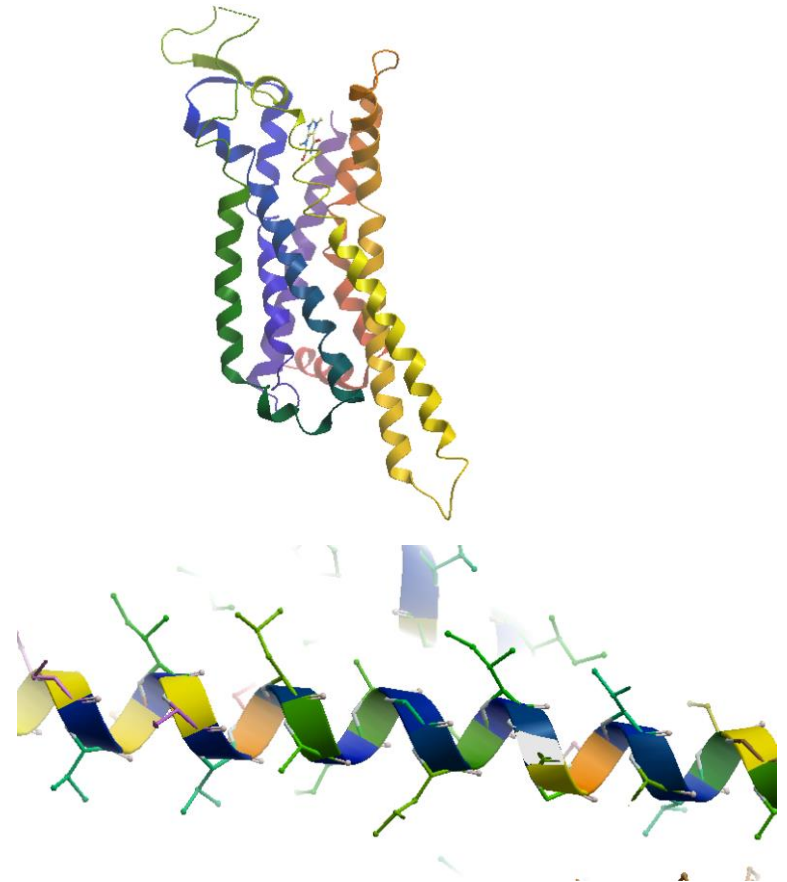
N1CCN(CC1)C(C(F)=C2)=CC(=C2C4=O)N(C3CC3)C=C4C(=O)O



Molecular and protein grammar

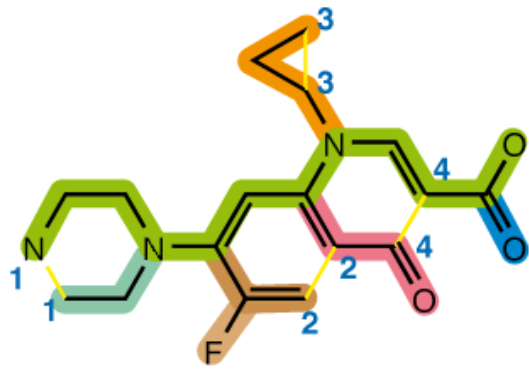
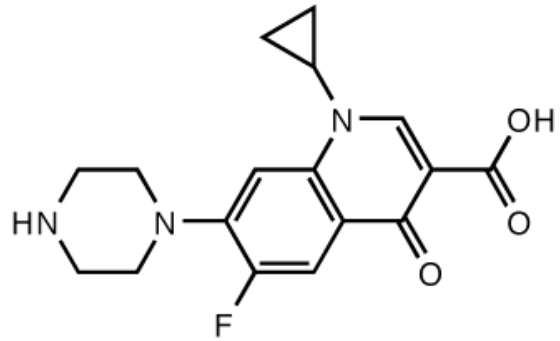


N1CCN(CC1)C(C(F)=C2)=CC(=C2C4=O)N(C3CC3)C=C4C(=O)O



MPIMGSSVYITVEL

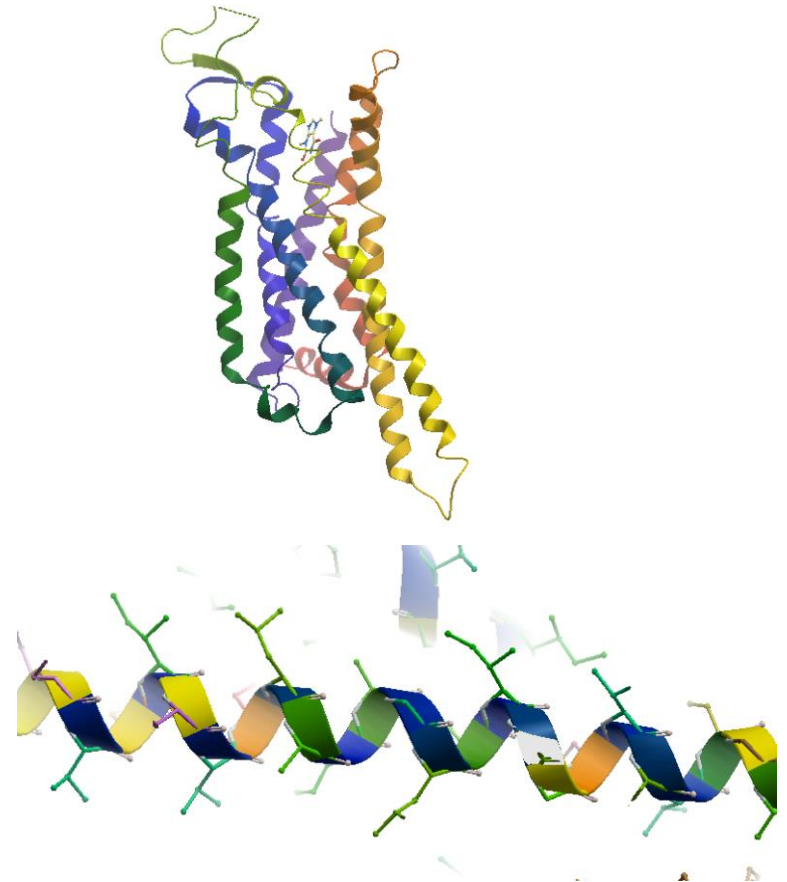
Molecular and protein grammar



N1CCN(CC1)C(C(F)=C2)=CC(=C2C4=O)N(C3CC3)C=C4C(=O)O



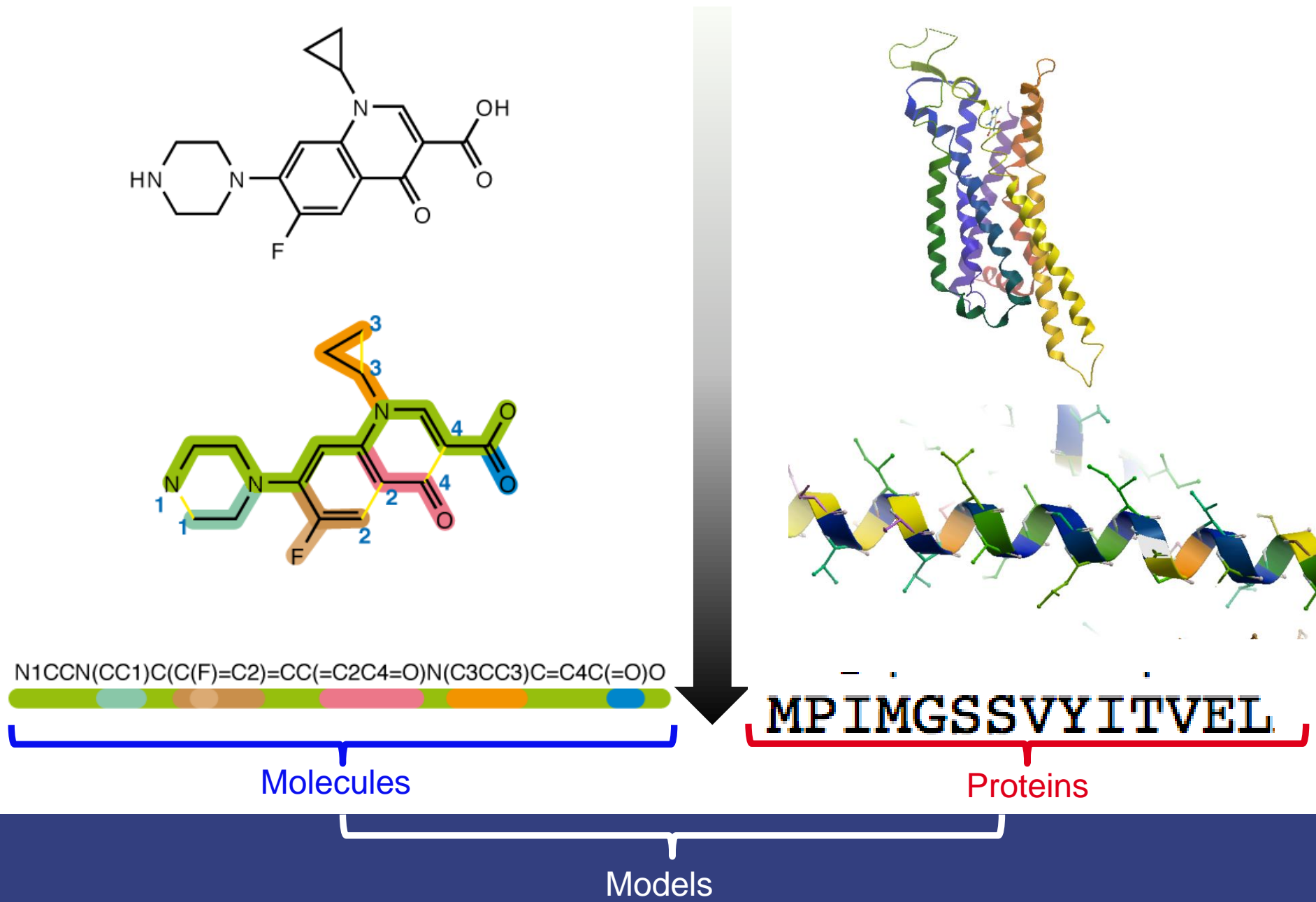
Molecules



MPIMGSSVYITVEL

Proteins

Molecular and protein grammar



Outlook

- Use machine learning to learn the grammar of a language
 - 'Google translate'



Outlook

Malfunctioning protein



Outlook

Malfunctioning protein



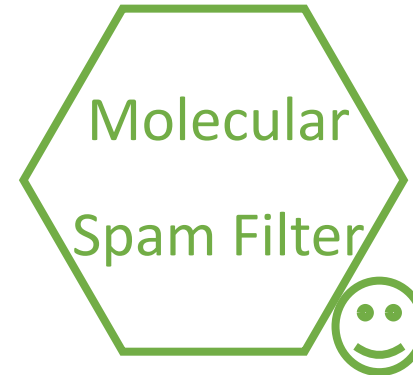
→ New molecules

Outlook

Malfunctioning protein



→ New molecules



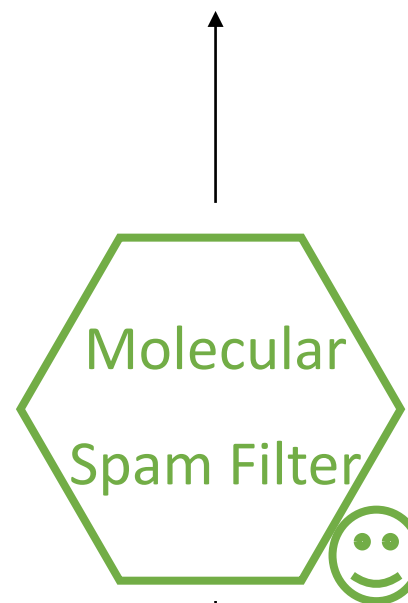
Outlook

Malfunctioning protein



New molecules

Working Molecules
(validation)



Inactive Molecules

Outlook

Malfunctioning protein

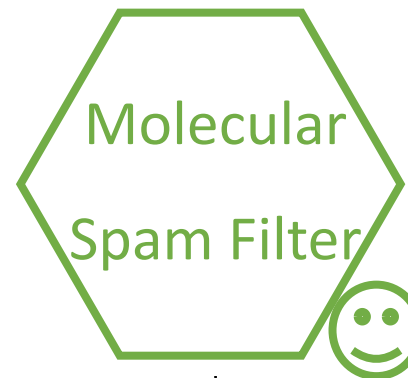


Round two

New molecules

Round two

Working Molecules
(validation)



Inactive Molecules

Outlook

Malfunctioning protein

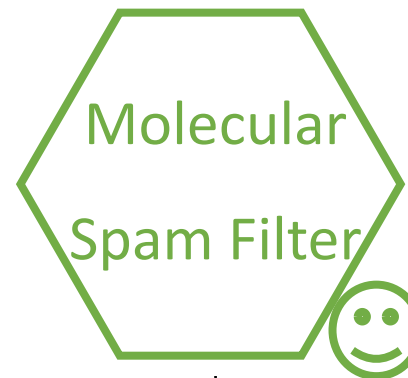


Round three
Round two

New molecules

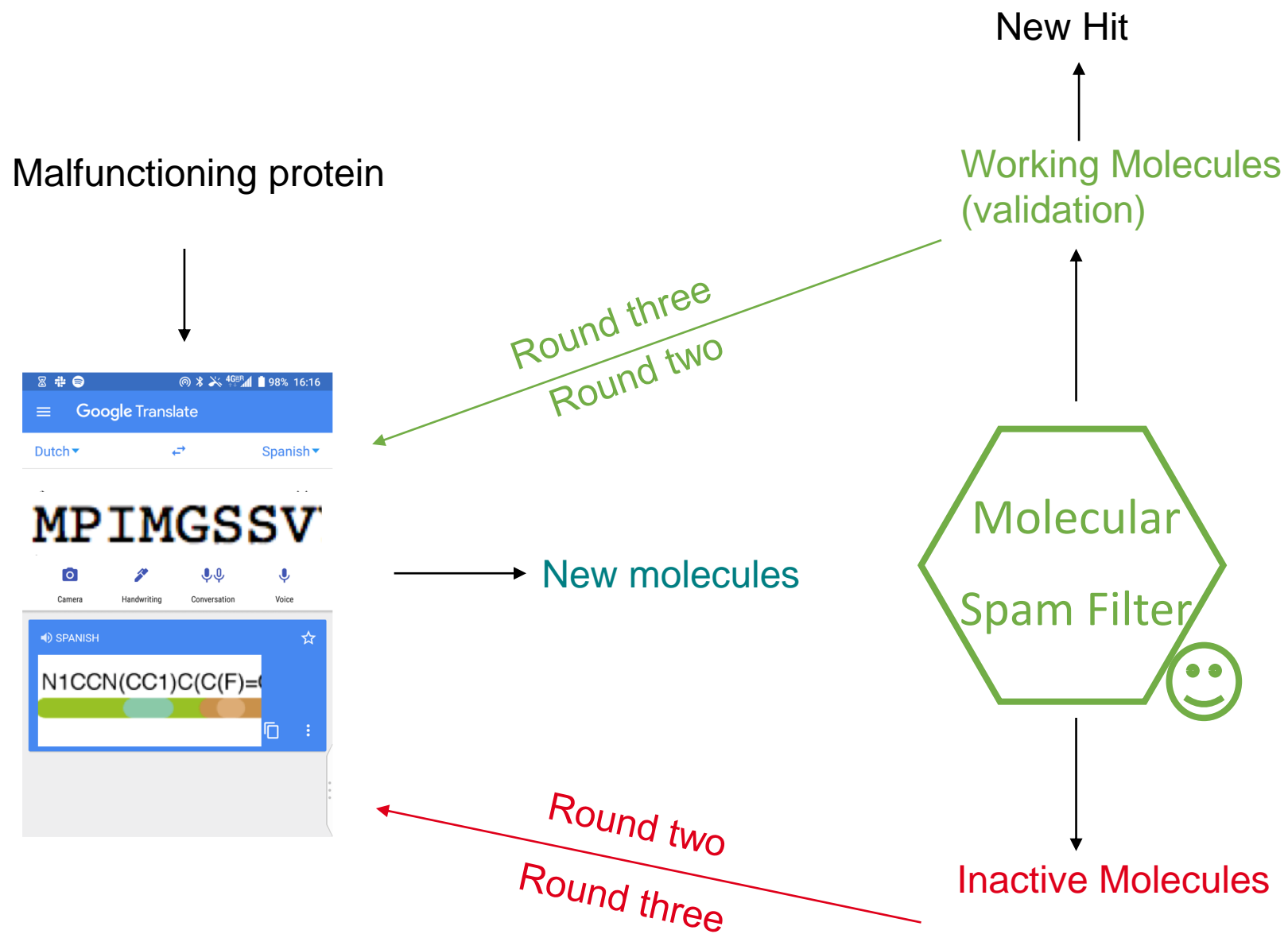
Round two
Round three

Working Molecules
(validation)



Inactive Molecules

Outlook



Outlook

Malfunctioning protein



Round three
Round two

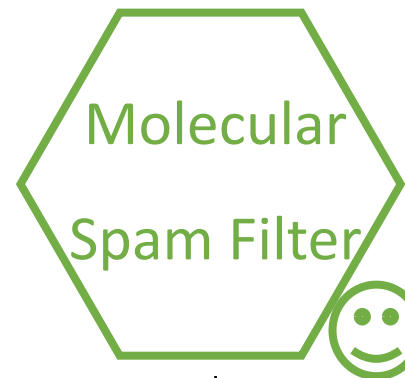
New molecules

Round two
Round three

New Hit

Experimental Validation

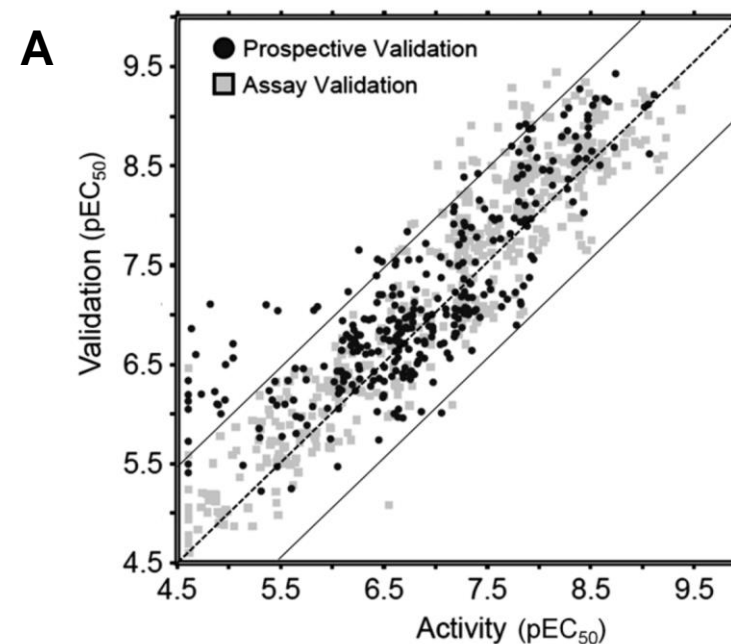
Working Molecules
(validation)



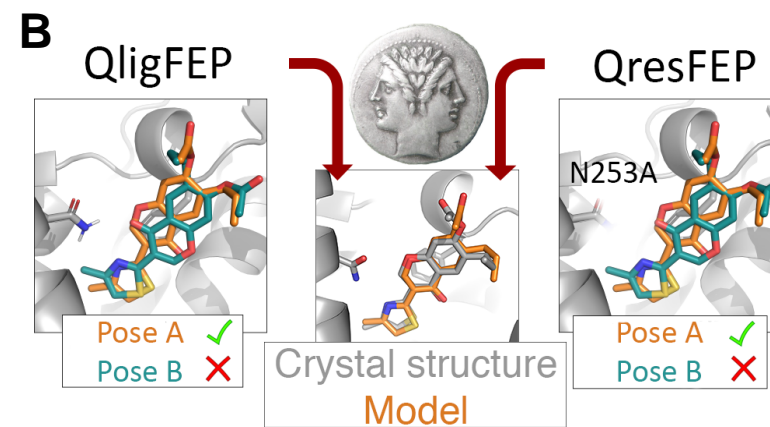
Inactive Molecules

In general two flavors of computational drug design

- Ligand based methods
 - Quantitative Structure-Activity Relationship (QSAR)
 - Artificial Intelligence
 - Property prediction (2d chemical structures)
 - *de novo* chemical structure generation

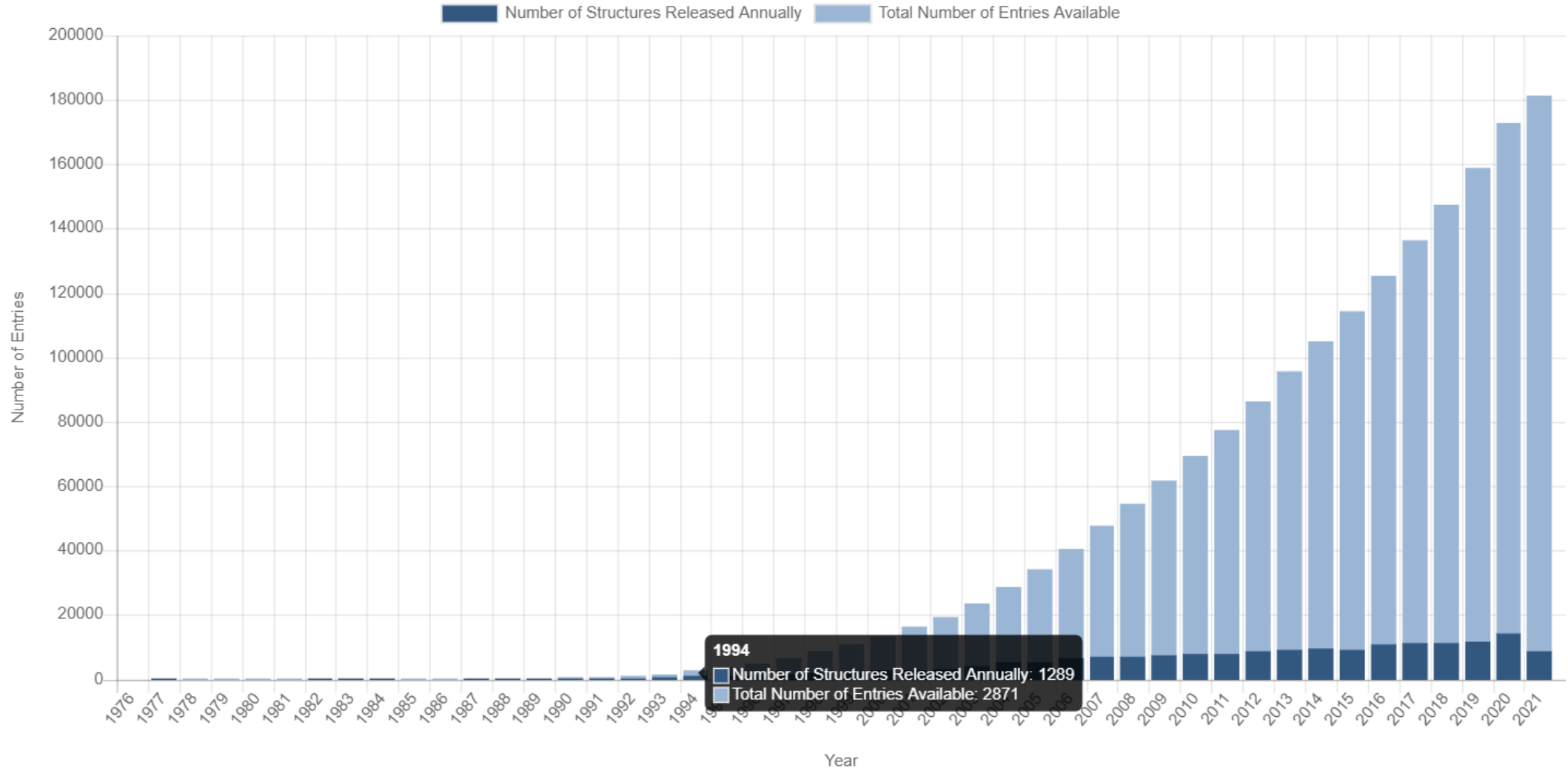


- Structure-based methods
 - Docking and scoring
 - Artificial Intelligence
 - 3D protein structure generation
 - Trajectory analysis

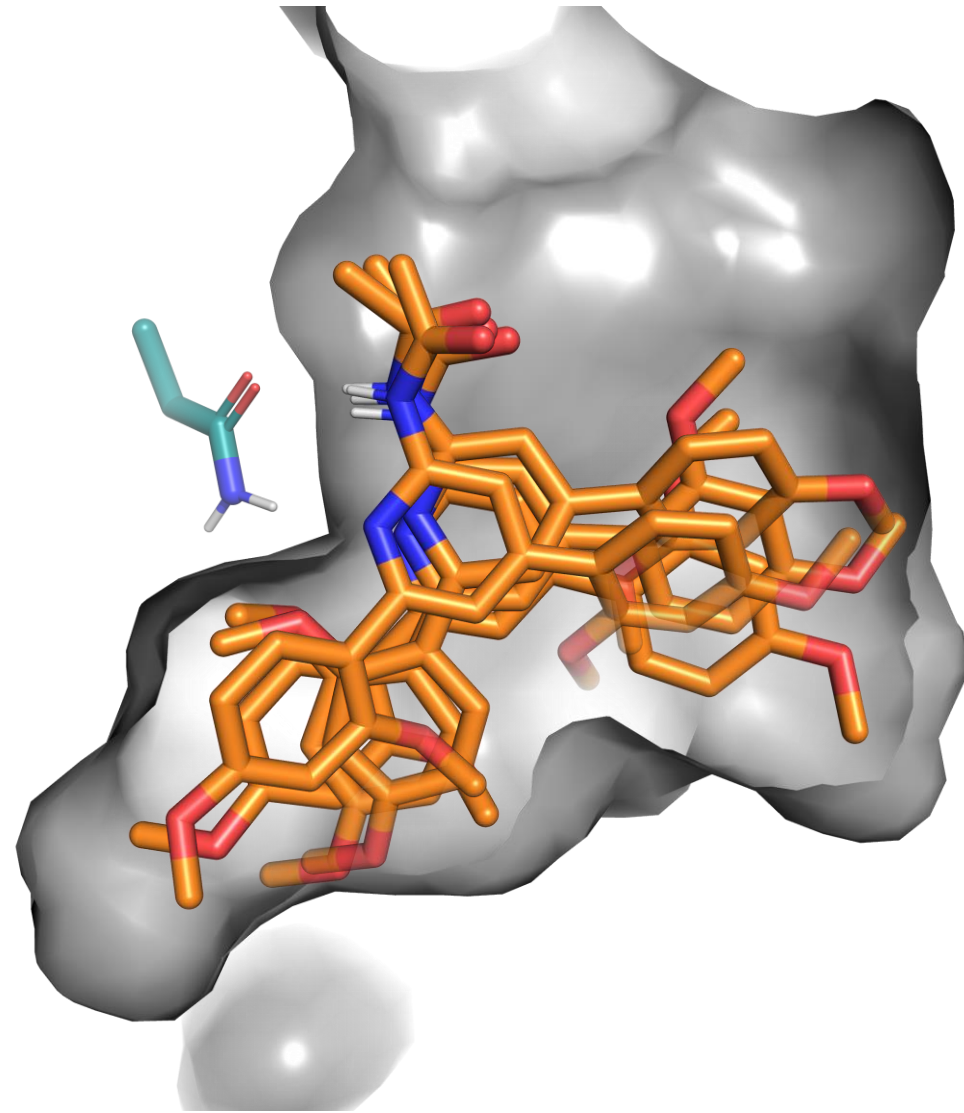


Structure Based Modelling

- Consistent increase of data over the last years (rcsb.org)



Docking



$$\Delta G_{bind} = C_0 + C_{lipo} \sum f(r_{lr}) + C_{hbond} \sum g(\Delta r)h(\Delta a) + C_{metal} \sum f(r_{lm}) + C_{rotb} + H_{rotb}$$

Molecular Dynamics

Docking is based on a static representation of the protein

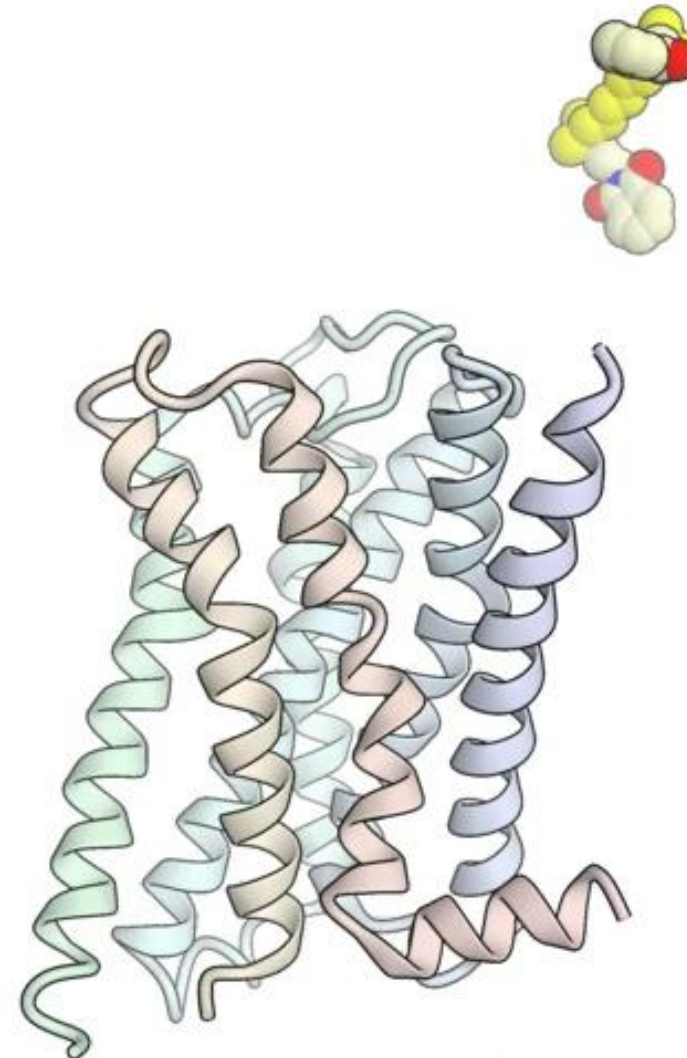
However, binding is a dynamic process

Molecular Dynamics simulates this process, by iteratively solving Newton's laws of motion

Typical one MD timestep is 2 fs, which means we need to do $5 \cdot 10^9$ steps to get this video

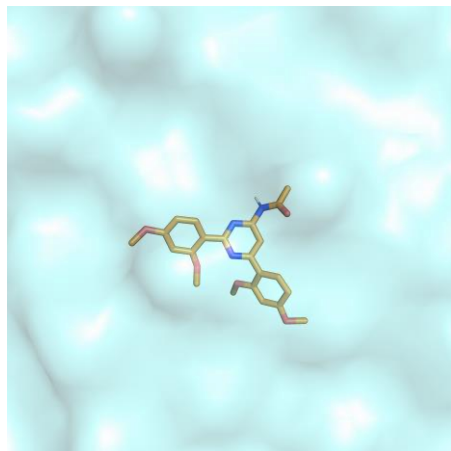
This is not feasible for larger number of molecules

0.0 us

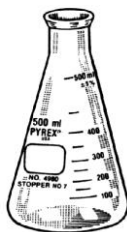


Dror et al., *Nature* 2013

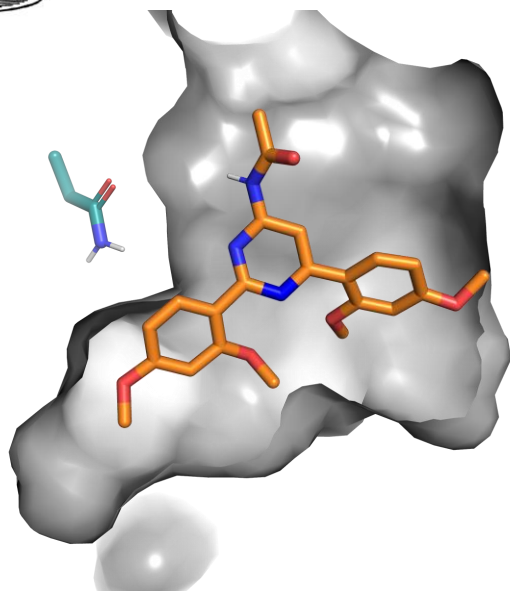
Ligand A
in water



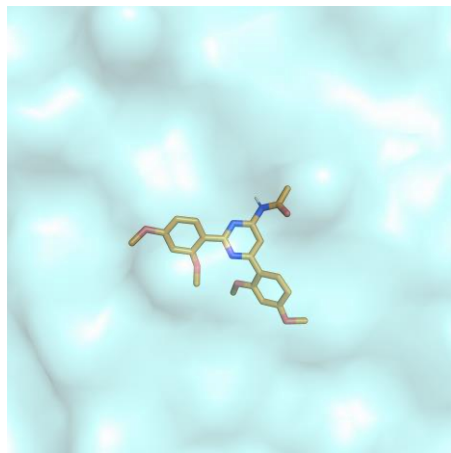
Affinity
(pKi)



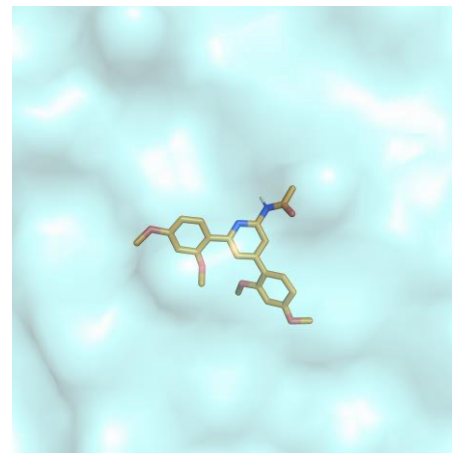
Ligand A
in protein



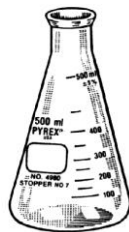
Ligand A
in water



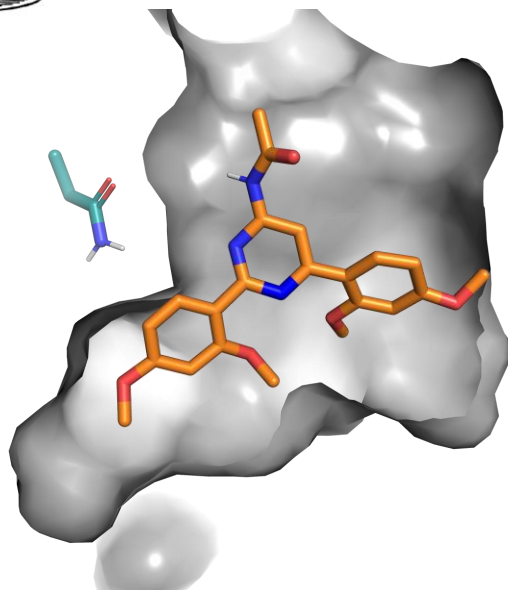
Ligand B
in water



Affinity
(pKi)



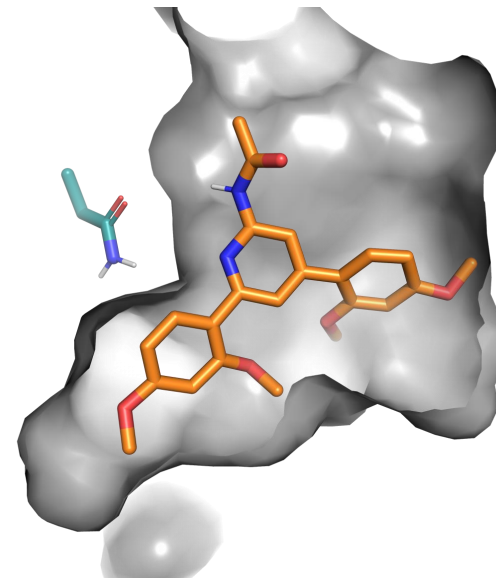
Ligand A
in protein



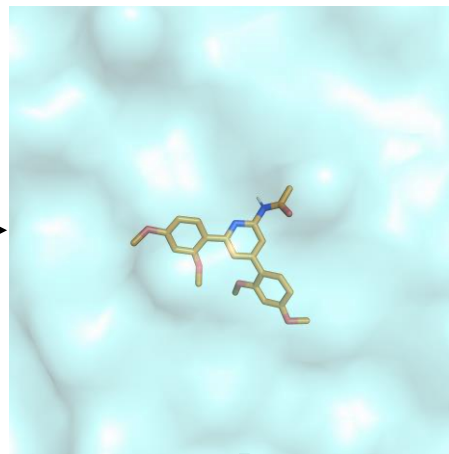
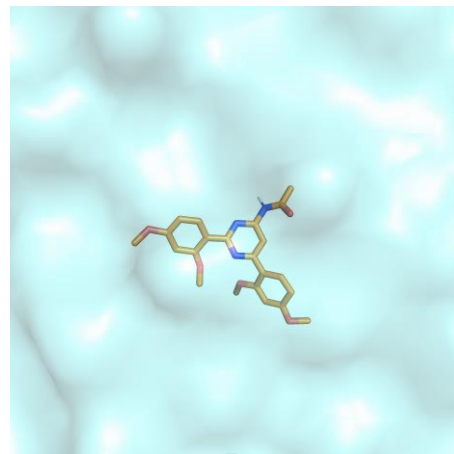
Affinity
(pKi)



Ligand B
in protein



Ligand A
in water

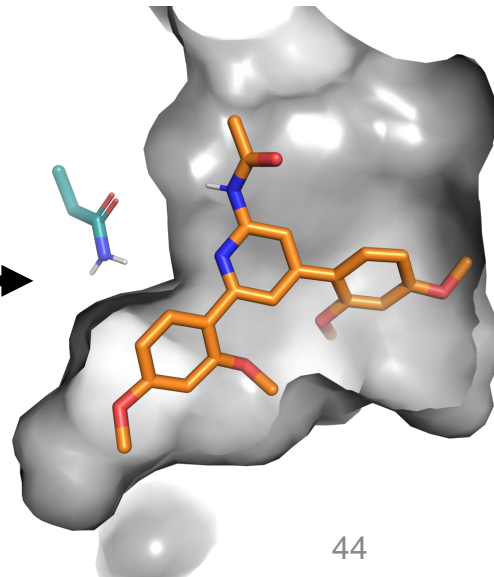
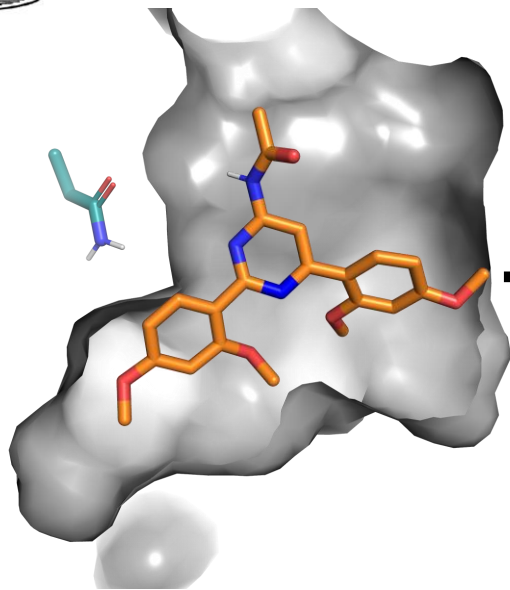


Ligand B
in water

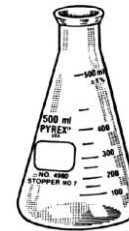
Affinity
(pKi)



Ligand A
in protein



Affinity
(pKi)



Ligand B
in protein

LACDR

Computational Drug Discovery

Olivier Bequignon

Brandon Bongers

Xuhan Liu

Marina Gorostiola Gonzalez

Hein vd Wall

Anthe Janssen

Willem Jaspers

Helle vd Maagdenberg

Rosan Kuin

Sohvi Luukkonen

Colin Bournez

Roelof vd Kleij

Drug Discovery & Safety

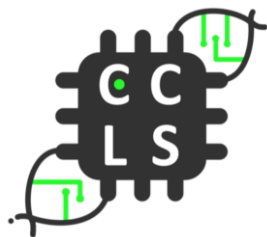
Bob vd Water

Giulia Callegaro



Scientific Director

Hubertus Irth



Hugo Gutiérrez
de Terán

Marc Willuhn

Janssen
Herman v Vlijmen



Mario van der Stelt
Hermen Overkleeft



Galápagos

Bart Lenselink
Pieter Stouten



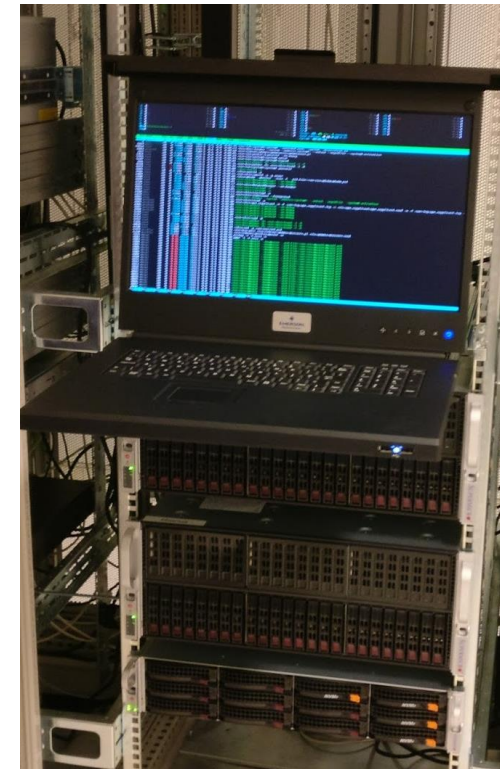
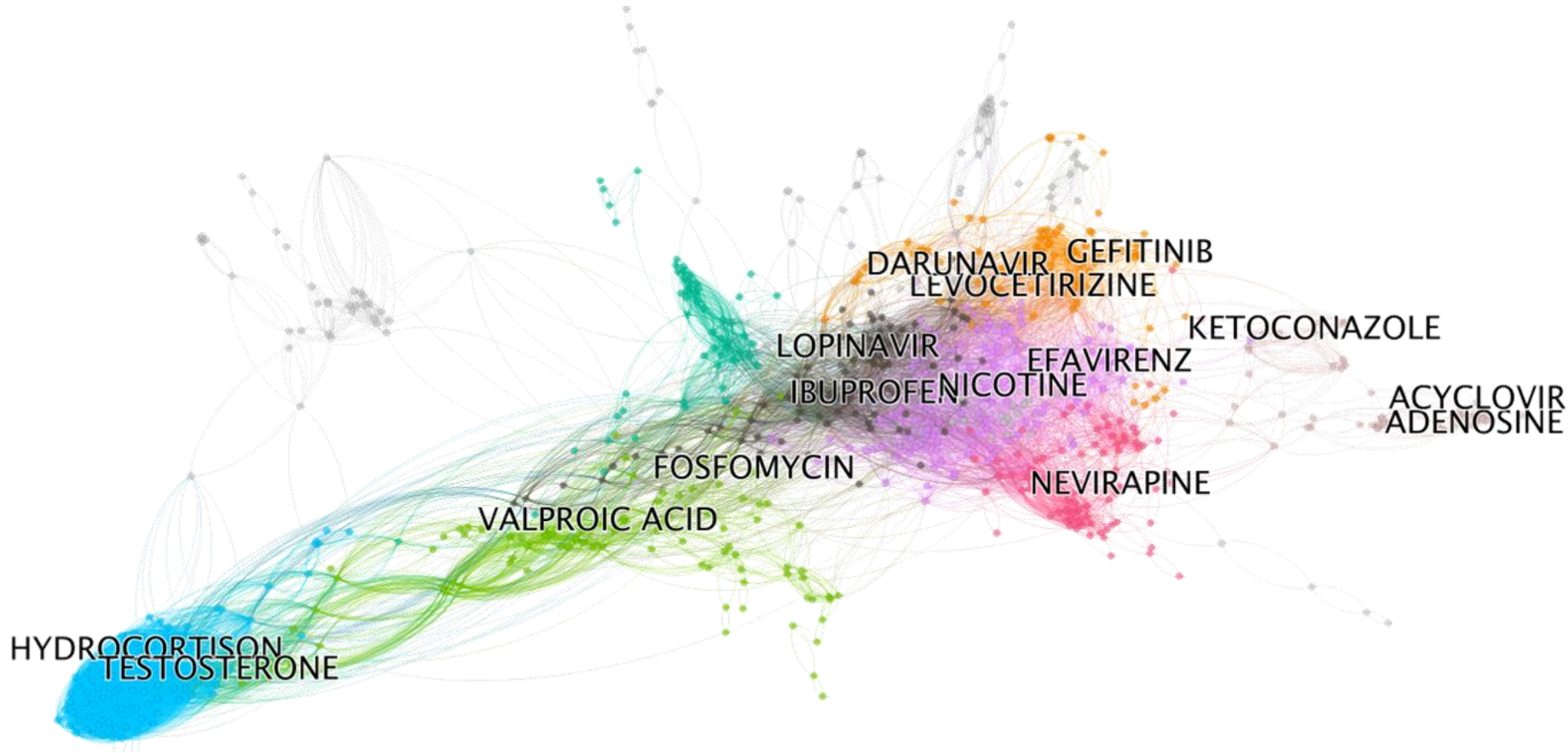
NVIDIA
Mark Berger

liacs Leiden Institute of
Advanced
Computer
Science

Michael Emmerich
Walter Kosters
Wojtek Kowalczyk
Holger Hoos
Aske Plaat
Joost Batenburg

Computational Drug Discovery

CCLS Matchmaking Event



Gerard JP van Westen

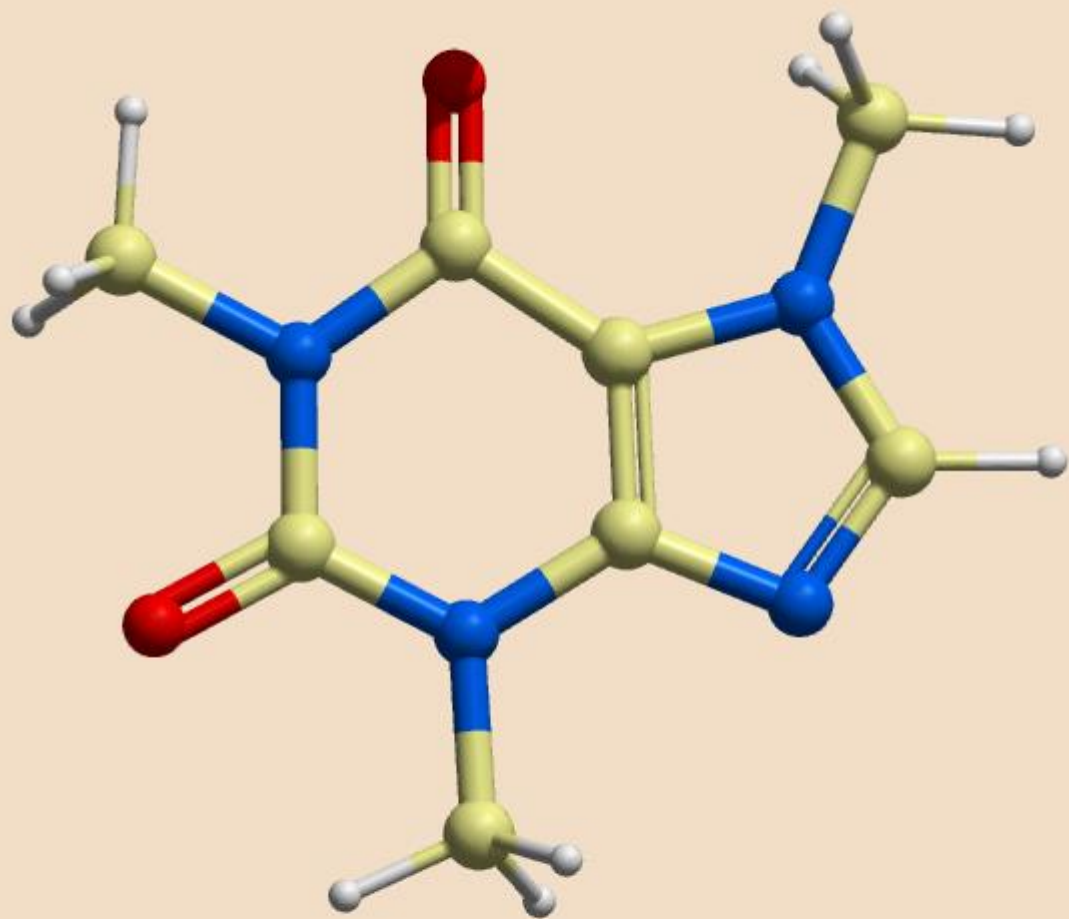
Willem Jaspers

LACDR



Applied and Engineering Sciences

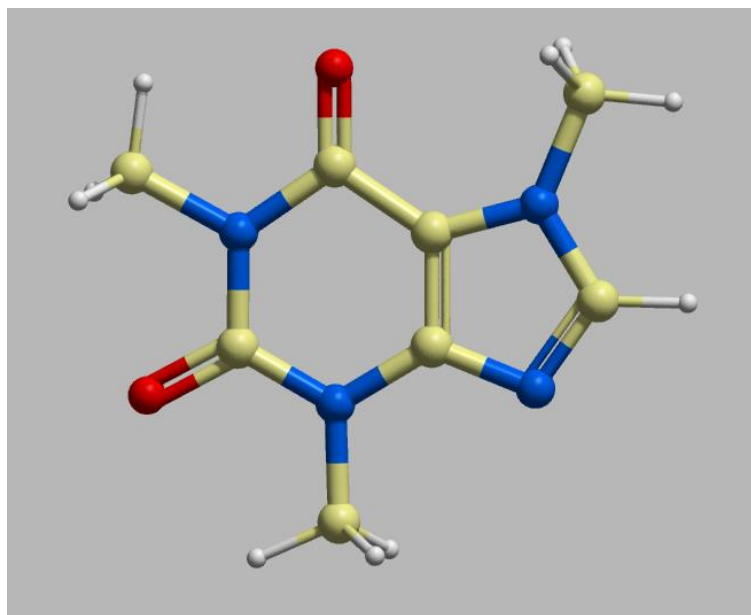




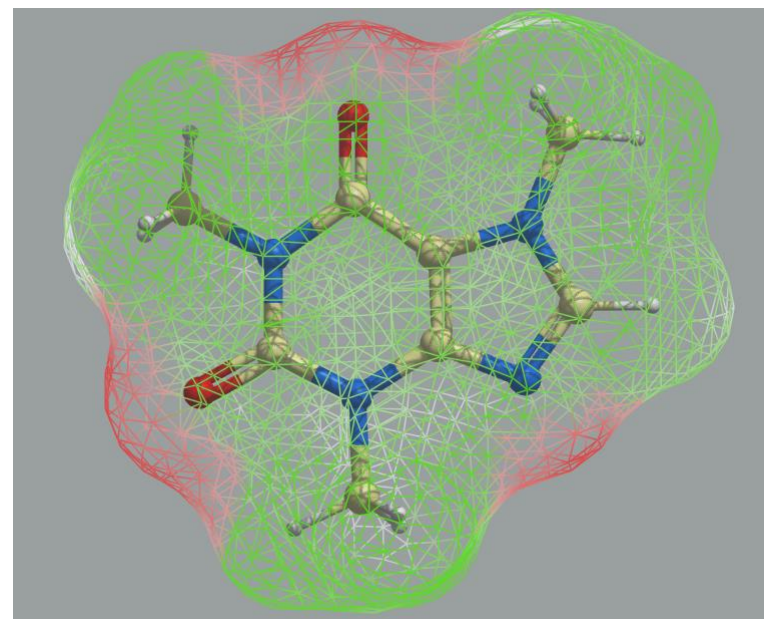
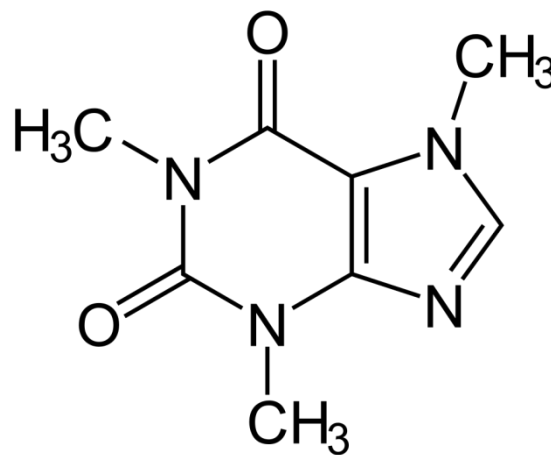
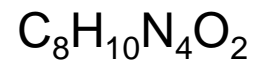
Ceci n'est pas une molécule

Molecules

- What is a molecule?

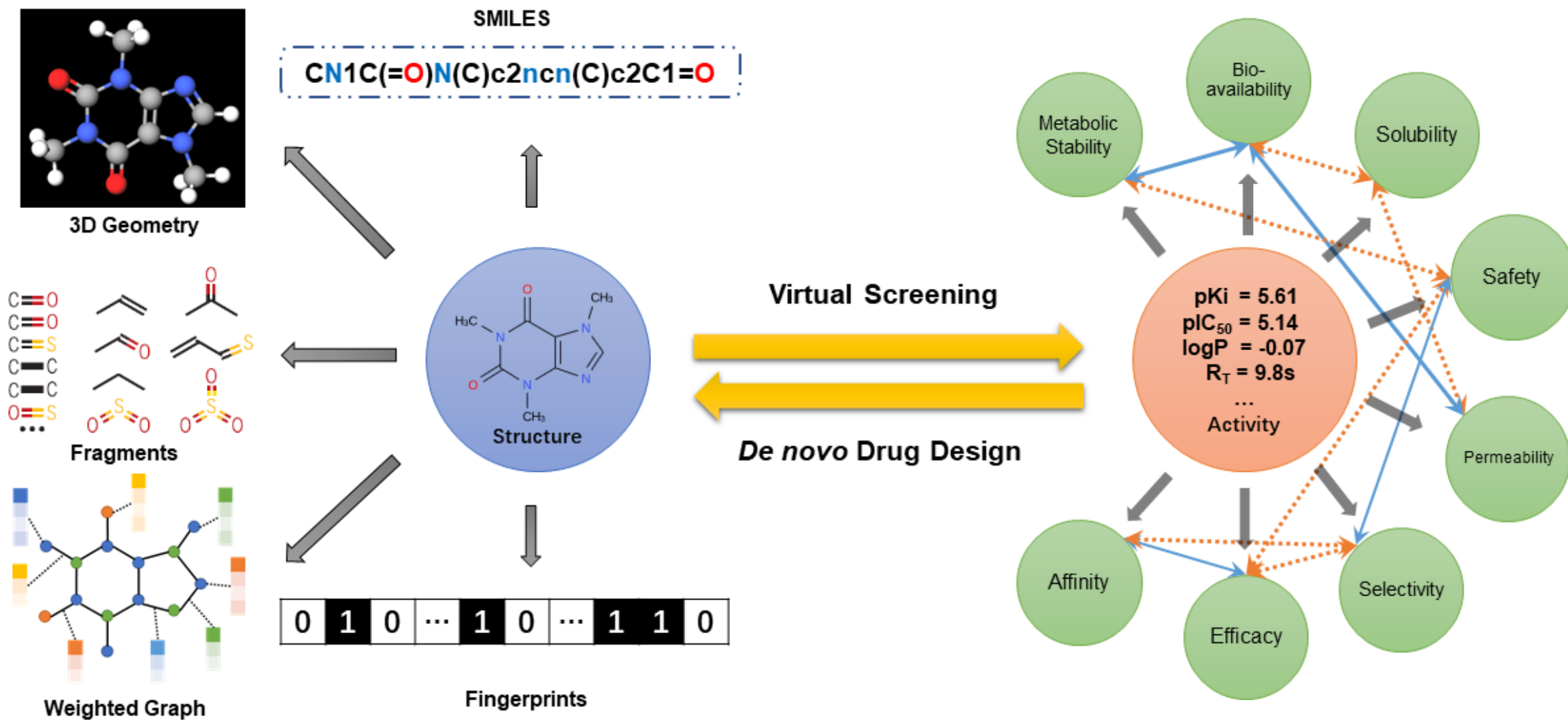


Caffeine



- Depending on the application a given representation may make sense..

AI approaches in a ligand based world..



Chemical Standardization

- Molecular structure is never the absolute truth..
 - Is it a salt form (i.e. used to improve poor solubility)
 - At which pH (is there a charge)?
 - Acids / Bases protonated?
 - Drawn the same way (double bonds / aromatic bonds)
 - Tautomers
 - Stereochemistry
 - ..etc

Molecules

- How to *store* a molecule?

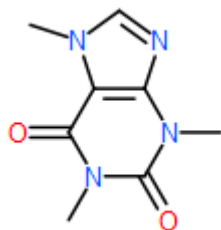
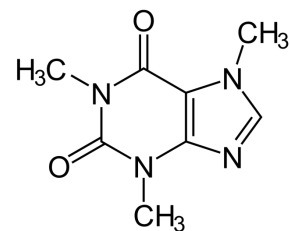
```
Caffeine
Comment Line
14 15 0 0 0 0 0 0 0 0999 V2000
-1.4765 -1.4521 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
-1.2216 -0.6674 0.0000 N 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
-1.7065 0.0000 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
-1.2216 0.6674 0.0000 N 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
-0.4369 0.4125 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
-0.4369 -0.4125 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0.2775 -0.8250 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0.2775 -1.6500 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0.9920 -0.4125 0.0000 N 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0.9920 0.4125 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1.7065 0.8250 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0.2775 0.8250 0.0000 N 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0.2775 1.6500 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1.7065 -0.8250 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 2 1 0
2 3 1 0
3 4 2 0
4 5 1 0
5 6 2 0
2 6 1 0
6 7 1 0
7 8 2 0
7 9 1 0
9 10 1 0
10 11 2 0
10 12 1 0
5 12 1 0
12 13 1 0
9 14 1 0
M END
$$$$
```

1,3,7-Trimethylpurine-2,6-dione

1,3,7,-Trimethylxanthine

CAS 58-08-2

$C_8H_{10}N_4O_2$



InChI=1S/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3

CN1C(=O)N(C)c2ncn(C)c2C1=O

RYYVLZVUVIJVGH-UHFFFAOYSA-N

Molecules

- How to *store* a molecule?

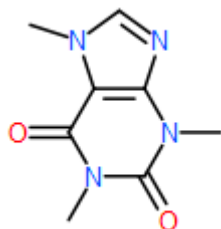
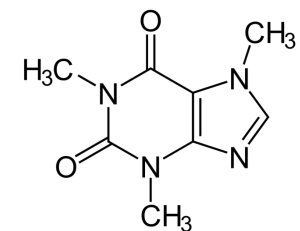
```
Caffeine
Comment Line
14 15 0 0 0 0 0 0 0 0999 V2000
-1.4765 -1.4521 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
-1.2216 -0.6674 0.0000 N 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
-1.7065 0.0000 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
-1.2216 0.6674 0.0000 N 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
-0.4369 0.4125 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
-0.4369 -0.4125 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0.2775 -0.8250 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0.2775 -1.6500 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0.9920 -0.4125 0.0000 N 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0.9920 0.4125 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1.7065 0.8250 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0.2775 0.8250 0.0000 N 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0.2775 1.6500 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1.7065 -0.8250 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 2 1 0
2 3 1 0
3 4 2 0
4 5 1 0
5 6 2 0
2 6 1 0
6 7 1 0
7 8 2 0
7 9 1 0
9 10 1 0
10 11 2 0
10 12 1 0
5 12 1 0
12 13 1 0
9 14 1 0
M END
$$$$
```

1,3,7-Trimethylpurine-2,6-dione

1,3,7,-Trimethylxanthine

CAS 58-08-2

$C_8H_{10}N_4O_2$



InChI=1S/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3

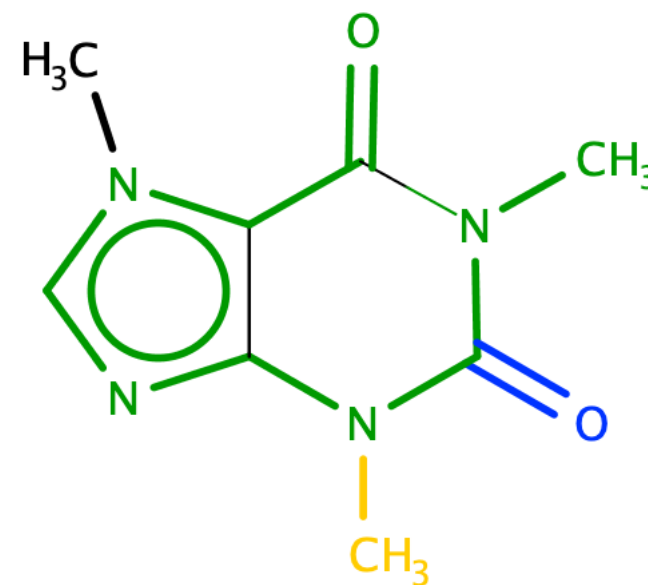
CN1C(=O)N(C)c2ncn(C)c2C1=O

RYYVLZVUVIJVGH-UHFFFAOYSA-N

SMILES

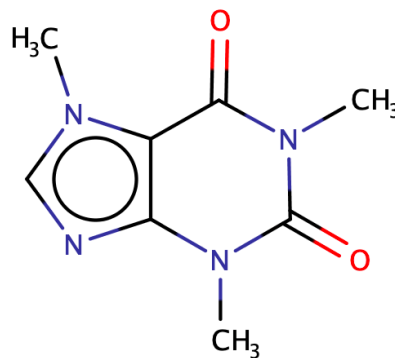
- Simplified molecular-input line-entry system (SMILES)
- Line format (describing the chemical graph)
 - Supports stereochemistry but hardly used..
 - Branch: ()
 - Rings : Number at start and closure

CN1C(=O)N(C)c2nncn(C)c2C1=O



InChI: International Chemical Identifier

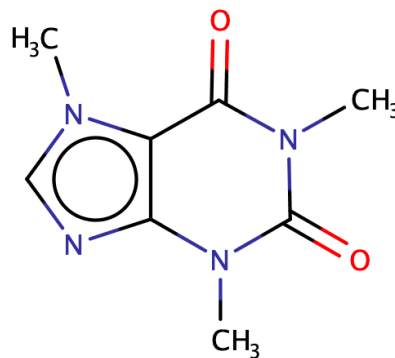
- Built up of layers and sublayers of information
 - the atoms, their bond connectivity, tautomeric information, isotope information, stereochemistry, electronic charge information
 - IUPAC
- Unique



InChI=1S/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3

InChI: International Chemical Identifier

- Built up of layers and sublayers of information
 - the atoms, their bond connectivity, tautomeric information, isotope information, stereochemistry, electronic charge information
 - IUPAC
- Unique

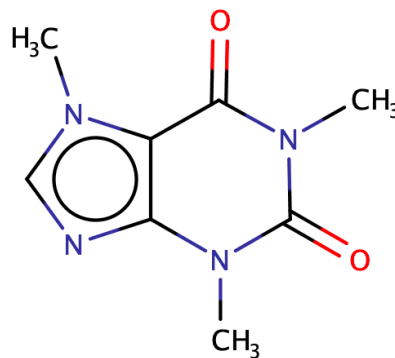


InChI=1S/**C8H10N4O2**/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3

Start, version

InChI: International Chemical Identifier

- Built up of layers and sublayers of information
 - the atoms, their bond connectivity, tautomeric information, isotope information, stereochemistry, electronic charge information
 - IUPAC
- Unique

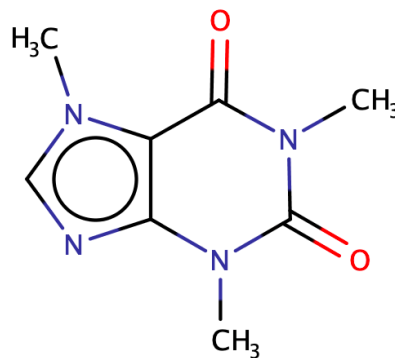


InChI=1S/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3

Chemical formula

InChI: International Chemical Identifier

- Built up of layers and sublayers of information
 - the atoms, their bond connectivity, tautomeric information, isotope information, stereochemistry, electronic charge information
 - IUPAC
- Unique



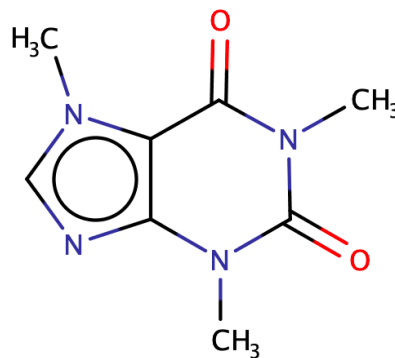
InChI=1S/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3

Heavy atoms and connectivity

InChI: International Chemical Identifier

- Built up of layers and sublayers of information
 - the atoms, their bond connectivity, tautomeric information, isotope information, stereochemistry, electronic charge information
 - IUPAC

- Unique



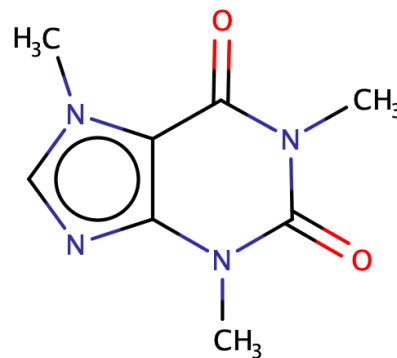
InChI=1S/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3

Placement of hydrogens

InChI: International Chemical Identifier

- Built up of layers and sublayers of information
 - the atoms, their bond connectivity, tautomeric information, isotope information, stereochemistry, electronic charge information
 - IUPAC

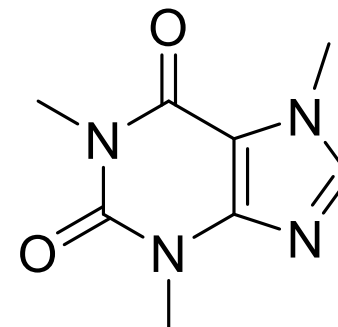
- Unique



InChI=1S/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3/.....
Chirality

InChIKey

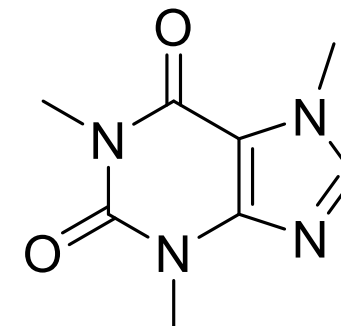
- Hashed version of InChI: InChiKey
- Fixed length: 27 characters
- SHA-256 cryptographic hash
- Structure based lookup-identifier
 - Generated directly from chemical structure
- Clashes estimated $1:10^{11}$



RYYVLZVUVIJVGH-UHFFFAOYSA-N

InChIKey

- Hashed version of InChI: InChiKey
- Fixed length: 27 characters
- SHA-256 cryptographic hash
- Structure based lookup-identifier
 - Generated directly from chemical structure
- Clashes estimated $1:10^{11}$

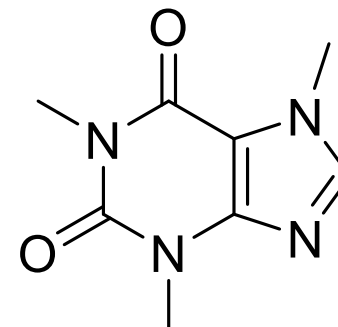


RYYVLZVUVIJVGH-UHFFFAOYSA-N

Core molecular
scaffold

InChIKey

- Hashed version of InChI: InChiKey
- Fixed length: 27 characters
- SHA-256 cryptographic hash
- Structure based lookup-identifier
 - Generated directly from chemical structure
- Clashes estimated $1:10^{11}$

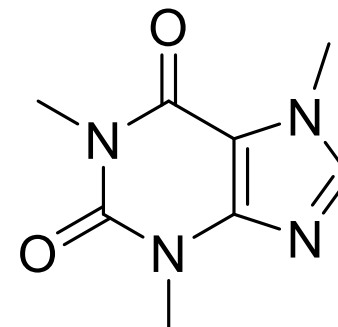


RYYVLZVUVIJVGH-UHFFFAOYSA-N

UHFFFAOYSA
All other layers

InChIKey

- Hashed version of InChI: InChIKey
- Fixed length: 27 characters
- SHA-256 cryptographic hash
- Structure based lookup-identifier
 - Generated directly from chemical structure
- Clashes estimated $1:10^{11}$

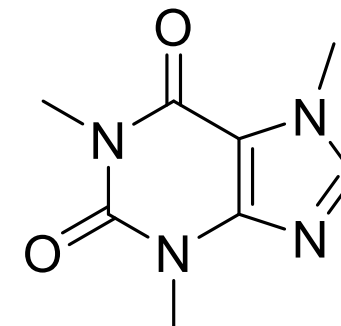


RYYVLZVUVIJVGH-UHFFFAOYSA-N

S: standard
A: version 1

InChIKey

- Hashed version of InChI: InChIKey
- Fixed length: 27 characters
- SHA-256 cryptographic hash
- Structure based lookup-identifier
 - Generated directly from chemical structure
- Theoretical clashes..



RYYVLZVUVIJVGH-UHFFFAOYSA-N

Protonation
N: Neutral

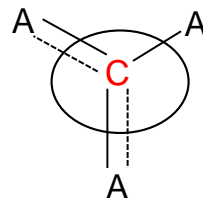
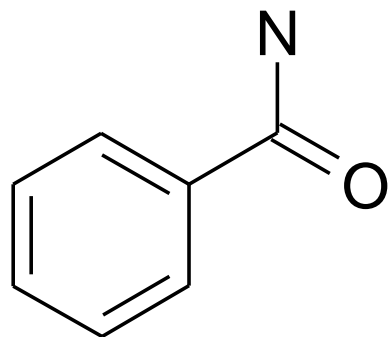
Molecular similarity

Molecular similarity

- In cheminformatics methods rely on the similarity principle which states that ‘similar molecules are expected to have similar bioactivities’
 - “If it looks like a duck, swims like a duck, and quacks like a duck, then it probably is a duck.”
- A comparable principle exists for protein targets. Similar proteins are expected to interact with similar molecules.

Descriptors

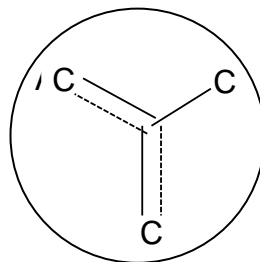
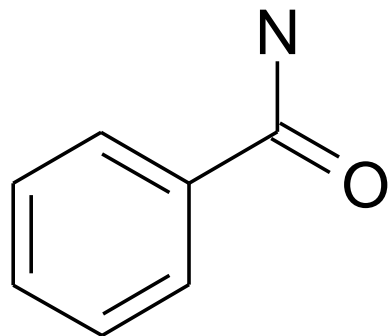
- Fingerprints convert chemical features to a bit string



Feature
16

Descriptors

- Fingerprints convert chemical features to a bit string



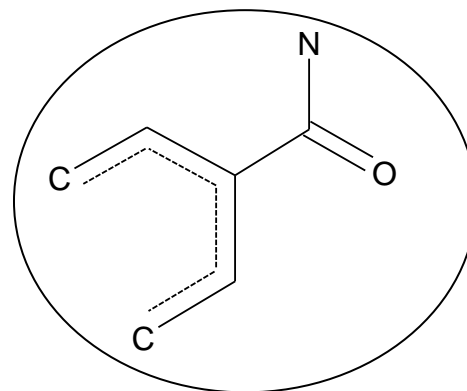
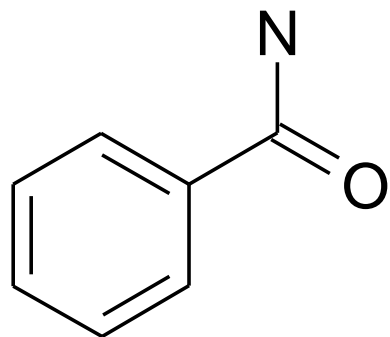
Feature

16

1618154665

Descriptors

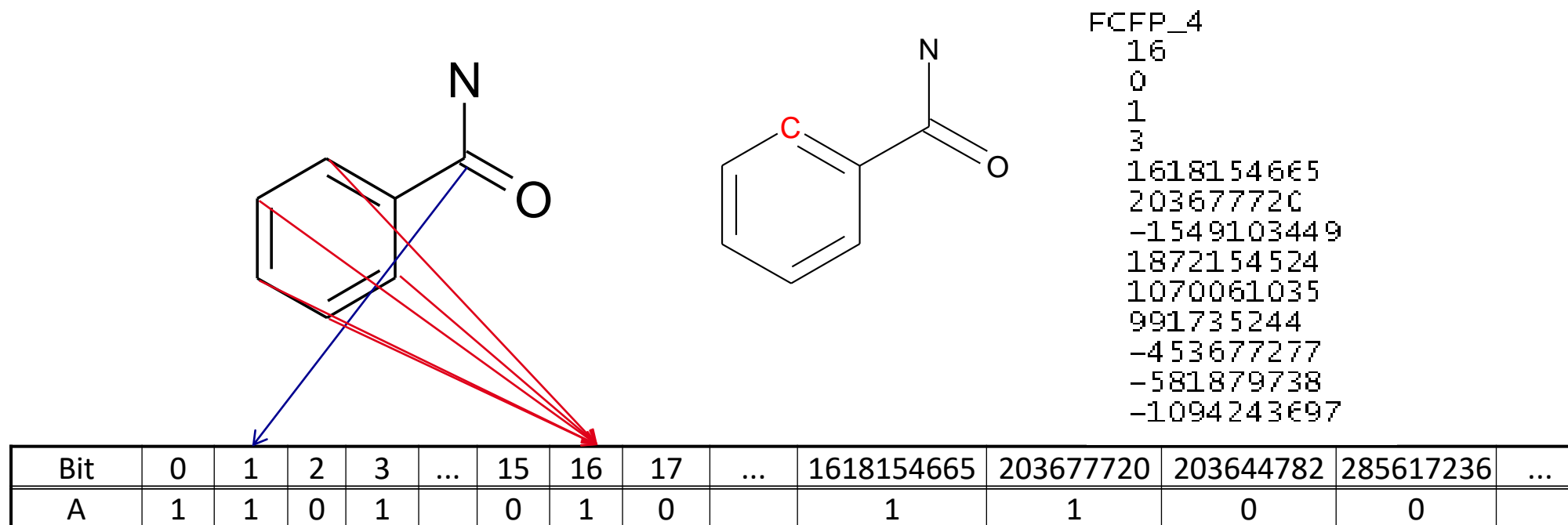
- Fingerprints convert chemical features to a bit string



```
FCFP_4  
16  
0  
1  
3  
1618154665  
203677720  
-1549103449  
1872154524  
1070061035  
991735244  
-453677277  
-581879738  
-1094243697
```

Descriptors

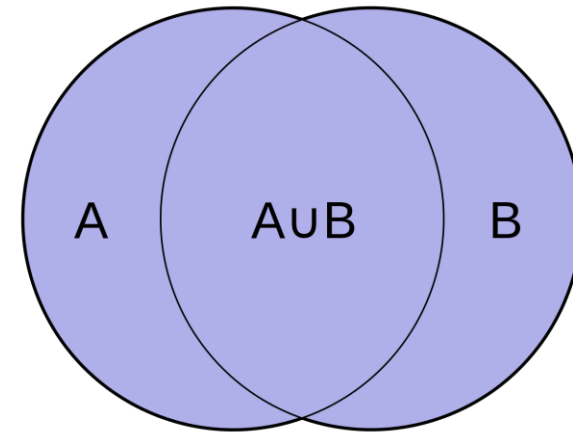
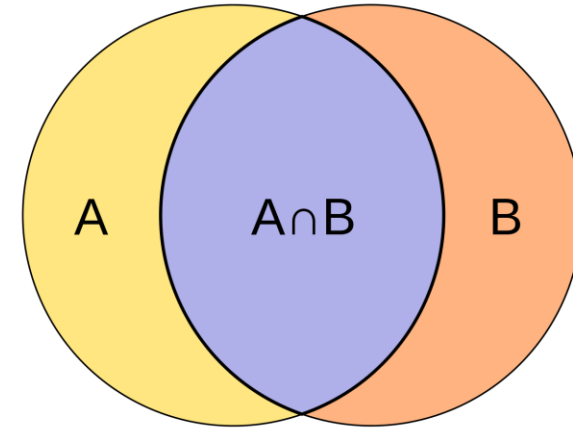
- Fingerprints convert chemical features to a bit string



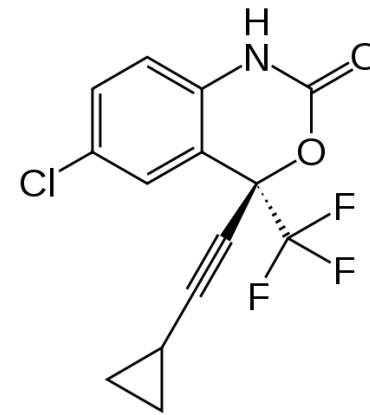
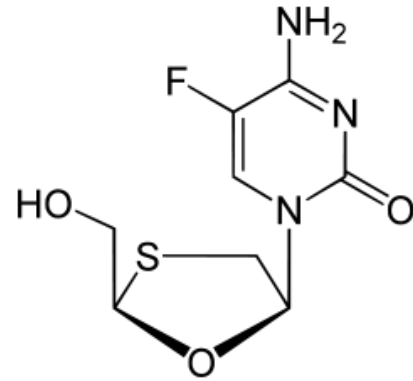
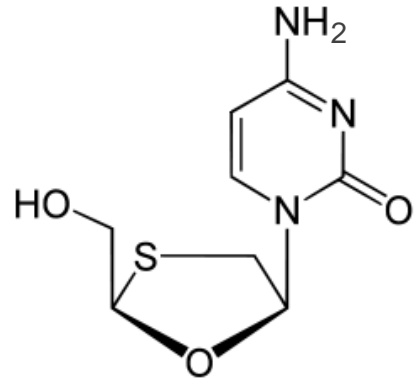
Jaccard or Tanimoto index & distance

- Tanimoto similarity (index)
 - Count # bits set in both A & B (intersection)
 - Count total # bits set in either A & B (union)
 - Divide
- Tanimoto distance
 - 1-(Tanimoto similarity)

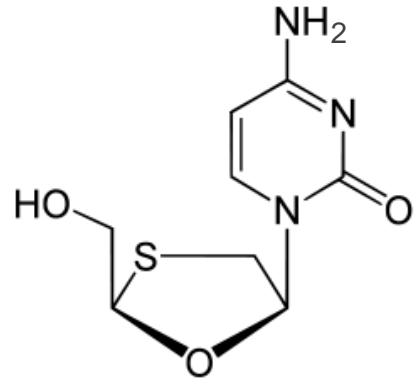
$$index = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$



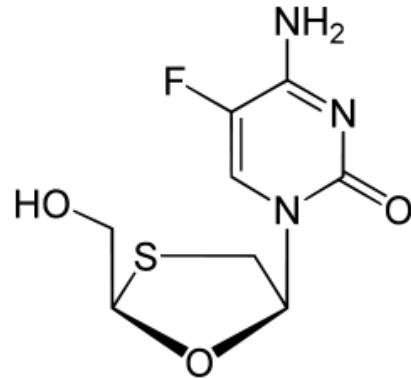
Molecular Similarity



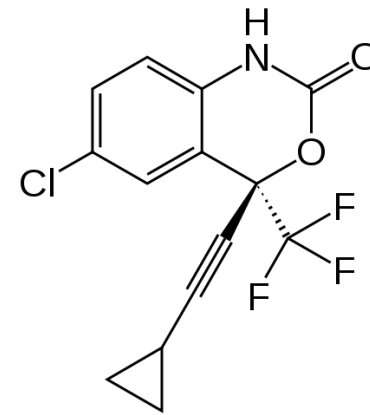
Molecular Similarity



Lamivudine, 3TC
(NRTI)



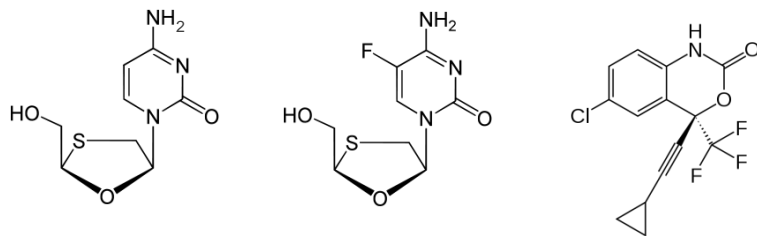
Emtricitabine, FTC
(NRTI)

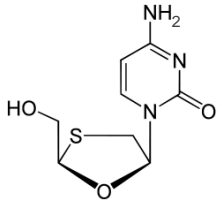
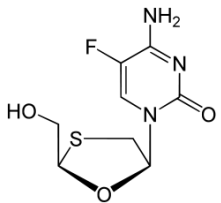
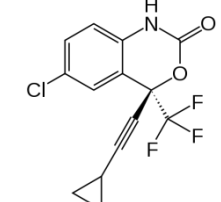


Efavirenz, EFV
(NNRTI)

Similar compounds have similar properties

Molecular Similarity



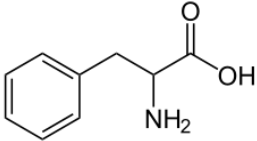
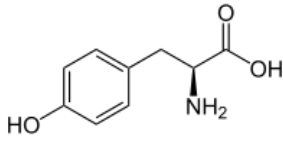
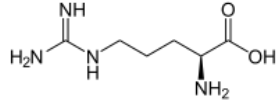
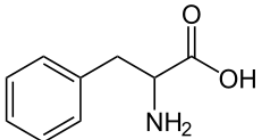
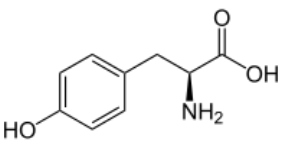
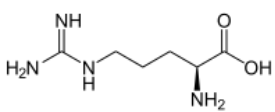
	1.0	0.9	0.3
	0.9	1.0	0.4
	0.3	0.4	1.0

AI approaches in a ligand based world..

AI – Property Prediction

- Classical approach (training)
 - Retrieve data (chemical structures + biological activity)
 - Standardize chemistry + convert to fingerprints
 - Train a machine learning model
 - Random Forests, Gradient Boosting, Support Vector Machines, Deep Neural networks
- Classical approach (application)
 - Apply to a chemical vendor database to identify novel compounds (virtual screening)
 - Use ML to generate list of probable protein targets for a given molecule (target prediction ; mode-of-action for natural products)

More similarity

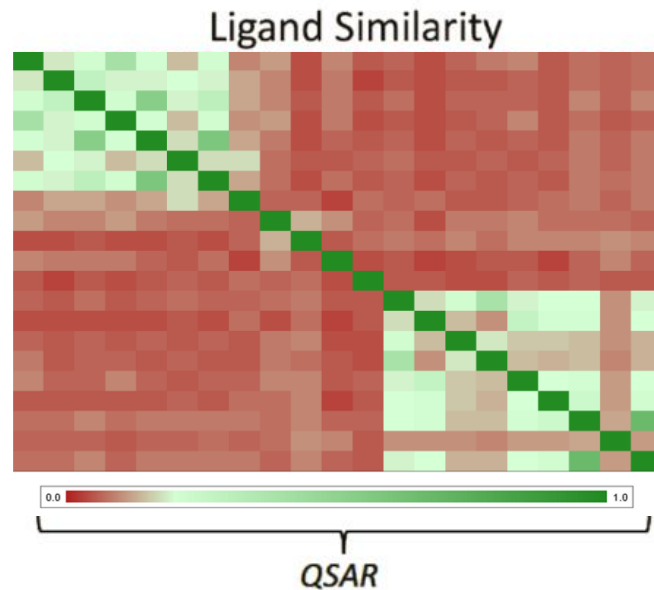
			
Phenylalanine	 <chem>N[C@@H](Cc1ccccc1)C(=O)O</chem>	 <chem>N[C@@H](Cc1ccc(O)cc1)C(=O)O</chem>	 <chem>N[C@@H](CCCCNC(=O)N)C(=O)O</chem>
	1.0	0.9	0.3
Tyrosine	0.9	1.0	0.4
Arginine	0.3	0.4	1.0

Sequence Similarity

	FYI	IYF	WTF
FYI	1.0	0.9	0.3
IYF	0.9	1.0	0.4
WTF	0.3	0.4	1.0

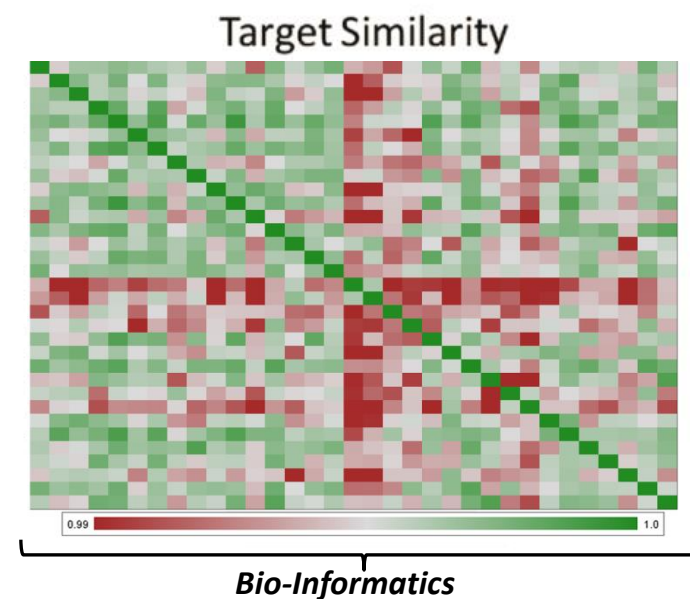
The how... *what is PCM ?*

- Proteochemometric modeling combines both a ligand descriptor and target descriptor



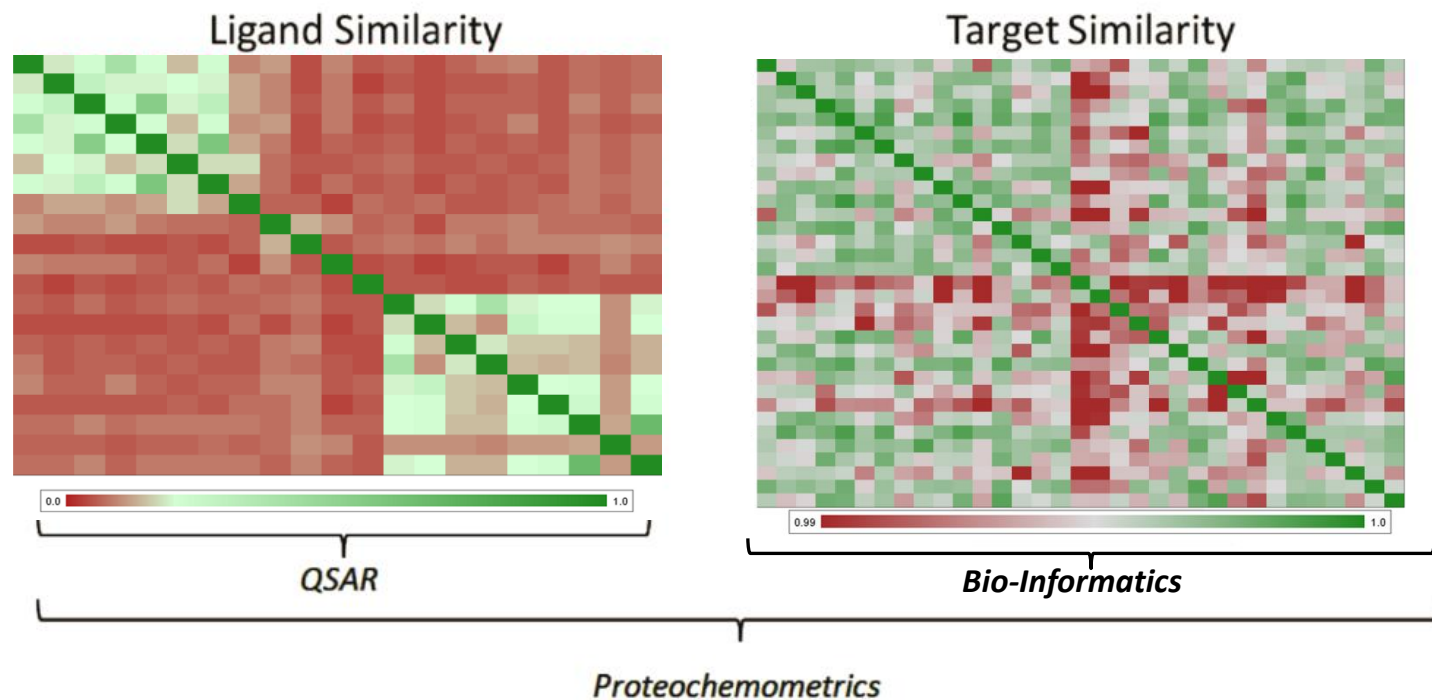
What is PCM ?

- Proteochemometric modeling combines both a ligand descriptor and target descriptor

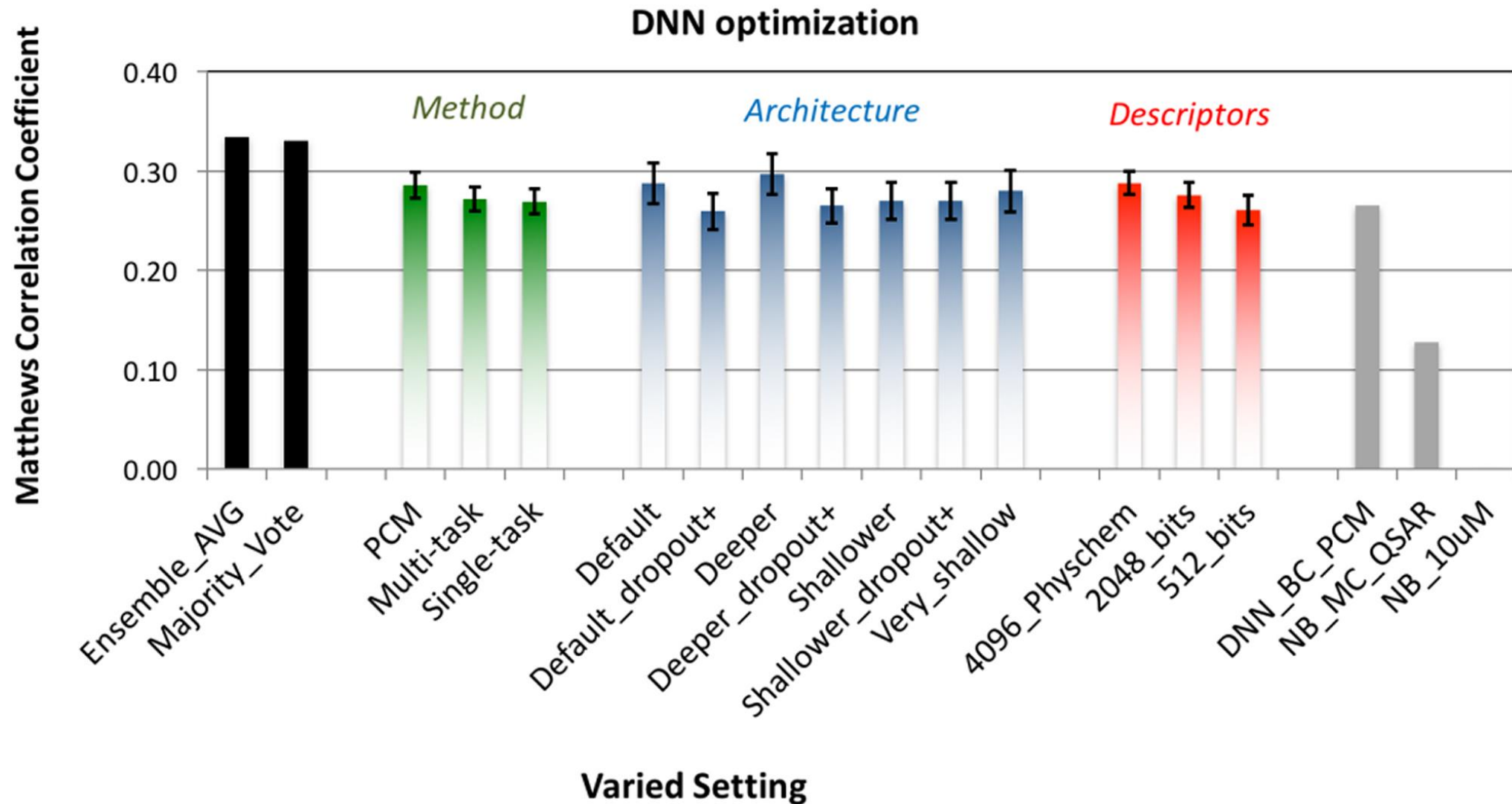


What is PCM ?

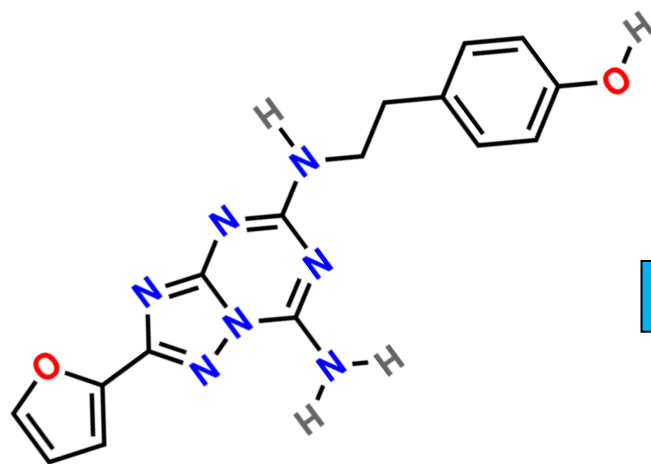
- Proteochemometric modeling combines both a ligand descriptor and target descriptor



Able to extrapolate to unseen compd / target combinations

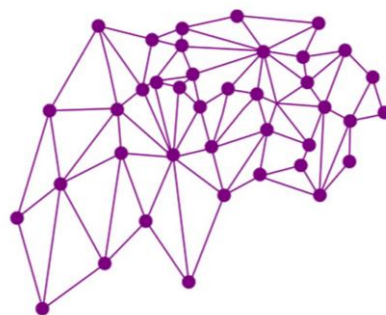
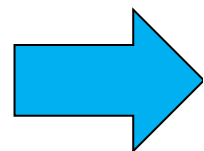


De novo generation
Learning a machine to suggest new molecules..

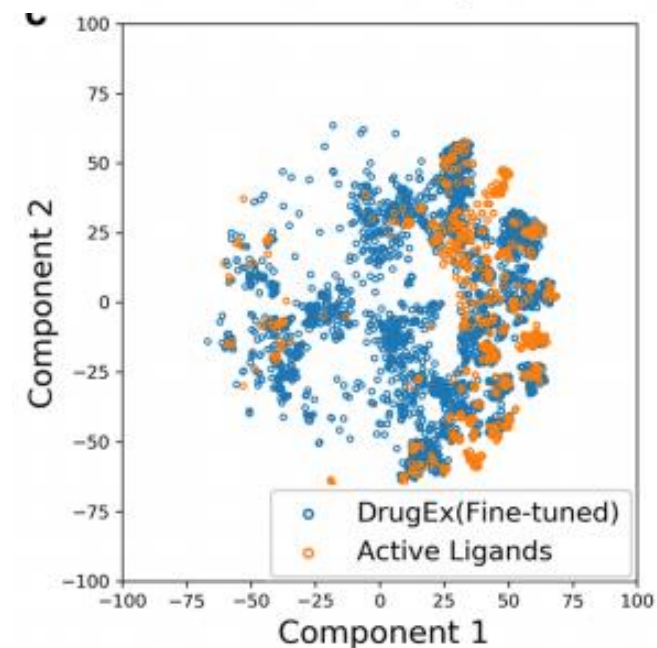
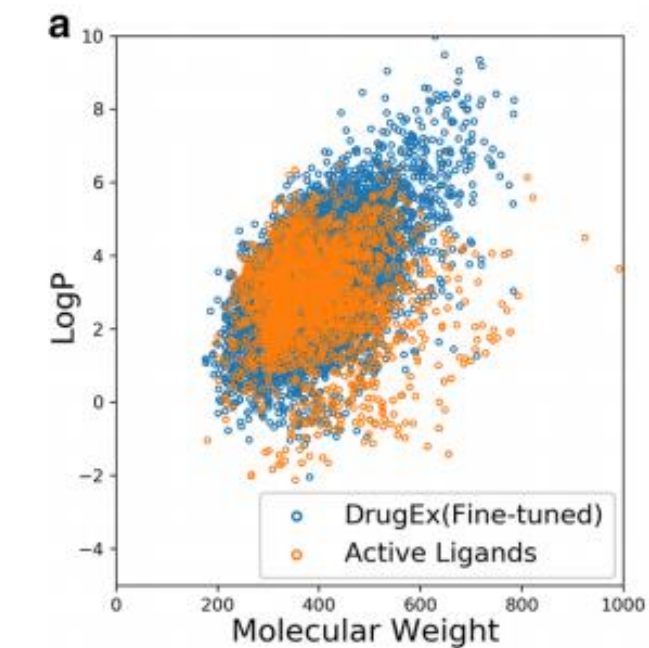
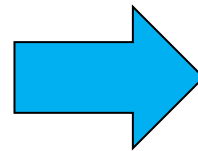


ZM241385

Nc1nc(NCCc2ccc(O)cc2)nc3nc(nn13)c4occc4

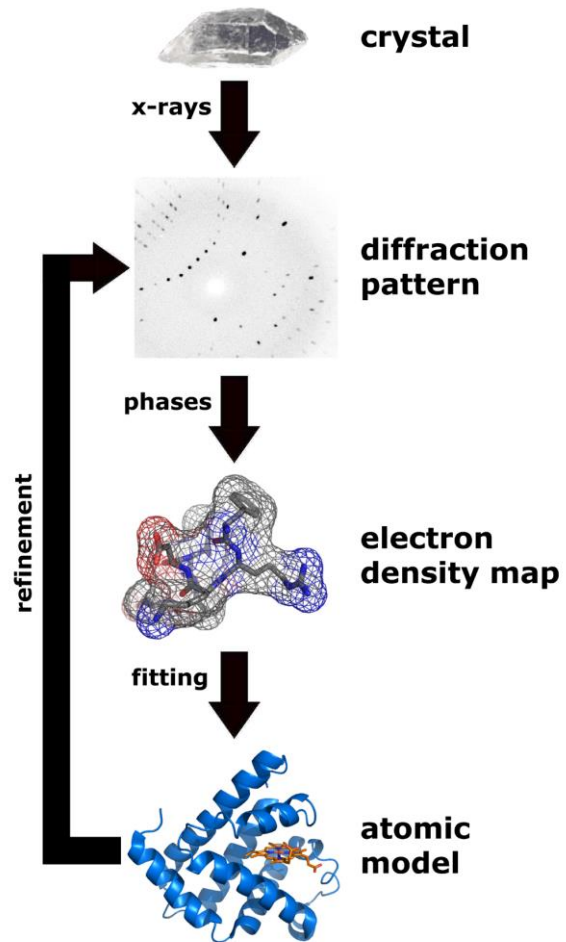


Machine Learning

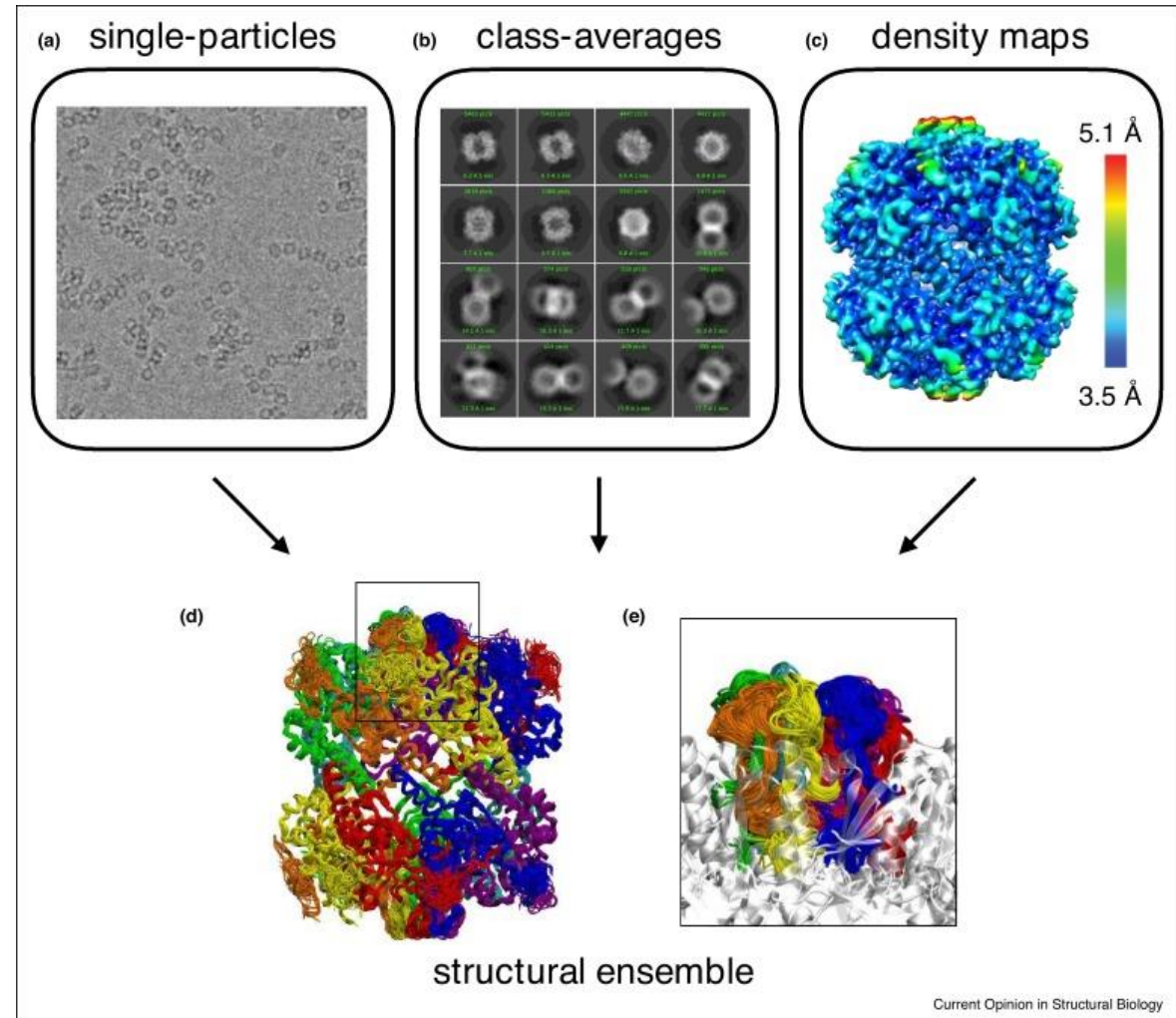


So far for ligand based...

Structure Based Modelling



[1] X-Ray Crystallography



[2] Cryo-EM

[1] Protopedia.org, T Splettstoesser.

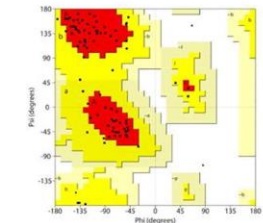
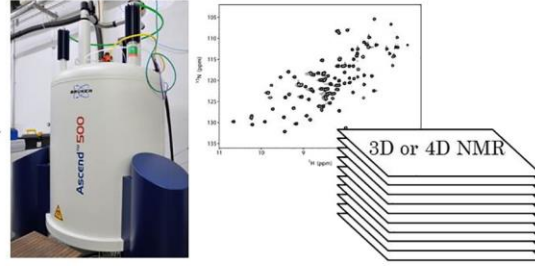
[2] Bonomi, M et al. (2019) Cur. Op. Struct. Bio., 37-45

Structure Based Modelling

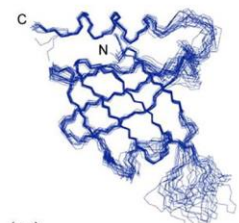
(A) Isotope-labeled protein preparation (section 2)



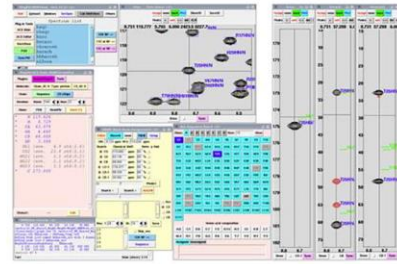
(B) NMR data collection (section 3)



(E) Validation of precision and accuracy of the structure models (section 6)

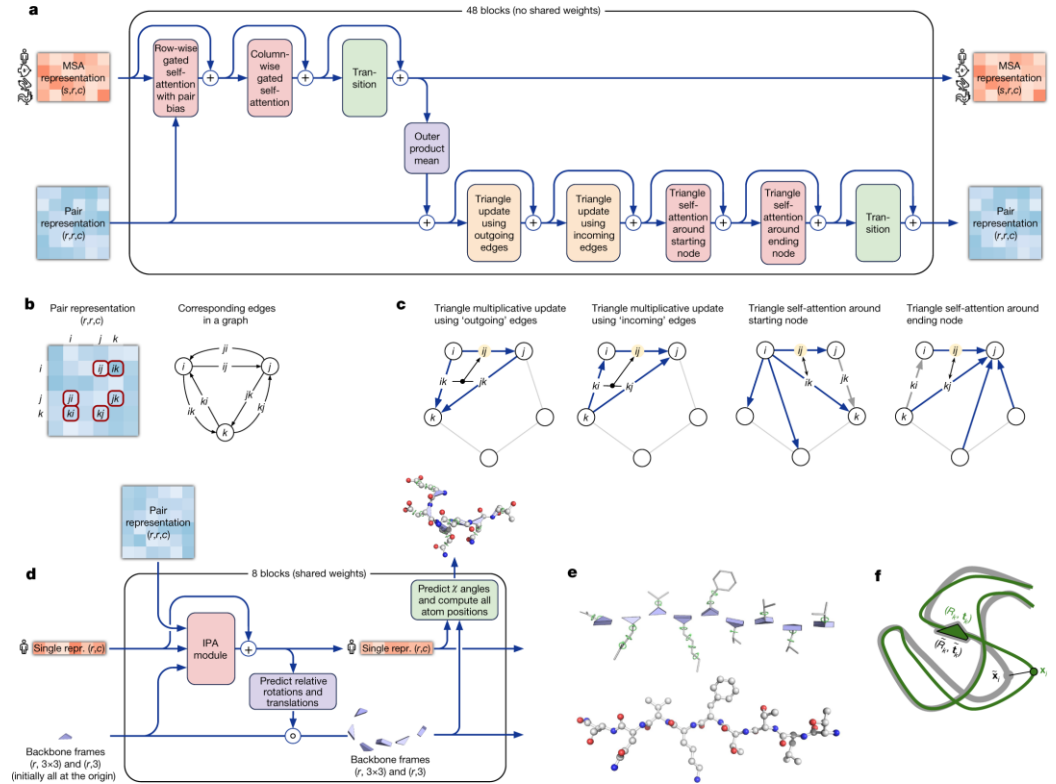


(D) Simulated annealing with NMR-based restraints e.g. ^1H - ^1H distance Dihedral angle (section 4, 5)



(C) Signal assignments (section 3)

[1] NMR



[2] AlphaFold2

[1] Sugiki T et al. (2017), *Comp. Struct. Bio. Jour.*, 15: 328-339

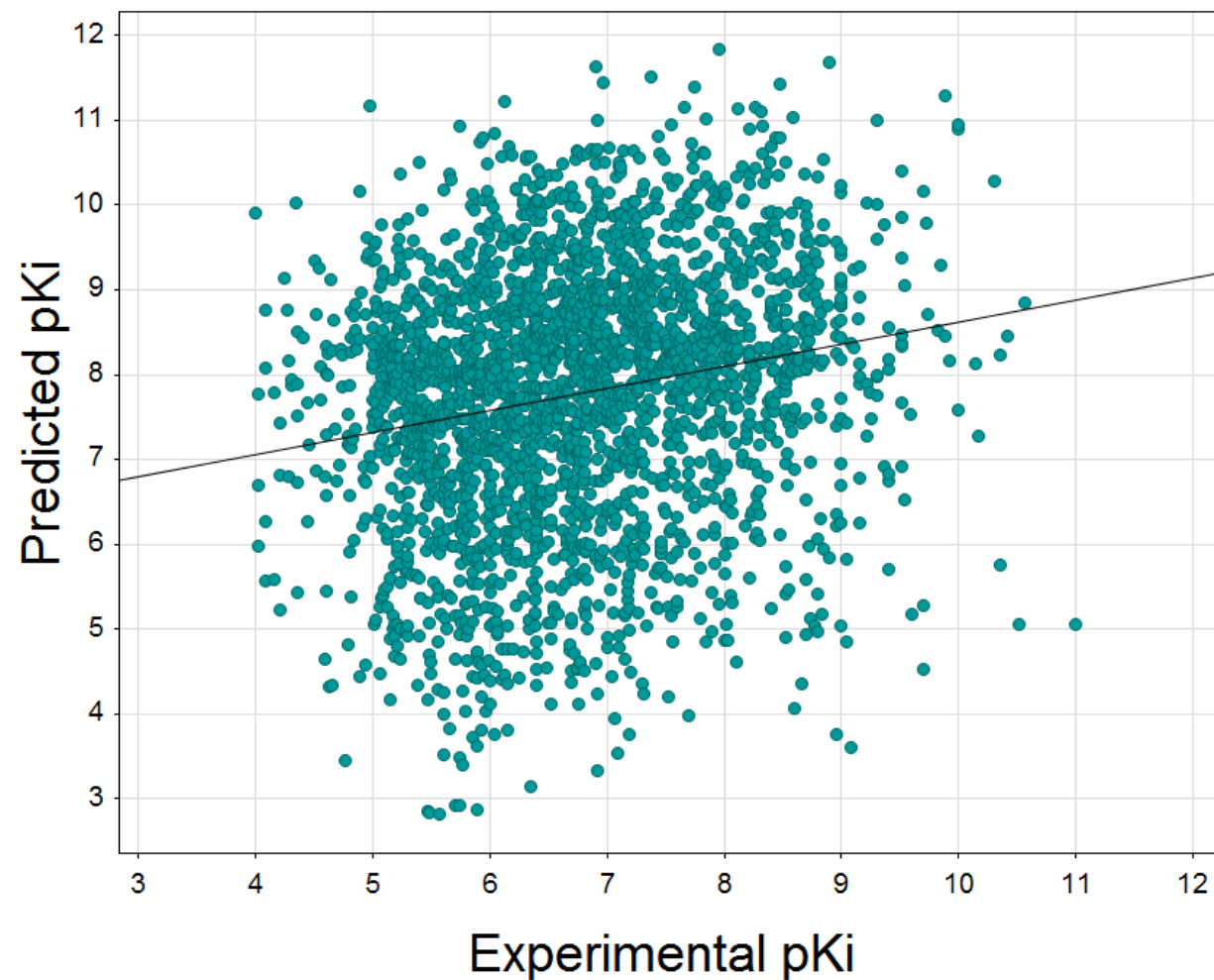
[2] Jumper J et al. (2021) *Nature* 596, 583-589

Structure Based Modelling

- So we can reliably predict the structures of a significant number of proteins
- However we only have data of a few hundred protein-ligand complexes (remember there are $10^{33} * 10^5$ potential combinations)
- We need a way to predict protein-ligand complexes if we want to screen compounds based on structures

Docking

Docking is great at sampling potential conformation, but really quite bad at predicting affinities

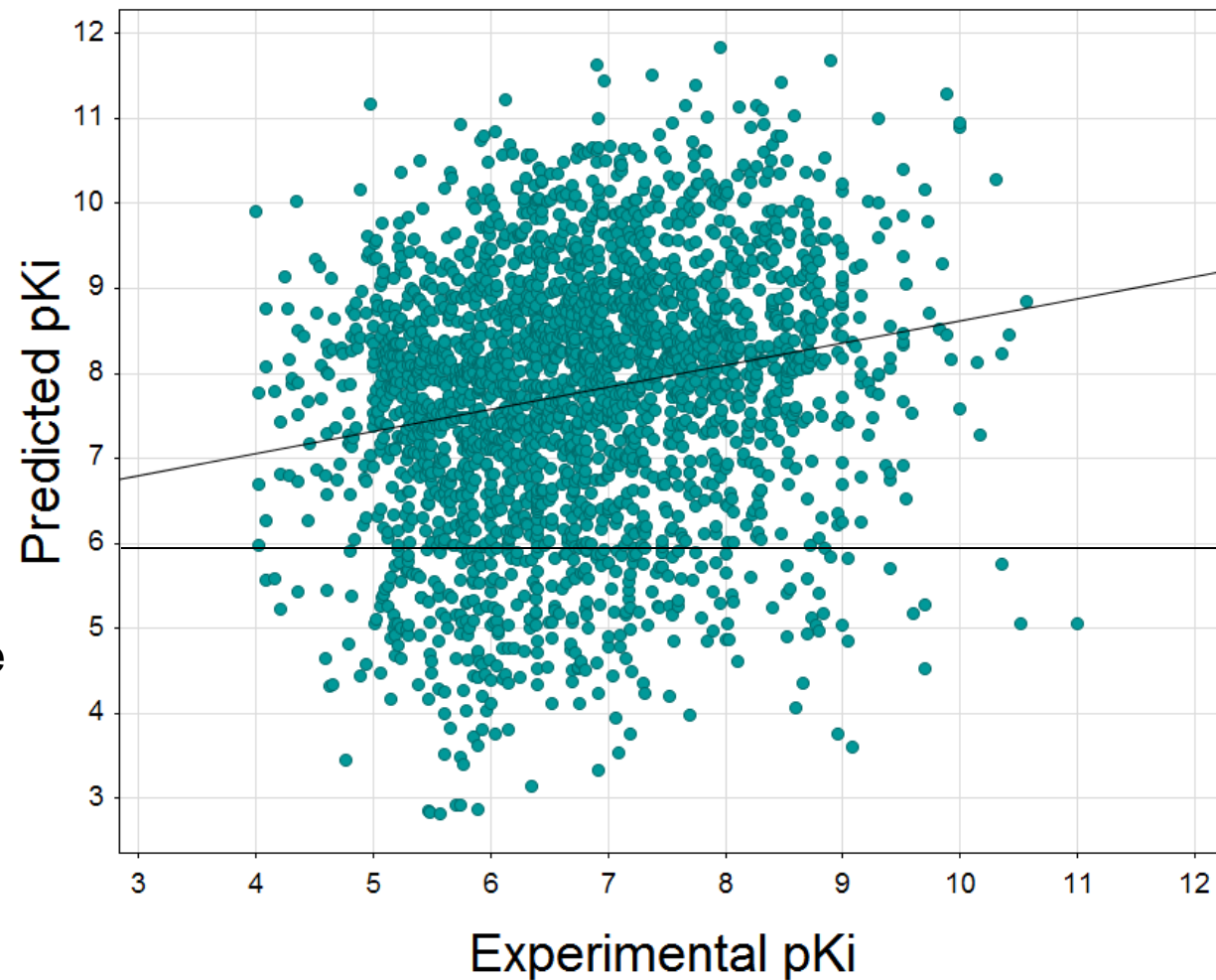


$R = 0.01$

Docking

Docking is great at sampling potential conformation, but really quite bad at predicting affinities

But we can at least use it to remove some of the inactive compounds



$R = 0.01$

Free Energy Perturbation (FEP)



Jhonny:
High affinity for wet lab



Willem:
Low affinity for wet lab



Jhonny:
High affinity for wet lab



Jhilly:
50/50



Willem:
Low affinity for wet lab



Jhonny:

High affinity for wet lab



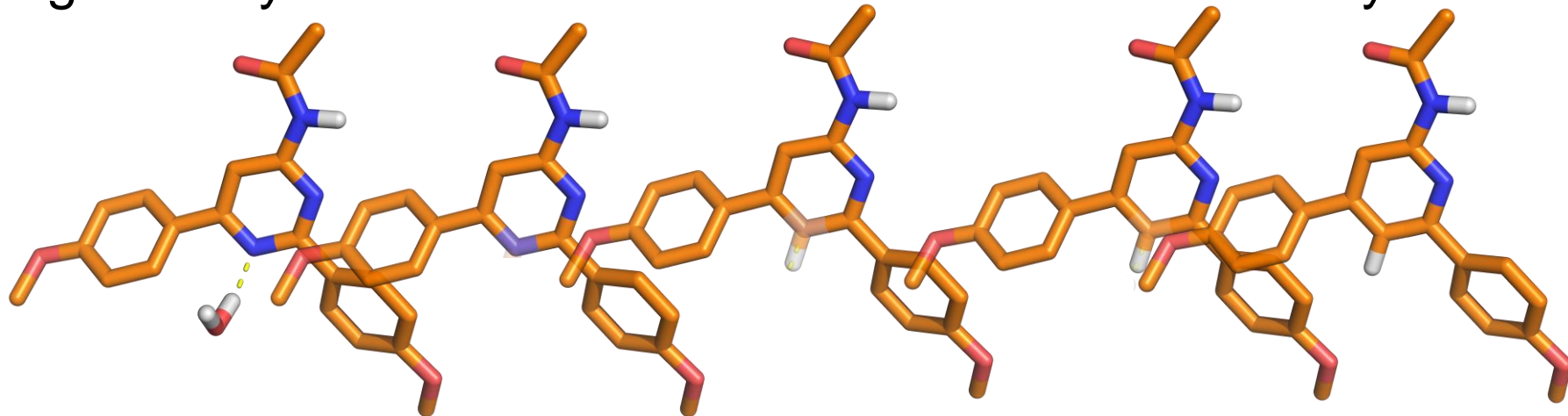
Jhilly:

50/50



Willem:

Low affinity for wet lab



High affinity for receptor

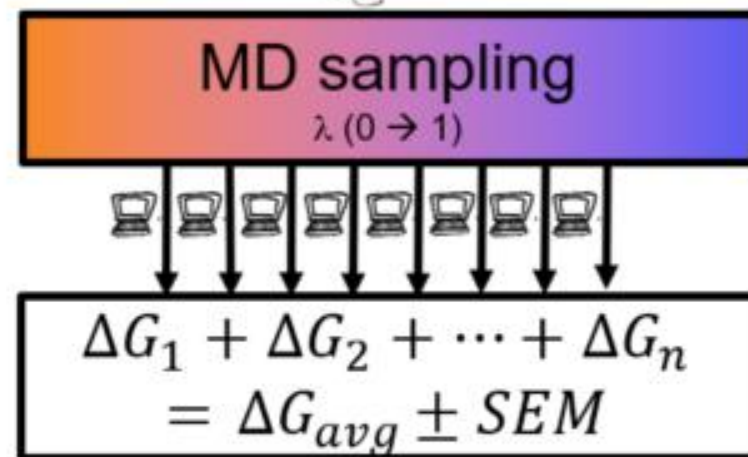
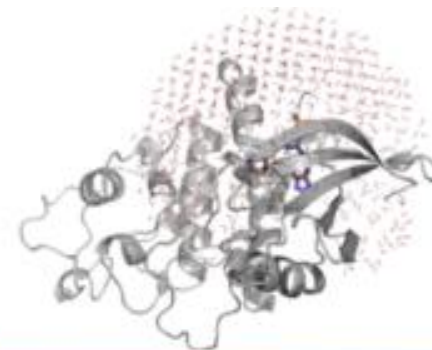
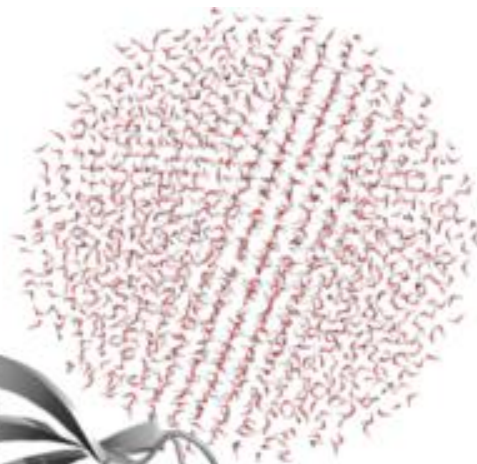
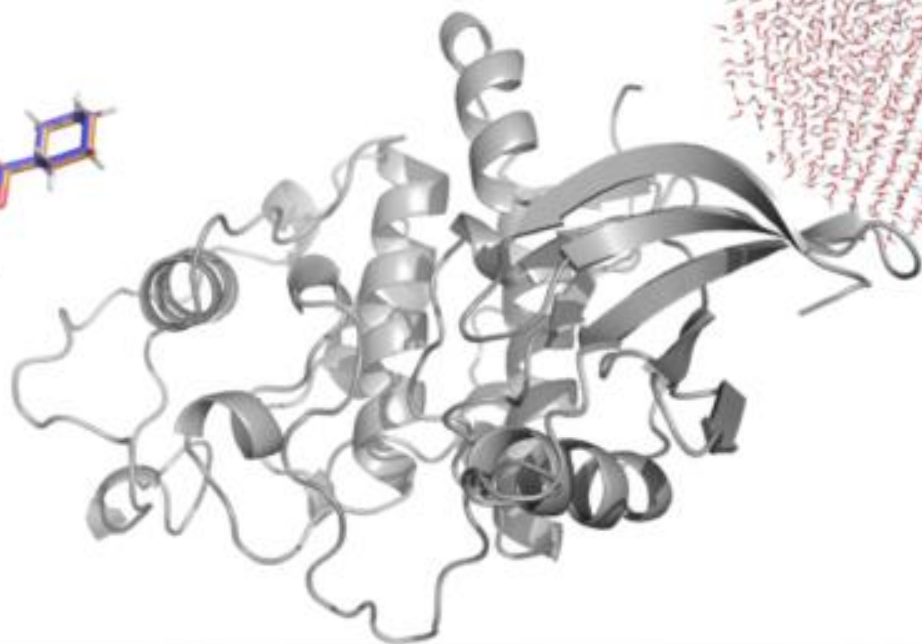
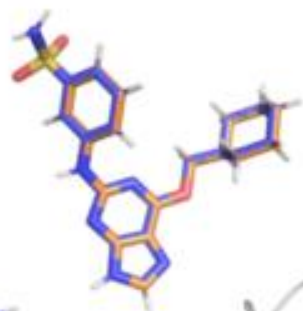


Low affinity for receptor



QligFEP

High-throughput ligand FEP



Dual topology
ligand FEP

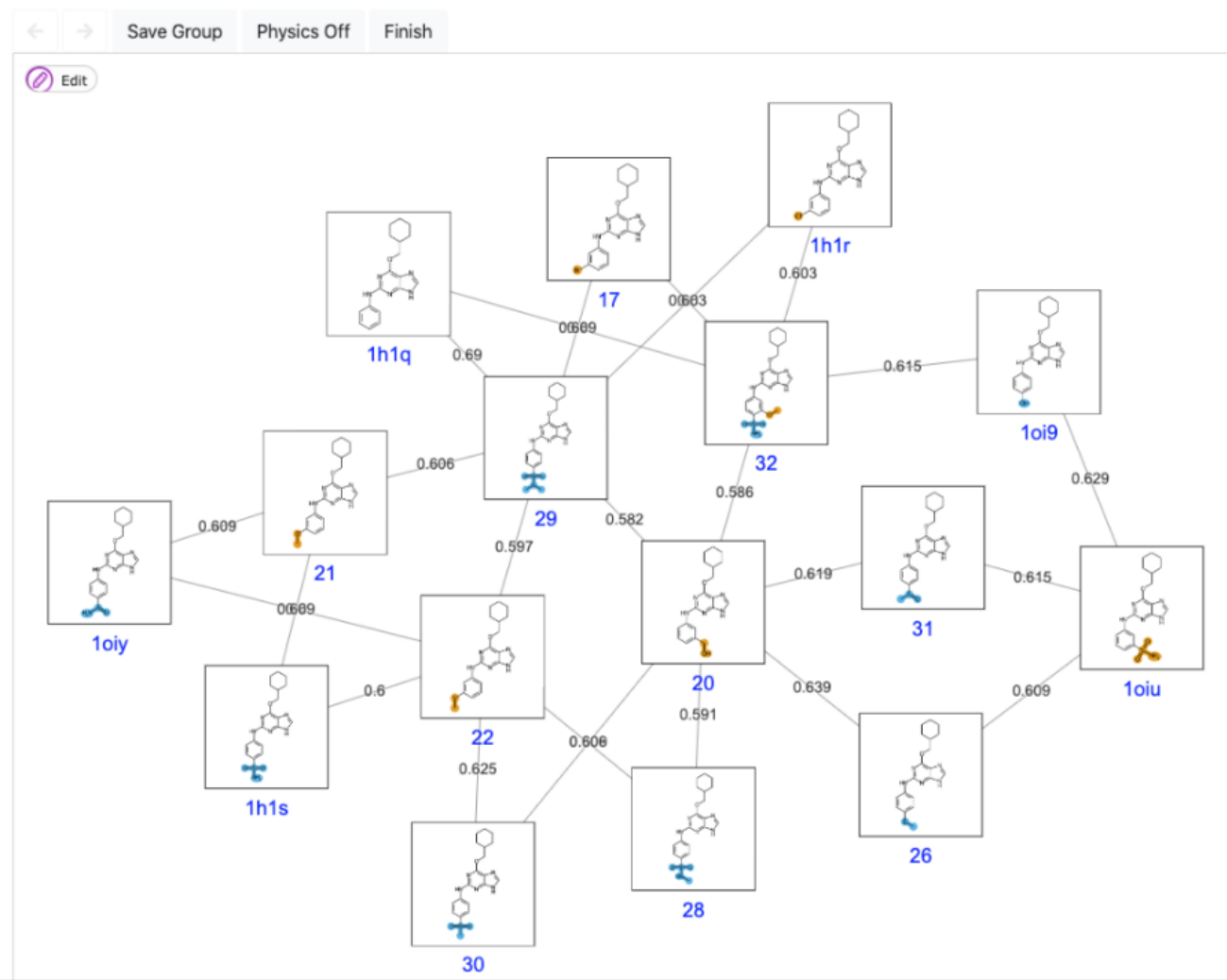
Binding site
simulations

Robust,
automated



QligFEP

High-throughput ligand FEP



QmapFEP

Pairwise perturbations that cover a dataset

Fingerprints, similarity, Cycle closure correction

~1.5 FEP / ligand

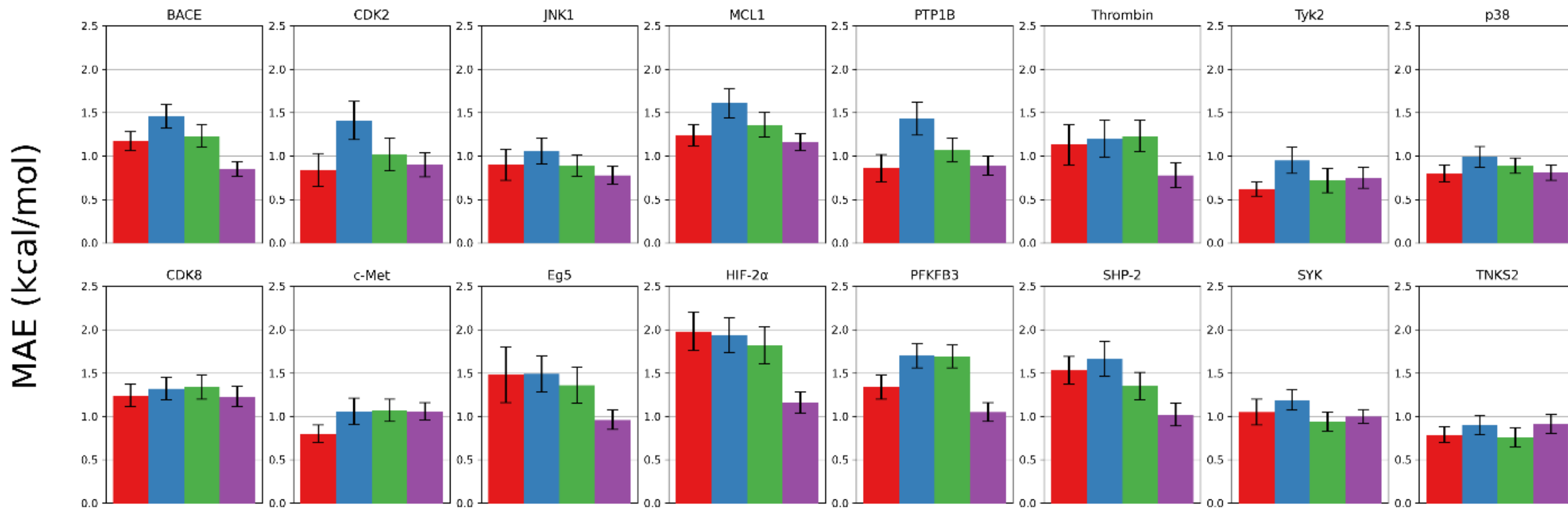


QligFEP

Benchmarking



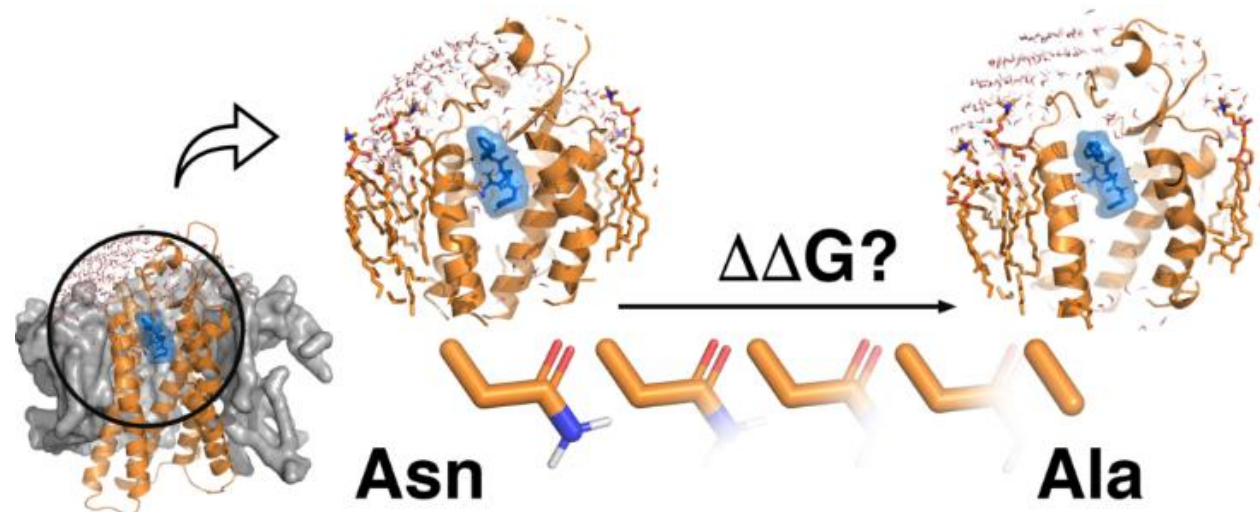
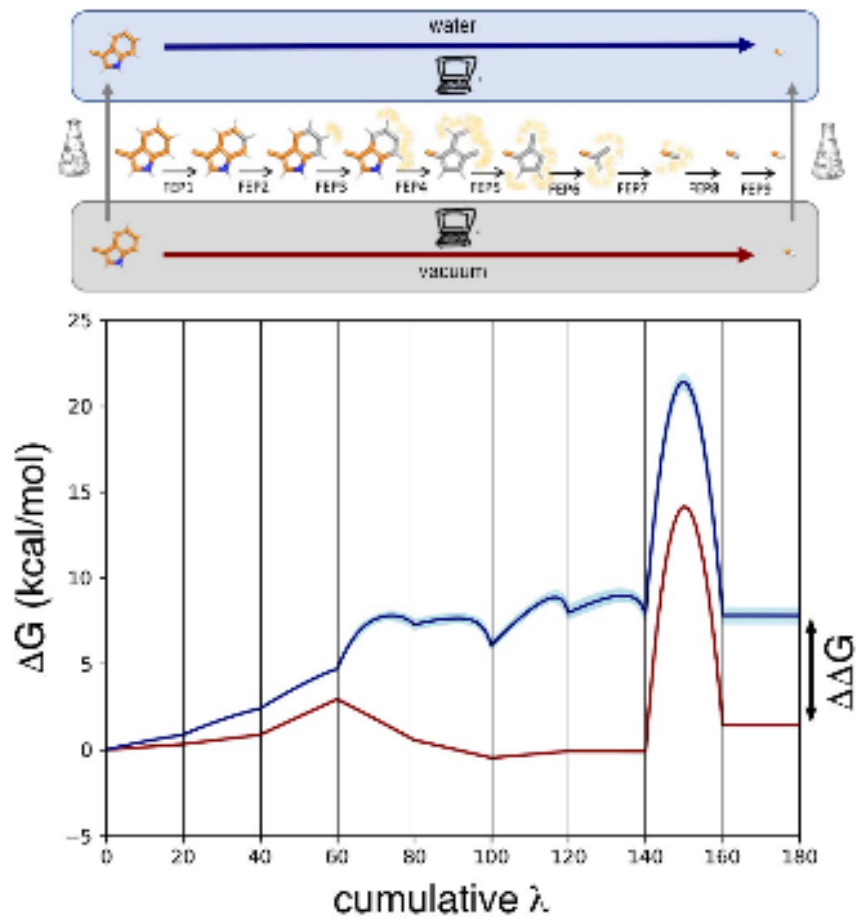
 QligFEP (openFF1.2)  QligFEP (openFF2.0)  QligFEP (OPLS2015)  FEP+ (OPLS3e)





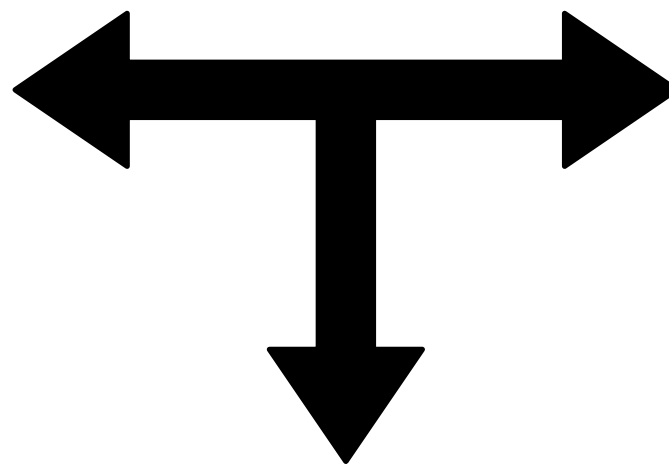
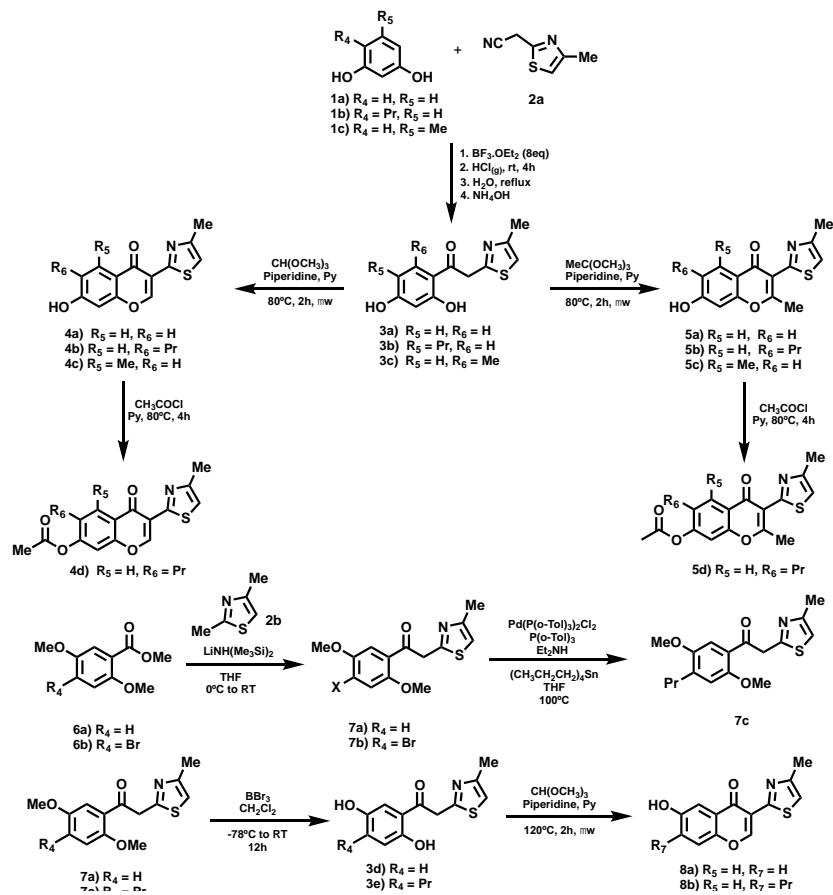
QresFEP

In silico mutagenesis

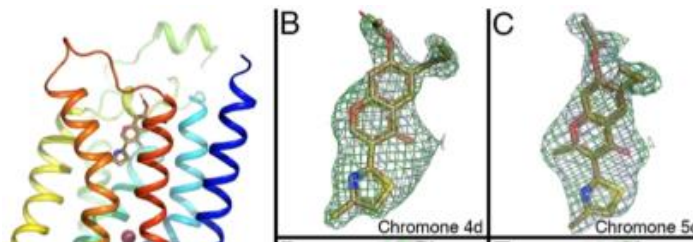


First principle FEP to evaluate mutagenesis effects

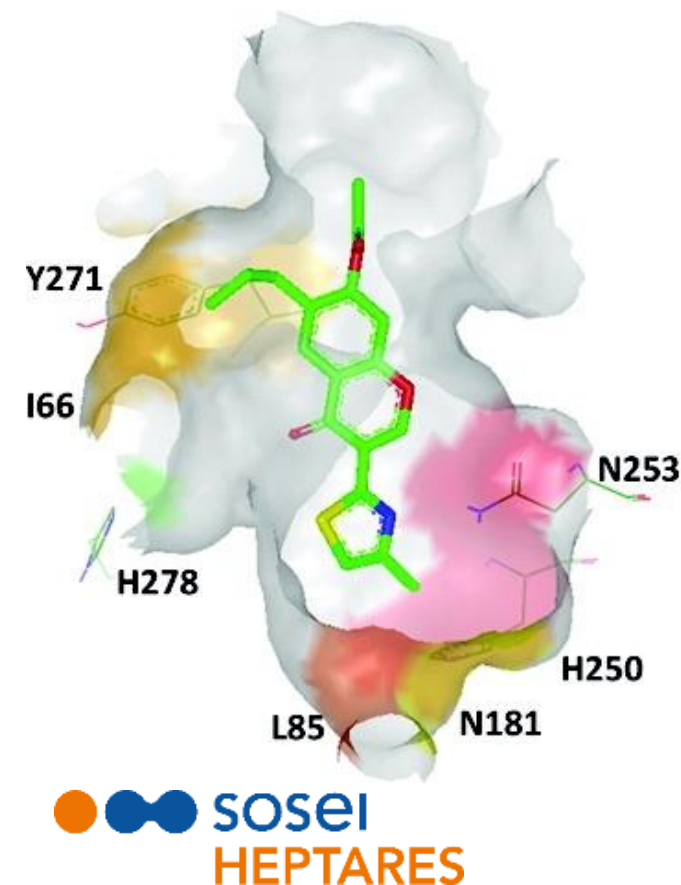
Ligand synthesis



Crystal structure



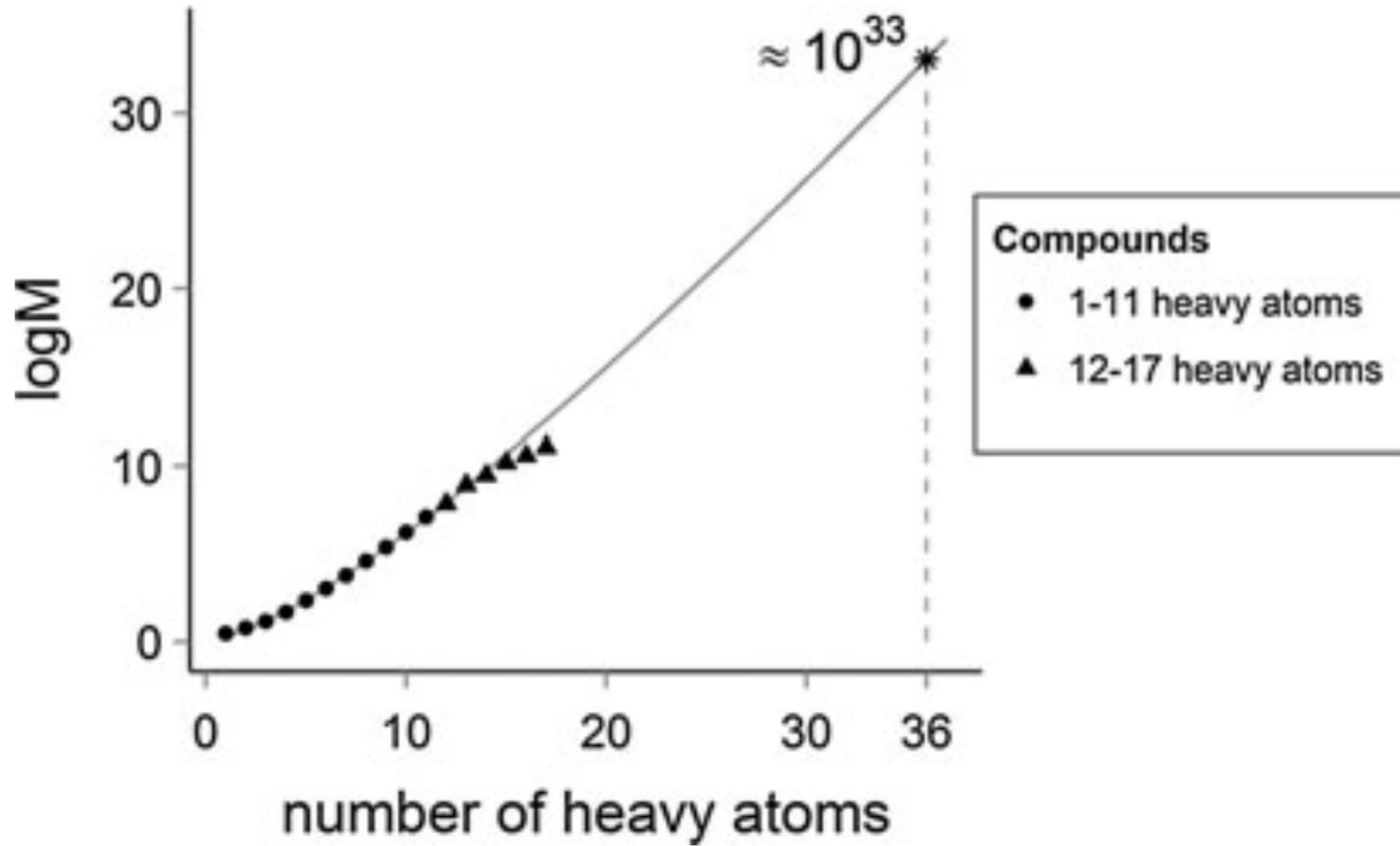
Protein mutations



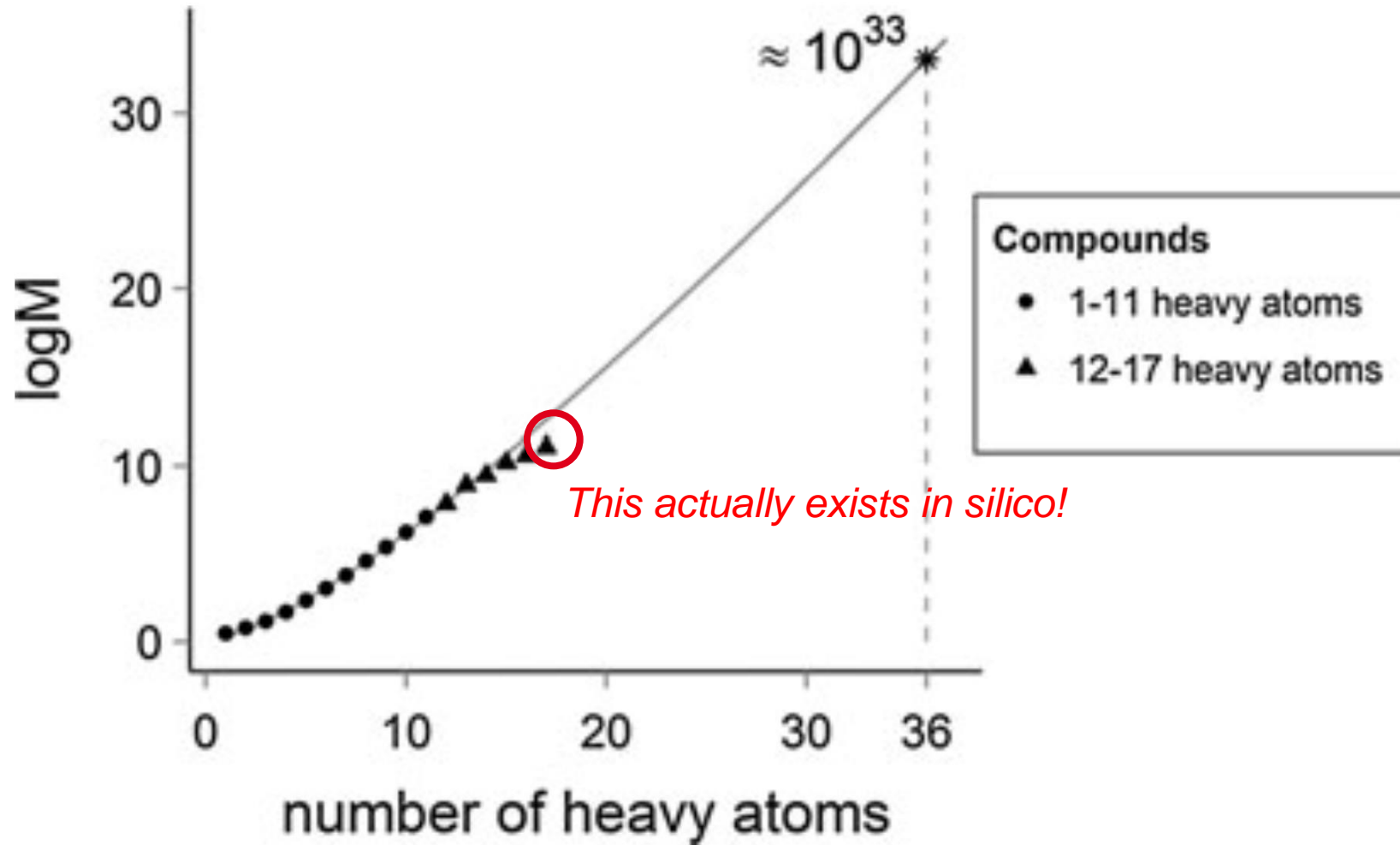
The data is out there ...



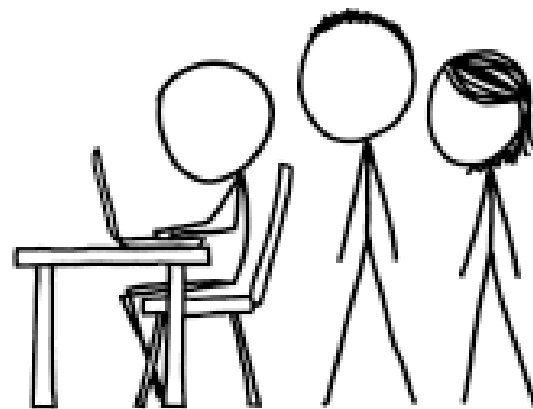
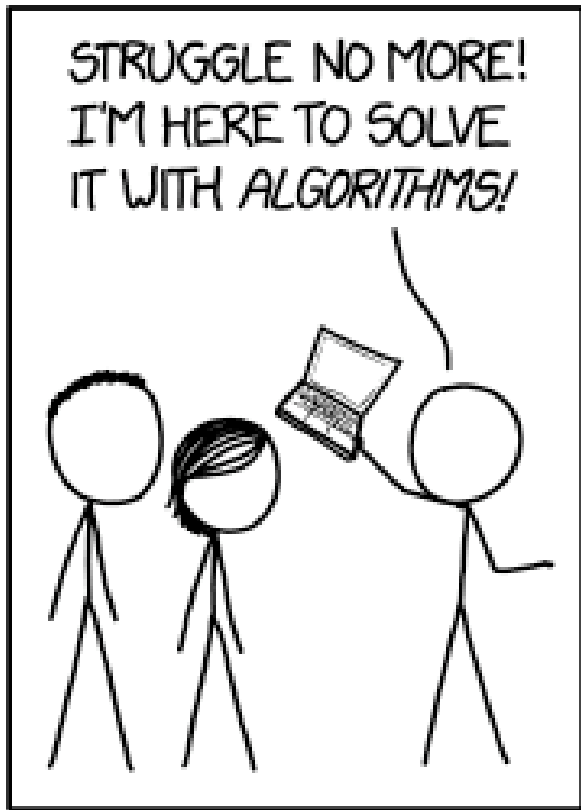
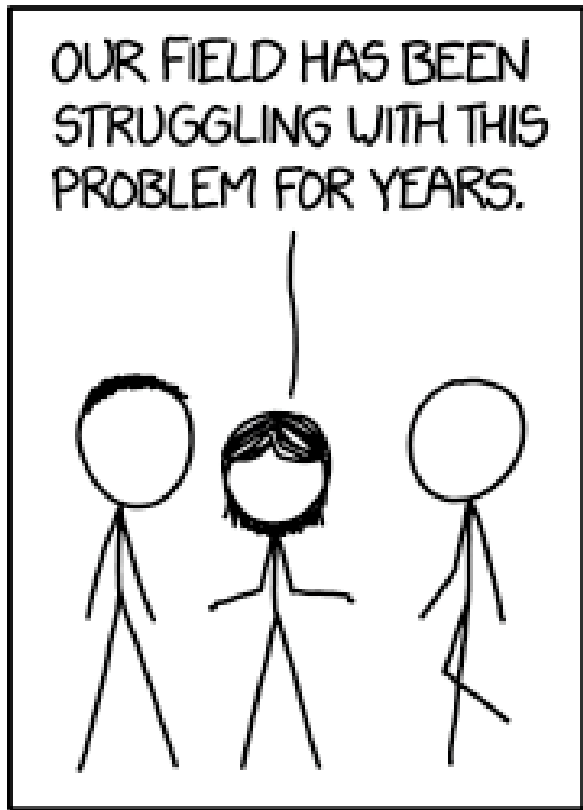
Why 10^{33} ?



Why 10^{33} ?



GDB-17, 166 billion molecules... 10^{11}



So where do we get the data?

Synthesis, Pharmacological Characterization, and Docking Analysis of a Novel Family of Diarylisoaxazoles as Highly Selective Cyclooxygenase-1 (COX-1) Inhibitors

Paola Vitale,^{†,¶} Stefania Tacconelli,^{§,⊥,¶} Maria Grazia Perrone,^{†,¶} Paola Malerba,[†] Laura Simone,[†] Antonio Scilimati,^{*,†} Antonio Lavecchia,^{*,‡} Melania Dovizio,^{§,⊥} Emanuela Marcantoni,^{§,⊥} Annalisa Bruno,^{||,⊥} and Paola Patrignani^{*,§,⊥}

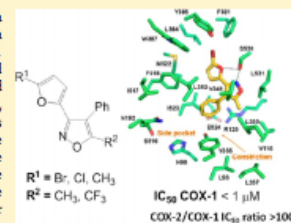
[†]Dipartimento di Farmacia-Scienze del Farmaco, Università degli Studi di Bari "A. Moro", Via Orabona 4, 70125 Bari, Italy

[‡]Dipartimento di Farmacia, "Drug Discovery" Laboratory, Università di Napoli "Federico II", Via D. Montesano 49, 80131 Napoli, Italy

[§]Department of Neuroscience and Imaging, ^{||}Department of Medicine and Aging, "G. d'Annunzio" University, and [⊥]Center of Excellence on Aging (CeSI), Chieti, Italy

Supporting Information

ABSTRACT: 3-(5-Chlorofuran-2-yl)-5-methyl-4-phenylisoaxazole (P6), a known selective cyclooxygenase-1 (COX-1) inhibitor, was used to design a new series of 3,4-diarylisoaxazoles in order to improve its biochemical COX-1 selectivity and antiplatelet efficacy. Structure–activity relationships were studied using human whole blood assays for COX-1 and COX-2 inhibition *in vitro*, and results showed that the simultaneous presence of 5-methyl (or -CF₃), 4-phenyl, and 5-chloro(-bromo or -methyl)furan-2-yl groups on the isoaxazole core was essential for their selectivity toward COX-1. **3g**, **3s**, and **3d** were potent and selective COX-1 inhibitors that affected platelet aggregation *in vitro* through the inhibition of COX-1-dependent thromboxane (TX) A₂. Moreover, we characterized their kinetics of COX-1 inhibition. **3g**, **3s**, and **3d** were more potent inhibitors of platelet COX-1 and aggregation than P6 (named **6**) for their tighter binding to the enzyme. The pharmacological results were supported by docking simulations. The oral administration of **3d** to mice translated into preferential inhibition of platelet-derived TXA₂ over protective vascular-derived prostacyclin (PGI₂).



INTRODUCTION

Cyclooxygenase-1 (COX-1) and cyclooxygenase-2 (COX-2) catalyze the first step of the biosynthesis of prostanoids from arachidonic acid (AA).¹ Different from COX-2 gene that is mainly inducible, COX-1 is a housekeeping gene constitutively expressed in almost all mammalian tissues and cells.² However, the expression of COX-1 can be regulated in some circumstances, such as during development.² In physiological conditions, COX-1 is highly expressed in the gastrointestinal (GI) tract and platelets,^{3,4} where it is involved in the generation of cytoprotective prostaglandin (PG)E₂ and platelet proaggregatory thromboxane (TX) A₂, respectively.¹

COX-1 plays a role in several pathological conditions such as thrombosis, atherosclerosis, and tumorigenesis.^{5–10} Importantly, platelet COX-1 is the target of one of the most efficacious antithrombotic agents used for prevention of vascular occlusive events, *i.e.*, aspirin.¹¹

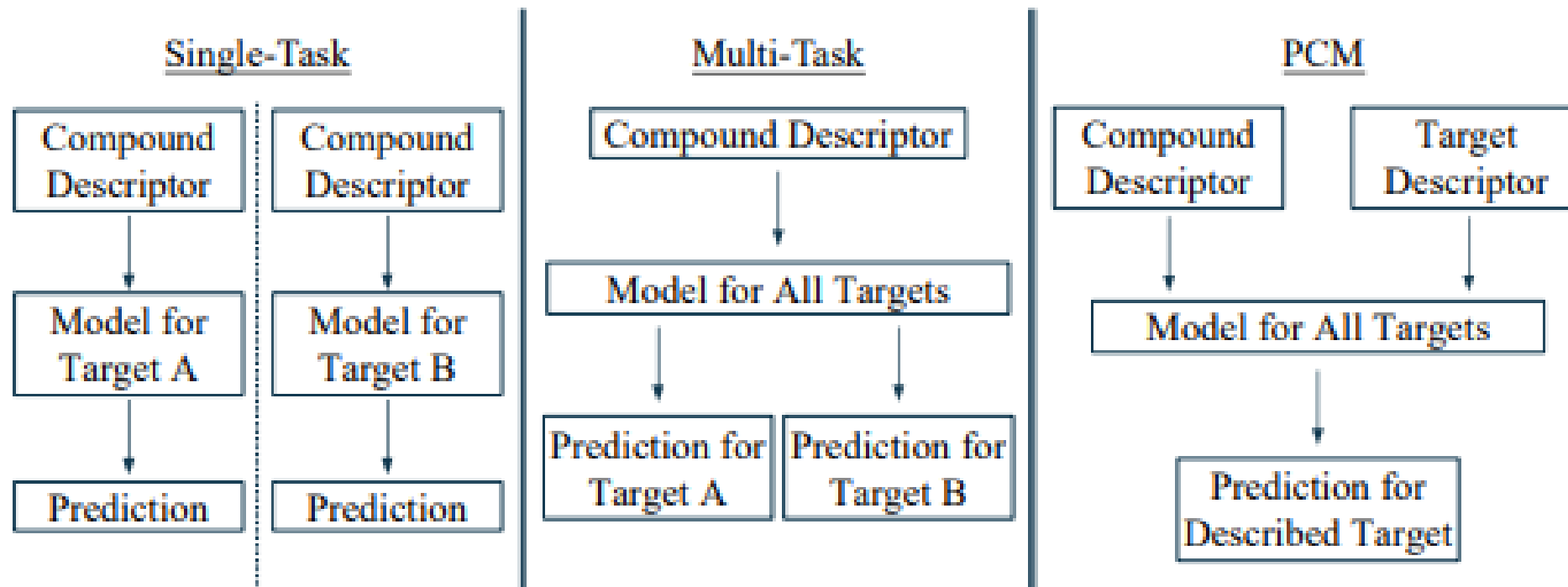
Aspirin irreversibly acetylates Ser529 and Ser516 of COX-1 and COX-2, respectively,¹² leading to irreversible enzyme inactivation. Because of pharmacokinetics features (*i.e.*, short half-life of 20 min) and pharmacodynamics features (higher

potency to inhibit COX-1 than COX-2), low doses of aspirin (75–100 mg once daily) act by affecting platelet COX-1 activity while they cause only a marginal and transient inhibitory effect on COX-2 and extra-platelet cellular COX-1. Low doses of aspirin cause an almost complete suppression (≥95%) of platelet TXA₂ generation *ex vivo*, persisting throughout the dosing interval (*i.e.*, 24 h).^{11,13} This is a fundamental requisite to obtain an antiplatelet effect,¹⁴ since even tiny concentrations of TXA₂ may activate platelets and importantly they synergize with other platelet agonists.¹⁵ The antiplatelet effect of aspirin is strictly related to platelet turnover.^{11,12} Thus, enhanced platelet turnover rate detected in some cardiovascular (CV) conditions, such as diabetes, might decrease the efficacy of the drug to halt almost completely platelet TXA₂ generation.¹⁶ Moreover, we have recently shown that platelets contain COX-1 mRNA and the enzymatic machinery for protein synthesis; thus, these anucleated megakaryocyte fragments are able to synthesize *de novo* COX-1 in response to platelet activation *in vitro*.¹⁷ Thus,

Received: December 28, 2012

Published: May 7, 2013

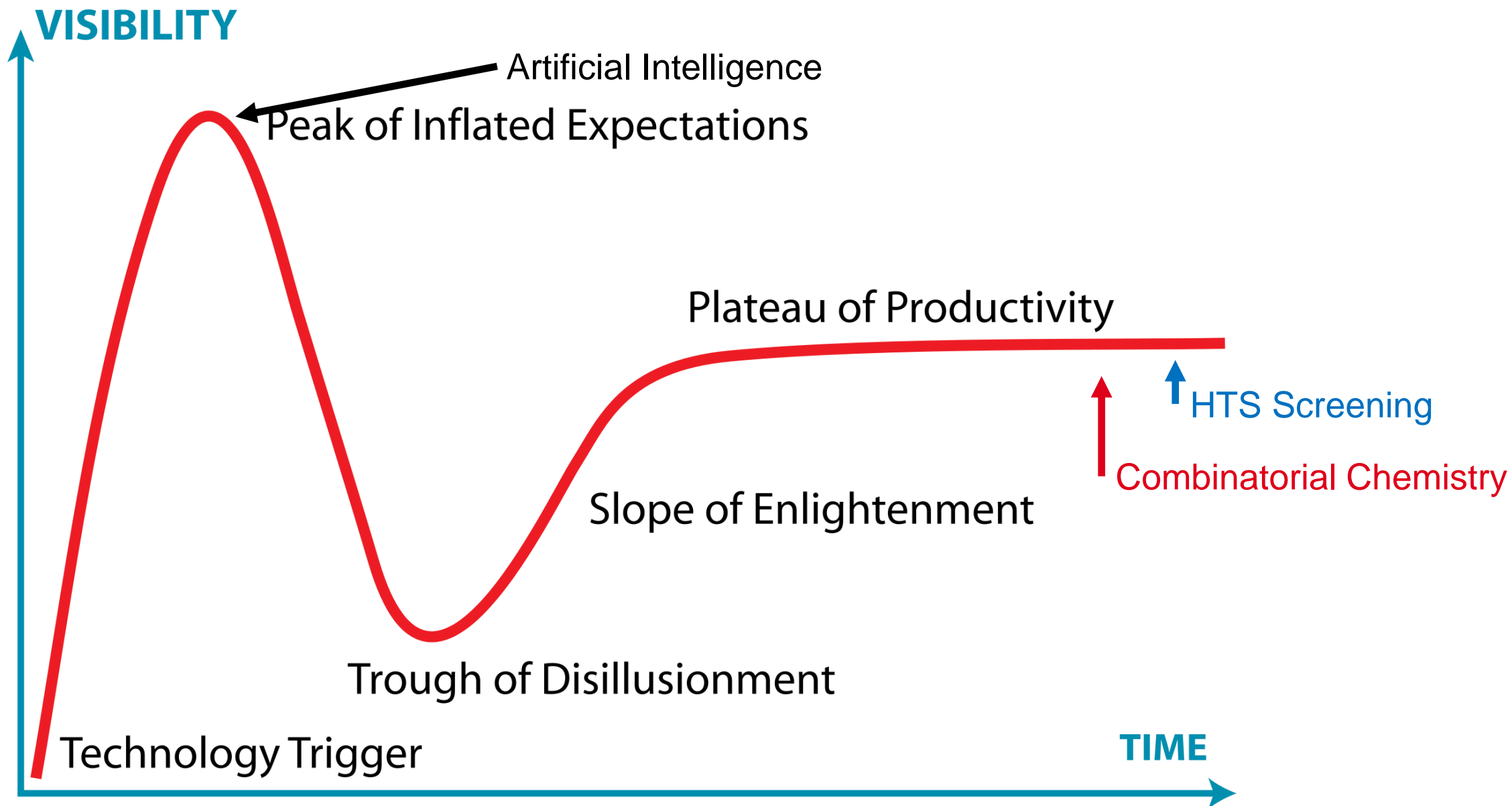
Multiple ways to model single bioactivity set..



What is 'Artificial Intelligence'?

- Artificial Intelligence is like teenage sex:
 - Everybody is talking about it
 - Nobody really knows how to do it
 - Everyone thinks everyone else is doing it
 - ...so everyone claims they are doing it...

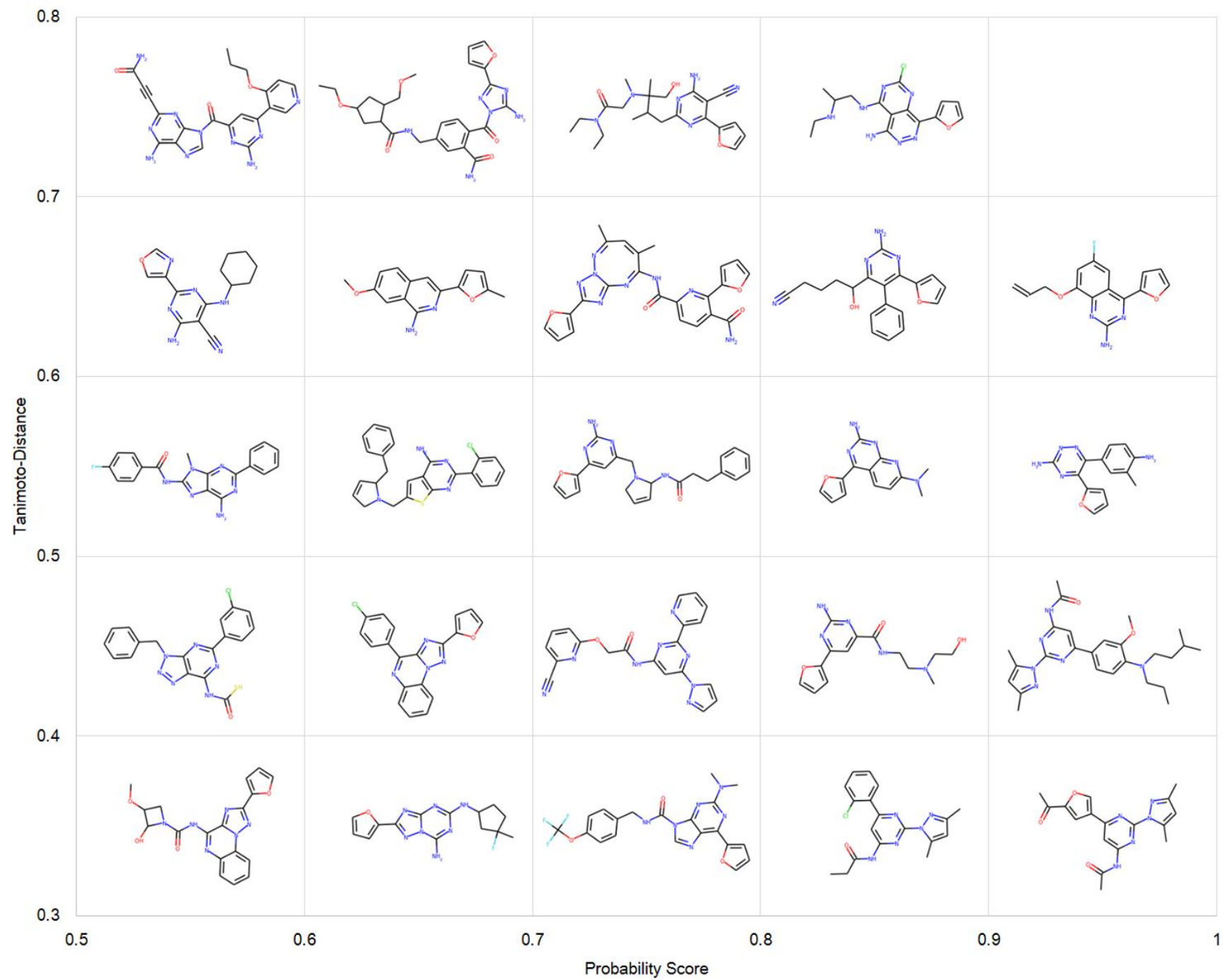
-Dan Ariely



Approximation, not real data...

Adaptation of Jeremy Kemp, Wikimedia Commons, 'Gartner Hype Cycle.svg', CC-BY

Sample



Jaccard or Tanimoto index & distance

- Tanimoto similarity (index)

$$\frac{3}{8} = 0.375$$

- Tanimoto distance

$$1 - 0.375 = 0.625$$

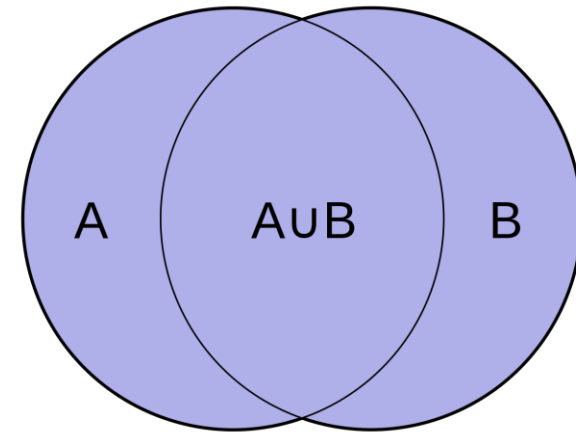
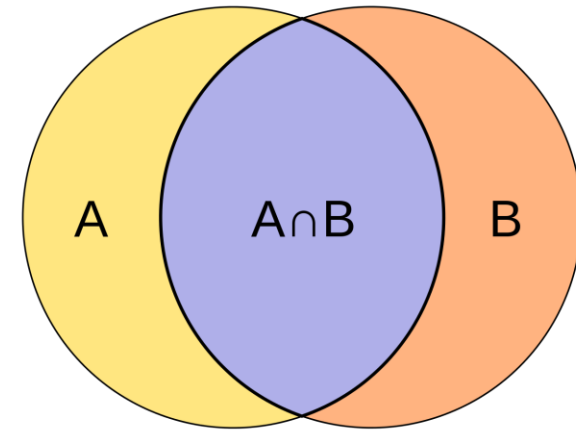
A: 10101001101

B: 10010101100

\cap : 3

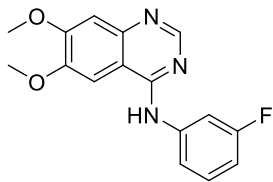
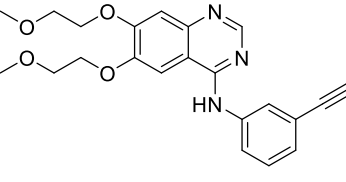
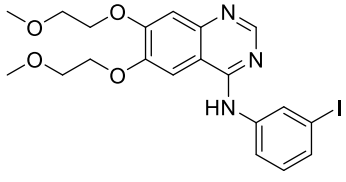
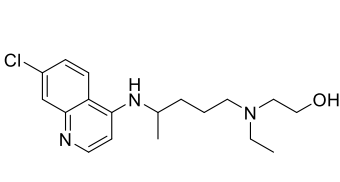
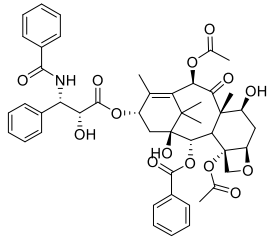
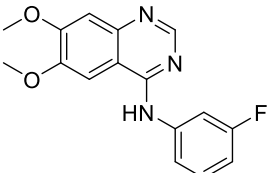
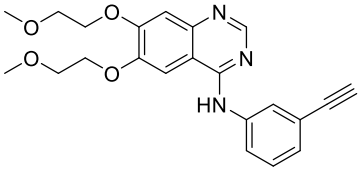
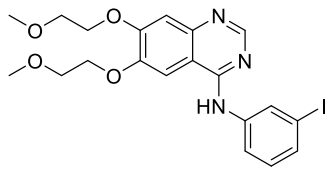
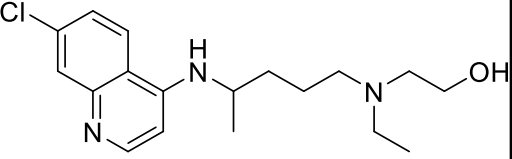
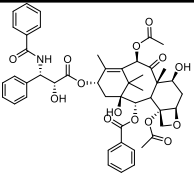
\cup : 8

$$index = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$



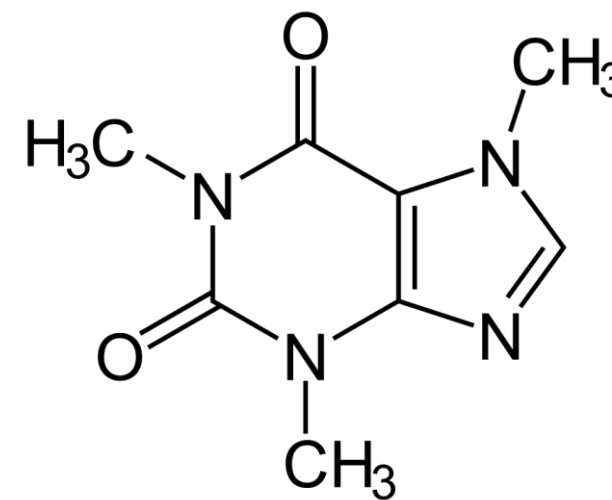
Similarities: examples

ECFP4, r=2

Molecule					
	1,00	0,54	0,57	0,14	0,06
	0,54	1,00	0,71	0,14	0,06
	0,57	0,71	1,00	0,15	0,07
	0,14	0,14	0,15	1,00	0,08
	0,06	0,06	0,07	0,08	1,00

IUPAC Naming

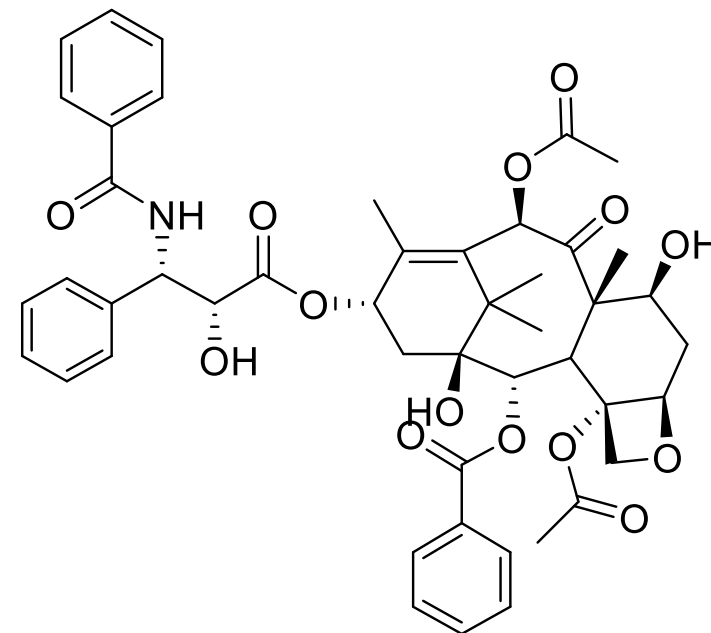
- Systematic naming convention
- Subject to changes
- Inefficient
- Complicated encoding and decoding



1,3,7-Trimethylpurine-2,6-dione

IUPAC Naming

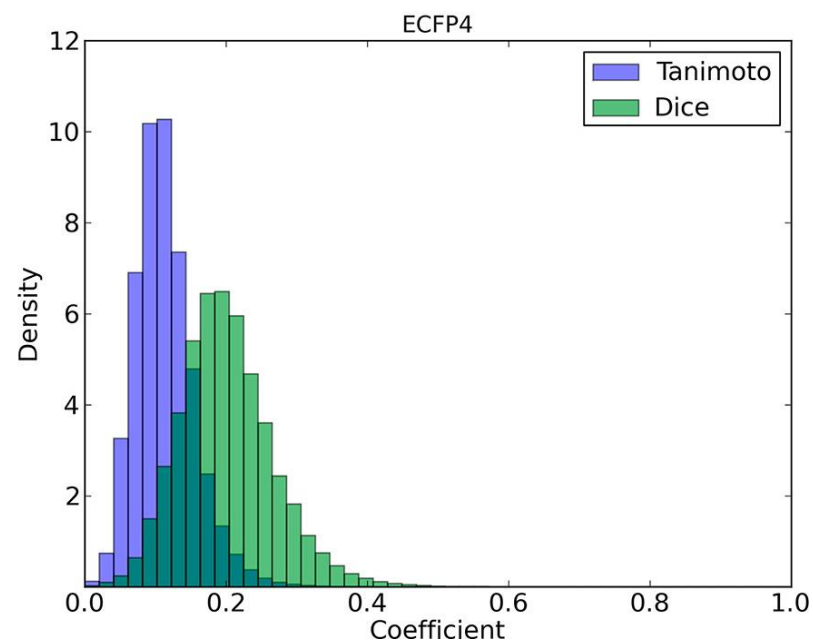
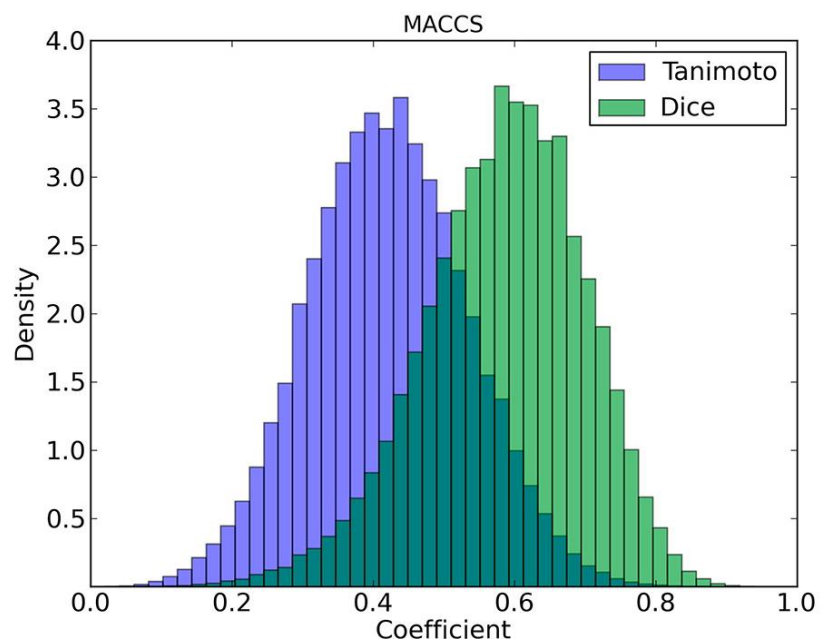
- Systematic naming convention
- Subject to changes
- Inefficient
- Complicated encoding and decoding



(2aR,4S,4aS,6R,9S,11S,12S,12bS)-9-(((2R,3S)-3-benzamido-2-hydroxy-3-phenylpropanoyl)oxy)-12-(benzoyloxy)-4,11-dihydroxy-4a,8,13,13-tetramethyl-5-oxo-3,4,4a,5,6,9,10,11,12,12a-decahydro-1*H*-7,11-methanocyclodecal[3,4]benzo[1,2-b]oxete-6,12b(2aH)-diyl diacetate

Fingerprint dependent

- Coefficients between 1 mln randomly selected molecules
- MACCS vs ECFP4



$$\begin{aligned} \text{Tanimoto similarity} &= \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \end{aligned}$$

$$\begin{aligned} \text{Dice similarity} &= \frac{2|A \cap B|}{|A| + |B|} \end{aligned}$$

What is wrong

- IUPAC provides API to translate graph to InChI string
- Encoding is thus 'flawless': everyone uses same system
- Decoding is software dependant: small differences possible
- Canonicalization is where things go south

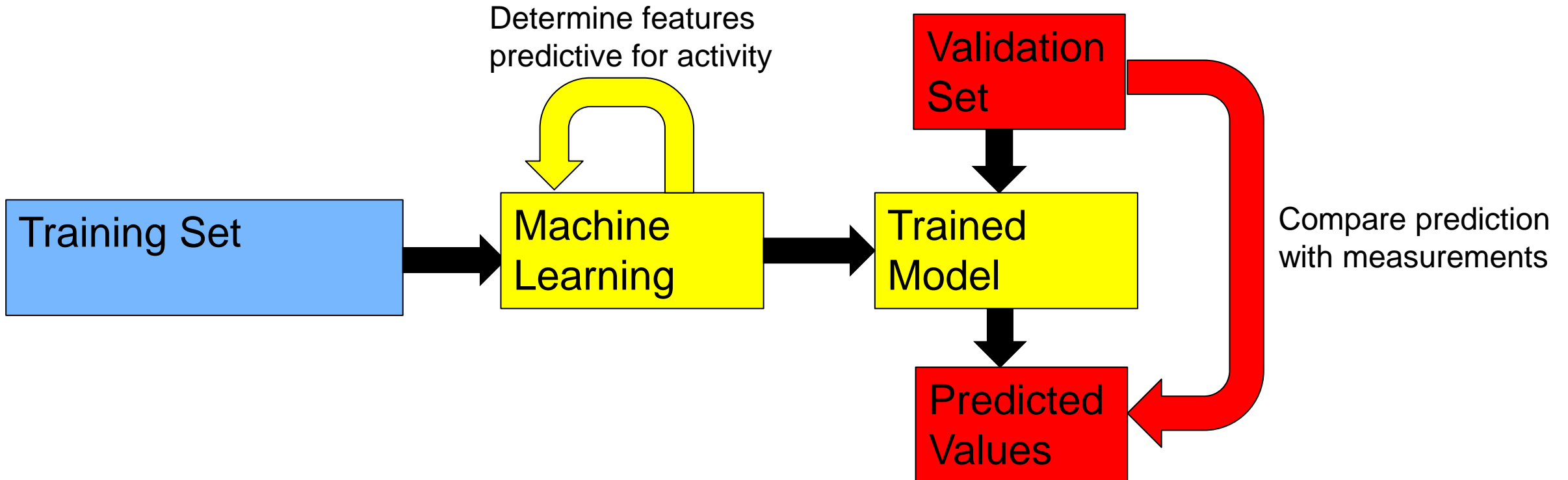
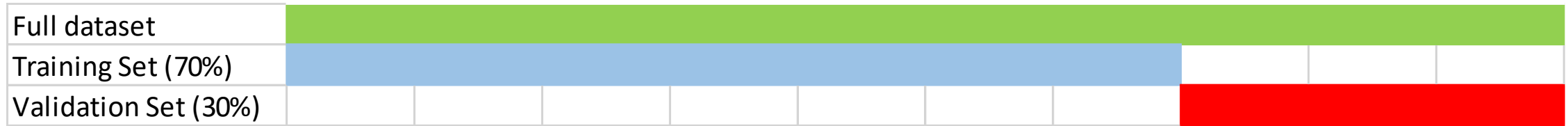
Downsides

- Indices are symmetrical
 - A vs. B is same as B vs. A
- Comparison with third molecule is difficult
 - $A \langle \rangle B = 0.71$
 - $B \langle \rangle C = 0.64$
 - $A \langle \rangle C = ??$
- Unintuitive decline (sharp drop)

Molecular Similarity in Medicinal Chemistry

<https://doi.org/10.1021/jm401411z>

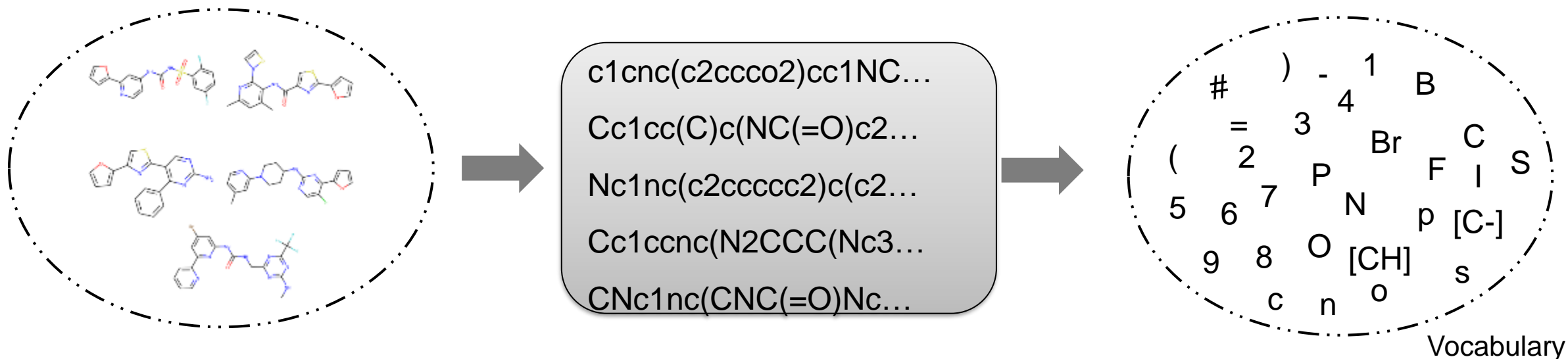
Training and validation



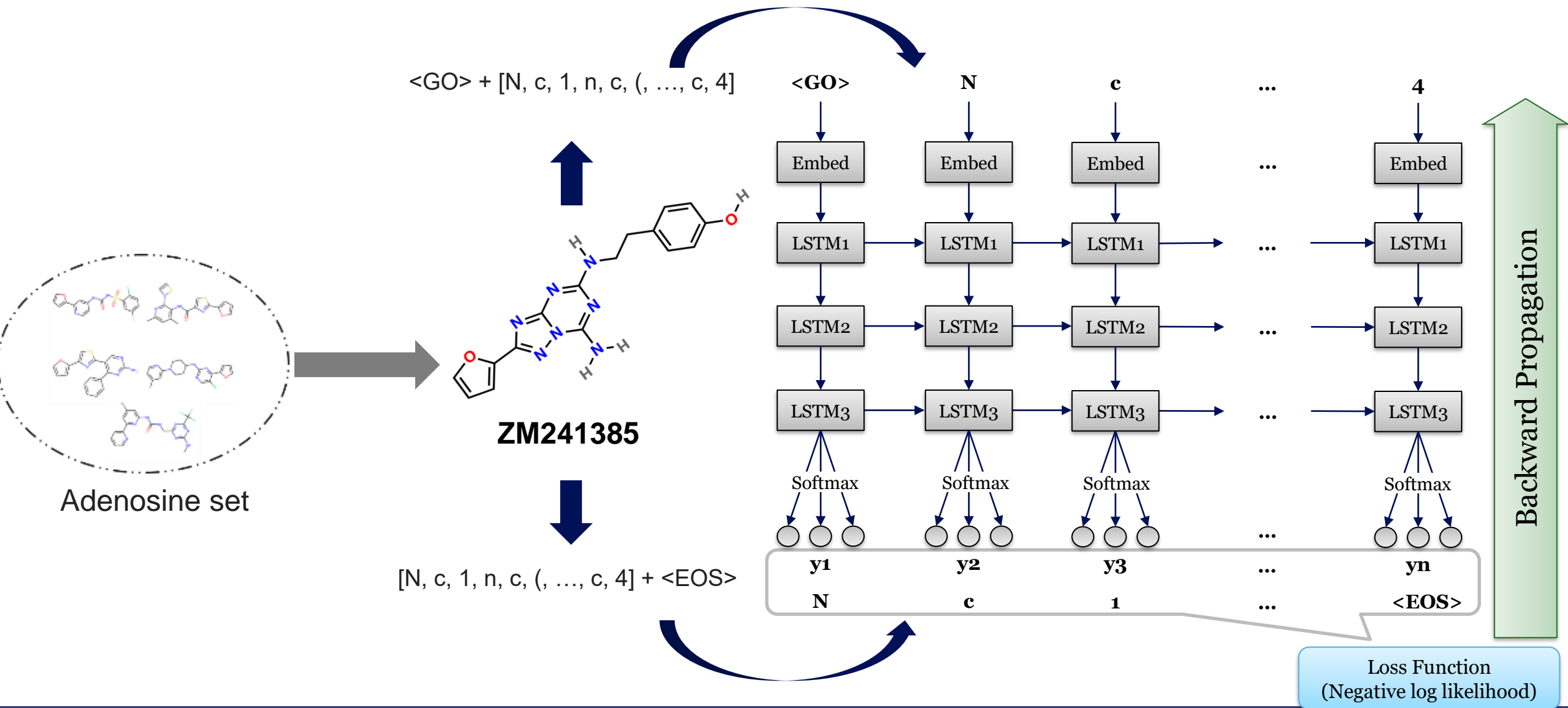
Dataset

Adenosine dataset

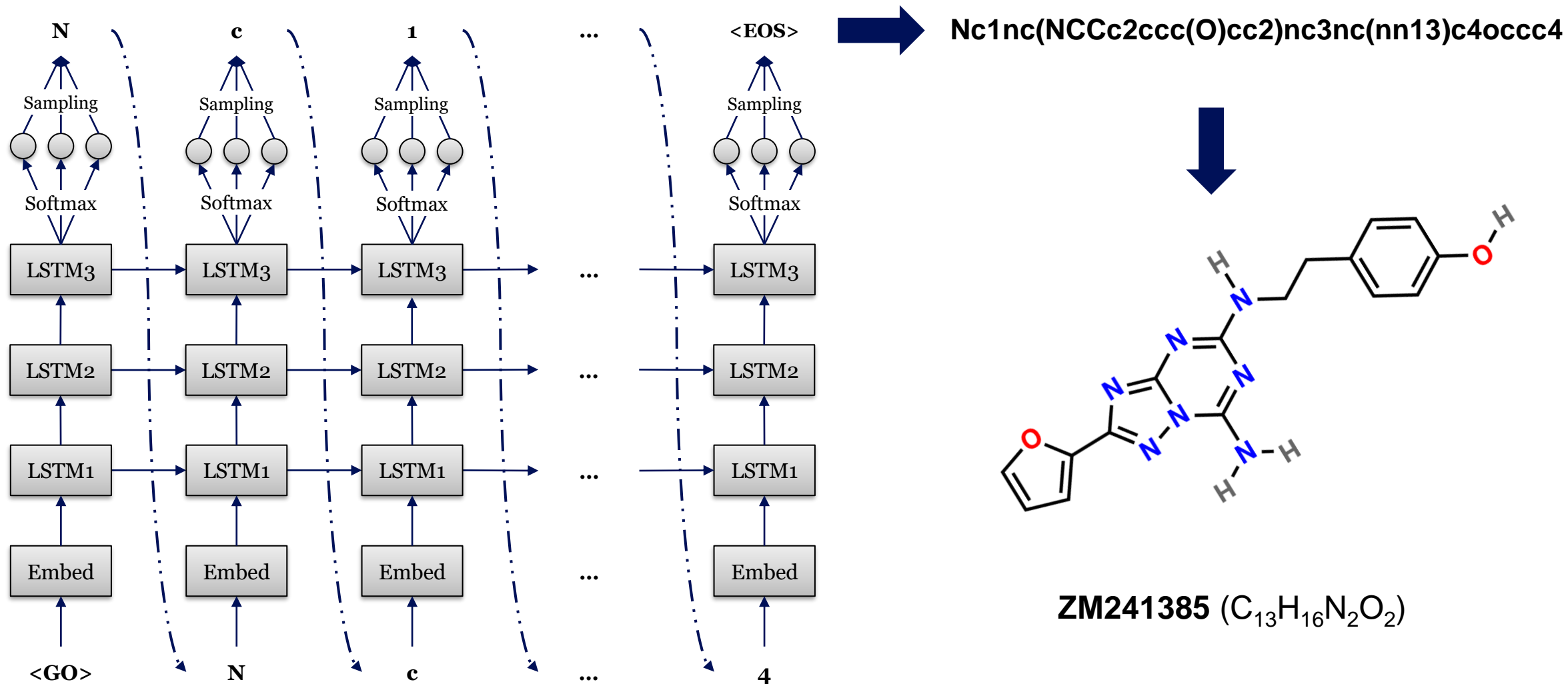
- All public compounds tested on the adenosine receptors ChEMBL (v 24).



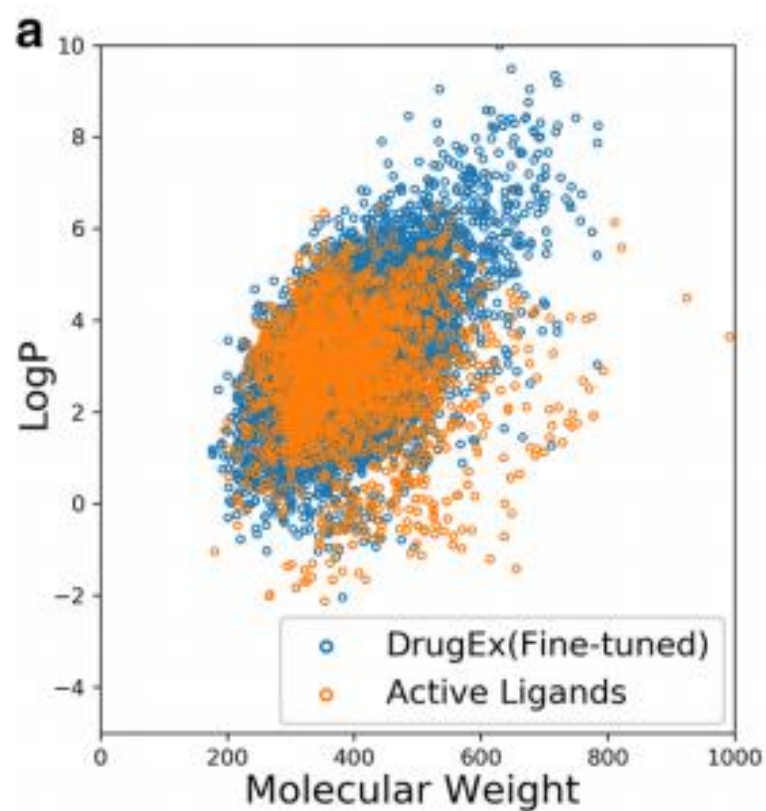
RNN Training



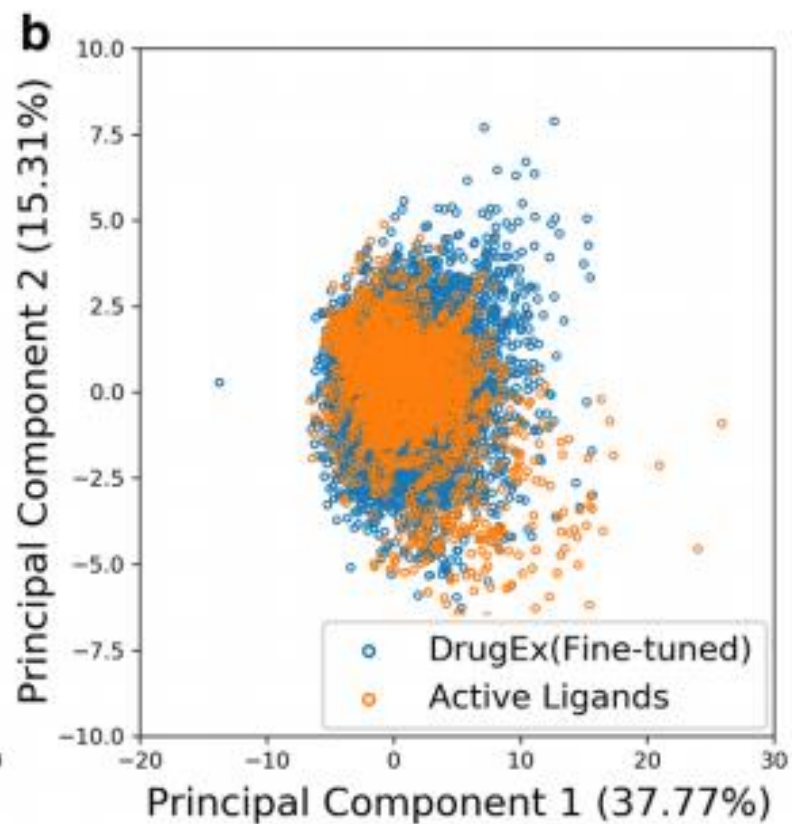
Molecule Generation



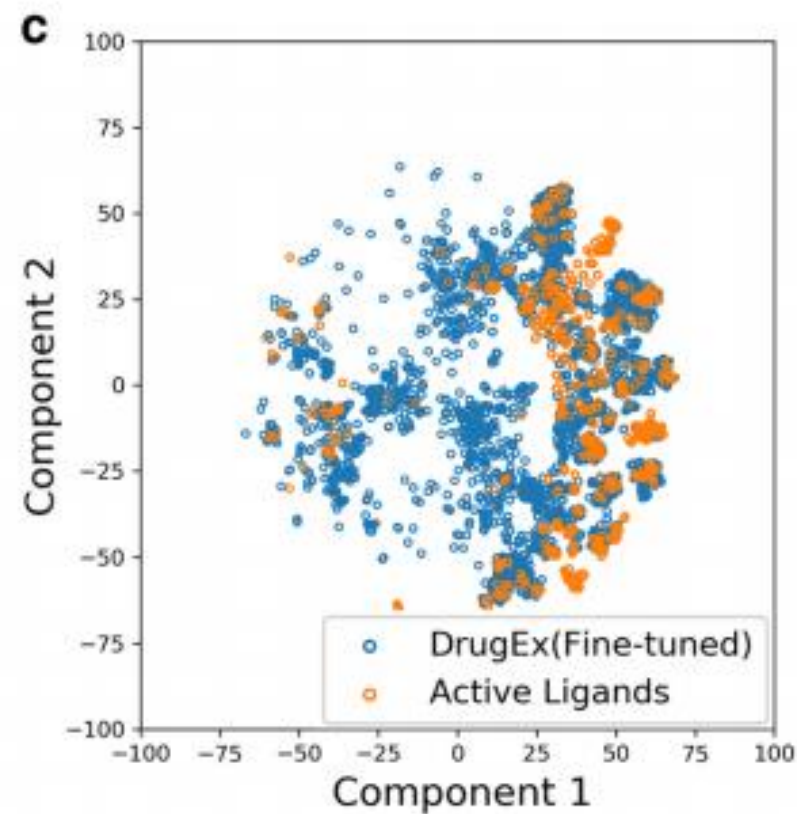
New A2A ligands



logP~MW



PCA (PhysChem)



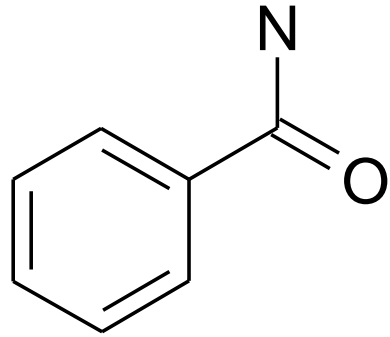
**t-SNE
(Fingerprints)**

Also more complex chemical features are generated

		Fused Ring	Furan Ring	Benzene Ring
DrugEx (Pre-trained)		9.12%	82.32%	61.48%
DrugEx (Fine-tuned)		60.69%	66.35%	65.62%
REINVENT		0.20%	95.26%	61.98%
ORGANIC		0.02%	99.96%	39.45%
Pre-trained		24.22%	4.51%	63.31%
Fine-tuned		76.33%	23.82%	72.85%
ZINC		26.66%	3.86%	63.97%
A2AR	Active	79.09%	40.29%	75.33%
	Inactive	76.73%	9.33%	70.88%

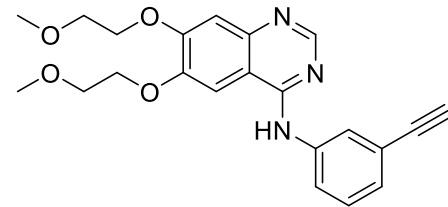
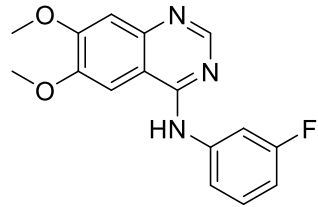
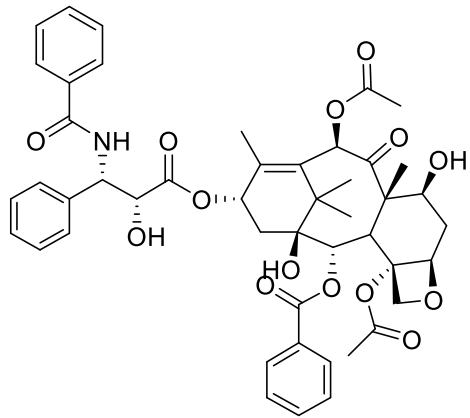
Descriptors

- Physiochemical properties

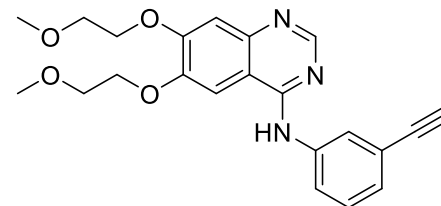
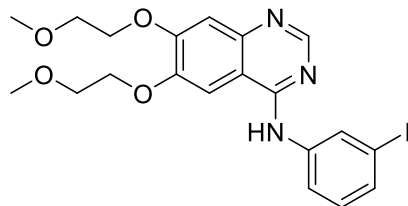
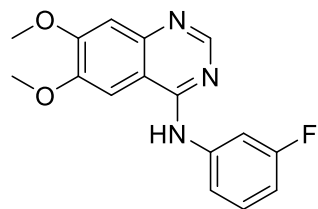


Molecular Weight	ALogP	Hydrogenbond Donors	Hydrogenbond Acceptors	Polar Surface Area
121.1	0.83	1	1	43.09

Molecular Similarity



Molecular Similarity



Molecular Similarity

