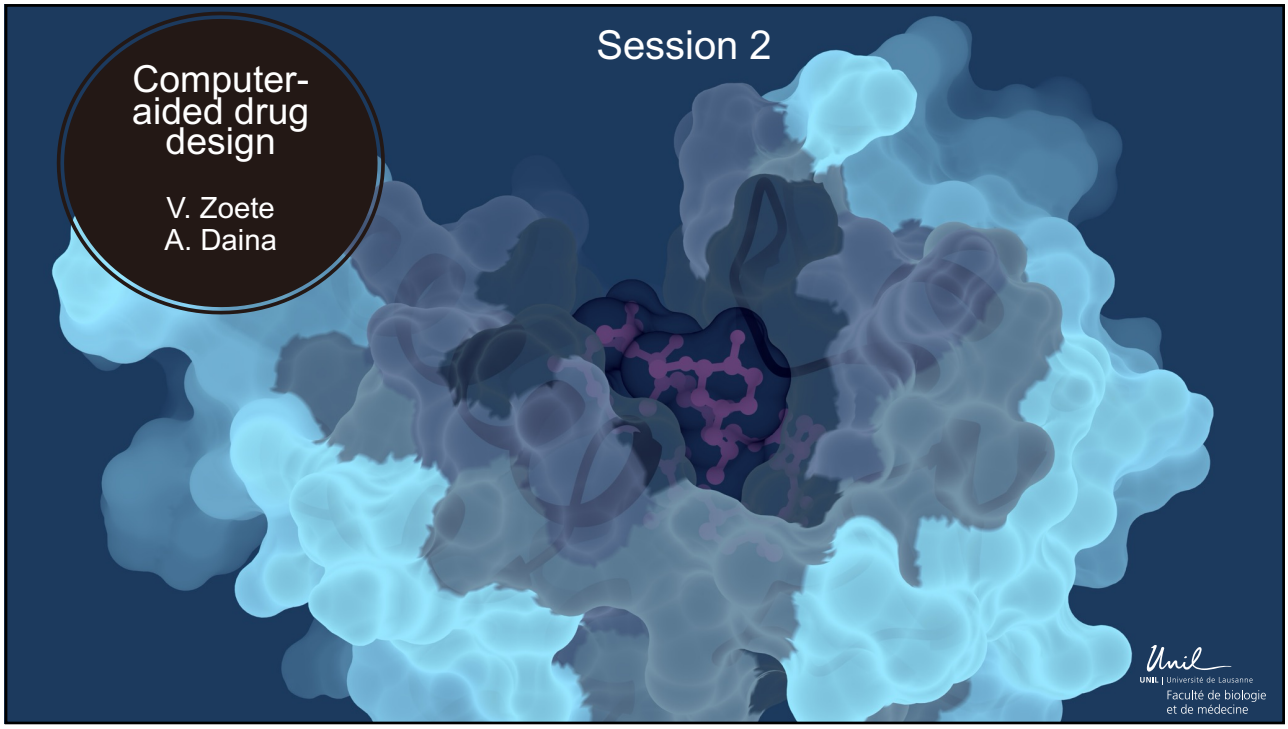


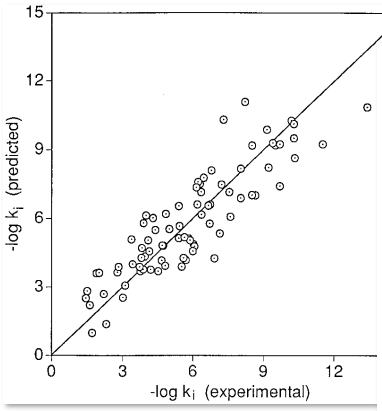
1

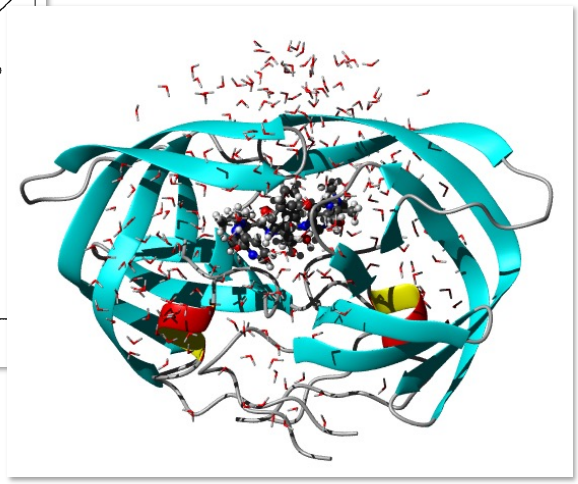
Lectures & Practices Agenda

Session	Lecture	Practice
1	Prologue: molecular representation	
	Introduction to (computer-aided) drug design	
	Origin of 3D structures	
	Molecular recognition	Use of UCSF chimera to analyze protein-ligand complexes
2	Binding free energy estimation	
	Introduction to molecular docking	Ligand-protein docking with AutoDock Vina
3	Introduction to molecular (virtual) screening	Ligand-based virtual screening with **SwissSimilarity**
4	Short introduction on target prediction of small molecules	Use of **SwissTargetPrediction** to perform reverse screening.
5	Introduction to ADME, pharmacokinetics, druglikeness	Estimate physicochemical, pharmacokinetic, druglike and related properties with **SwissADME**
6	Short introduction to bioisosterism	Use of **SwissBioisostere** to perform bioisosteric design

2

Binding free energy estimation

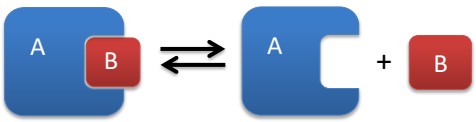




3

Binding free energy – Link between in silico and experimental worlds

Link between experiment and modeling



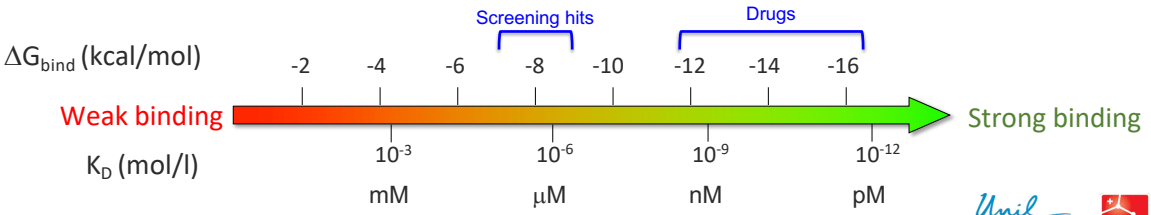
$$K_D = \frac{[A][B]}{[AB]}$$

K_D : dissociation constant

Experimentally measured

Accessible by computer-aided methods

$$\Delta G_{bind} = RT \ln(K_D) = \Delta H - T\Delta S$$



4

Binding free energy – The computational methods

Ligand-based

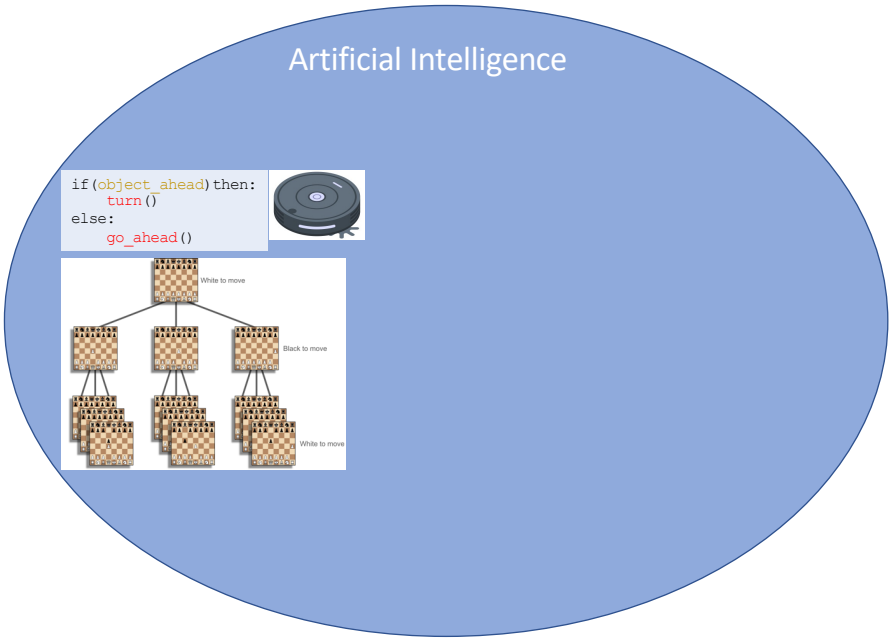
- **Machine Learning approaches**:
- 2D QSAR. Ex: Hansch equations
- 3D QSAR. Ex: CoMFA

Structure-based

- **Force field** methods:
  - Free energy simulation (FEP, TI)
  - MM-PBSA, MM-GBSA
  - Linear interaction energy (LIE)
- **Empirical** scoring functions (regression based approaches). Ex: LUDI score
- **Knowledge**-based approaches (Potential of Mean Force). Ex: PMF score

5

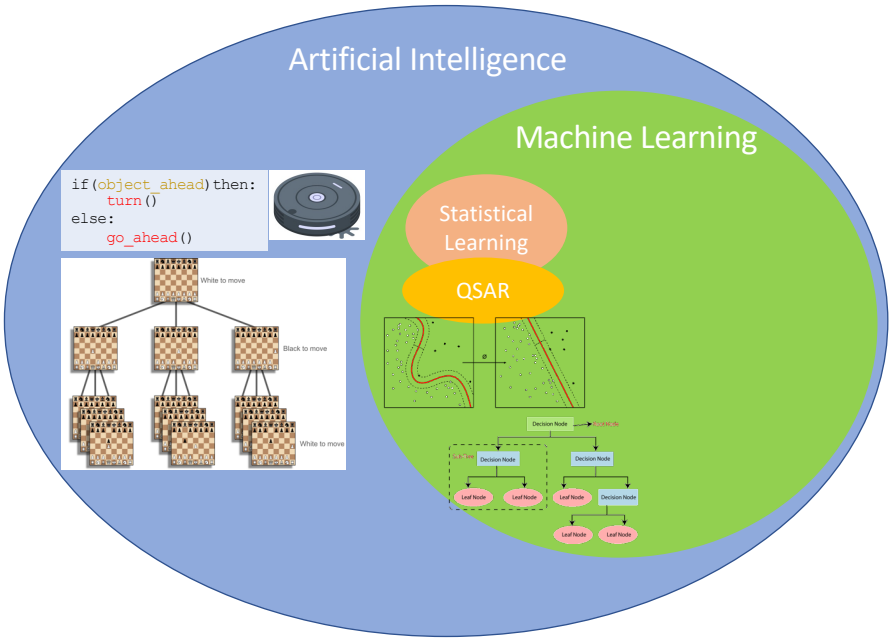
Binding free energy – The computational methods – Artificial intelligence?



Artificial intelligence:
Theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making…

6

Binding free energy – The computational methods – Artificial intelligence?

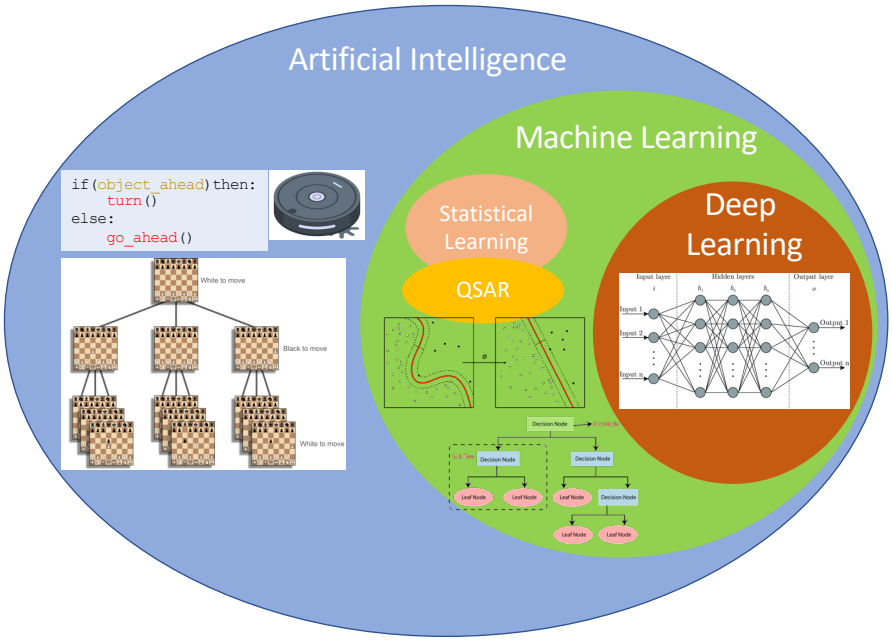


Artificial intelligence:
Theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making…

Machine learning:
Branch of Artificial Intelligence which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy

7

Binding free energy – The computational methods – Artificial intelligence?



Artificial intelligence:
Theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making…

Machine learning:
Branch of Artificial Intelligence which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy

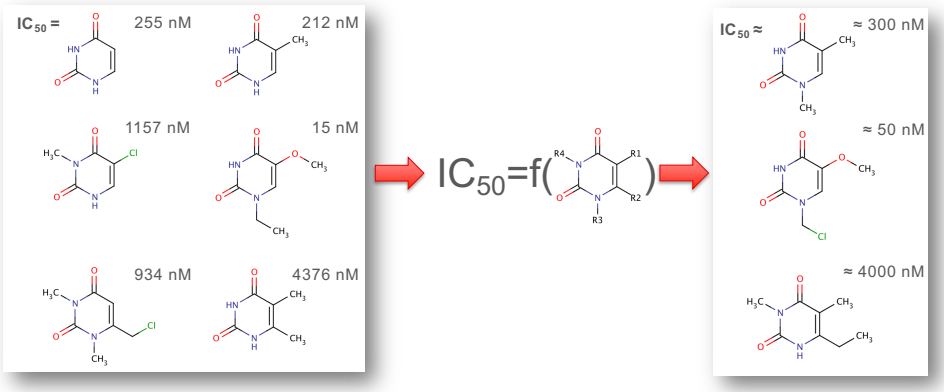
Deep learning:
Type of Machine Learning based on artificial neural networks in which multiple layers of processing are used to extract progressively higher-level features from data

8

Binding free energy – Ligand-based - QSAR

Quantitative Structure-Activity Relationship :

1. Built a set of **molecules** with **known experimental affinities** (activities).
2. Define a **mathematical relationship** between the **structure** (properties) of molecules and their **activities**.
3. Use this equation to **predict the binding affinities** (activity) of new molecules.



Binding free energy – Ligand-based - QSAR

Quantitative Structure-Activity Relationship :

1. Built a set of **molecules** with **known experimental affinities** (activities).
2. Define a **mathematical relationship** between the **structure** (properties) of molecules and their **activities**.
3. Use this equation to **predict the binding affinities** (activity) of new molecules.

Assumptions

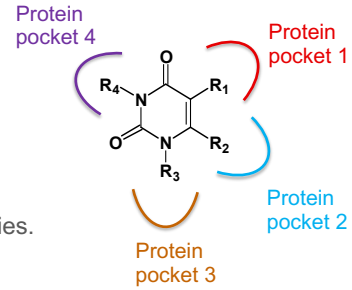
Chemical similarity of ligands ➡ Similarity of biological response
Affinity is a function of the differences in the ligand properties.
The binding mode is similar.

Advantages

No need for structural information about the target
Once established, extremely quick calculation. Suitable for very large libraries.

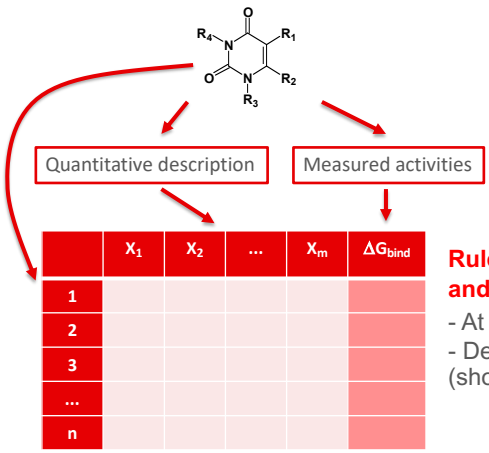
Limitations

The set of molecules need to be **large**, with a broad **spread of activity**.
Limited to **structurally similar** molecules (applicability domain).



Binding free energy – Ligand-based – 2D QSAR

n structurally related molecules characterized experimentally (i.e. the training set)



	X_1	X_2	...	X_m	ΔG_{bind}
1					
2					
3					
...					
n					

Global or substituent "2D" Descriptors (X_i):

- molecular weight
- log *P*
- polar surface area (PSA)
- simple count of atoms
- electronegativity
- charges

Rules for statistical relevance and model quality:

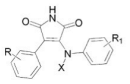
- At least **5 molecules per descriptor**
- Descriptors should **not be intercorrelated** (should not contain the same information)

MLR → $\Delta G_{bind} = k_0 + \sum k_i X_i$ **Multiple Linear Regression**, for instance.

C. Hansch and T. Fujita, *JACS*, **1964**, *86*, 1616

Binding free energy – Ligand-based – 2D QSAR

Ex: **Probing the physicochemical and structural requirements for glycogen synthase kinase-3alpha inhibition: 2D-QSAR for 3-anilino-4-phenylmaleimides.**
Sivaprakasam, P.; Xie, A.; Doerksen, R. J.; *Bioorg Med Chem* **2006**, *14*, 8210–8218.



Compound	R	R_1	X	IC_{50}[a]	pIC_{50}[b]	HA_R	f_{ortho}	σ_{metaR}	I_{3ClR1}	I_{4ClR1}	I_{3OHR1}	I_{xCH3}	E^s_{orthoR}	π_{metaR1}
1	H	H	H	529	6.28	0	0	0	0	0	0	0	0	0
2	2-Cl	H	H	216	6.67	0	0.41	0	0	0	0	0	−0.97	0
3	2-OCH_3	H	H	216	6.67	1	0.26	0	0	0	0	0	−0.55	0
4	3-NO_2	H	H	141	6.85	1	0	0.71	0	0	0	0	0	0
5	4-Cl	H	H	514	6.29	0	0	0	0	0	0	0	0	0
6	4-OCH_3	H	H	390	6.41	1	0	0	0	0	0	0	0	0
7	H	3-Cl	H	301	6.52	0	0	0	1	0	0	0	0	0.71
8	2-Cl	3-Cl	H	195	6.71	0	0.41	0	1	0	0	0	−0.97	0.71
9	2-OCH_3	3-Cl	H	114	6.94	1	0.26	0	1	0	0	0	−0.55	0.71
10	2-NO_2	3-Cl	H	104	6.98	1	0.67	0	1	0	0	0	−1.01	0.71

...

Table 4. Statistical parameters of 2D-QSAR models for 3-anilino-4-phenylmaleimides

QSAR model	*n*	*r*	r^2	*F*	*s*	q^2
1	64	0.815	0.665	23.0	0.216	0.590
2	64	0.860	0.739	26.9	0.192	0.675
3	64	0.862	0.743	27.5	0.191	0.679
4	67	0.847	0.718	31.1	0.231	0.671
5	67	0.873	0.762	32.0	0.214	0.706
6	67	0.901	0.812	36.5	0.191	0.761
7	67	0.902	0.814	36.9	0.190	0.762
8	67	0.806	0.650	39.1	0.253	0.608
9	67	0.852	0.726	41.0	0.226	0.688
10	67	0.862	0.743	35.2	0.220	0.702
11	67	0.863	0.745	35.7	0.219	0.705
12	67	0.880	0.774	41.8	0.206	0.732
13	67	0.909	0.827	47.6	0.182	0.788
13a[a]	66	0.922	0.850	55.6	0.165	0.814
14	67	0.911	0.829	48.6	0.181	0.789
14a[a]	66	0.922	0.850	55.9	0.165	0.813

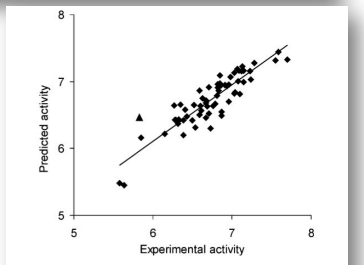


Figure 1. Correlation between the observed and predicted GSK-3α inhibitory activity from QSAR model 14 showing compound **19** as outlier (triangle).

Binding free energy – Ligand-based – 2D QSAR

Limited to structurally related molecules

Needs the experimental activity of a series ligands

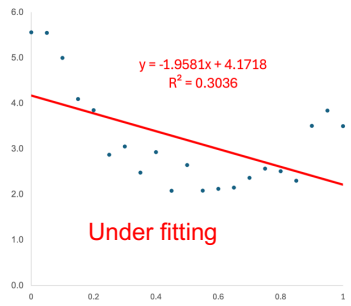
➡ Not for *ab initio* studies

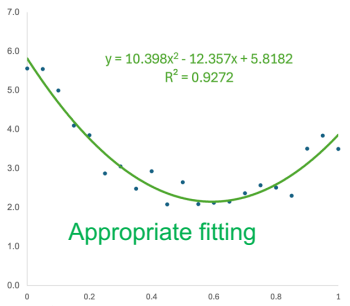
Overfitting

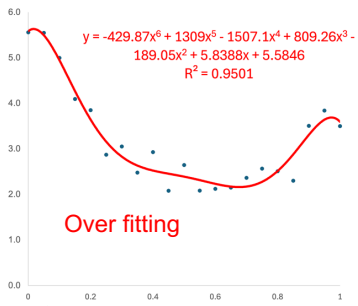
➡ - Method for selecting the descriptors (genetic algorithm)
- Estimation of the predictive ability
(external test set, Y-randomization, Cross-validation, ...)

Binding free energy – Ligand-based – 2D QSAR

Exercise:
- Create a polynomial function y=f(x), of degree 2 (i.e. a parabola), and add noise to generate 'experimental datapoints'
- Then, make models to fit the datapoints, and see how polynomial function of degrees 1, 2, 3 and more perform in terms of underfitting, appropriate fitting and overfitting



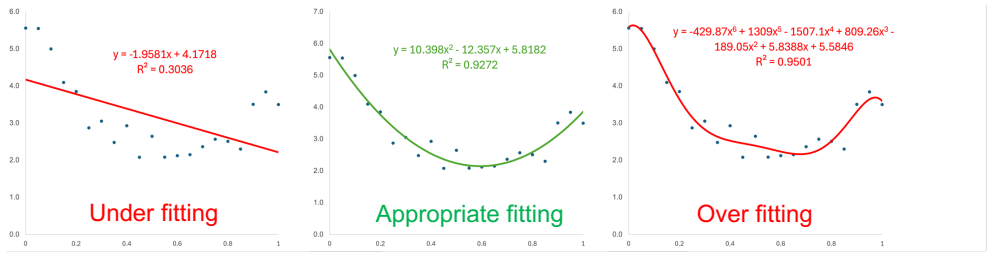




● "Experimental" datapoints — Models

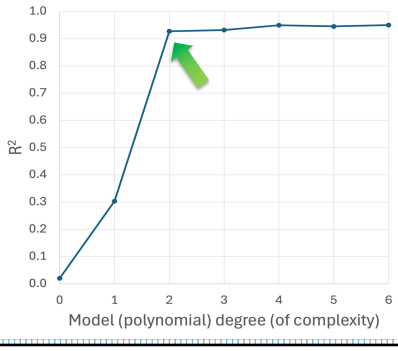
Underfitted and overfitted models are higly unlikely to make correct predictions, notably far from the training set (<0 or >1 in this example)

Binding free energy – Ligand-based – 2D QSAR



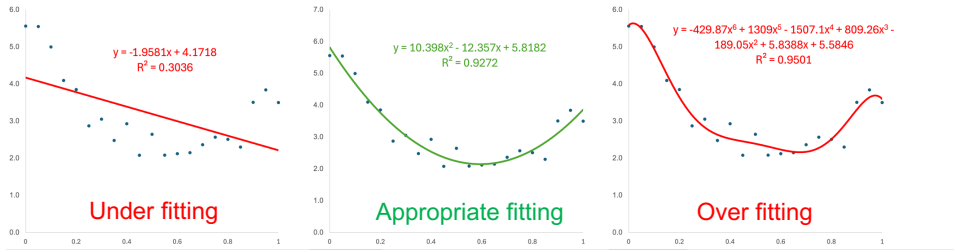
- "Experimental" data — Models

Adequate level of complexity (just enough to capture the essence of the phenomenon)



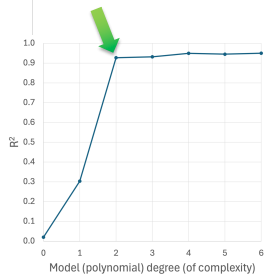
15

Binding free energy – Ligand-based – 2D QSAR

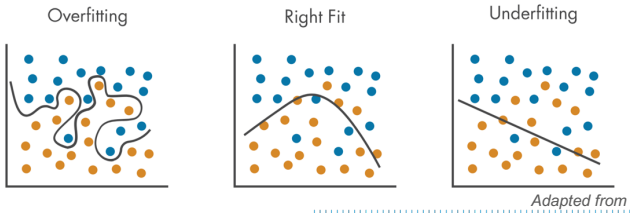


- "Experimental" data — Models

Adequate level of complexity (just enough to capture the essence of the phenomenon)



Also applies to classification:



Adapted from MathWorks

16

Binding free energy – Ligand-based – 2D QSAR

Principle of Y-randomization

True pIC_{50}

Molecule	MW	logP	tPSA	pIC50
1	110	2.5	211	9.50
2	232	2.4	152	8.93
3	254	5.6	111	8.86
4	451	4.3	101	8.38
5	312	2.8	194	8.70
6	291	6.1	75	8.74
7	185	3.5	145	9.08
8	345	4.6	167	8.61
9	267	3.4	130	8.82
10	435	5.2	94	8.41
11	511	3.7	173	8.27
12	254	6.2	93	8.85
13	267	2.6	172	8.81
14	273	2.5	110	8.81
15	318	4.5	145	8.68

Randomized pIC_{50}

Molecule	MW	logP	tPSA	pIC50
1	110	2.5	211	8.86
2	232	2.4	152	9.08
3	254	5.6	111	9.50
4	451	4.3	101	8.70
5	312	2.8	194	8.27
6	291	6.1	75	8.41
7	185	3.5	145	8.38
8	345	4.6	167	8.93
9	267	3.4	130	8.68
10	435	5.2	94	8.81
11	511	3.7	173	8.74
12	254	6.2	93	8.61
13	267	2.6	172	8.82
14	273	2.5	110	8.85
15	318	4.5	145	8.81

Search for models: pIC_{50} = f(MW, logP, tPSA)

Best model has R_{true}^2=0.95

Search for models: pIC_{50} = f(MW, logP, tPSA)

Best model has R_{rand}^2=0.01

Do pIC_{50} randomization thousands of times.
True model is trustable if R_{true} is always better than all R_{rand}

Binding free energy – Ligand-based – 2D QSAR

Limited to structurally related molecules

Needs the experimental activity of a series ligands

➡ Not for *ab initio* studies

Overfitting

➡ - Method for selecting the descriptors (genetic algorithm)
- Estimation of the predictive ability
(external test set, Y-randomization, Cross-validation, ...)

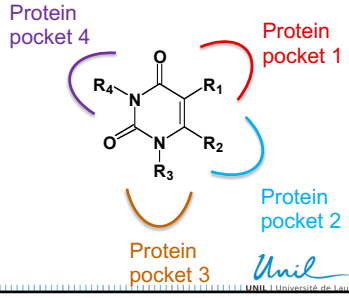
Validity domain. Interpretability strictly linked to the descriptor and training set:

If only hydrophobic groups at R_1 in the training set

➡ Influence of a hydrophilic group at R_1 ?

If only methyl, ethyl, propyl, butyl at R_1 in the training set

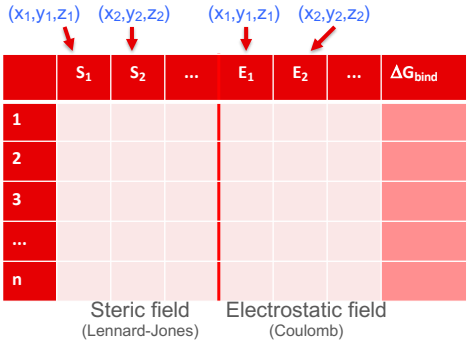
➡ Contribution of pentyl, hexyl, etc... ?



Binding free energy – Ligand-based – 3D QSAR

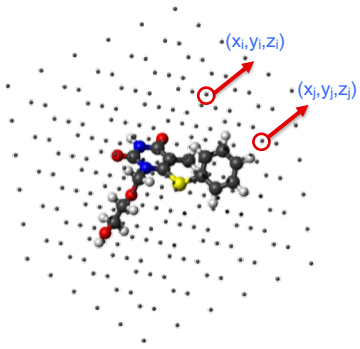
Example : Comparative molecular field analysis (CoMFA)

R.D. Cramer *et al.*, *JACS*, **1988**, *110*, 5959



Molecules superposed in a 3D grid

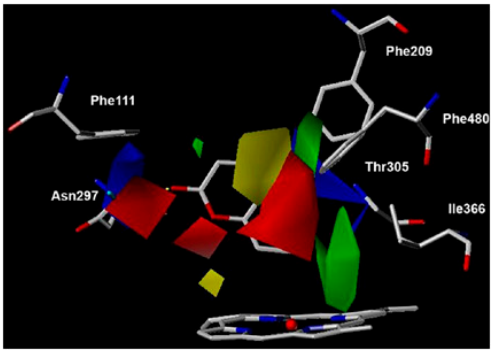
"3D" Descriptor = Molecular Fields



MLR $\Delta G_{bind} = k_0 + \sum \alpha_i S_i + \sum \beta_i E_i$ **Multiple Linear Regression**, for instance.

Binding free energy – Ligand-based – 3D QSAR

→ "inverse" image of the receptor binding site



Identification of inhibitors of the nicotine metabolising CYP2A6 enzyme: an *in silico* approach
M. Rahnasto, C. Wittekindt, R. O. Juvonen, M. Turpeinen, A. Petsalo, O. Pelkonen, A. Poso, G. Stahl, H-D Hotje and H. Raunio
Nature, 2008, 8(5), 328-338

CoMFA **electrostatic** fields:
- **blue**, negative-charge disfavored area
- **red**, negative-charge detrimental area

CoMFA **steric** field:
- **green**, bulk favorable area
- **yellow**, bulk detrimental area

Binding free energy – Ligand-based – 3D QSAR

Requires the **experimental activity** of a series of ligands

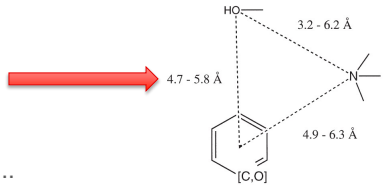
Risks of overfitting

Needs to respect the **domain of validity** when used for prediction

Not limited to structurally related molecules

Main limitation: **Alignment** of the molecules in their (guessed) **bioactive conformation**.
Possible help of:

- Structure of a protein-ligand complex available
  ➔ alignment over cocrystallized ligand or by docking.
- Set including rigid molecules
  ➔ alignment over rigid molecules
- Functional groups in agreement with a pharmacophore hypothesis
  ➔ alignment over pharmacophoric points.



Others : CoMSIA, HASL, Compass, APEX-3D, YAK, ...

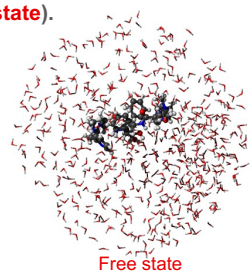
Binding free energy – Structure-based – Force field (Physics-based)

Example: Linear Interaction Energy (LIE)

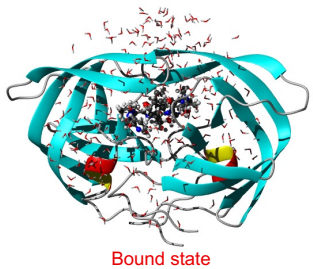
➔ **compute the van der Waals (Lennard-Jones) and electrostatic interaction (force field) energies of the ligand with water (free state) and of the ligand with protein and water (bound state).**

$\langle E_{l-s}^{vdw} \rangle_{free}$

$\langle E_{l-s}^{elec} \rangle_{free}$



Free state
environment = water



Bound state
environment = water + protein

$\langle E_{l-s}^{vdw} \rangle_{bound}$

$\langle E_{l-s}^{elec} \rangle_{bound}$

$$\Delta G_{bind} = \alpha \left(\langle E_{l-s}^{vdw} \rangle_{bound} - \langle E_{l-s}^{vdw} \rangle_{free} \right) + \beta \left(\langle E_{l-s}^{elec} \rangle_{bound} - \langle E_{l-s}^{elec} \rangle_{free} \right)$$

J. Åqvist, *J. Phys. Chem.*, **1994**, *98*, 8253
α=0.165 and β=0.5

T. Hansson *et al.*, *J. Comp.-Aided Molec. Design*, **1998**, *12*, 27
α=0.181 and β=0.5, 0.43, 0.37, 0.33

W. Wang, *Proteins*, **1999**, *34*, 395
α function of binding site hydrophobicity

Binding free energy – Structure-based – Force field (Physics-based)

Example: Linear Interaction Energy (LIE)

Modifications :

- Additional term proportional to buried surface upon complexation
  D.K. Jones-Hertzog and W.L. Jorgensen, *J. Med. Chem.*, **1997**, *40*, 1539
- Use of continuum solvent model instead of explicit solvent
  R. Zhou and W.L. Jorgensen *et al.*, *J. Phys. Chem.*, **2001**, *105*, 10388
- Replace molecular dynamics simulations by simple minimization
  Huang, D.; Caflisch, A. *J. Med. Chem.* **2004**, *47*, 5791

Advantages :

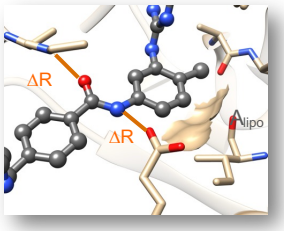
- Can treat more structurally different ligands than QSAR. But still generally restricted to rather similar ligands.

Shortcomings:

- Slower than scores based on a single conformation (LUDI, PMF, ...)
- Not really universal (α and β are system-dependent)
- Need experimental binding affinities of known complexes

Binding free energy – Structure-based – Empirical methods

Example: LUDI score



Evaluation of ΔG_{bind} : a simple count of various type of interactions between ligand and protein.

$$\Delta G_{bind} = \Delta G_0 + \Delta G_{polar} + \Delta G_{apolar} + \Delta G_{solv} + \Delta G_{flexi}$$

Developed using a 82 protein-ligand complexes dataset with known experimental ΔG_{bind}

Polar interactions

$$\Delta G_{polar} = \Delta G_{hb} \sum_{hb} f(\Delta R, \Delta\alpha) \times f(N_{neighb}) \times fpcs + \Delta G_{ion} \sum_{ion} f(\Delta R, \Delta\alpha) \times f(N_{neighb}) \times fpcs + \Delta G_{esrep} N_{esrep}$$

ΔG_{hb} = -0.81, ΔG_{ion} = -1.41 and ΔG_{esrep} = +0.10 kcal/mol

Apolar interactions

$$\Delta G_{apolar} = \Delta G_{lipo} A_{lipo} + \Delta G_{aro} \sum_{aro} f(R)$$

ΔG_{lipo} = -0.81 and ΔG_{aro} = -0.62 kcal/mol

Desolvation effect

Active site filled with water molecules → MD → Unbound water molecules

$$\Delta G_{solv} = \Delta G_{lipo\ wat} \sum \text{unbound water}$$

$\Delta G_{lipo\ water}$ = -0.33 kcal/mol

Ligand flexibility

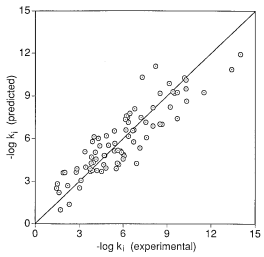
$$\Delta G_{flex} = \Delta G_{rot} N_{rot}$$

ΔG_{rot} = +0.26 kcal/mol N_{rot} : number of rotable bonds

H.J. Bohm, *J. Comput.-Aided Mol. Des.*, **1994**, *8*, 623
H.J. Bohm, *J. Comput.-Aided Mol. Des.*, **1998**, *12*, 309

Binding free energy – Structure-based – Empirical methods

Example: LUDI score



82 complexes of the training set

SD ~2 kcal/mol

Advantages :

- Allows identification of high affinity ligands
- **Rapid** estimation of the affinities
- Structurally **diverse ligands**
- Different proteins

➔ can be used routinely for **docking/virtual screening**

Others : ChemScore, VALIDATE

Shortcomings:

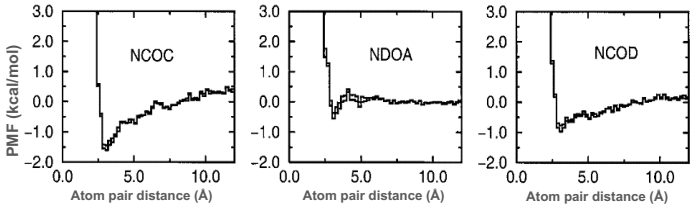
- Somewhat large errors
- Method **biased by training**:
  - certain type of proteins
  - only good complementarity protein/ligand
- Some interactions ignored:
  - cation – π
  - I...O, halogens
  - ...

Binding free energy – Structure-based – Knowledge-based

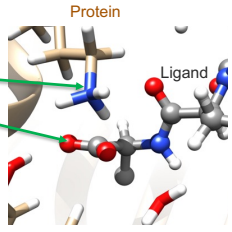
Example: PMF score

- Statistical **observations** of intermolecular **close contacts** in large 3D databases (e.g. the Protein Data Bank, **PDB**).
- The **more frequent** a protein-ligand contact between given atom types the **more favorable** to binding affinity.
- Trained on 697 PDB complexes.
- ➔ derivation of **potential energy** ("potential of mean force", **PMF**), no need to train on ΔG_{bind}

16 protein atom types, 34 ligand atom types



NC positively charged nitrogen
OC negatively charged oxygen
ND nitrogen as hydrogen bond donor
OA oxygen as hydrogen bond acceptor
OD oxygen as hydrogen bond donnor



$$\text{PMF score} = \sum_{ij} \sum_{kl \text{ of type } ij \atop r<r_{cutoff}} A_{ij}(r)$$

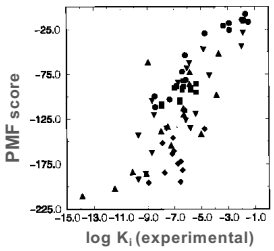
no unit, a « real » score

PMF between ligand atom (*i*) and atom protein (*j*)

I. Muegge *et al.*, *J. Med. Chem.*, **1999**, *42*, 791
I. Muegge *et al.*, *Persp. In Drug Disc. And Des.*, **2000**, *20*, 99

Binding free energy – Structure-based – Knowledge-based

Example: PMF score



77 complexes, 5 different proteins
SD ~2 kcal/mol

Advantages :

- Allows identification of high affinity ligands
- Rapid estimation of the affinities
- Structurally different ligands
- "Universal" (different ligand and protein types)
- No fitting parameters to measured ΔG_{bind}

Drawbacks :

- Somewhat large errors
- No measure of directionality of H-bonds
- Does not estimate directly binding in kcal/mol
- Still a bias if a rare interaction is not observed.

Others : DrugScore, GoldScore (trained on CSD)

Binding free energy – Conclusion

- Large variety of methods to estimate binding free energies

- None really satisfying in terms of predictive ability versus speed

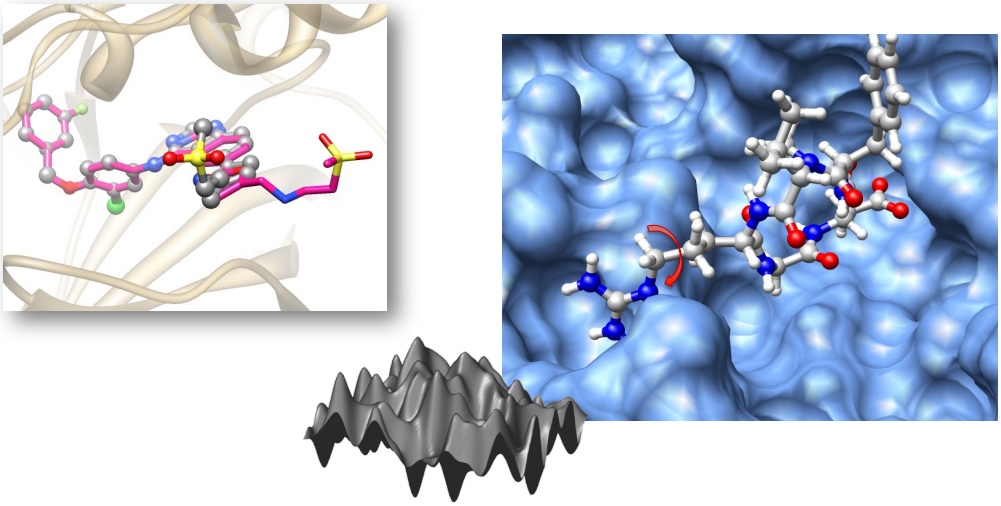
  ➡ Consensus scoring: evaluation of ΔG_{bind} with different scoring function, select virtual ligands that are predicted to be of high affinity by several scores.

  S.S. So and M. Karplus, *J. Comput. Aided. Mol. Des.*, **2001**, *15*, 613

- But, can be efficient to rank similar putative ligands

- Still a "limiting problem" of molecular docking methods and more generally for computational drug design

Small molecule docking

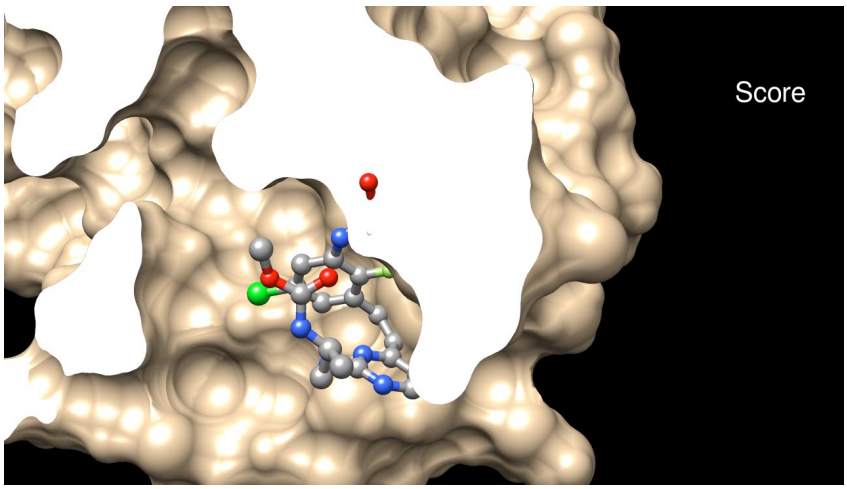


29

Docking – Objective

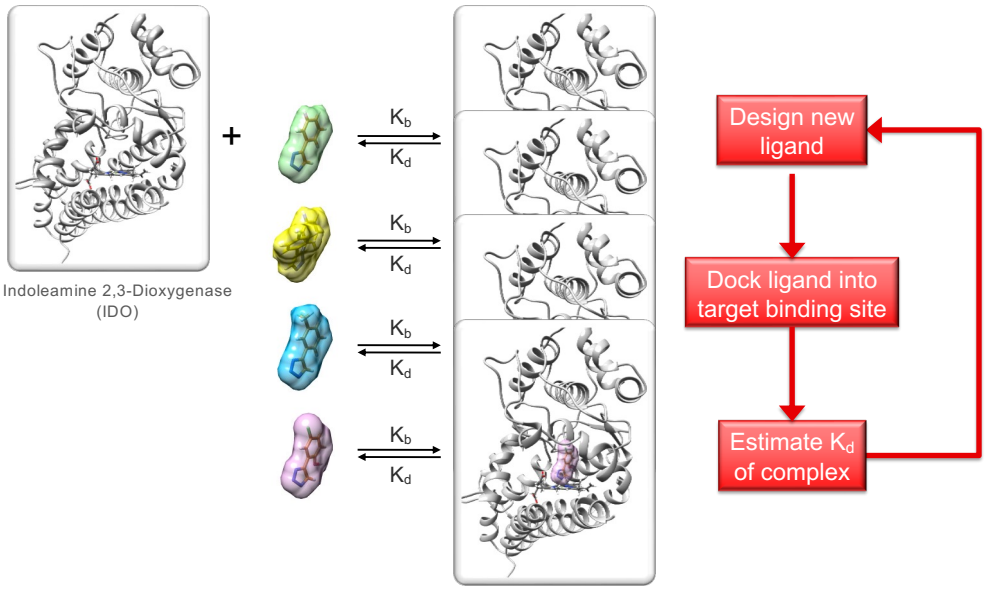
Docking small molecules into protein cavities:

Predict the **binding mode** (location, orientation and the geometry) of **the small molecule in the protein**
= "How the small molecule is **recognized by its macromolecular target**".



30

Docking – The cornerstone of structure-based ligand design



Courtesy of Dr. Ute Röhrig

Docking – Definitions

Pose: location, orientation and conformation of a small molecule on a macromolecule surface (cavity, pocket of groove) ~ tentative binding modes.

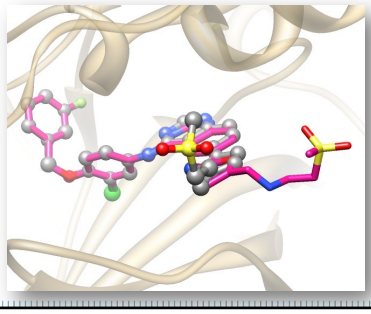
Native binding mode: experimentally defined binding mode (X-ray, NMR). Expected to be the best binding mode in term of binding free energy.

Docking: predicting the (native) binding mode using molecular modeling approaches.

Re-docking : docking on the X-ray structure of the receptor obtained in complex with the studied ligand (i.e. perfect induced fit). Used for exercise or benchmark.

Cross-docking : docking on a X-ray structure of the receptor obtained without the studied ligand (*apo* protein ➔ no induced fit, or complex with another ligand ➔ different induced fit)

Success: ability to predict a binding mode close to the native binding mode (when known, *i.e.* exercise or benchmark of the approach). Generally, RMSD < 2 Å.



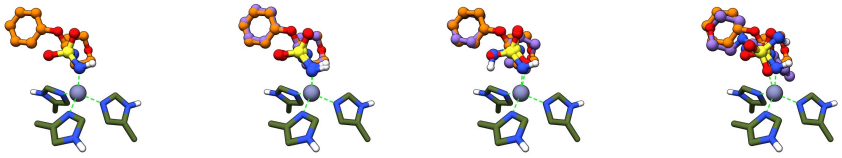
Docking – The root mean square deviation (RMSD)

Root Mean Square Deviation (RMSD, in Å [10^{-10} m]) :

- is the average distance between the pair of atoms (normally heavy atoms).
- is the measurement of superimposition of two poses of the same molecule.
- the greater the less superimposed. RMSD = 0 means perfect overlay.

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N} d_{ii}^2}$$

N : the number of atoms.
d_{ii} : the distance between atom i in the first pose and the same atom i in the second pose [Å]



X-ray Structure | RMSD = 0.2 Å | RMSD = 1.1 Å | RMSD = 4.5 Å

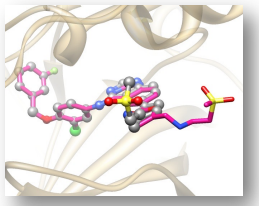
Docking – The root mean square deviation (RMSD)

Root Mean Square Deviation (RMSD, in Å [10^{-10} m]) :

- is the average distance between the pair of atoms (normally heavy atoms).
- is the measurement of superimposition of two poses of the same molecule.
- the greater the less superimposed. RMSD = 0 means perfect overlay.

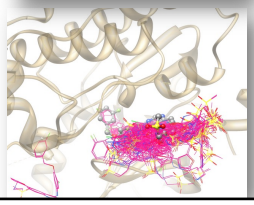
$$RMSD = \sqrt{\frac{1}{N}\sum_{ij=1}^{N} d_{ij}^2}$$

N : the number of pairs of atoms.
i : atom of pose 1
j : atom of pose 2
d_{ij} : distance between atom i and atom j [Å]



Measure of success (redocking):

- Measure all distances between heavy atom pairs of the crystallographic pose (C in grey ball & stick) and the docking pose (C in pink stick).
- Calculate RMSD.
- Rule-of-thumb: if RMSD < 2Å: SUCCESS!
- Here: ~2Å



Clustering (help at selecting the numerous poses)

- Each pose is compared to all poses by computing RMSDs
- Poses are classified into clusters (e.g. 2Å).
- All members of a given cluster have not more than 2Å RMSD to any member of the same cluster.
- A cluster represents all poses with similar binding mode.

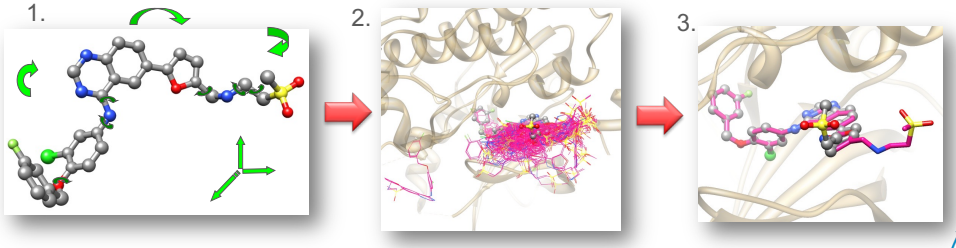
Docking – General approach

The "standard" methodology:

1. **Generate a large number of poses**. Sampling the posing space (orientational/translational/conformational) of the ligand into the protein binding site
2. **Assess the binding strength**. Scoring each possible ligand pose (~ fast evaluation of the ligand affinity)
3. **Selecting** the pose with the most favorable binding (**best score**)
   ➔ **predicted binding mode**

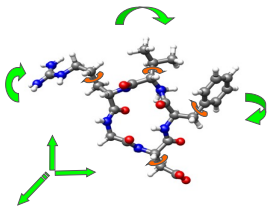
Different levels of approximation:

- protein and ligand are rigid (the past!)
- the protein is rigid, the ligand is flexible (today, except HTVS)
- protein and ligand are flexible (possible today at computational cost)



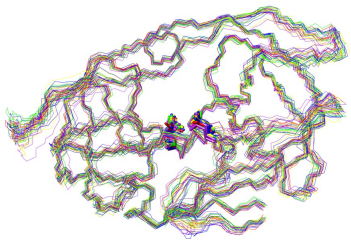
35

Docking – Handling ligand and protein flexibility



Ligand flexibility

- Pre-generated conformational libraries.
  Generation of several conformers, and rigid docking of each conformer
- Conformational search "**on the fly**" by the **posing algorithm.**



Protein flexibility

- Selection of **several** different protein **conformations** (from experimental structures, or MD simulations) and parallel dockings (**ensemble docking**).
- Use of an "**averaged protein structure**": Incorporating multiple receptor structures into a **grid** (energy-weighted grid, evt with reduced atom size).
- Conformational search "**on the fly**". Selected side chains can take preferential known conformations (**rotamer libraries**).

36

Docking – Existing programs

Many programs exist. They differ in:
- the **posing** algorithm
- the handling of ligand and protein **flexibility**
- the **scoring** function

Program	Posing algorithm	Scoring function	Protein flexibility
Autodock	EA	Force field	Flexible side chains
UCSF Dock	Incremental build	Force field / contact score	Protein side chain and backbone flexibility
Autodock Vina (swissdock.ch)	MC + local search	Empirical + knowledge-based	Flexible side chains
FlexX	Incremental build	Empirical score	Ensemble of protein structures
Gold	EA	Empirical / Knowledge-based	Selected side chain / ensemble docking
Glide	Exhaustive search	Empirical score	-
EADock 2	EA	Force field	Protein side chain and backbone flexibility
EADock DSS (old.swissdock.ch)	Incremental build	Force field	Protein side chain and backbone flexibility
Attracting Cavities (swissdock.ch)	Energy minimizations	Force field	Protein side chain and backbone flexibility

37

Docking – Posing algorithms

Three types of sampling algorithms:

Systematic search
- Combinatorial exploration of degrees of freedom
- Incremental construction

Stochastic search
- Monte Carlo (MC)
- Evolutionary algorithm (EA)
- Particle swarm optimization

Deterministic search
- Minimization
- Molecular dynamics simulation (MD)

38

Docking – Posing algorithms – Systematic search

- Anchor and grow (Ex: **EADock DSS [Swissdock.ch]**, FlexX, DOCK)

A. The ligand is divided into rigid (core fragments) and flexible parts (side chains)

B. An anchor is selected among the rigid fragments and docked into the target

C. The ligand is rebuilt incrementally, starting from the anchor, through systematic dihedral angle exploration. Unfavorable dihedral angles are removed (pruning algorithm)

JCAMD, 2001, **15**, 411-428

Methods differ in the docking of the anchor and in the pruning algorithm

- Fragment growing (Ex: Hammerhead)

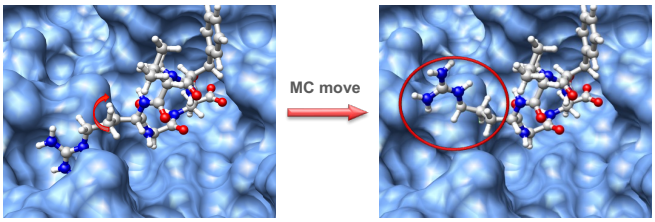
1. The ligand is divided into rigid (core fragments) and flexible parts (side chains)
2. All rigid fragments are docked
3. The ligand is rebuild from fragments that have acceptable initial scores

More “reconstruction algorithms” than “systematic search”

Docking – Posing algorithms – Stochastic search

The monte Carlo algorithm:

1. Generate an **initial pose** (ligand random conformation, translation and rotation) and **score it**
2. Generate a **new pose from the previous one** (through random conformational change, translation, rotation) and **score it**
3. Use **metropolis criterion**(*) to determine whether the new pose is retained
4. Repeat steps 2-3 until the number of desired poses is obtained (typically >100,000)

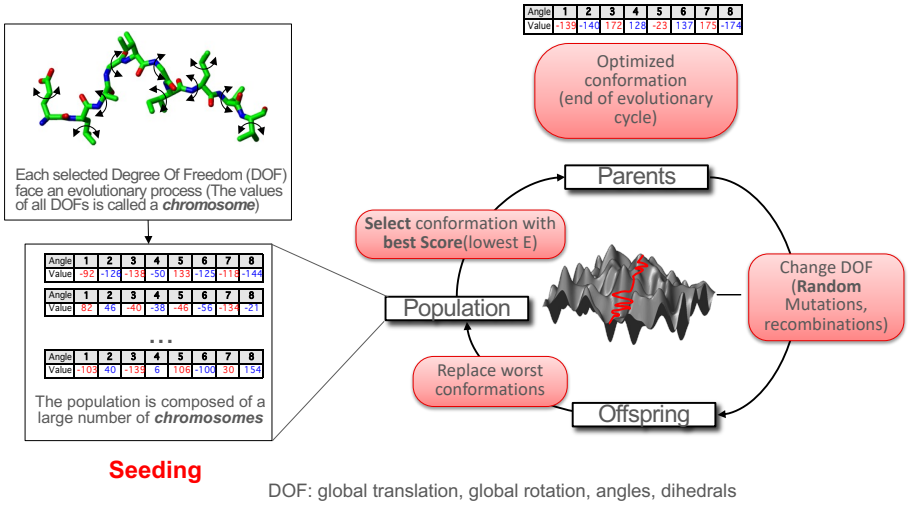


(*) Metropolis criterion:
If the difference in energy between the new and previous pose (ΔE) is negative (i.e. the new pose has better interactions with the protein), then the new pose is accepted.
If ΔE is positive, a random number between 0 and 1, $0<X<1$, is generated and the new pose is accepted only if $\exp(-\Delta E/RT)>X$.

Docking – Posing algorithms – Stochastic search

Genetic or evolutionary algorithms:

Ex: Gold, Autodock, EADock 2

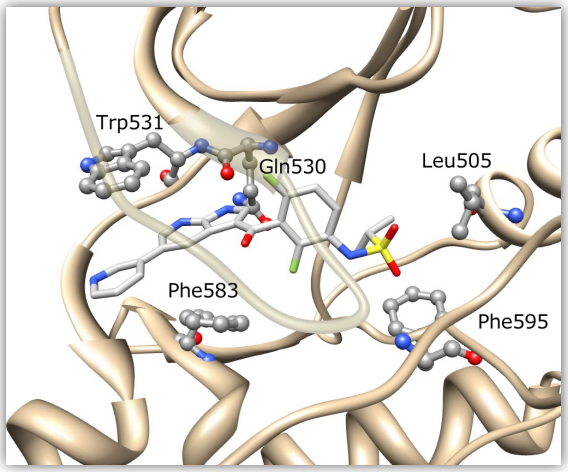


DOF: global translation, global rotation, angles, dihedrals

Docking – Posing algorithms – Deterministic search

Minimizations and Molecular Dynamics (MD) simulations cannot cross easily energy barriers.

➡ Generally restricted **to local search around the starting pose**
May be useful to **after** systematic or stochastic process to refine poses
More useful in **post-processing**, computationally demanding!



MD simulation of a BRAF/inhibitor complex.

Docking – Scoring

The two roles of scoring functions:

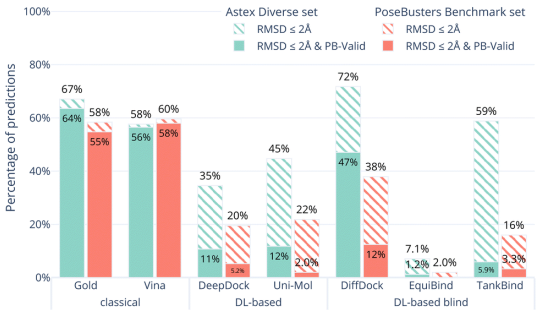
1. Rank poses for one ligand in a given target (docking ➔ predict binding mode)
2. Rank the binding modes of different ligands for a given target (compounds **selection** in **lead optimization** or **virtual screening** in hit-finding in **large databases**).

➔ must be a quick estimate of the binding.

- **Force-field** (physics-based) functions:
  - the most descriptive and comprehensive, potentially the most accurate,
  - slower,
  - **Simplified Force field**: minimal description of interactions ➔ sum of vdW and electrostatic interaction energies for each atoms (e.g. AutoDock function).
  - **Exception: EADock DSS/Attracting Cavities**, affinity evaluated by computation of ΔG_{bind} ➔ CHARMM force field, MMFF force field and FACTS solvation model.
- **Empirical** functions, most common for docking
  - excellent balance fastness/predictive power
  - only hard-coded interactions are accounted for e.g. ChemScore, LUDI
- **Knowledge-based** scoring functions
  - good speed/accuracy balance
  - a "real" score: without unit nor true physical meaning. e.g. DrugScore, GoldScore, PMF-Score

Docking – Beyond Physics-Based Docking: Deep Learning

Method	Authors	Date	Search space
DeepDock[6]	Méndez-Lucio *et al.*	Dec 2021	Pocket
DiffDock[7]	Corso *et al.*	Feb 2023	Blind
EquiBind[8]	Stärk *et al.*	Feb 2022	Blind
TankBind[9]	Lu *et al.*	Oct 2022	Blind
Uni-Mol[10]	Zhou *et al.*	Feb 2023	Pocket

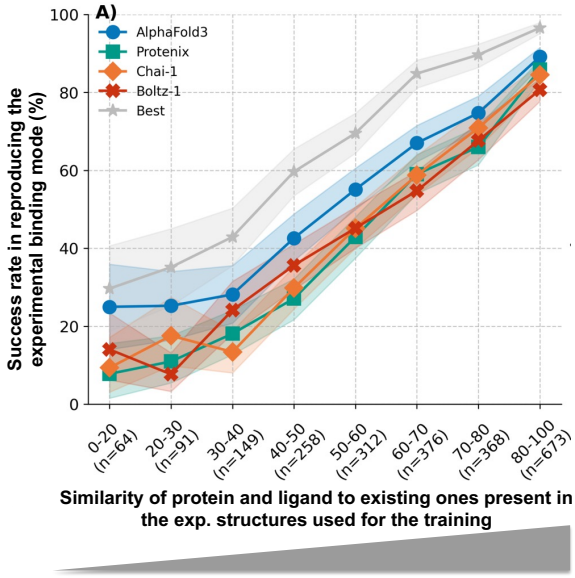


- PoseBusters: plausibility checks for generated molecule poses (i.e. are predicted binding modes follow th physics behind molecular interactions?)
- Co-folding: predict protein and ligand structure simultaneously (Umol, AlphaFold 3)

M. Buttenschoen, G. M. Morris, C. Deane, **PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences**, *Chem. Sci.* **15**, 3130 (2024)
Abramson, J., Adler, J., Dunger, J. et al. **Accurate structure prediction of biomolecular interactions with AlphaFold 3**. *Nature* (2024)
Bryant, P., Kelkar, A., Guljas, A. *et al.* **Structure prediction of protein-ligand complexes from sequence information with Umol**. *Nat Commun* **15**, 4536 (2024)

Courtesy of Dr. U. Röhrig

Docking – Beyond Physics-Based Docking: Deep Learning



DL cofolding approaches are overfitted: they predict/reproduce well known complexes, but fail to generalize to new, unseen, complexes

Have protein-ligand co-folding methods moved beyond memorisation?
Peter Škrinjar, Jérôme Eberhardt, Janani Durairaj, Torsten Schwede, BioRXiv, Feb 2025

45

Docking – Beyond Physics-Based Docking: Deep Learning

Current Shortcomings of Deep Learning approaches

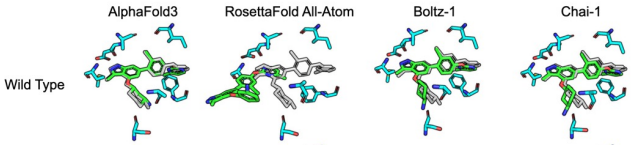


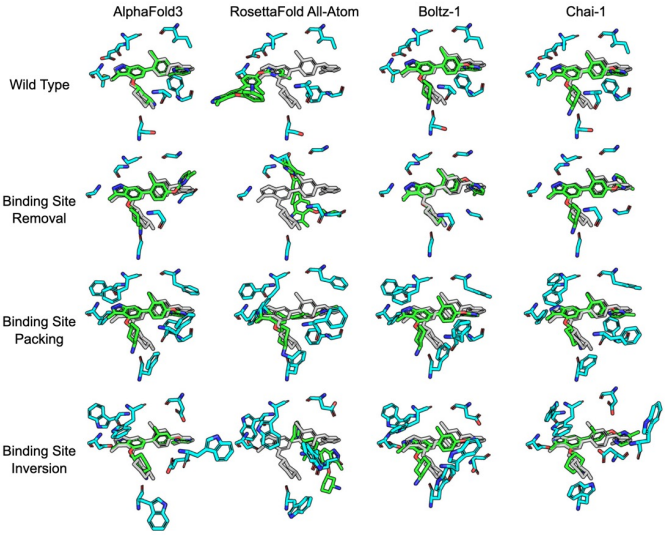
Fig. 2 | Binding site mutagenesis challenges against co-folding models using the MEK1 system (PDB: 7XLP). Predicted binding site residues are shown as cyan sticks, predicted ligand poses are shown as green sticks, and the original co-crystallized ligand pose is shown as gray sticks. Results are presented in the same manner as Fig. 1.

M.R. Masters, A.H. Mahmoud, M.A. Lill, Investigating whether deep learning models for co-folding learn the physics of protein- ligand interactions, Nat Comm. 2025

46

Docking – Beyond Physics-Based Docking: Deep Learning

Current Shortcomings of Deep Learning approaches



DL cofolding approaches do not consider correctly molecular interactions

Fig. 2 | Binding site mutagenesis challenges against co-folding models using the MEK1 system (PDB: 7XLP). Predicted binding site residues are shown as cyan sticks, predicted ligand poses are shown as green sticks, and the original co-crystallized ligand pose is shown as gray sticks. Results are presented in the same manner as Fig. 1.

M.R. Masters, A.H. Mahmoud, M.A. Lill, Investigating whether deep learning models for co-folding learn the physics of protein- ligand interactions, Nat Comm. 2025

47

Docking – Success

Success: ability to predict a binding mode close to the native binding mode (redocking, i.e. exercise or test of the approach). Generally, RMSD < 2 Å.

Success rate : ~ 50 to 90 % in benchmarks (re-docking)

in “real” application (cross docking) the success rate decreases by at least 30%
➔ room for improvement.
➔ need for high precision docking programs, handling protein flexibility (but also water, ion, ...).

48

Binding free energy estimation and molecular docking

Contacts: vincent.zoete@unil.ch , antoine.daina@sib.swiss