고효율 효능검색 기술

가상검색을통한 약물설계와 검색방법론

한국보건복지인력개발원

제1기 의약품 후보물질발굴 심화과정

2020. 9. 1.

한국화학연구원 의약바이오연구본부 채종학

(chchae@krict.re.kr; 010-9509-8740)

Bio

- BA Chemistry (Natural Science) : Seoul National University
- MSc Physical Chemistry : Seoul National University
- PhD Computational Chemistry : Seoul National University
- Postdoc Computational Chemistry : Seoul National University
- Currently : Principle Researcher in KRICT

- Area of expertise:
  - Molecular modeling : CADD/VS, SBDD, LBDD, Kinases; ~20 targets/year
  - Cheminformatics : Database, Web; Programming (C/C++, Python, Fortran), Machine Learning

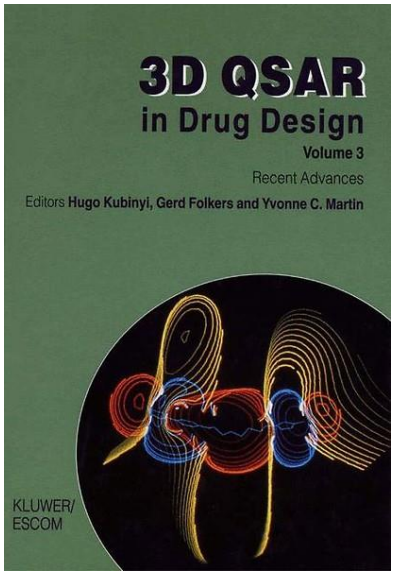
Table of Contents

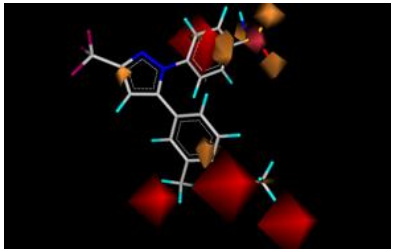
- CADD / ChemInformatics / Virtual Screening
  - Introduction
  - Compound structure, SMILES, SDF, PDB
  - Descriptors, Fingerprints, Similarity
- Protein (3D) Structure-based Prediction
  - Protein structure : Homology modeling
  - Docking : Scoring 문제
  - Demo : Schrodinger, Homology modeling & Docking
- Ligand-Based Prediction
  - Fingerprint (2D/3D)
  - Pharmacophore (3D/2D)
  - Machine Learning / QSAR
  - Demo : Python, RDKit, TensorFlow, DeepChem ML Prediction Model

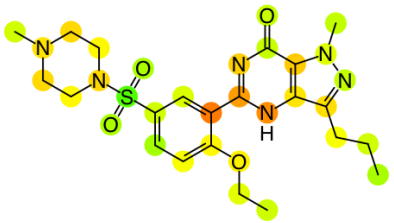
History of CADD

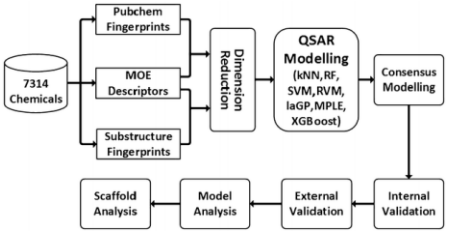
1981 → 1988 → 2000 → 2015

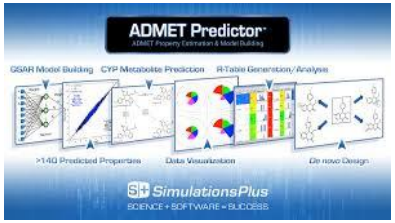
Modeling | **QSAR** | **QSPR** | **AI/BigData**

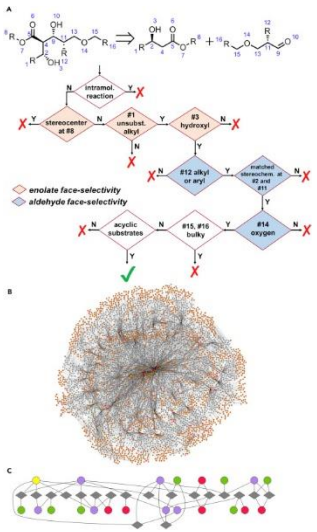




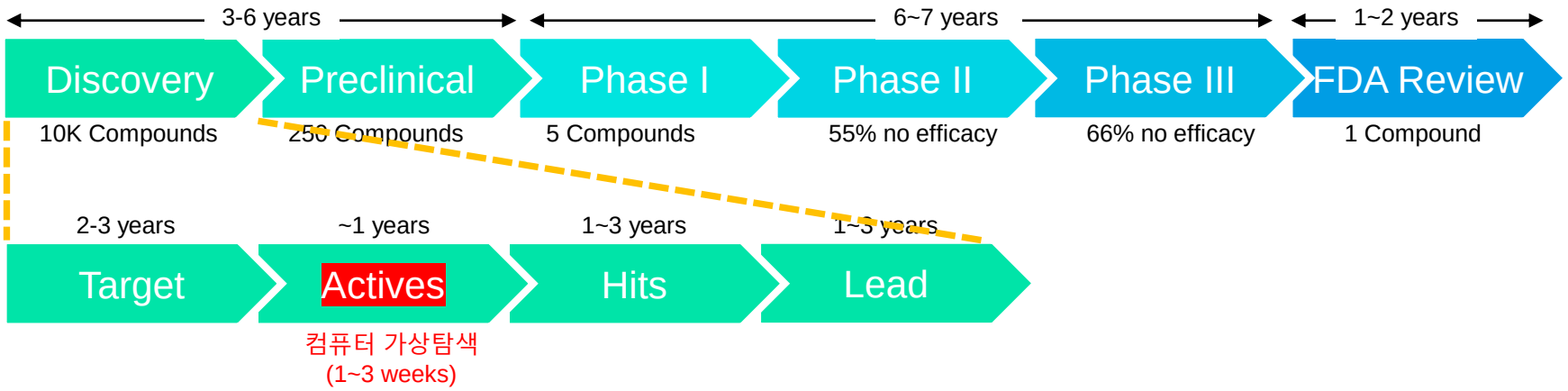


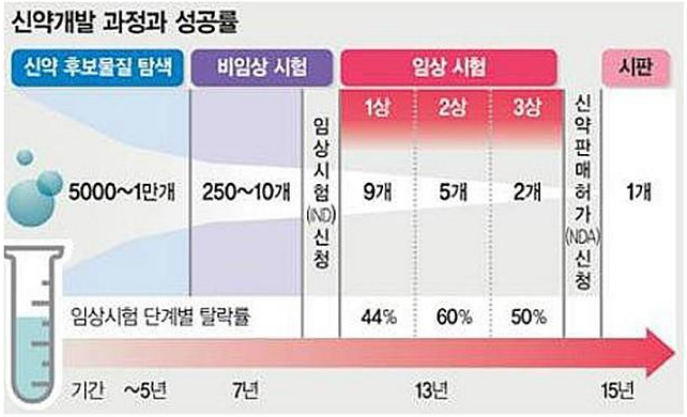


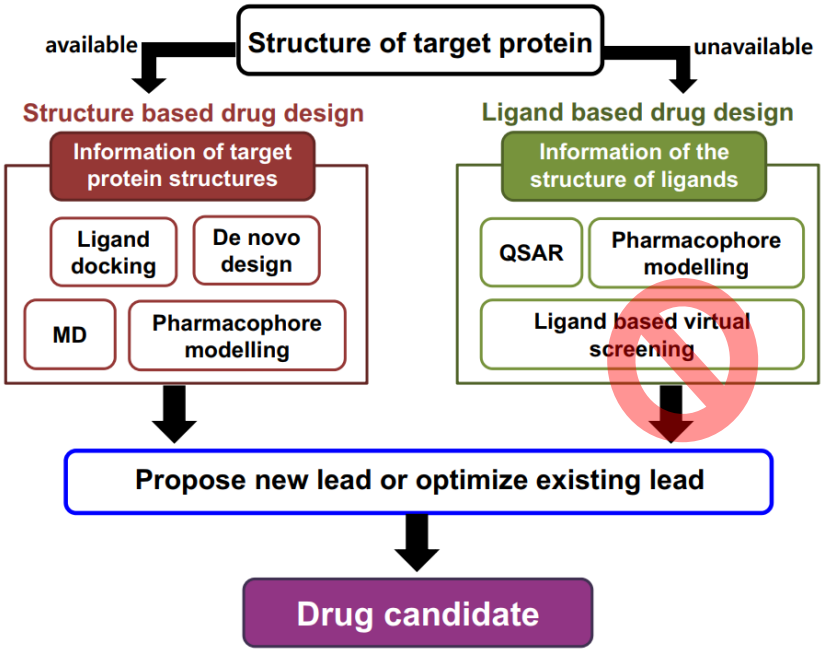




Why CADD ?





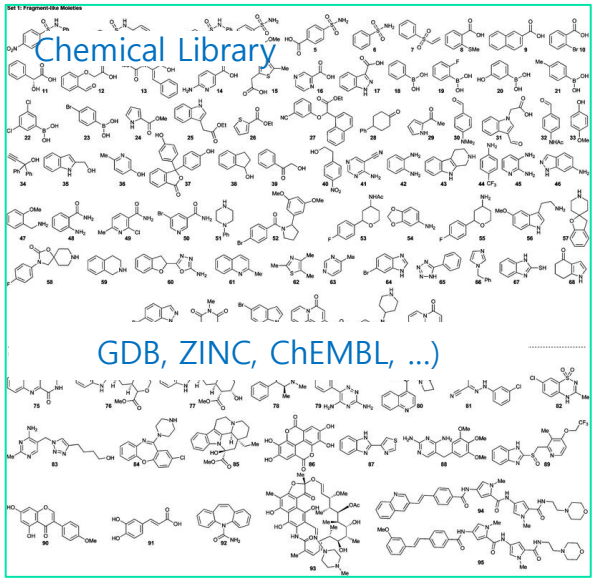


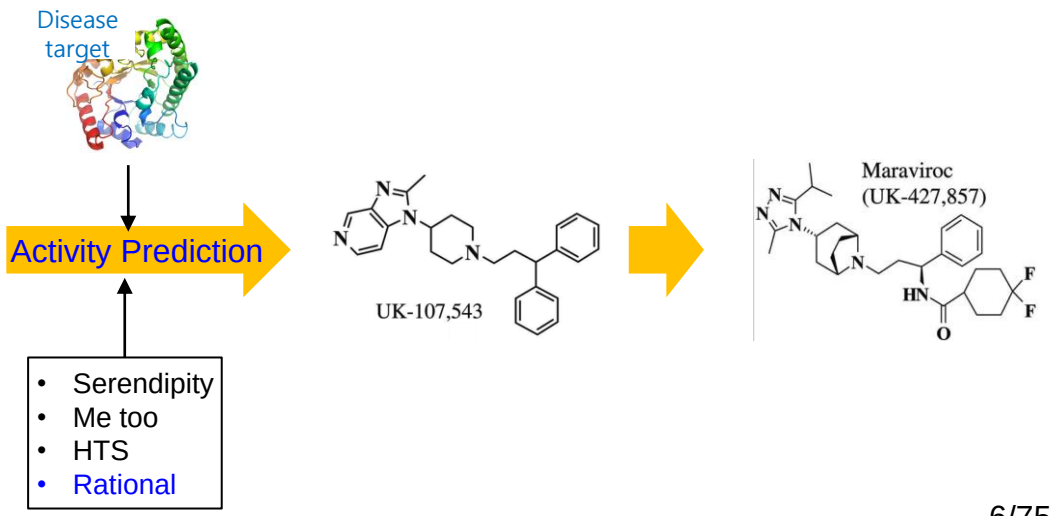
- **평균 15년, 1조원 이상의 R&D 비용 투입**
- **그러나, 성공율이 매우 낮음**
- **돈, 시간 관점에서 매우 값비싼 과정**

➔ **약을 만드는 효율적인 방법 탐색은 매우 중요**

Virtual Screening : Choosing the Right Molecules

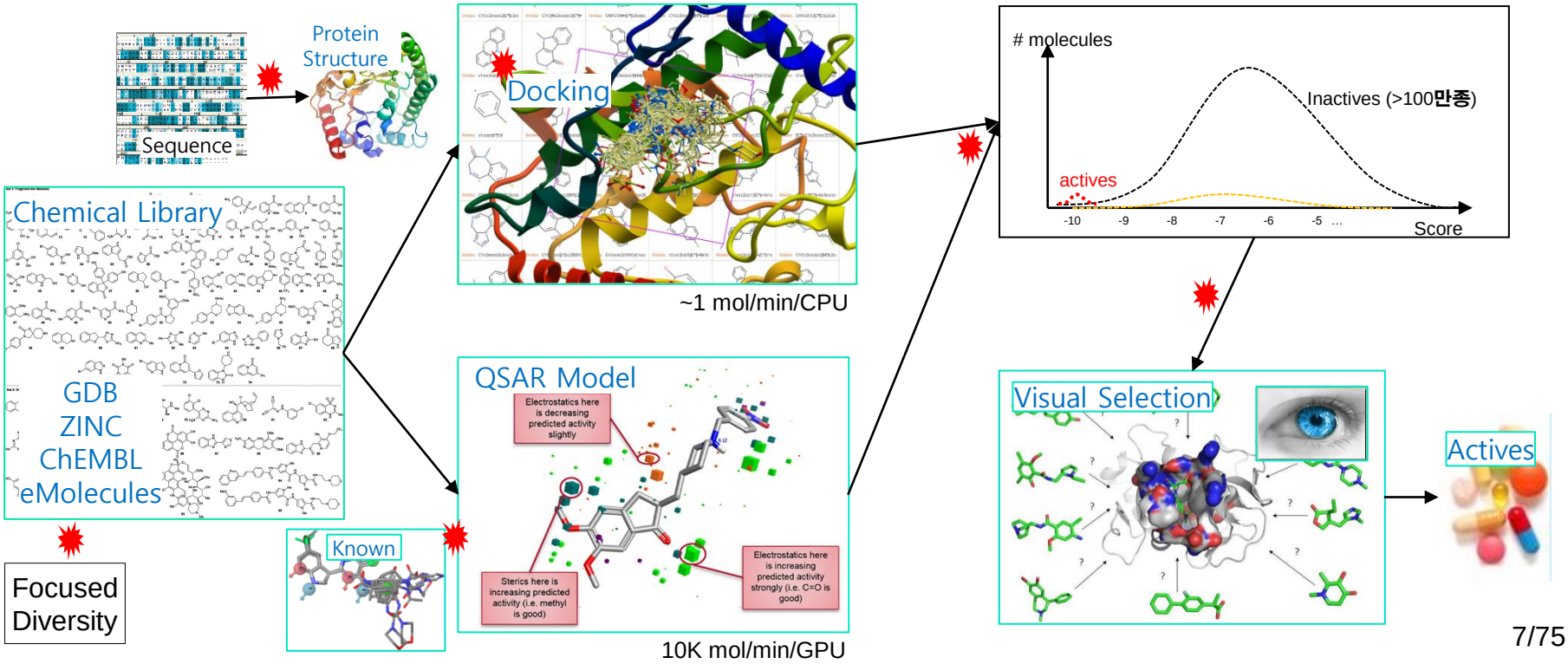
- Goal : to find actives that can be optimized to give drug candidates
  - *in-silico* techniques used to search large compound databases to select a smaller number for biological testing
- Challenge : chemical space is vast
  - ~ 30 M known compounds; ~1 billion compounds up to 13 heavy atoms
  - ~ x M compounds per pharma companies
- HTS allows ~ 1 M compounds to be tested
  - But very small portion of available compounds
  - Large scale screening is expensive



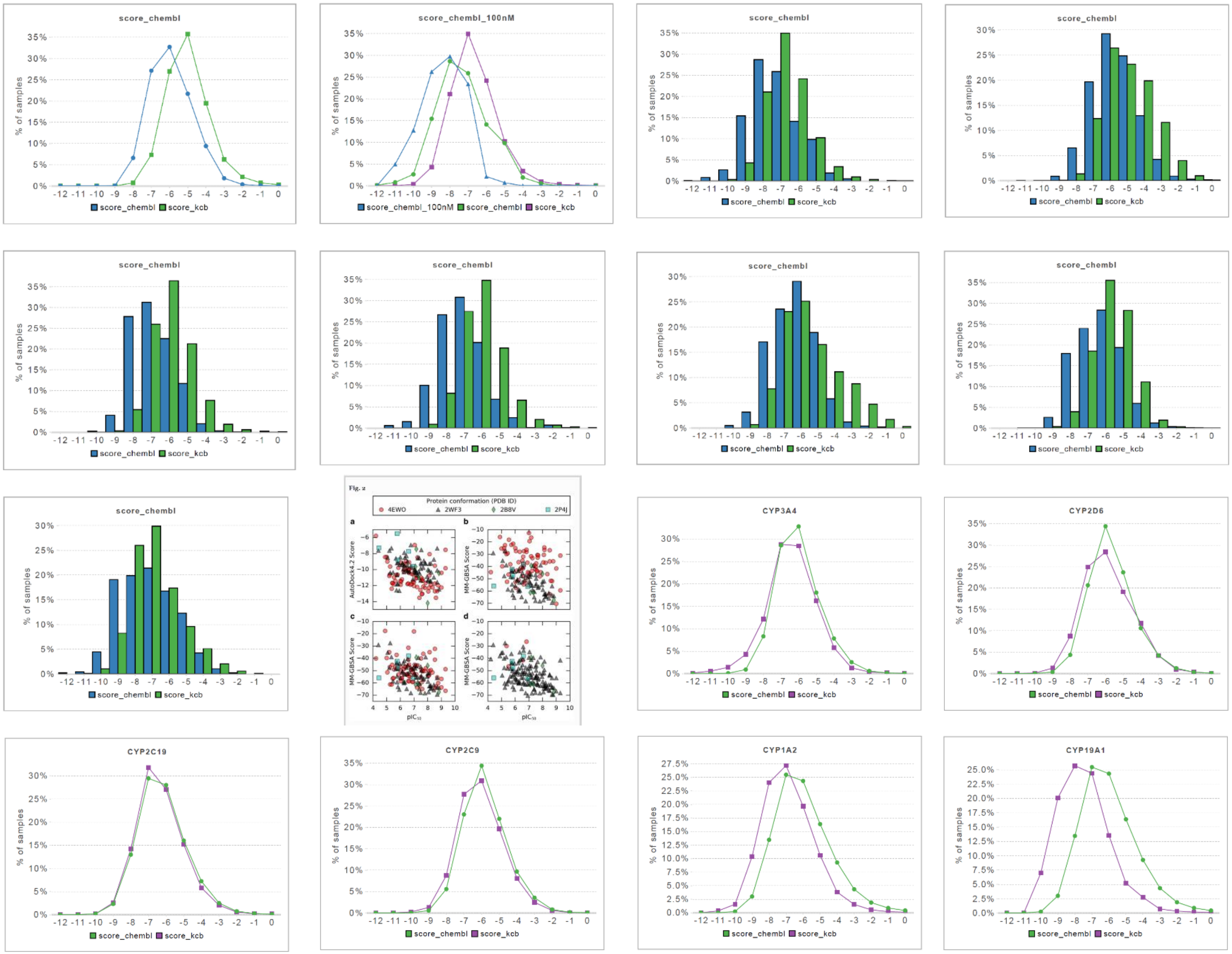


Virtual Screening

- What for?
  - Select compounds for screening from in-house databases
  - Choose compounds to purchase from external suppliers
  - Decide which compounds to synthesize next
- The technique applied depends on the amount of information available about the particular disease target (3D structure)

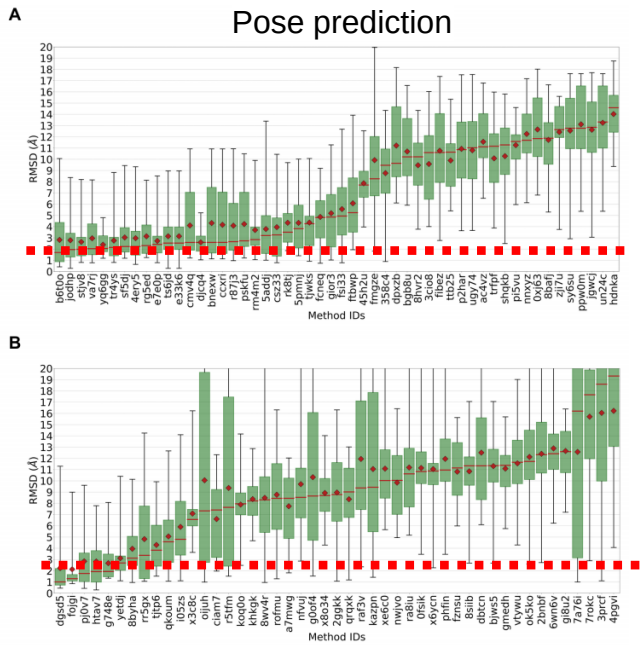


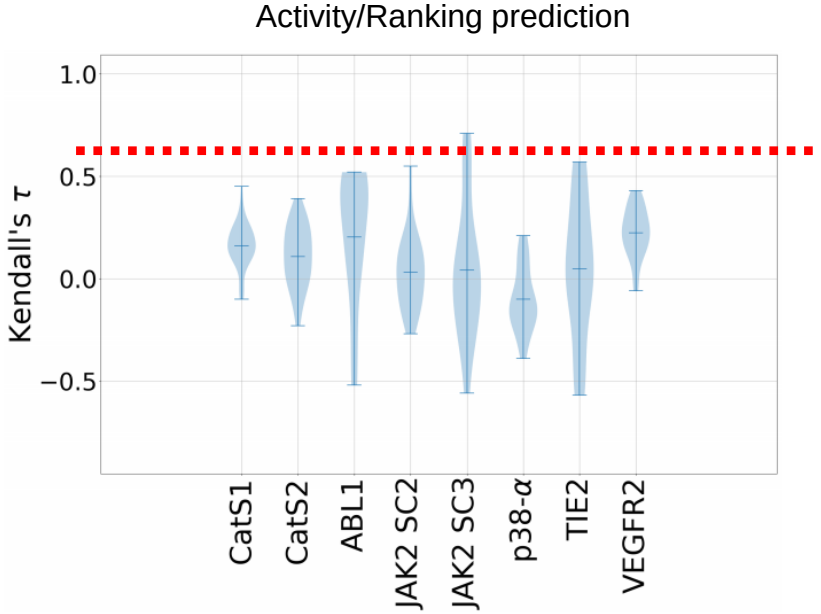
Discrimination of Docking Scores



D3R Grand Challenge

- UCSD; To share high quality protein-ligand complex data and workflow
- Data set
  - GC4 (2018) : Cathepsin S (460 from Janssen), BACE-1 (154 from Novartis)
  - GC3 (2017) : Cathepsin S (136 from Janssen), 5 kinases (from SGC-UNC)
  - GC2 (2016) : FXR (102 from Roche),
  - GC1 (2015) : HSP90 (174 from Abbvie)





https://drugdesigndata.org/about/grand-challenge

Enhancement of Docking VS with ML : vScreenML

- 기존 scoring은 active compound에 너무 training 되어 있음
- Active 화합물과 최대한 가까운 inactive를 이용하여 scoring function을 training하여 변별력을 높임
- Performance : 10/23 actives; up to 0.28 uM for AChE
- AChE : 136 crystal structures, 14,000 activity data !!!

Machine learning classification can reduce false positives in structure-based virtual screening

Yusuf O. Adeshina[a,b], Eric J. Deeds[b,c], and John Karanicolas[a,1]

[a]Program in Molecular Therapeutics, Fox Chase Cancer Center, Philadelphia, PA 19111; [b]Center for Computational Biology, University of Kansas, Lawrence, KS 66045; and [c]Department of Molecular Biosciences, University of Kansas, Lawrence, KS 66045

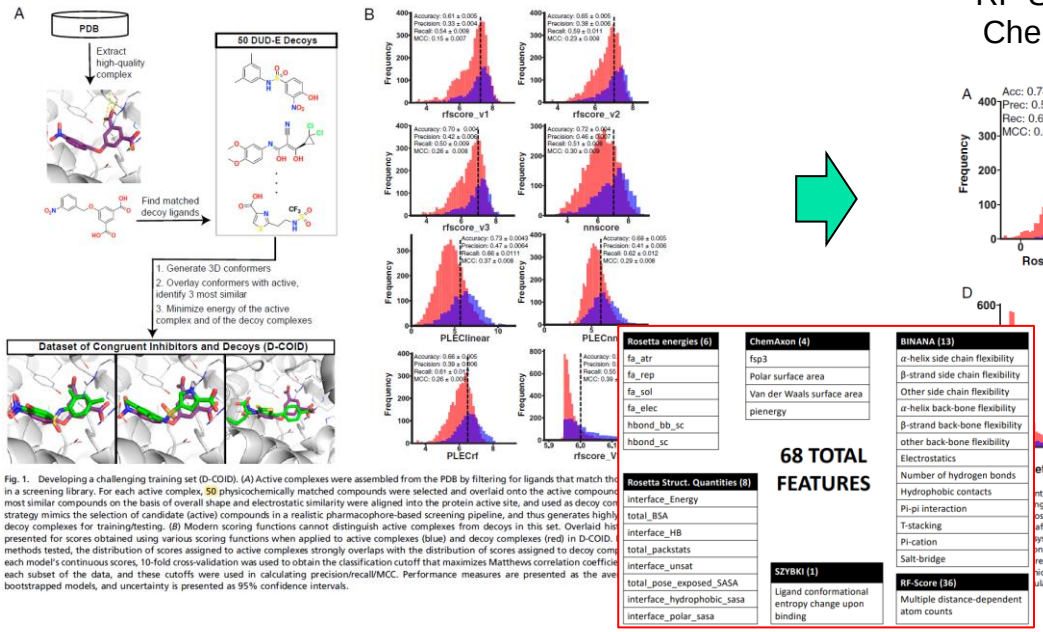
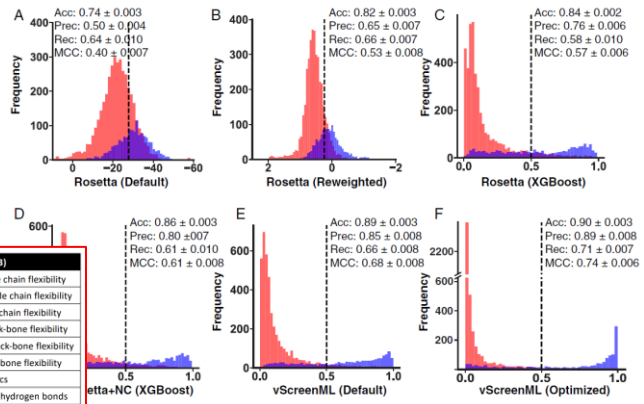


Fig. 1. Developing a challenging training set (D-COID). (A) Active complexes were assembled from the PDB by filtering for ligands that match the... in a screening library. For each active complex, 50 physicochemically matched compounds were selected and overlaid onto the active compound... most similar compounds on the basis of overall shape and electrostatic similarity were aligned into the protein active site, and used as decoy com... strategy mimics the selection of candidate (active) compounds in a realistic pharmacophore-based screening pipeline, and thus generates highly... decoy complexes for training/testing. (B) Modern scoring functions cannot distinguish active complexes from decoys in this set. Overlaid his... presented for scores obtained using various scoring functions when applied to active complexes (blue) and decoy complexes (red) in D-COID. ... methods tested, the distribution of scores assigned to active complexes strongly overlaps with the distribution of scores assigned to decoy com... each model's continuous scores, 10-fold cross-validation was used to obtain the classification cutoff that maximizes Matthews correlation coeffici... each subset of the data, and these cutoffs were used in calculating precision/recall/MCC. Performance measures are presented as the ave... bootstrapped models, and uncertainty is presented as 95% confidence intervals.

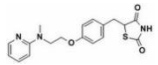
Rosetta energy function = classical MM + empirical terms
- reweight terms
- component energy**로** decision tree, XGBoost
- structural quality assessment term **추가**
- RF-Score (contact),BINANA (contact), ChemAxon (descriptor), Szybki (descriptor)



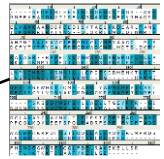
nt of vScreenML. Overlaid histograms are presented for scores obtained when scoring active complexes (blue) and decoy complexes (red) ng functions used were: (A) default Rosetta energy function, (B) linearly reweighted Rosetta energy terms, (C) Rosetta energy terms ost, (D) Rosetta energy terms plus structural assessments, (E) Rosetta terms plus additional diverse descriptors (nonoptimized vScreenML), after hyperparameter tuning. Over the course of this sequence, the overlap between the active and decoy complexes is progressively ystematically increases. For the first two panels, 10-fold cross-validation was used to obtain the classification cutoff that maximizes on coefficient (MCC) on each subset of the data, and these cutoffs were used in calculating precision/recall/MCC. Because the remaining results from classification models, their thresholds are fixed at 0.5. Performance measures are presented as the average of 100 trained hich derived from 10-fold cross-validation (*Methods*). Uncertainty is presented as 95% confidence intervals. In all cases, performance ulated for a subset of the data that was held out from the training step.

www.pnas.org/cgi/doi/10.1073/pnas.2000585117

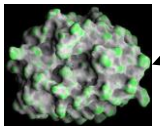
Virtual Screening



SMILES
Fingerprint
Library filtering



Homology modeling



3D Structure of Target

No

Yes

Binding site detection

Ligand-Based Methods

Structure-Based Methods

Similarity search

Descriptors
Fingerprint

Ligand superposition (3D)

Protein-ligand docking

Similarity Searching

QSAR / Machine Learning

Pharmacophore Mapping

Protein-Ligand Docking

RDKit, DeepChem,
Keras/TensorFlow
(Free)

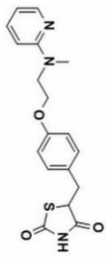
KINME, RapidMiner
PipelinePilot

Schrodinger
(Maestro, Glide, Prime)
($$$)

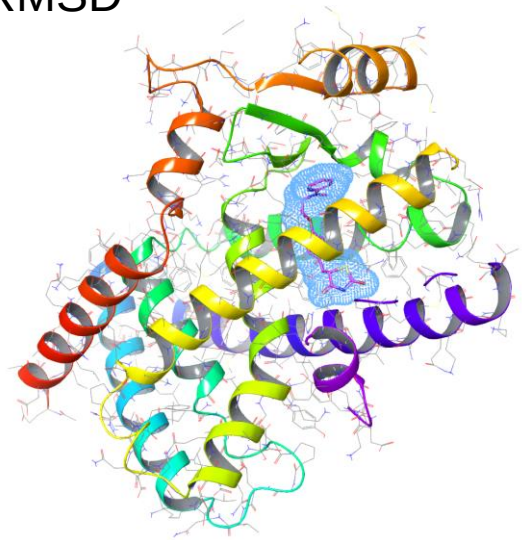
AutoDock
Gromacs

Protein-Ligand Docking

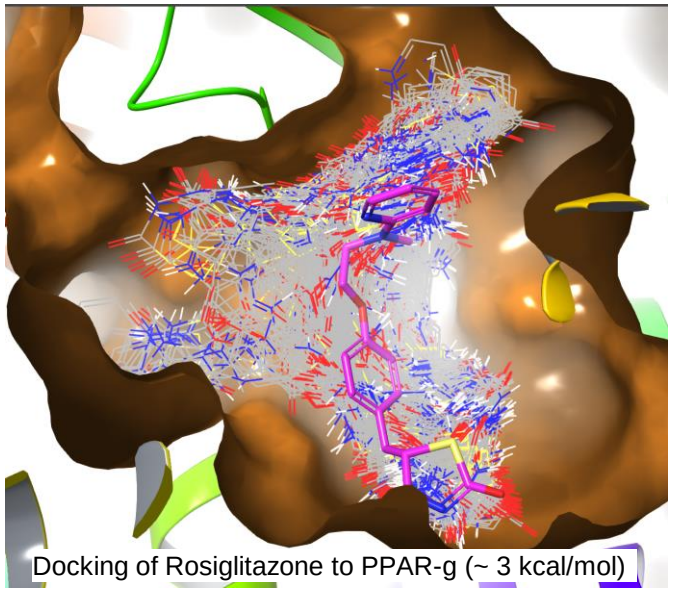
- Goal
  - given a protein structure, predict its ligand binding (optimal pose, energetics)
- Questions
  - Where will the ligand bind?
  - Which ligand will bind
  - How will the ligand bind
  - When, Why, etc
- Metric
  - RMSD



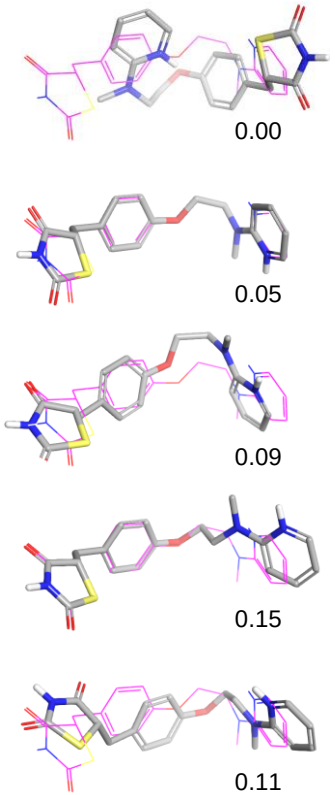
+



=



Docking of Rosiglitazone to PPAR-g (~ 3 kcal/mol)

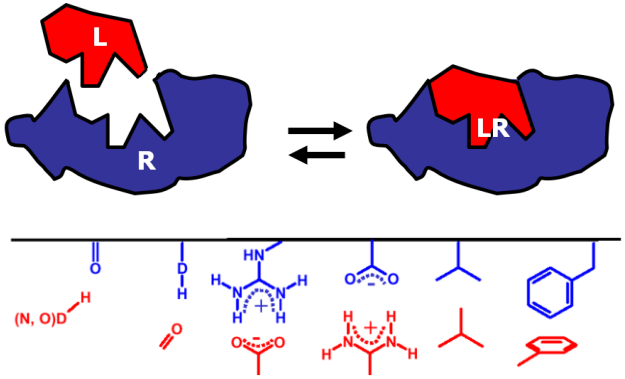


Docking : Challenges

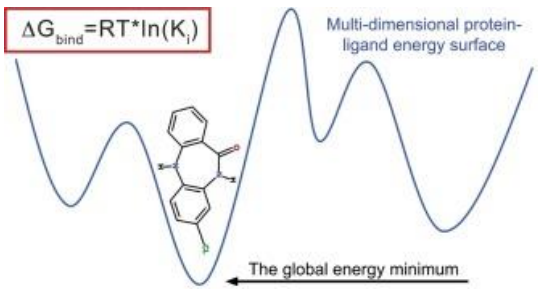
- Docking : placing a ligand into a receptor cavity

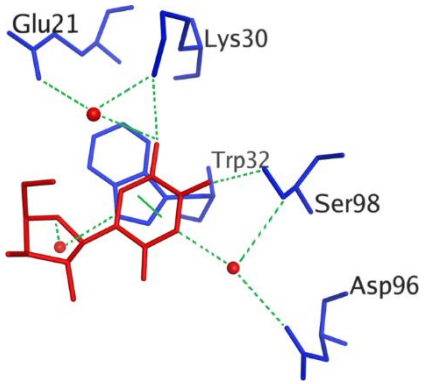
- Druggable binding site
  - Cavity : depth, volume, lipophilic surface
  - ~ 6,000 druggable sites in PDB

- Challenges
  - Scoring function
  - Protein / Ligand flexibility ; Water



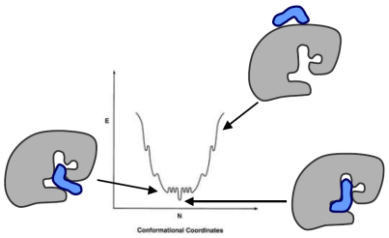
$$dG = -RT \ln KA; \quad K_A = \frac{[LR]}{[L][R]}$$





Scoring Function

- Aim
  - Given protein-ligand poses
  - Estimate binding affinity



	Protein–ligand	Internal ligand
G-Score	$E_{vdW}+E_{H\text{-}bond}=\sum_{prot}\sum_{lig}\left[\left(\frac{A_{ij}}{d_{ij}^{12}}-\frac{B_{ij}}{d_{ij}^{6}}\right)+(E_{vdw}+E_{hb})-(E_{vdw}+E_{hb})\right]$	$E_{vdw}+E_{torsion}=\sum_{lig}\left(\frac{C_{ij}}{d_{ij}^{12}}-\frac{D_{ij}}{d_{ij}^{6}}\right)+\sum_{lig}\frac{1}{2}V\left[1+\frac{n}{|\pi|}\cos(|\pi|\varphi)\right]$
D-Score	$E_{vdW}+E_{electrostatic}=\sum_{prot}\sum_{lig}\left[\left(\frac{A_{ij}}{d_{ij}^{12}}-\frac{B_{ij}}{d_{ij}^{6}}\right)+332.0\frac{q_iq_j}{\epsilon(d_{ij})d_{ij}}\right]$	
Gold	$E_{vdW}+E_{electrostatic}=\sum_{prot}\sum_{lig}\left[\left(\frac{A_{ij}}{d_{ij}^{12}}-\frac{B_{ij}}{d_{ij}^{6}}\right)+332.0\frac{q_iq_j}{\epsilon(d_{ij})d_{ij}}\right]$	$E_{vdW}+E_{electrostatic}=\sum_{lig}\left[\left(\frac{A_{ij}}{d_{ij}^{12}}-\frac{B_{ij}}{d_{ij}^{6}}\right)+332.0\frac{q_iq_j}{\epsilon(d_{ij})d_{ij}}\right]$ + optional $E_{H\text{-}bond}$
AutoDock	$E_{vdW}+E_{H\text{-}bond}+E_{electrostatic}=\sum_{prot}\sum_{lig}\left[\left(\frac{A_{ij}}{d_{ij}^{12}}-\frac{B_{ij}}{d_{ij}^{6}}\right)+E(t)\times\left(\frac{C_{ij}}{d_{ij}^{12}}-\frac{D_{ij}}{d_{ij}^{10}}\right)+332.0\frac{q_iq_j}{\epsilon(d_{ij})d_{ij}}\right]$ $E(t)$ = angular weight factor	$E_{vdW}+E_{H\text{-}bond}+E_{electrostatic}=\sum_{lig}\left[\left(\frac{A_{ij}}{d_{ij}^{12}}-\frac{B_{ij}}{d_{ij}^{6}}\right)+E(t)\left(\frac{C_{ij}}{d_{ij}^{12}}-\frac{D_{ij}}{d_{ij}^{10}}\right)+332.0\frac{q_iq_j}{4(d_{ij})d_{ij}}\right]$ $E(t)$ = angular weight factor
DOCK (v4.0)	$E_{vdW}+E_{electrostatic}=\sum_{prot}\sum_{lig}\left[\left(\frac{A_{ij}}{d_{ij}^{12}}-\frac{B_{ij}}{d_{ij}^{6}}\right)+332.0\frac{q_iq_j}{\epsilon(d_{ij})d_{ij}}\right]$	

- Scoring functions
  - Force field (MM) :
    - Physics-based : E(R-L) + E(L) + E(R)
    - Enthalpy in gas phase only, but no solvation and entropy
  - Empirical :
    - Fit to reproduce experimental data
  - Knowledge-based :
    - Experimental structure
    - Simple

	Functional form
PMF	Parametrized pairwise potential PMF score : $PMF=\sum_{prot}\sum_{lig}A_{ij}(d_{ij})\quad A_{ij}(d_{ij})=-k_BT\ln\left[f^{i}_{Vol\_corr}(r)\frac{\rho^{ij}_{obs}(r)}{\rho^{ij}_{bulk}}\right]$ where k_B is the Boltzmann constant, $f^{i}_{Vol\_corr}(r)$ is a ligand volume correction factor and $\frac{\rho^{ij}_{obs}(r)}{\rho^{ij}_{bulk}}$ indicates a radial distribution function for a protein atom i and a ligand atom j.
DrugScore (v1.2)	$\Delta W=\gamma\sum_{prot}\sum_{lig}\Delta W_{ij}(r)+(1-\gamma)\times\left[\sum_{lig}\Delta W_i(SAS,SAS_0)+\sum_{prot}\Delta W_j(SAS,SAS_0)\right]$ SAS = Solvent accessible surface area terms, W_{ij} = distance dependent pairwise potential
SMoG	$G=\sum_{ij}g_{ij}\Delta_{ij},\qquad \Delta_{ij}=\begin{cases}0 & \text{(i, j more than 5 Å)}\\ 1 & \text{(i, j within 5 Å)}\end{cases},\qquad g_{ij}=-kT\log\left[\frac{p_{ij}}{\bar{p}}\right]$; p_{ij} an $\bar{p}$ are interatomic and averaged interactomic interactions

	Functional form
LUDI	$\Delta G_{bind}=\Delta G_{H\text{-}bond}\sum_{H\text{-}bond}f(\Delta R,\Delta\alpha)+\Delta G_{ionic}\sum_{ionic}f(\Delta R,\Delta\alpha)+\Delta G_{hydrophobic}\sum_{hydrophobic}\left\lvert A_{hydrophobic}\right\rvert+\Delta G_{rotor}N_{rotor}+\Delta G_0$ $A_{hydrophobic}$ = molecular surface area
F-Score	$\Delta G_{bind}=\Delta G_{H\text{-}bond}\sum_{H\text{-}bond}f(\Delta R,\Delta\alpha)+\Delta G_{ionic}\sum_{ionic}f(\Delta R,\Delta\alpha)+\Delta G_{aromatic}\sum_{aromatic}f(\Delta R,\Delta\alpha)+\Delta G_{contact}\sum_{contact}f(\Delta R,\Delta\alpha)+\Delta G_{rotor}N_{rotor}+\Delta G_0$
Chem-Score	$\Delta G_{bind}=\Delta G_{H\text{-}bond}\sum_{H\text{-}bond}f(\Delta R,\Delta\alpha)+\Delta G_{metal}\sum_{metal}f(\Delta R,\Delta\alpha)+\Delta G_{lipo}\sum_{lipo}f(\Delta R)+\Delta G_{rotor}\sum_{rotor}f(P_{ai},P_{aj})+\Delta G_0$

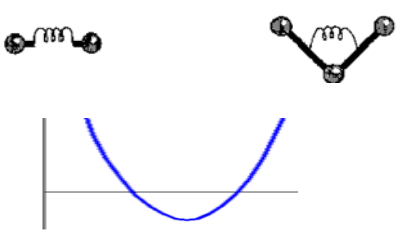
The free energy of binding, ΔG_{bind} is approximated as a sum of contributing free energy terms of hydrogen bonding (*H-bond*), ionic (*ionic*), hydrophobic (*hydrophobic*), ligand rotational entropy (*rotor*), contact (*contact*), lipophilic (*lipo*) and metal (*metal*) components. The scoring functions differ in the terms included and the functional forms of the contributing free energy terms. $\Delta G_{H\text{-}bond}$, ΔG_{ionic}, $\Delta G_{hydrophobic}$, ΔG_{rotor}, $\Delta G_{aromatic}$, $\Delta G_{contact}$, ΔG_{metal}, ΔG_{lipo} are regression coefficients for each corresponding free energy term. ΔG_0 is a regression constant. The free energy terms are calculated with a function, f, which can depend on an angular ($\Delta\alpha$) and/or a b distance (ΔR) term.

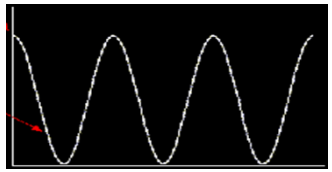
- Consensus scoring
  - X-CSORE = GOLD + DOCK + ChemScore + PMF +FlexX

Energy function : Forcefield

- Functional form and parameter sets
- To describe the potential energy of a system of particles (atoms)
- AMBER, CHARMm, CVFF, CFF, MMFF, OPLS, …

$$E_{pot} = \underbrace{\sum_b K_2 (b - b_0)^2}_{\text{length}} + \underbrace{\sum_\theta H_\theta (\theta - \theta_0)^2}_{\text{angle}} + \underbrace{\sum_\phi \frac{V_n}{2} [1 + \cos(n\phi - \phi_0)]}_{\text{torsion}} + \underbrace{\sum \varepsilon [(r^*/r)^{12} - 2(r^*/r)^6]}_{\text{vdw}} + \underbrace{\sum q_i q_j / \varepsilon_{ij} r_{ij}}_{\text{electrostatic}} + \underbrace{\sum \left[\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}}\right]}_{\text{H-bonding}}$$

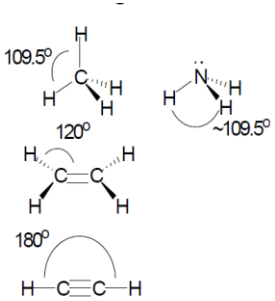


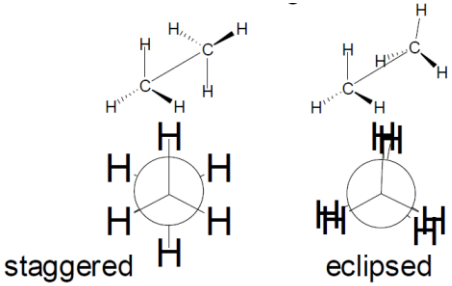


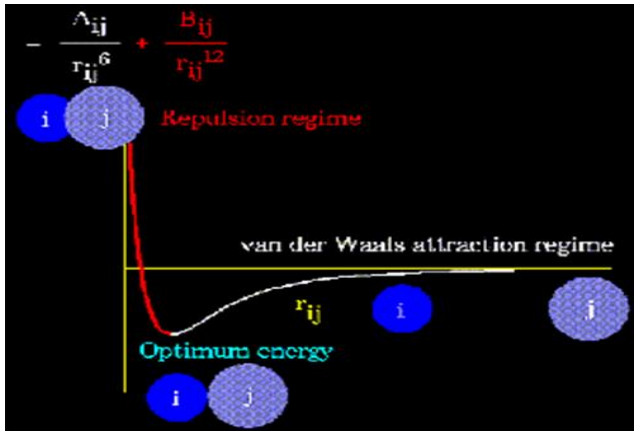
types		kb	r0
1	1	4.258	1.508

atom types				V1	V2	V3
1	1	1	3	0.066	-0.156	0.143

types			ka	theta0
1	1	3	0.851	108.9



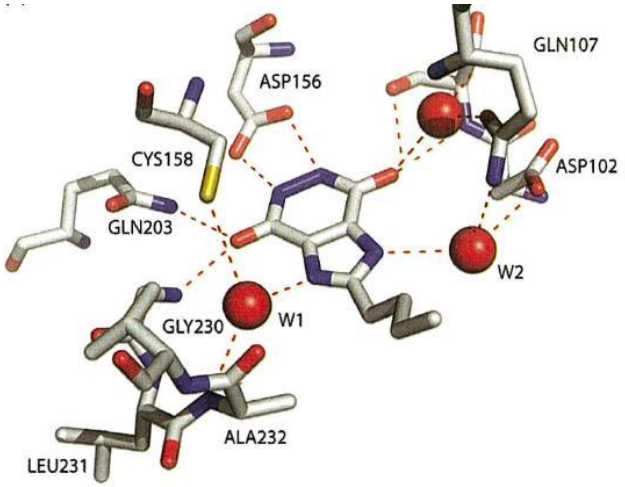


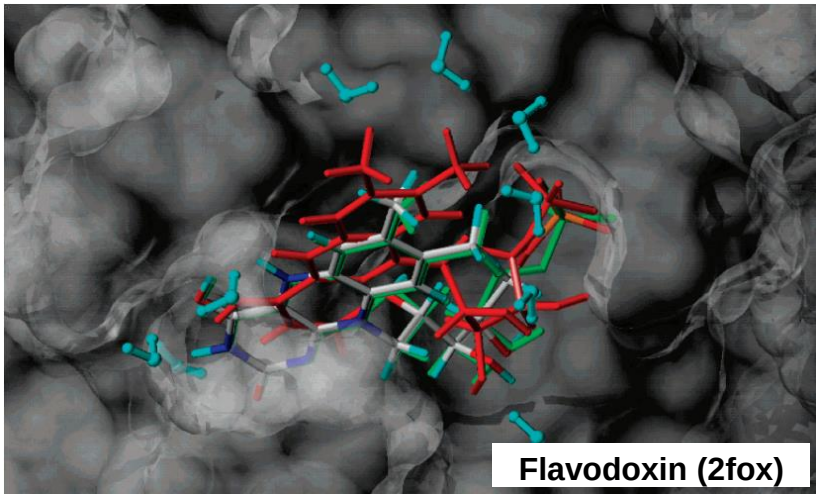


atom type	R^{*j}	ϵ^k
C'	1.9080	0.0860
CA	1.9080	0.0860
CM	1.9080	0.0860
Cs	3.3950	0.0000806
CT	1.9080	0.1094
F	1.75	0.061
H	0.6000	0.0157
H1	1.3870	0.0157

Water Hydration

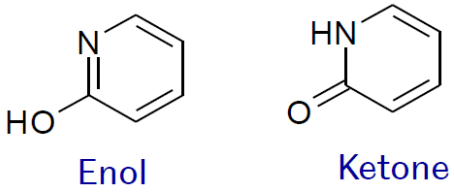
- Important role in protein - ligand interaction : hydrogen bonding, size, shape, polarity of binding site, binding affinity
- Explicit consideration
  - Water positions are from X-ray crystallography, MD, …
- Some waters are displaceable for binding affinity
- The number and position of water depend on the protein conformation and ligand structure



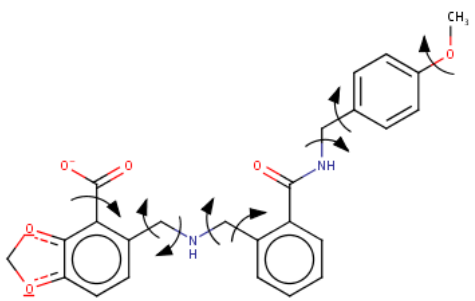


Ligand Protonation & Flexibility

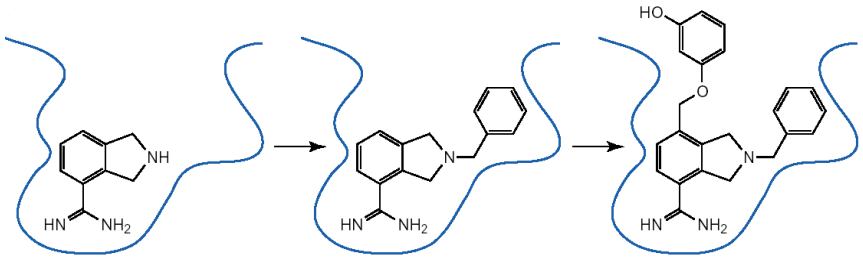
- Protonation state / Tautomeric form influence H-bonding



- Conformation
  - Bioactive conformation ← Sufficient sampling of ligand conformation space
  - Systematic search : combinatorial explosion problem; Hammerhead method; pre-generated conformation DB
  - Random search : Monte Carlo, Genetic, Tabu
  - Simulation : MD / Simulated annealing
  - Incremental construction

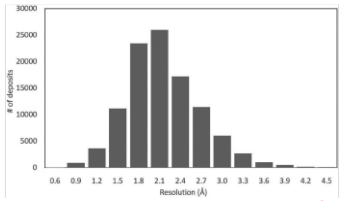


Incremental construction

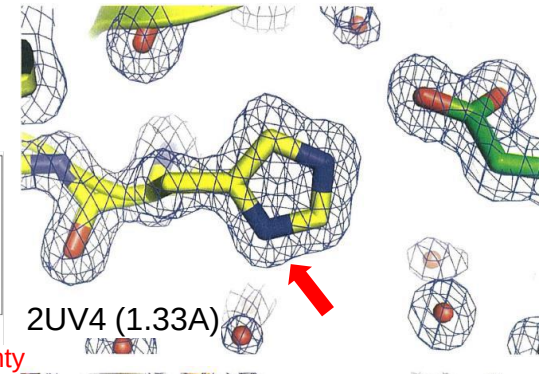


Protein Flexibility & Protonation

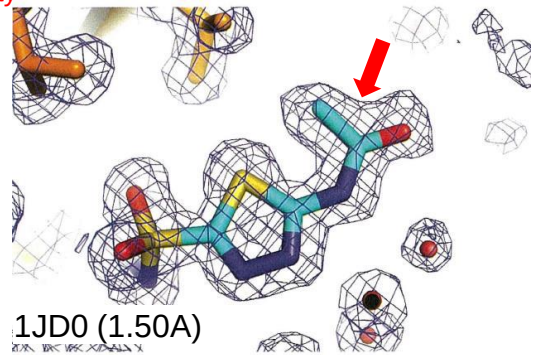
- Problem
  - No single representative protein structure
  - Experimental uncertainty
- Methods
  - MD, MC
  - Rotamer library
  - Induced fit docking
  - Ensemble docking, Consensus scoring



Uncertainty

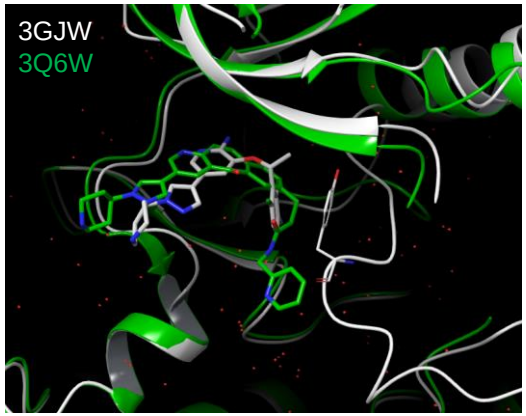


2UV4 (1.33A)



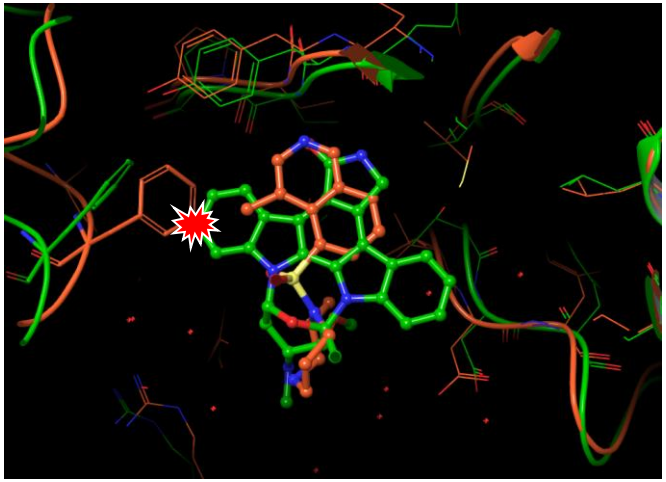
1JD0 (1.50A)

Electron density



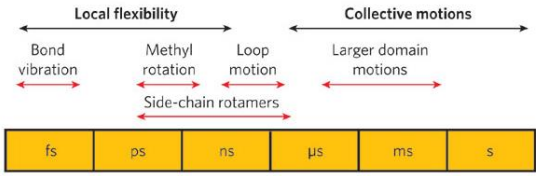
Crizotinib vs MK2461

Large scale movement



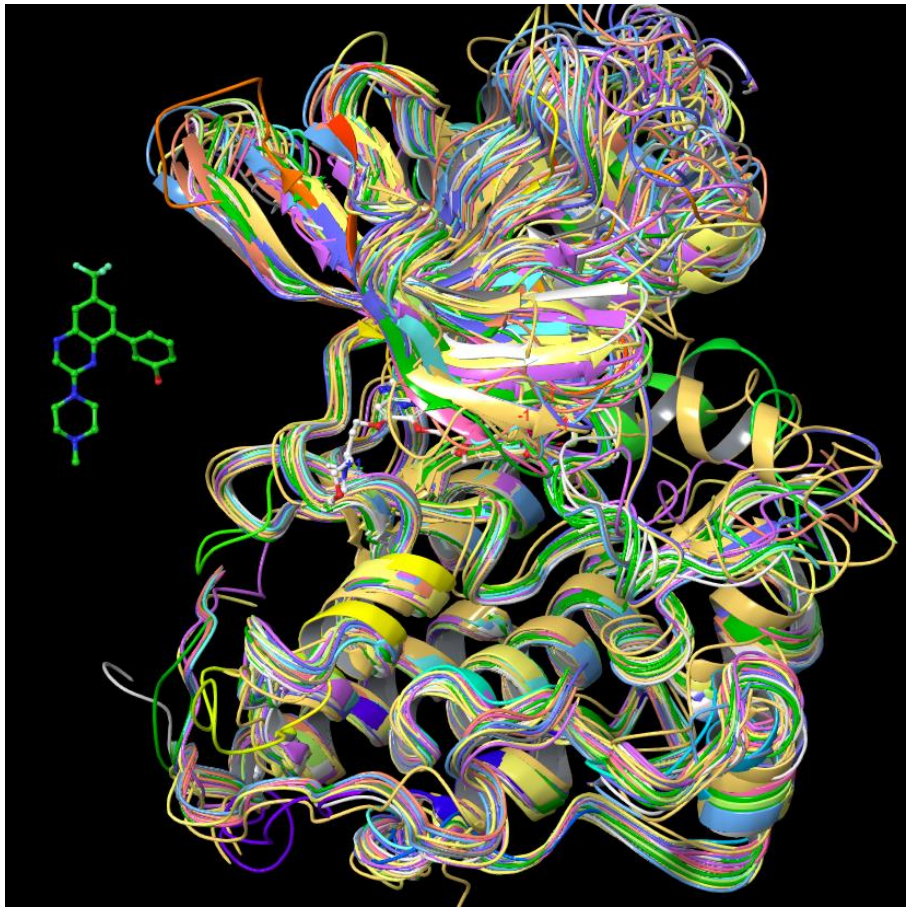
Staurosporine in kinases (1STC, 1Q8U)

Sidechain movement

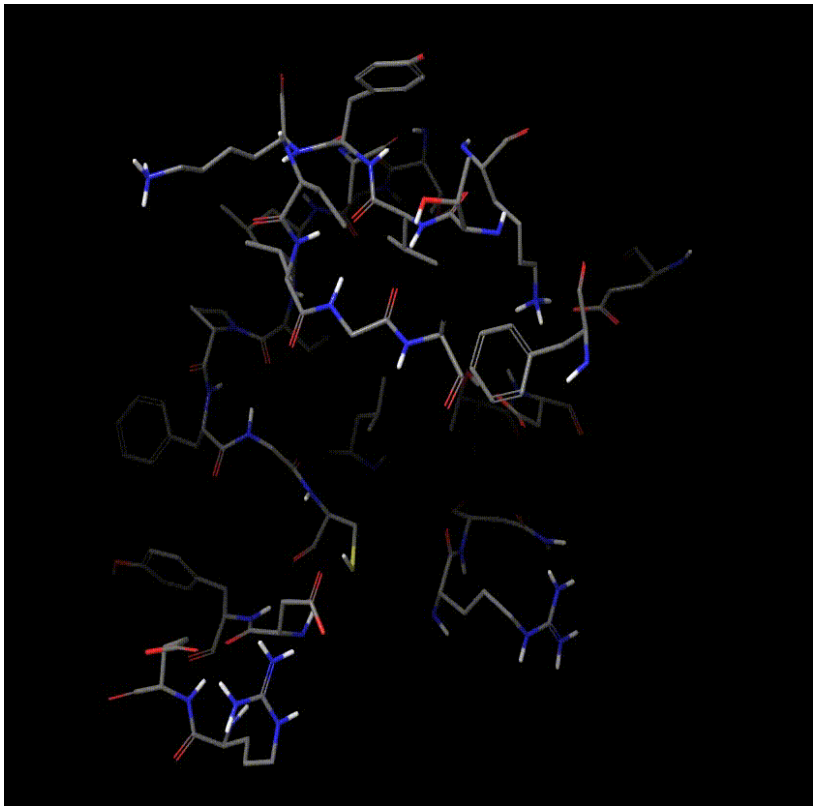


Nature 450, 964-972 (2007)

Dynamics of Proteins (Kinases)



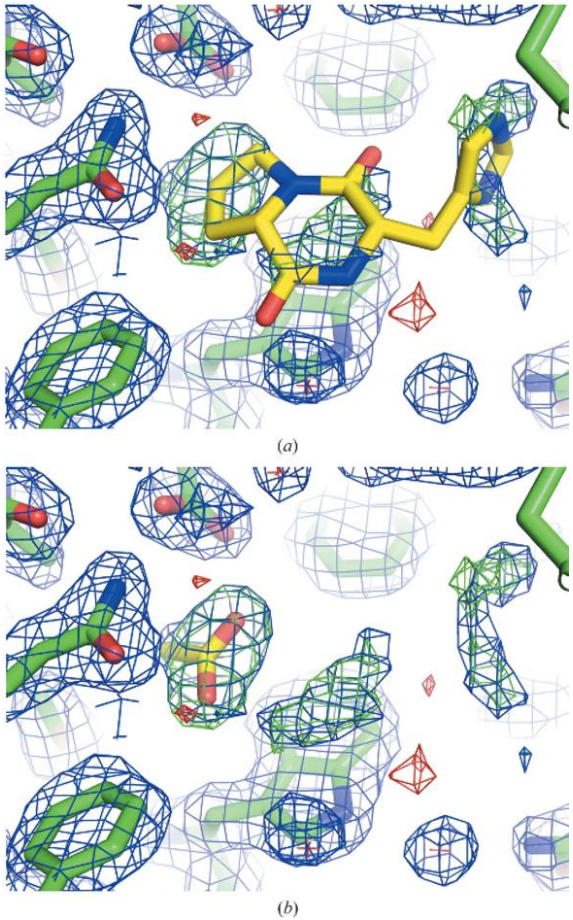
Overlap of 43 cMet Kinases

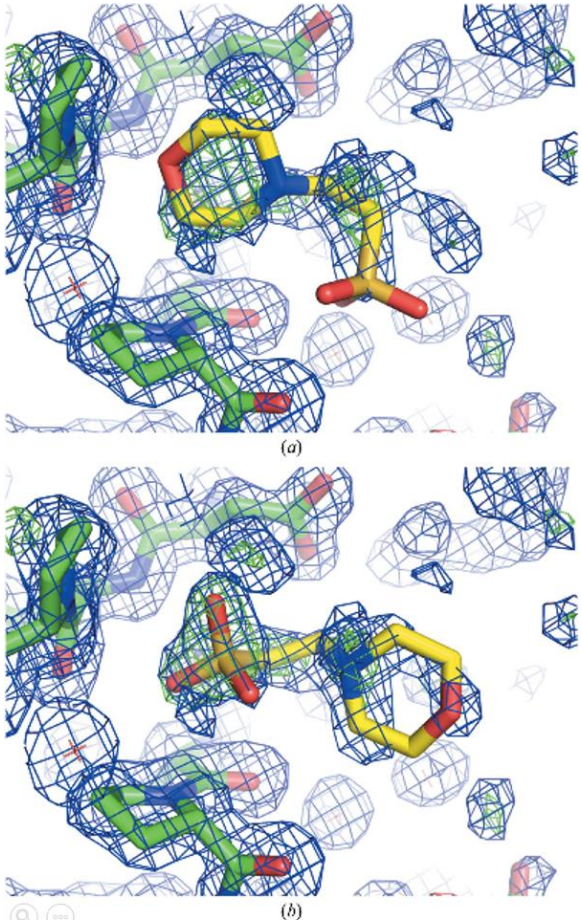


Dynamics of Protein Structure

Uncertainty in Protein Structure

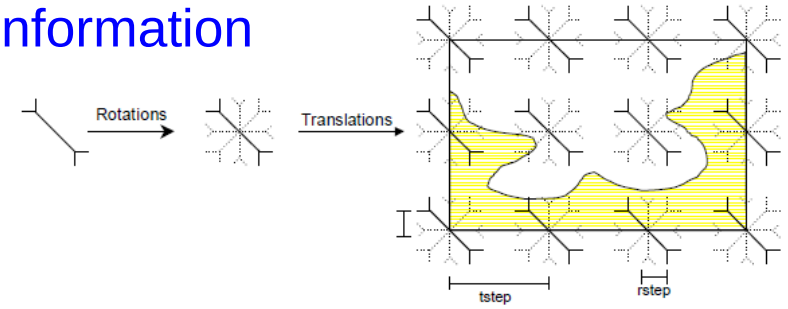
- Ligand identification in twilight electron density
- Mistake in this step cannot be rescued!





Search : Systematic

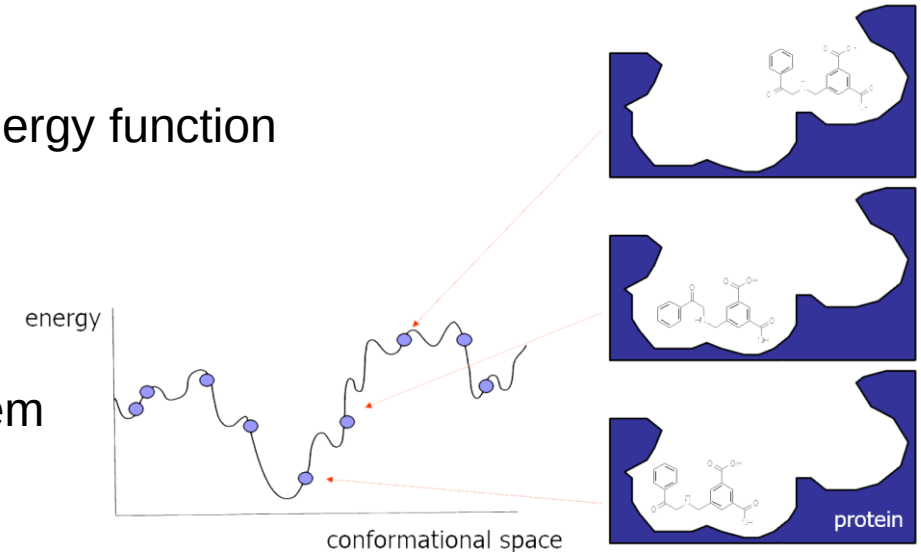
- Search : position, orientation, ligand conformation



- Systematic Search
  - Uniform sampling of search space
  - Exhaustive, deterministic
  - Quality depends on granularity of sampling
  - Feasible only for low-dimensional problems

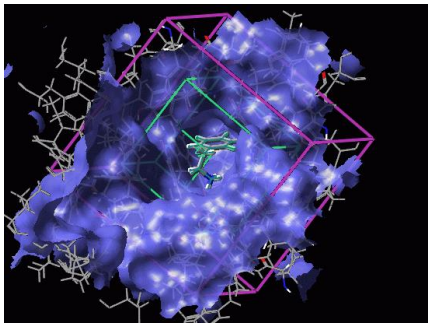
- Local minimization
  - Start from specific pose
  - Move directed by derivative of energy function
  - Stop at local minima

- MC/MD
  - To overcome local minima problem
  - Controlled by temperature



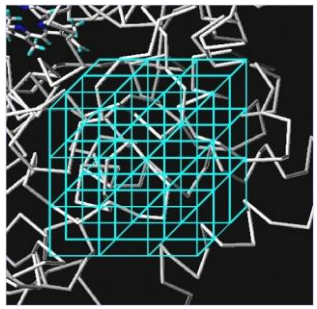
Receptor Energy Grid

- Computation of binding energy at binding site is very costly
  - Sum terms computed at positions of ligand atoms vs protein atoms
  - Precompute “force field” for each term of scoring function
  - Sample force fields at positions of ligand atoms
- Accelerate calculation of scoring function by 100X
- ↔ Fixed Receptor structure : protonation, flexibility, water



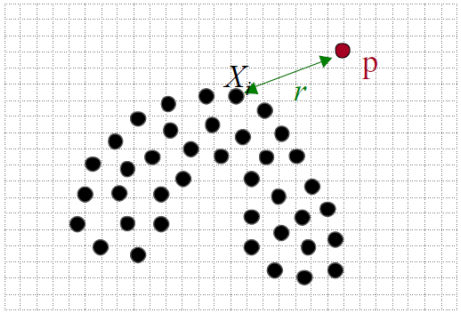
Set up boundaries for grid

Generate energy grids

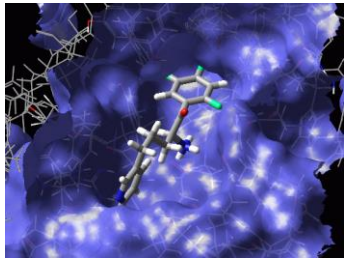


Grids

Calculate energy at atom positions



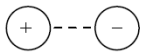
Dock a molecule into the grid



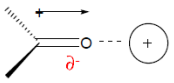
Important Factors in Docking Analysis

Electrostatic

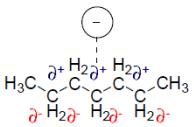
Charge-charge
- pH dependence



Charge-dipole



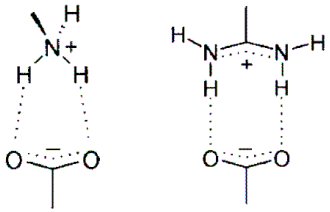
Charge-induced dipole



increasing distance dependence

Salt bridge

4-7 kca



energetics similar to H-bonds (2-10 kcal/mol)

H-bonding

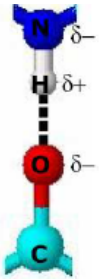
2.5-3.2 Å

Donor

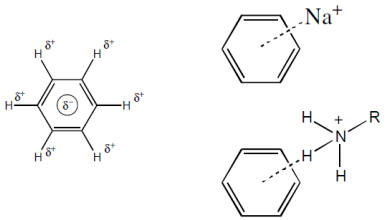
highly directional
(130° ≤ Θ ≤ 180°)

2-10 kca'

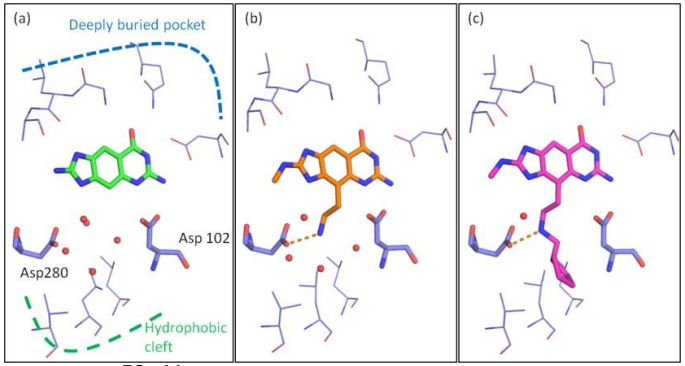
Acceptor



Cation-π

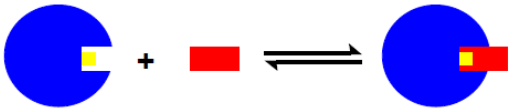


Water-mediated



58 nM 55 nM 2 nM

Solvation

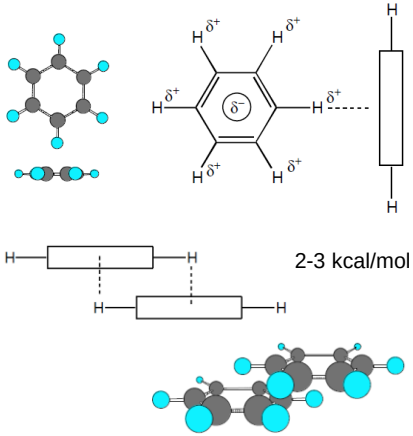


Binding pocket becomes "interior" upon complexation with ligand

Big penalty: charged or polar groups buried but unpaired

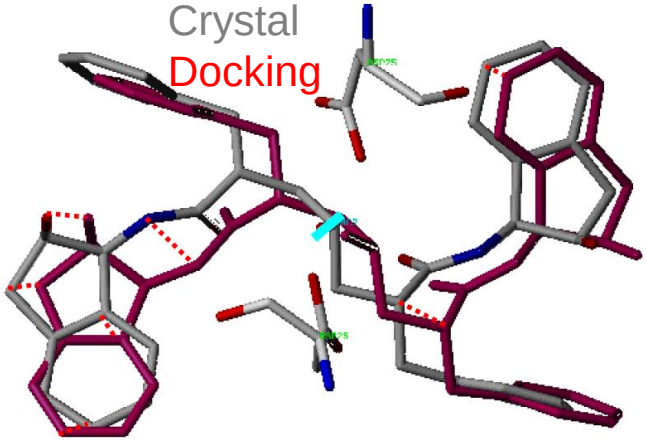
Energetic contribution is propotional to the size of the surface buried upon ligand binding
- e.g. $-CH_3$ group (25 Å^2): 3 to 6 kcal/mol

Π-Stacking (Face to face)

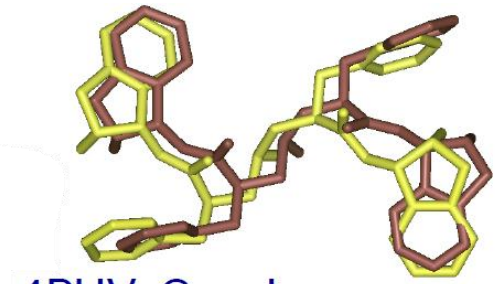


Evaluation of Docking Performance

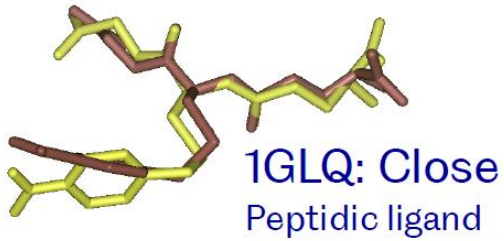
- Extract the ligand from a known protein-ligand complex from the PDB
- Minimize the conformation of the ligand (outside protein)
- Dock back into the protein
- Compare the docked pose with the experimental data
- GOLD experiment with 224 protein structures :
  - 72% Crystal structure were reproduced as top rank within 2Å RMSD



$$RMSD = \sqrt{\frac{\sum_{N} (x_a - x_b)^2 + (y_a - y_b)^2 + (z_a - z_b)^2}{N}}$$

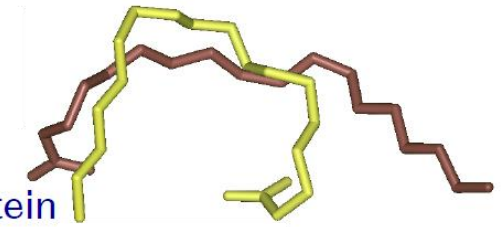


4PHV: Good
HIV Protease
15 rotatable bonds



1GLQ: Close
Peptidic ligand

1CIN: Wrong
Fatty acid binding protein



Proteins 2002, 49, 457-471

Current Status of Docking

- Docking is widely used, with many success story

- Scoring function
  - Performance varies target, scoring function, …
  - Reasonable for pose prediction, not for binding energy estimation

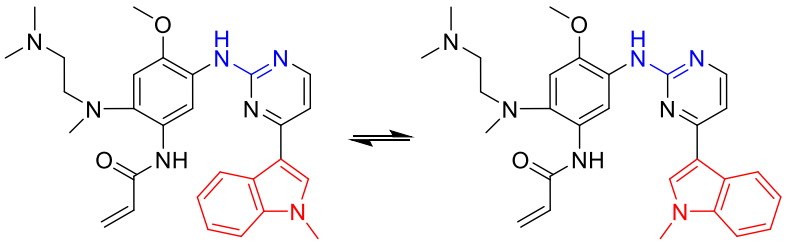
- Ligand flexibility : conformer library, incremental docking
- Protein flexibility : sidechain only, for small number of ligands
- Water : switch on/off

- Protein preparation is most important : conformation, protonation, …

- Variety of post-processing : to minimize false positives

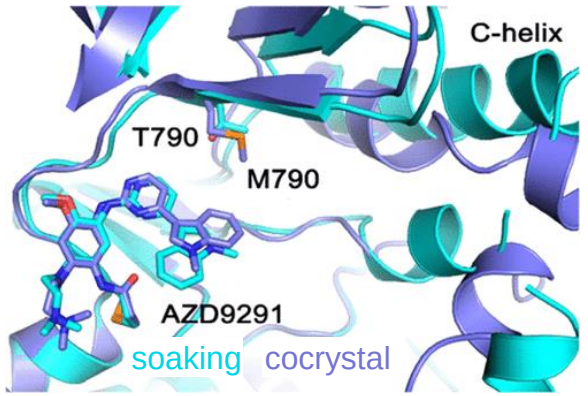
Binding of Osimertinib to EGFR WT/T790M

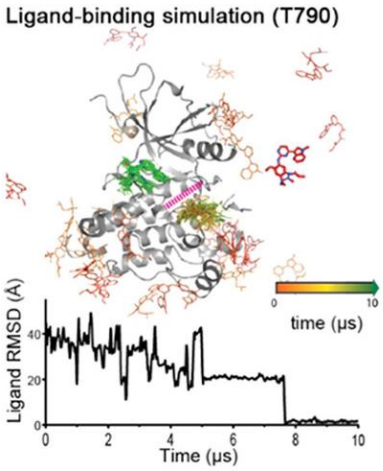
- AZD binds as un-flipped form, and flip at binding site
- WT prefers un-flipped form, while T790M prefers flipped form due to vdw interaction (~0.4 kcal)
- Crystallization & unconstrained molecular dynamics simulation (10 us !)

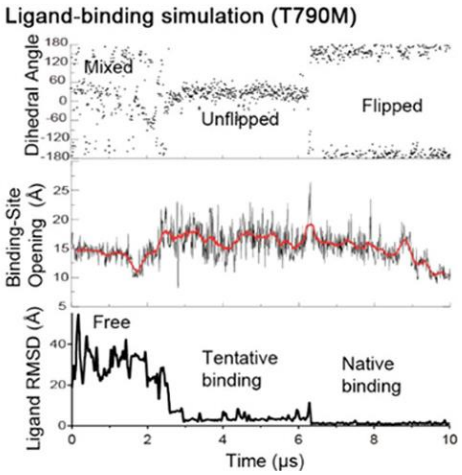


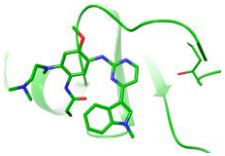
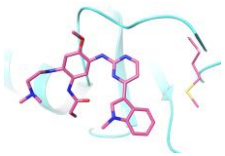
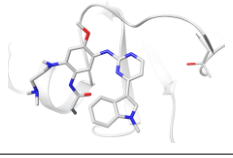
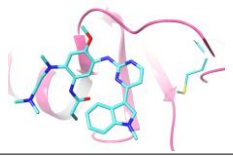
Un-flipped, solution

Flipped, T790M



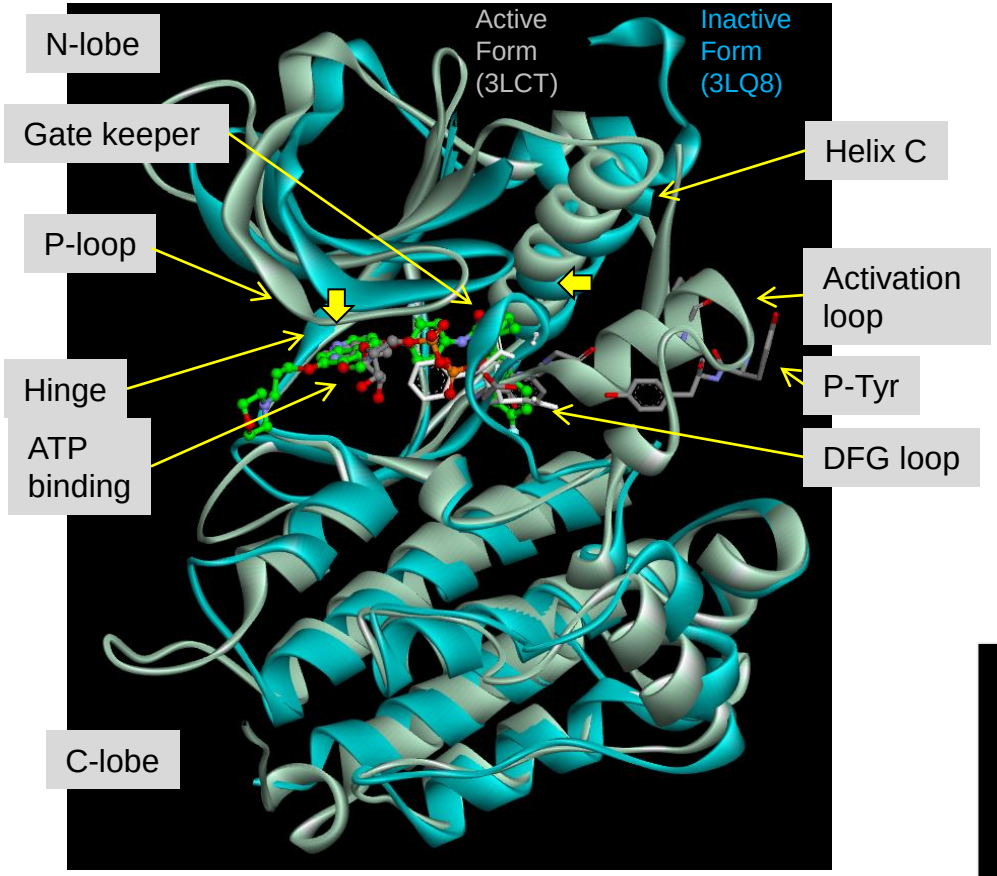




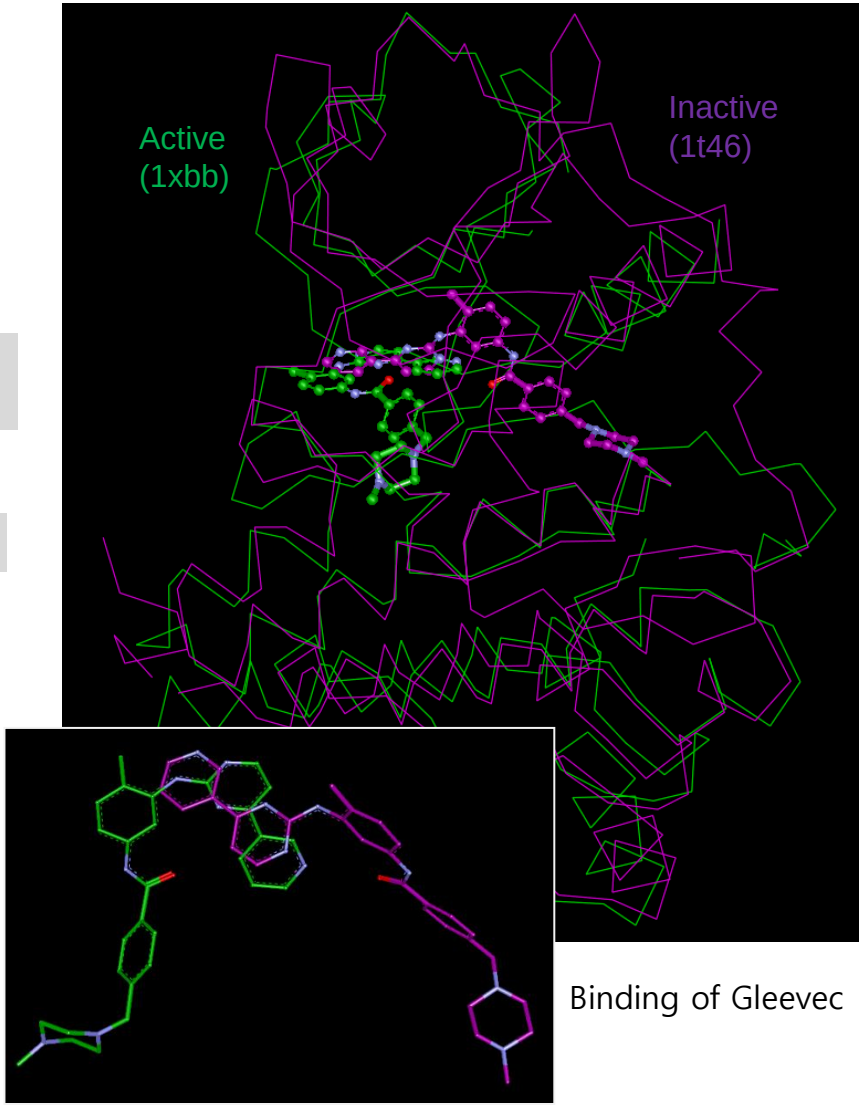
	WT	T790M
Co-crystal		
soaking		

https://pubs.acs.org/doi/10.1021/acs.jmedchem.0c00891

Protein Structure : Kinase

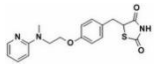


Comparison of active/inactive form



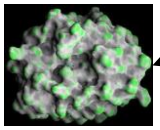
Binding of Gleevec

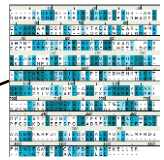
Virtual Screening



SMILES
Fingerprint
Library filtering

3D Structure of Target





Homology modeling

Binding site detection

Ligand-Based Methods

Structure-Based Methods

Similarity search

Descriptors
Fingerprint

Ligand superposition

Protein-ligand docking

Similarity Searching

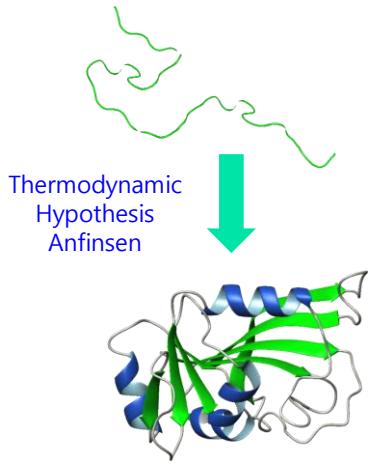
QSAR / Machine Learning

Pharmacophore Mapping

Protein-Ligand Docking

Protein Structure Prediction

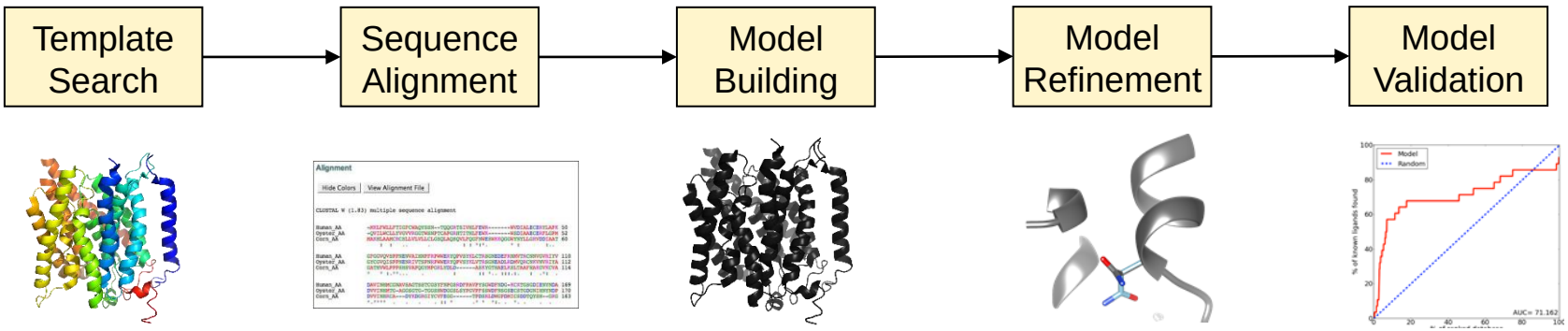
- Limit of current experimental techniques : 167K PDB / 10M proteins
- Assumptions
  - Increase in sequence identity correlates with increase in structural similarity
  - RMSD of core Cα ~1Å when 50% sequence identity; >70% recommended
  - Quality of theoretical models depend on quality of input alignment!
- Structure of proteins
  - Primary : AA sequences
  - Secondary : a-helix, b-sheets, coil, turn, …
  - Tertiary : 3D structure of entire polypeptide
  - Quaternary : Spatial arrangement of subunits
- Methods
  - Ab initio; Energy-based simulation
    - Thermodynamic equilibrium with a minimum free energy
  - Comparative modeling; Knowledge-based simulation
    - Homology-based, threading, hierarchical



https://en.wikipedia.org/wiki/List_of_protein_structure_prediction_software

Homology Modeling Procedure

- Spend time on early steps
- Consider all structures and many sequences
- Optimize the region of interest, not a overall single score



SIMILAR?

- Similar globally?
- Similar local?
- Similar in function?
- Conformation?

Superposition

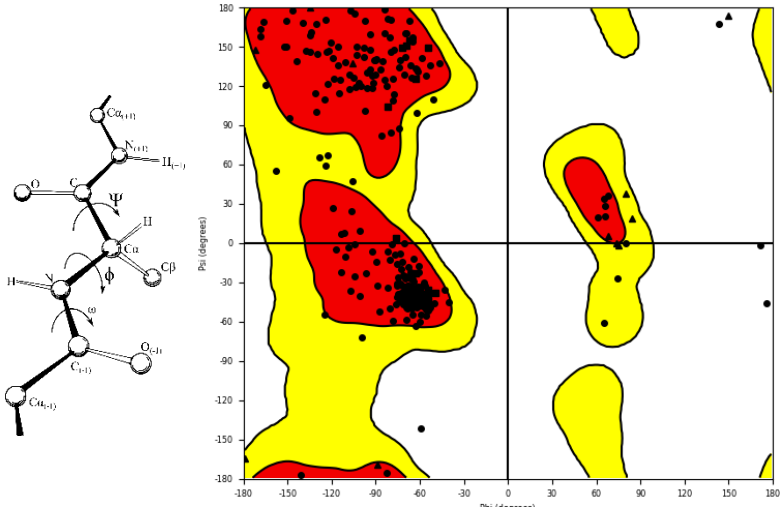
- Dynamic Programming, Pairwise Alignments
- Multiple Alignments, Profiles, HMMs
- Structure based approaches (Threading)
- Multi-sequence target alignment
- Don't apply rules blindly

https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1007449

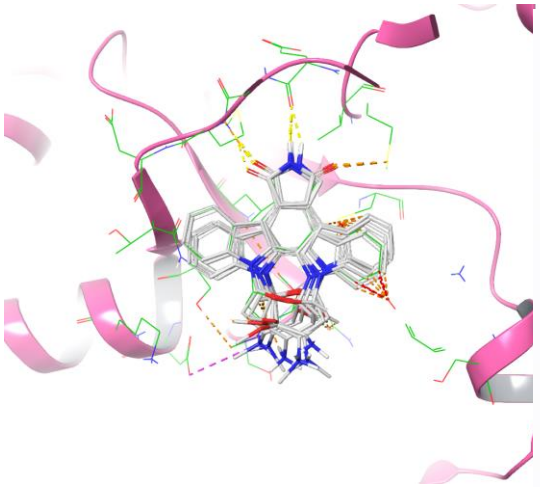
SBDD : Homology Modeling

- Template structure
- Sequence alignment

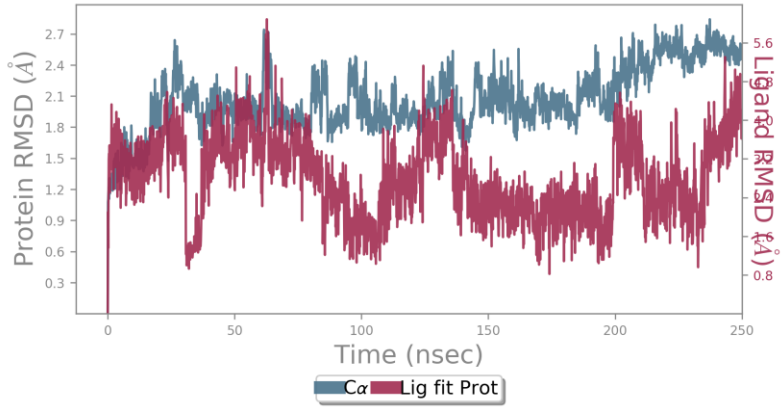
Homology Modeling : Validation



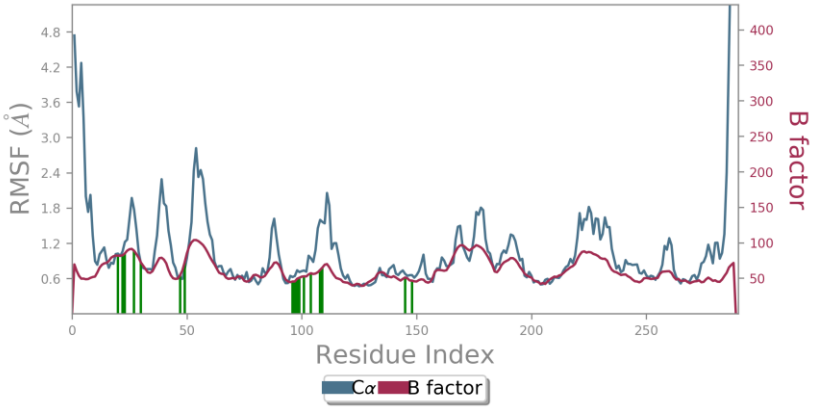
Ramachandran plot



Redocking of original/active ligand



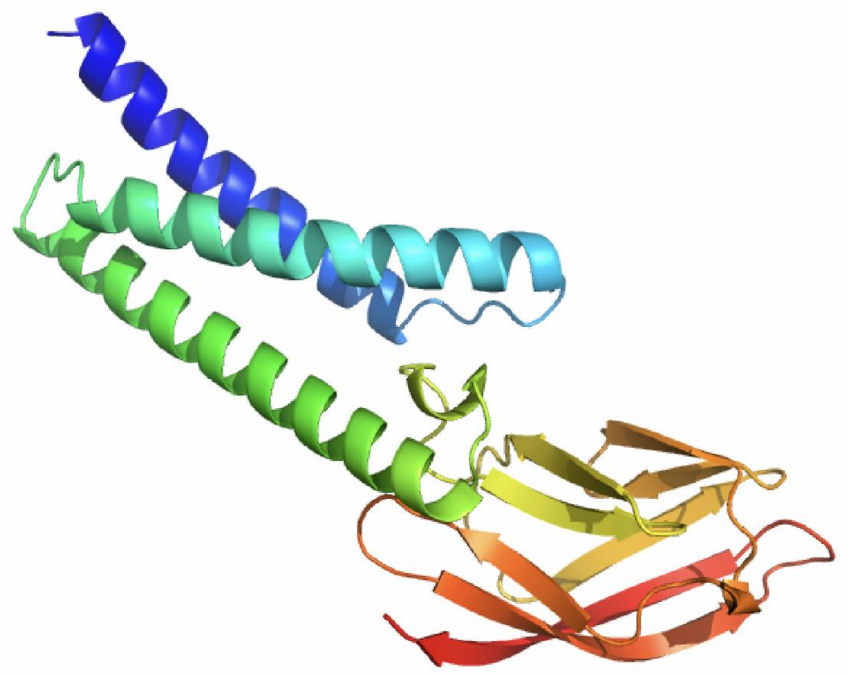
Protein stability via MD Simulation



Protein stability via MD Simulation

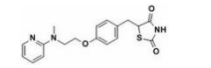
AlphaFold

- AlphaFold at DeepMind
- Source code Source code at GitHub
- COVID protein structure prediction under progress : 5-targets (SARS-CoV-2 membrane protein, Nsp2, Nsp4, Nsp6, and Papain-like proteinase)



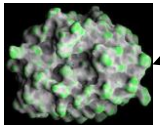
Structure of COVID-19 membrane protein created by AlphaFold

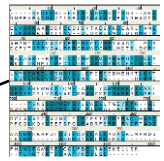
Virtual Screening



SMILES
Fingerprint
Library filtering

3D Structure of Target





Homology modeling

Binding site detection

Ligand-Based Methods

Structure-Based Methods

Similarity search

Descriptors
Fingerprint

Ligand superposition

Protein-ligand docking

Similarity Searching

QSAR / Machine Learning

Pharmacophore Mapping

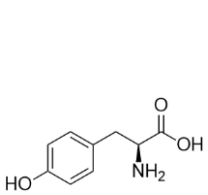
Protein-Ligand Docking

Ligand & LBDD

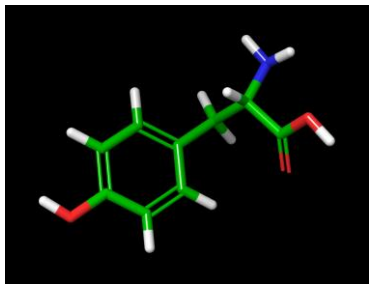
- Molecular representations
  - File formats : SMILES, SDF, PDB
  - For computers : Descriptors, fingerprints, pharmacophore
- Similarity search
- Pharmacophore search
- Library filtering
  - Physicochemical, drug-likeness (LO5, LO3, …)
  - MedChem, toxicity, warhead
  - PAINS

Molecular Representations

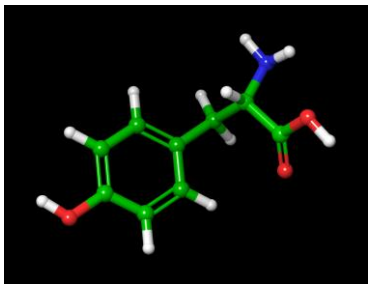
- Name : Tyrosine
- IUPAC Name : L-tyrosine
- SLN : NC[S=N]H(CC[4]=CC=C(O)C=C@5)C(O)=O
- SMILES : N[C@@H](CC1=CC=C(O)C=C1)C(O)=O
- InChI : InChI=1S/C9H11NO3/c10-8(9(12)13)5-6-1-3-7(11)4-2-6/h1-4,8,11H,5,10H2,(H,12,13)/t8-/m0/s1
- InChI key : OUYCCCASQSFEME-QMMMGPOBSA-N



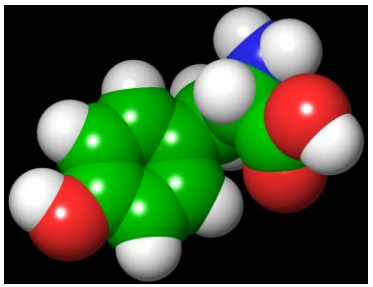
2D



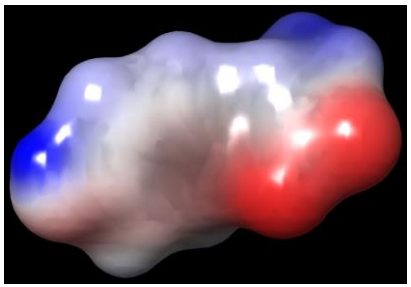
3D: Stick



3D: Ball-and-Stick

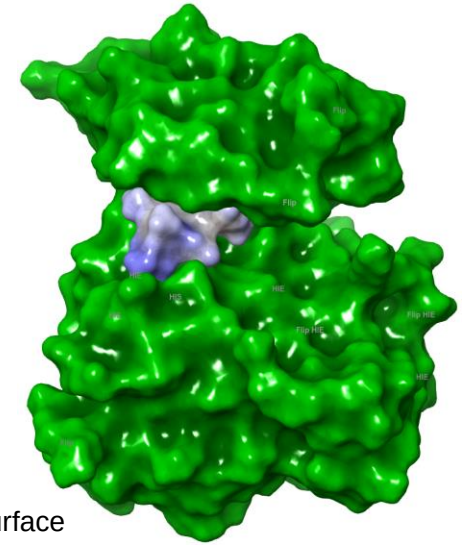


3D: CPK

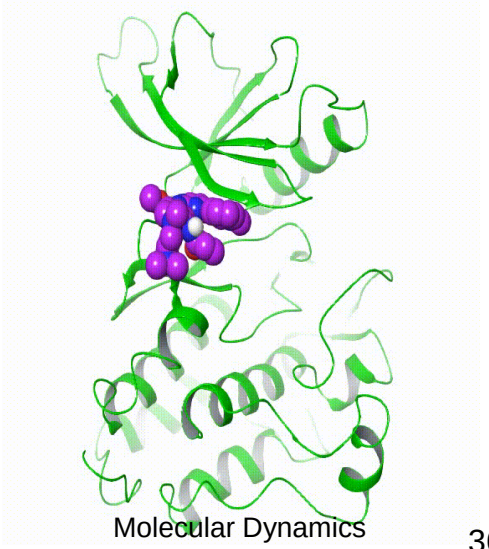


3D: Surface

Ribbon



Surface

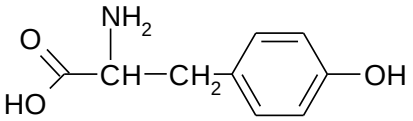


Molecular Dynamics

Linear Notations : SMILES, SMARTS, SMIRKS

- Representation of the molecule, search query, reaction (atoms, bonds, connectivity) as a linear text string
- SMILES(1980s, DayLight)
  - Atoms by normal chemical symbols (Aliphatics, aromatic)
  - Adjacent atoms imply single bonds, = for doubles, # for triples
  - Hydrogens usually implicit; Can make Hydrogens explicit
  - Parentheses imply branching; Ring closure indicated by numbers
  - Non-organic atoms are put in square brackets, e.g., [Xe]
  - Charges also in square brackets with a + or -, e.g., [Na+] or [O-]
  - Unknown atoms indicated by a *; Stereochemistry by /, \; @, @@
- SMARTS : Substructure query; *, A, a, ~; D, X;
- SMIRKS : Reaction; R>>P
- INCHI (2006, IUPAC/NIST);
- INCHIKEY (2009; NCI/PubChem/UniChem; 27-chars)

NC(Cc1ccc(O)cc1)C(=O)O

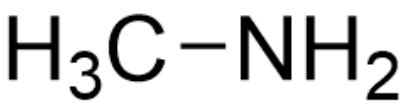


InChI=1S/C9H11NO3/c10-8(9(12)13)5-6-1-3-7(11)4-2-6/h1-4,8,11H,5,10H2,(H,12,13)/t8-/m0/s1

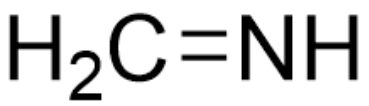
OUYCCCASQSFEME-QMMMGPOBSA-N

SMILES in Action

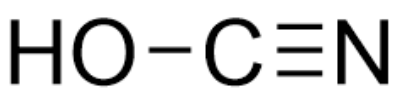
C



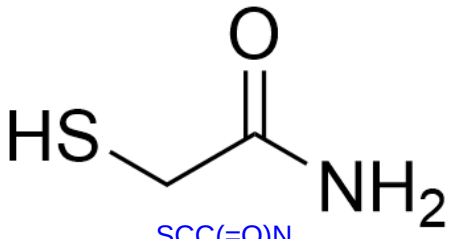
CN



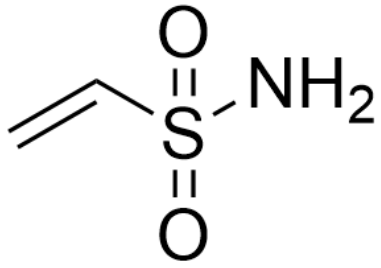
C=N



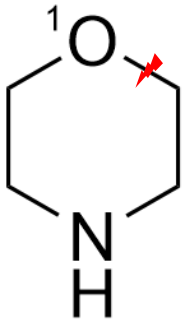
OC#C



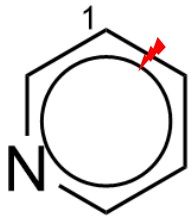
SCC(=O)N



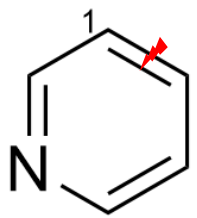
C=CS(=O)(=O)N



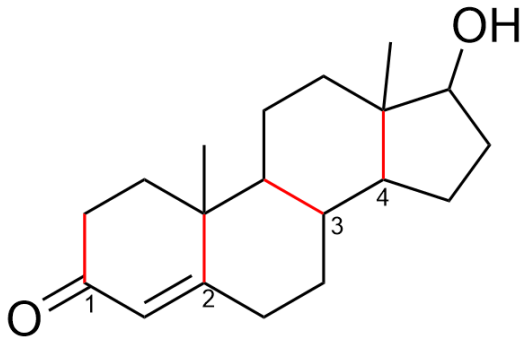
O1CCNCC1



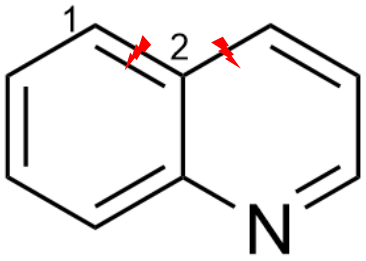
c1cnccc1



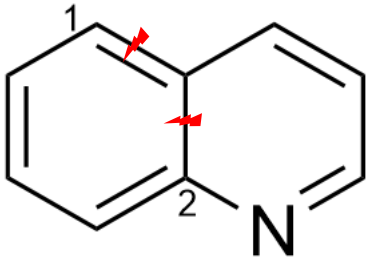
C1C=NC=CC=1



O=C1C=C2CCC3C4CCC(O)C4(C)CCC3C2(C)CC1



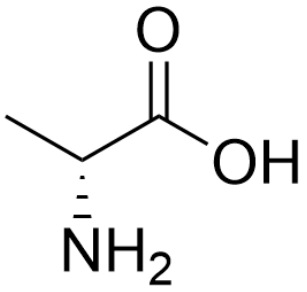
c1cccc(nccc2)c21



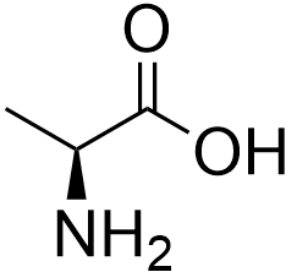
c1cccc2ncccc12

https://www.daylight.com/dayhtml_tutorials/languages/smiles/index.html

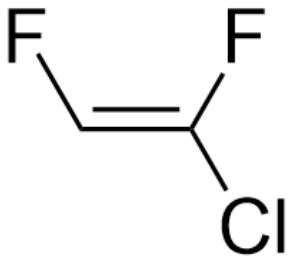
SMILES in Action (Stereochemistry)



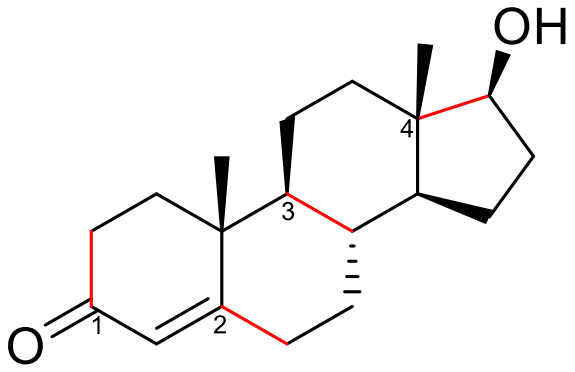
N[C@H](C)C(=O)O
C[C@H](C(=O)O)N



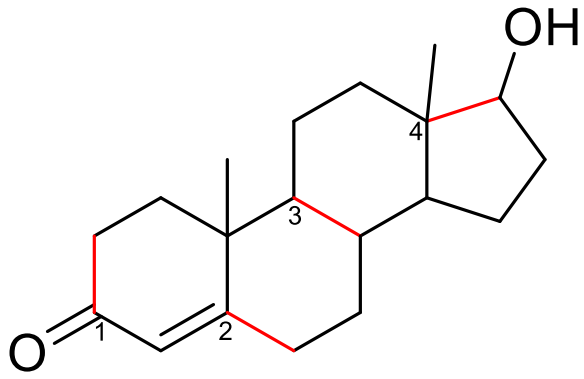
N[C@@H](C)C(=O)O
C[C@@H](C(=O)O)N



F/C=C(F)/Cl



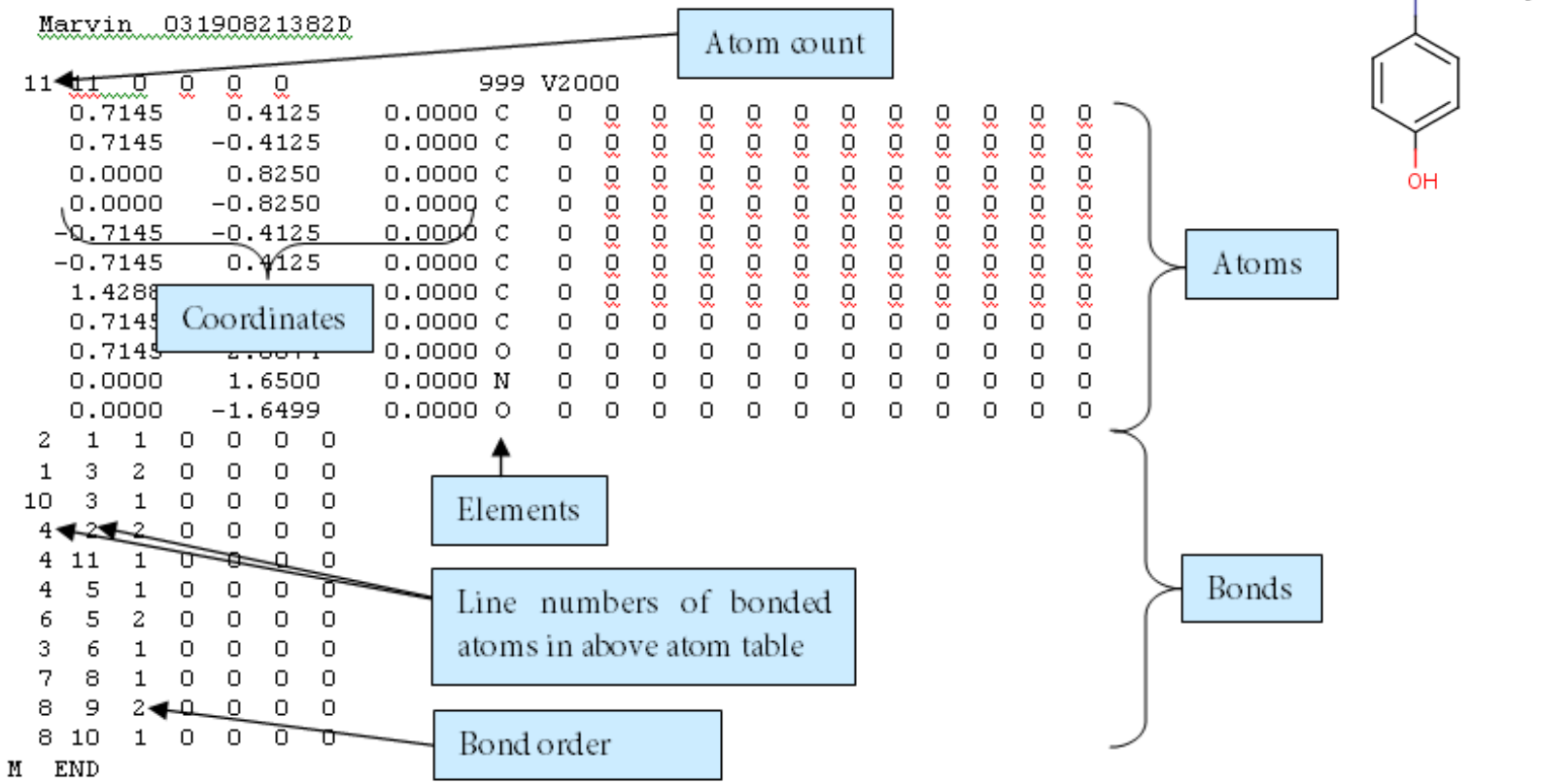
O=C1C=C2[C@](C)(CC1)[C@H]3CC[C@]4(C)[C@H](CC[C@H]4[C@@H]3CC2)O



O=C1C=C2CCC3C4CCC(O)C4(C)CCC3C2(C)CC1

File Formats : SDF

- SD, SDF file format for 3D coordinates (MDL)



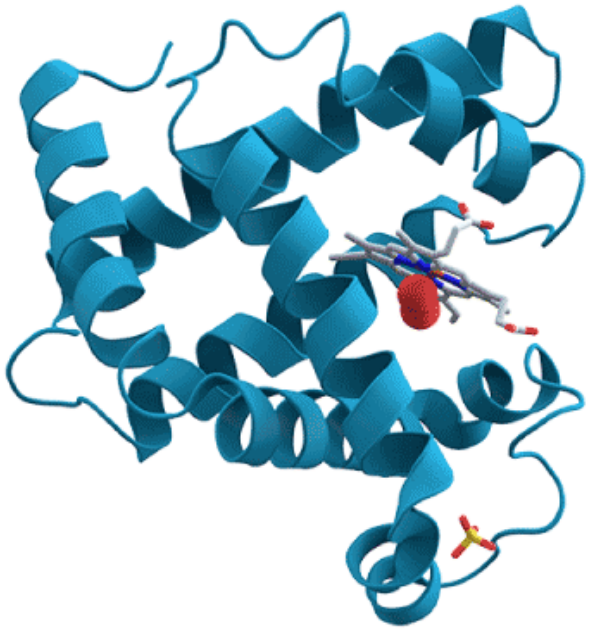

```
<activity>
7.65

$$$
```

File Formats : PDB

- Standard format of macromolecules in Protein Data Bank
- Annotation, sequence, atomic coordinate, connectivity, …
- Macromolecular mmCIF (2019)

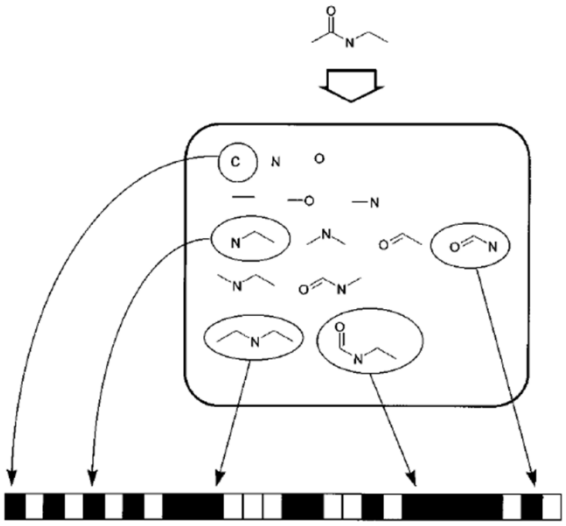
```
HEADER    SYNTHETIC PROTEIN MODEL                 02-JUL-90   1AL1      1AL1   2
COMPND    ALPHA - 1 (AMPHIPHILIC ALPHA HELIX)                             1AL1   3
SOURCE    SYNTHETIC                                                       1AL1   4
AUTHOR    C.P.HILL,D.H.ANDERSON,L.WESSON,W.F.DE*GRADO,D.EISENBERG         1AL1   5
REVDAT   2   15-JAN-95 1AL1A   1       HET                              1AL1A  1
REVDAT   1   15-OCT-91 1AL1    0                                        1AL1   6
JRNL        AUTH   C.P.HILL,D.H.ANDERSON,L.WESSON,W.F.DE*GRADO,          1AL1   7
JRNL        AUTH 2 D.EISENBERG                                          1AL1   8
JRNL        TITL   CRYSTAL STRUCTURE OF ALPHA=1=: IMPLICATIONS FOR      1AL1   9
REMARK   1                                                              1AL1  13
SEQRES   1     13  ACE GLU LEU LEU LYS LYS LEU LEU GLU GLU LEU LYS GLY  1AL1  39
HET    SO4     13       5     SULFATE ION                               1AL1A  5
FORMUL   2  SO4    O4 S1                                                1AL1  41
HELIX    1 HL1 ACE      0  LEU     10  1                                1AL1  42
CRYST1   62.350   62.350   62.350  90.00  90.00  90.00 I 41 3 2     48  1AL1  43
ORIGX1      1.000000  0.000000  0.000000        0.00000                 1AL1  44
ORIGX2      0.000000  1.000000  0.000000        0.00000                 1AL1  45
ORIGX3      0.000000  0.000000  1.000000        0.00000                 1AL1  46
SCALE1      0.016038  0.000000  0.000000        0.00000                 1AL1  47
SCALE2      0.000000  0.016038  0.000000        0.00000                 1AL1  48
SCALE3      0.000000  0.000000  0.016038        0.00000                 1AL1  49
ATOM      1  C   ACE     0      31.227  38.585  11.521  1.00 25.00      1AL1  50
ATOM      2  O   ACE     0      30.433  37.878  10.859  1.00 25.00      1AL1  51
ATOM      3  CH3 ACE     0      30.894  39.978  11.951  1.00 25.00      1AL1  52
ATOM      4  N   GLU     1      32.153  37.943  12.252  1.00 25.00      1AL1  53
ATOM      5  CA  GLU     1      32.594  36.639  11.811  1.00 25.00      1AL1  54
ATOM      6  C   GLU     1      32.002  35.428  12.514  1.00 25.00      1AL1  55
ATOM      7  O   GLU     1      32.521  34.279  12.454  1.00 25.00      1AL1  56
ATOM      8  CB  GLU     1      34.093  36.609  11.812  1.00 25.00      1AL1  57
…
ATOM    102  OXT GLY    12      20.888  27.022   1.650  1.00 25.00      1AL1 144
TER     103      GLY    12                                              1AL1 145
HETATM  104  S   SO4    13      31.477  38.950  15.821  0.50 25.00      1AL1 146
HETATM  105  O1  SO4    13      31.243  38.502  17.238  0.50 25.00      1AL1 147
HETATM  106  O2  SO4    13      30.616  40.133  15.527  0.50 25.00      1AL1 148
HETATM  107  O3  SO4    13      31.158  37.816  14.905  0.50 25.00      1AL1 149
HETATM  108  O4  SO4    13      32.916  39.343  15.640  0.50 25.00      1AL1 150
CONECT  104  105  106  107  108                                         1AL1 151
CONECT  105  104                                                        1AL1 152
CONECT  106  104                                                        1AL1 153
CONECT  107  104                                                        1AL1 154
CONECT  108  104                                                        1AL1 155
MASTER       29    0    1    1    0    0    0    6  100    1    5    1  1AL1A  6
END                                                                     1AL1 157
```

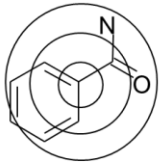
Filename extension	.pdb, .ent, .brk
Internet media type	chemical/x-pdb
Type of format	chemical file format



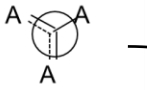
Molecular Representations : Descriptors

- Representations(number) of a molecular structure & property
  - Constitutional Descriptors : # atoms, functional groups, bonds, …
  - Topological Descriptors : Wiener, Randic, Kier shape index, …
  - Geometrical Descriptors : Surface area, Volume, SASA, …
  - Electrostatic Descriptors : partial charge, polarity, dipole moment, …
  - Quantum Chemical Descriptors : heat of formation, electronic energy, …
  - MO Related Descriptors : HOMO, LUMO, …
  - Thermodynamic Descriptors : Enthalpy, Entropy, Vibrational, …
  - DFT based Reactivity Descriptors : Chemical potential, …
- Structure : Fingerprint descriptors

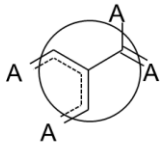




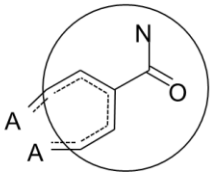
Each iteration adds bits that represent larger and larger structures



Iteration 0



Iteration 1

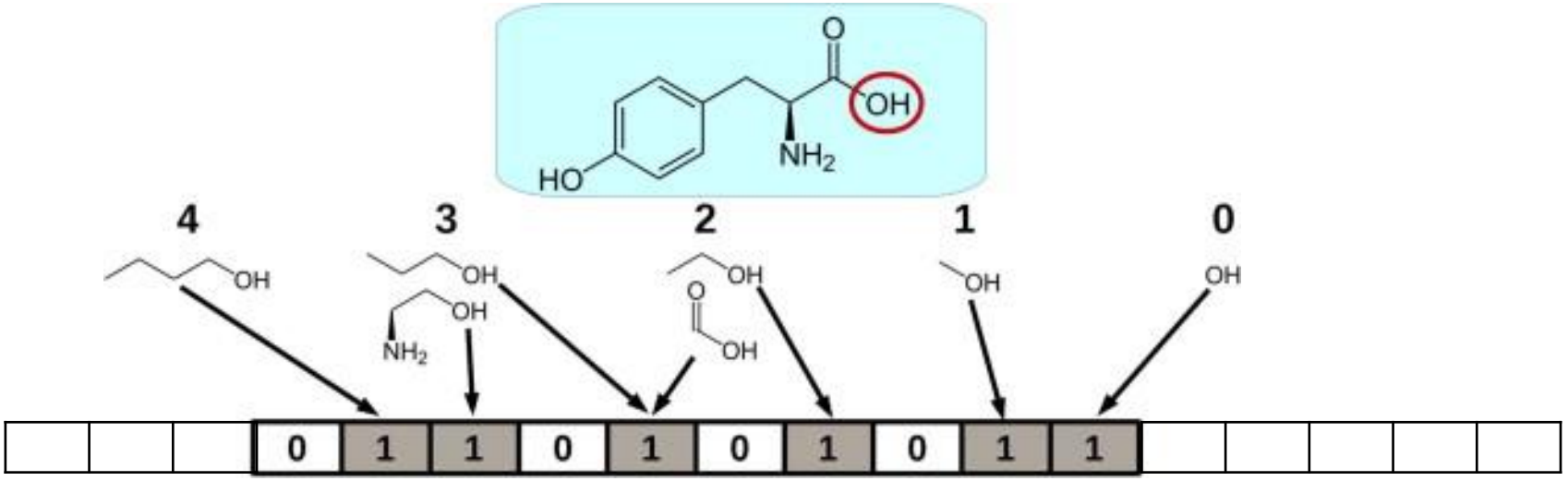


Iteration 2

	P1	P2	P3	P4	…	Pn
CPD_1	1	0	0	1	…	0
CPD_2	1	1	0	0	…	1
:	:	:	:	:	…	:
CPD_m	0	1	1	0	…	1

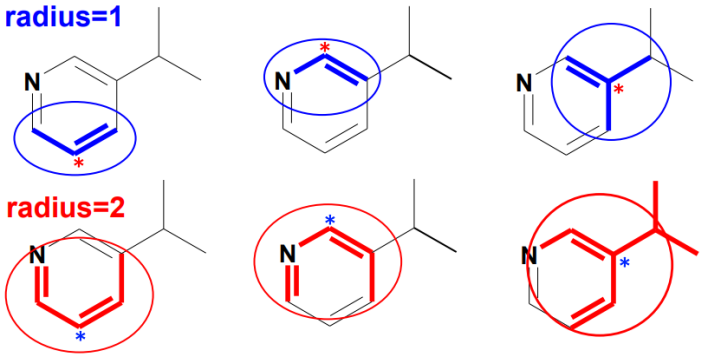
Fingerprints

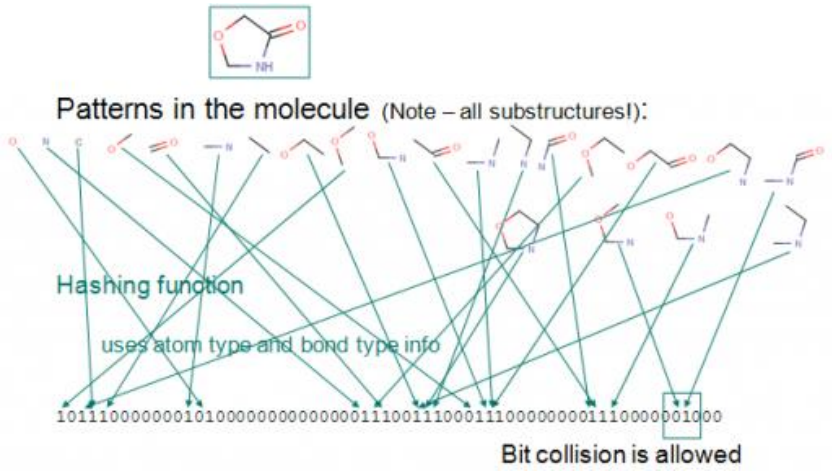
- 2D Fingerprints : molecules represented as binary vectors (~2048 bits)
  - Each bit in bit string represents one molecular fragments
  - Originally developed for speeding up substructure search
  - Similarity is based on the number of bits that are common to two structures
  - Dictionary-based (MACCS); Hashed (DayLight); Circular (Morgan)



Morgan Fingerprint

- Circular, extended-connectivity fingerprint
  - Connectivity: (Element, #heavy neighbors, #Hs, charge, isotope, inRing)
  - Chemical features: Donor, Acceptor, Aromatic, Halogen, Basic, Acidic
  - Fingerprint takes into account the neighborhood of each atom:



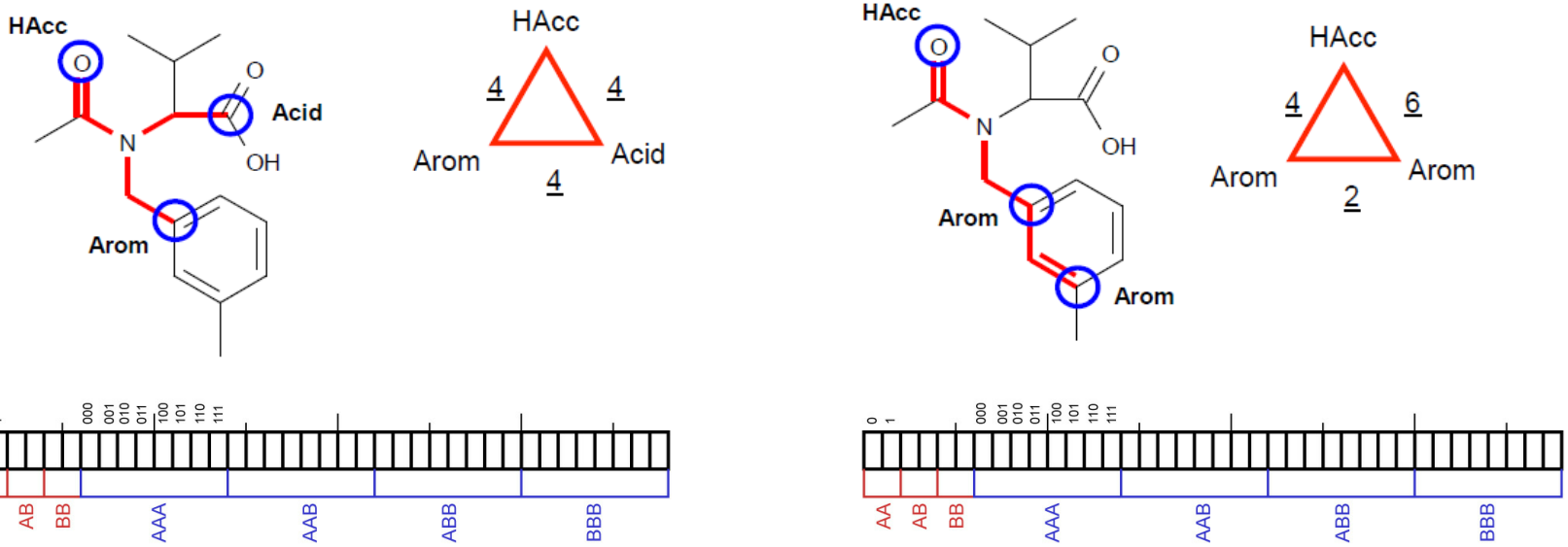


Donor	[$([N;!H0;v3,v4&+1]),\ $([O,S;H1;+0]),\ n&H1&+0]
Acceptor	[$([O,S;H1;v2;!$(*-*=[O,N,P,S])]),\ $([O,S;H0;v2]),\ $([O,S;-]),\ $([N;v3;!$(N-*=[O,N,P,S])]),\ n&H0&+0,\ $([o,s;+0;!$([o,s]:n);!$([o,s]:c:n))]
Aromatic	[a]
Halogen	[F,Cl,Br,I]
Basic	[#7;+,\ $([N;H2&+0][$([C,a]);!$([C,a](=O))]),\ $([N;H1&+0]([$([C,a]);!$([C,a](=O))])[$([C,a]);!$([C,a](=O))]),\ $([N;H0&+0]([C;!$(C(=O))])([C;!$(C(=O))])[C;!$(C(=O))])]
Acidic	[$([C,S](=[O,S,P])-[O;H1,-1])]

J. Chem. Inf. Model. 50, 742-754 (2010)

2D Pharmacophore Fingerprint

- Identify feature points in a molecule
- Calculate inter-feature topological distances
- Assign bit id to feature – distance combination
- Can be stored as counts or bits
- Feature definitions and distance bins are user-definable



Biotech and Bioeng **61**:47-54 (1998)

Molecular Representations : Similarity Measures

- Tanimoto coefficient

$$s(\mathbf{A},\mathbf{B}) = \mathrm{Tc}(\mathbf{A},\mathbf{B}) = \frac{c}{a+b-c}$$

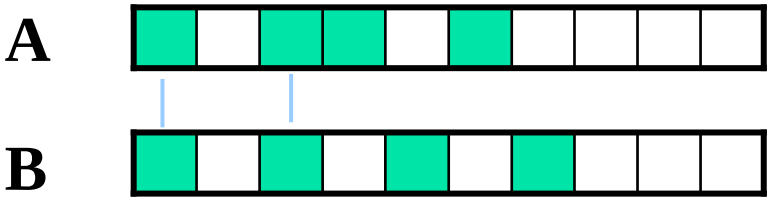
- Dice coefficient

$$s(\mathbf{A},\mathbf{B}) = \frac{2c}{a+b}$$

- Cosine coefficient

$$s(\mathbf{A},\mathbf{B}) = \frac{c}{\sqrt{ab}}$$

binary

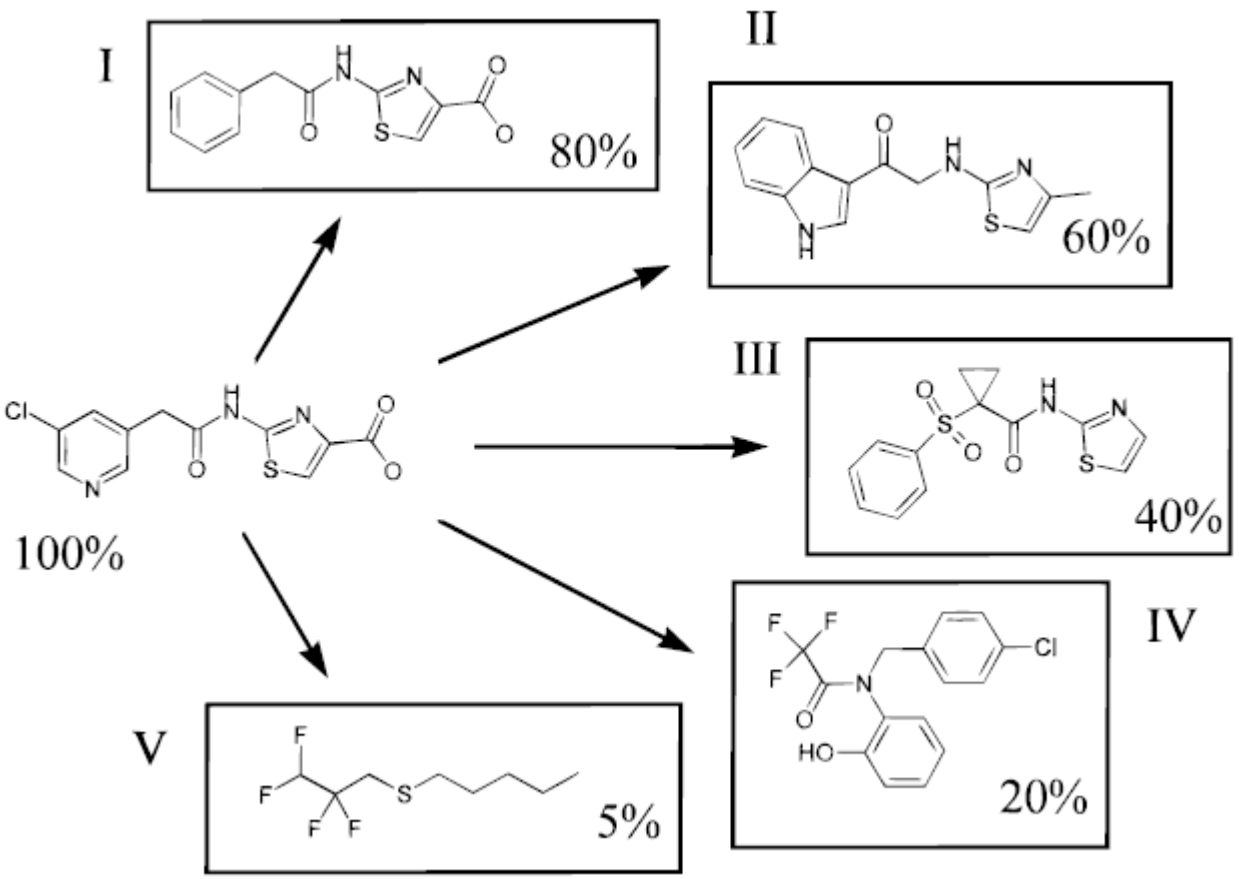


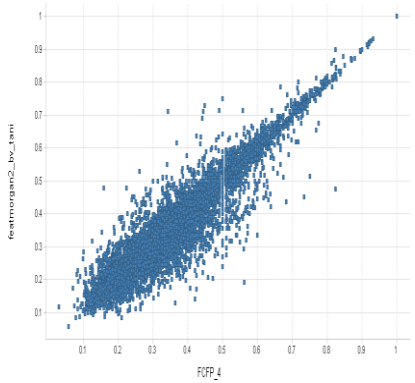
$a = 4,\ b = 4,\ c = 2$

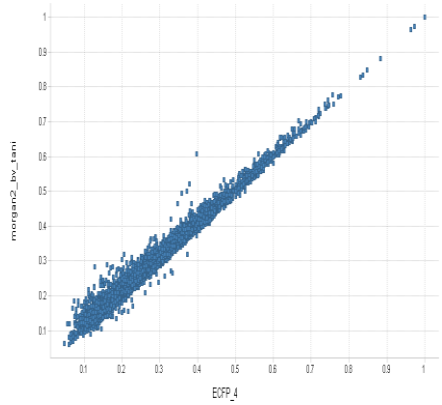
$$\mathrm{Tc}(\mathbf{A},\mathbf{B}) = \frac{2}{4+4-2} = \frac{2}{6} = \frac{1}{3}$$

Similarity Search

- Molecular similarity at a range of Tanimoto coefficients
- TC=0.6 ~ 0.8 can be a good criteria for similar molecules

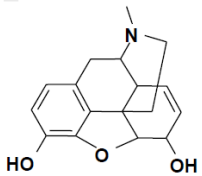




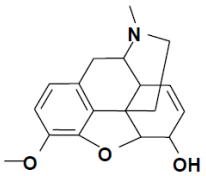


LBDD VS : Similarity-based VS

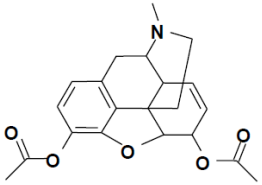
- Assumption : structurally similar molecules tend to have similar properties (neighborhood principle)
  - Basis of medicinal chemistry efforts and all LBDD methods
  - ↔ Activity cliffs



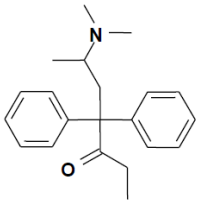
Morphine



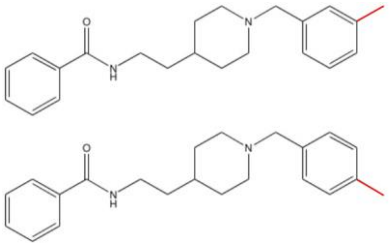
0.99 similar
Codeine



0.95 similar
Heroin



0.20 similar
Methadone



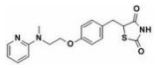
6.8

4.4

- Similarity-based VS
  - Given active reference structures, rank order database of compounds
  - Requires a way of similarity measure for ranking structures
  - No single measure of similarity : features(fingerprint, …), coefficients

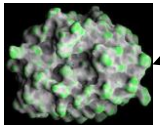
- QSAR, Machine learning
  - use a weighting function to ensure non-equal contributions from features

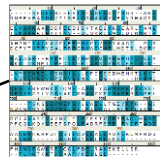
Virtual Screening



SMILES
Fingerprint
Library filtering

3D Structure of Target





Homology modeling

Binding site detection

Ligand-Based Methods

Structure-Based Methods

Similarity search

Descriptors
Fingerprint

Ligand superposition

Protein-ligand docking

Similarity Searching

QSAR / Machine Learning

Pharmacophore Mapping

Protein-Ligand Docking

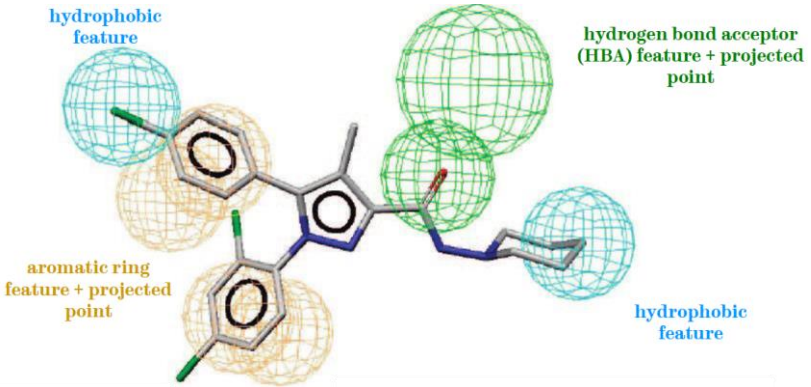
LBDD : 3D Similarity Search

- What the receptor sees?
  - 3D similarity search
  - 3D pharmacophore search

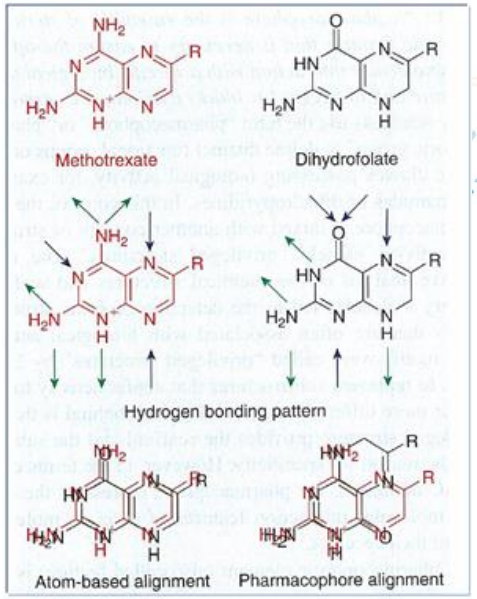
- Challenges
  - Conformation dependent : bioactive conformation?
  - Alignment dependent : fingerprint hashing algorithm
  - Alignment dependent : overlap of volume, surface, field

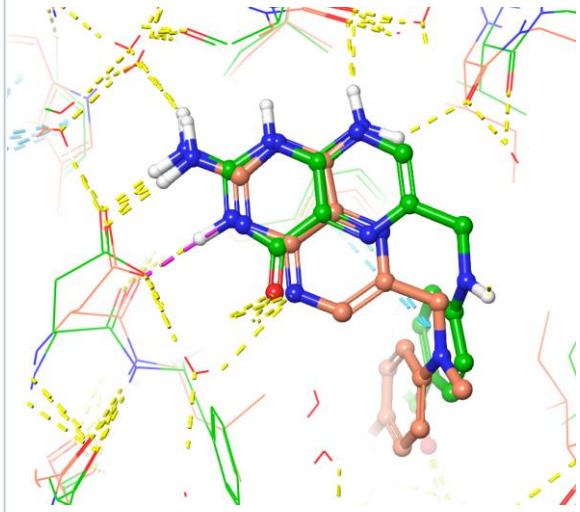
LBDD : Pharmacophore Modeling

- **Pharmacophore :** a set of features with their relative spatial orientations; To define distinct functional groups or substance classes possessing biological activity
- **Pharmacophoric features**
  - H-bond donors, acceptors
  - (+) or (-) charged groups
  - Hydrophobic regions
  - Aromatic rings

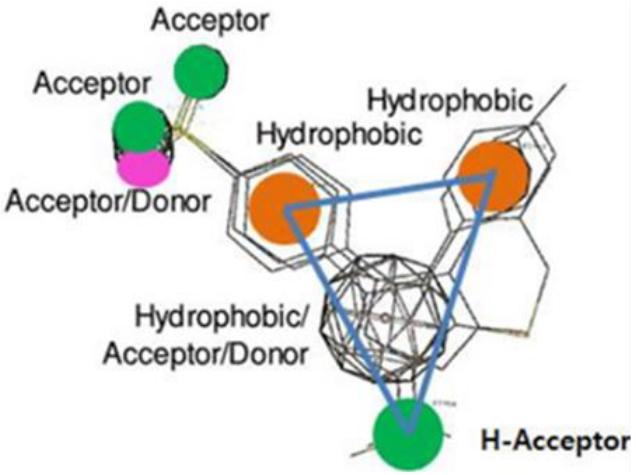


Rimonabant : CB1 receptor antagonist



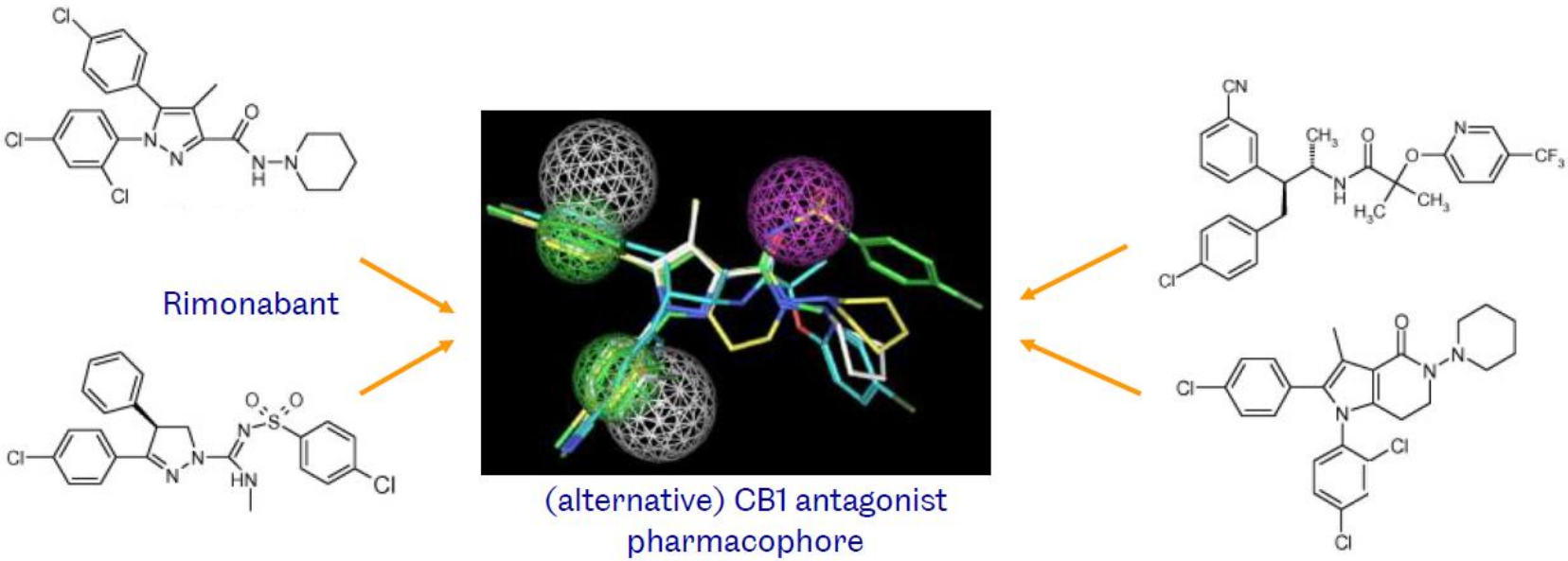


DHFR inhibitors : DHF(1DHF) vs MTX(4DFR)



Pharmacophore Generation

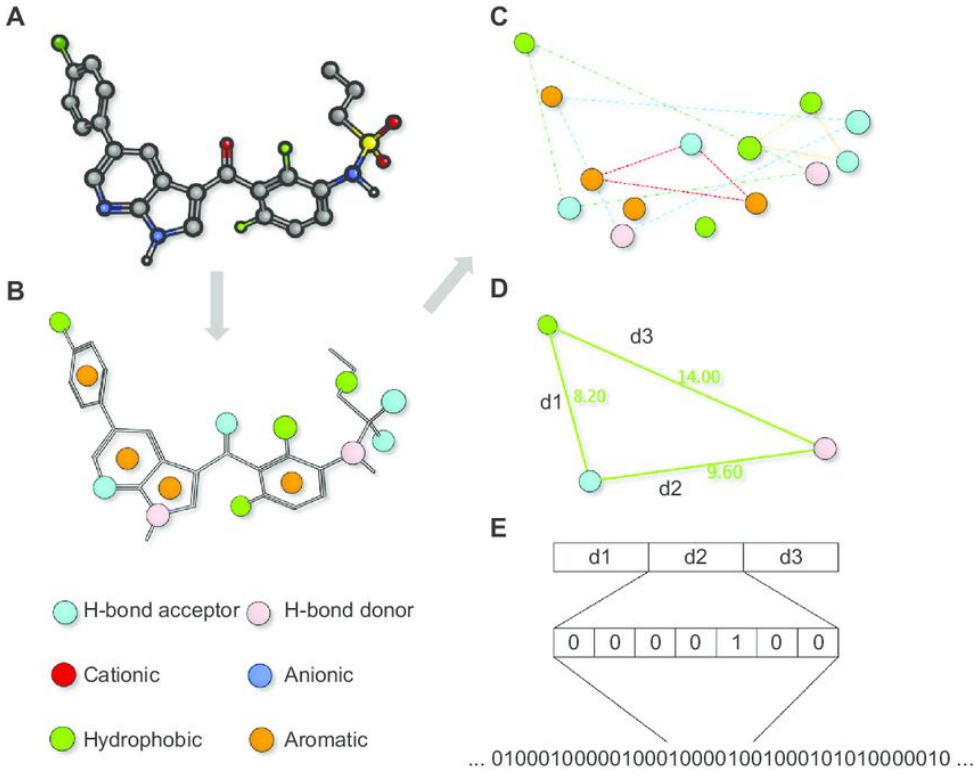
- Trying to predict how the ligands will bind to the receptor without knowing the structure of the receptor
- From the common features of known ligands (structure, activity)
- Requires proper superposition of bioactive conformation and corresponding phamacophores
- Selection of representative model from multiple hypothesis



Bioorg. Med. Chem. Lett. 2009, 19, 4183-4190

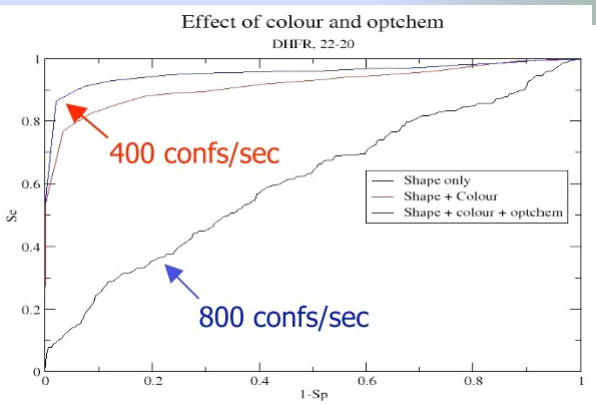
3D Pharmacophore Fingerprints

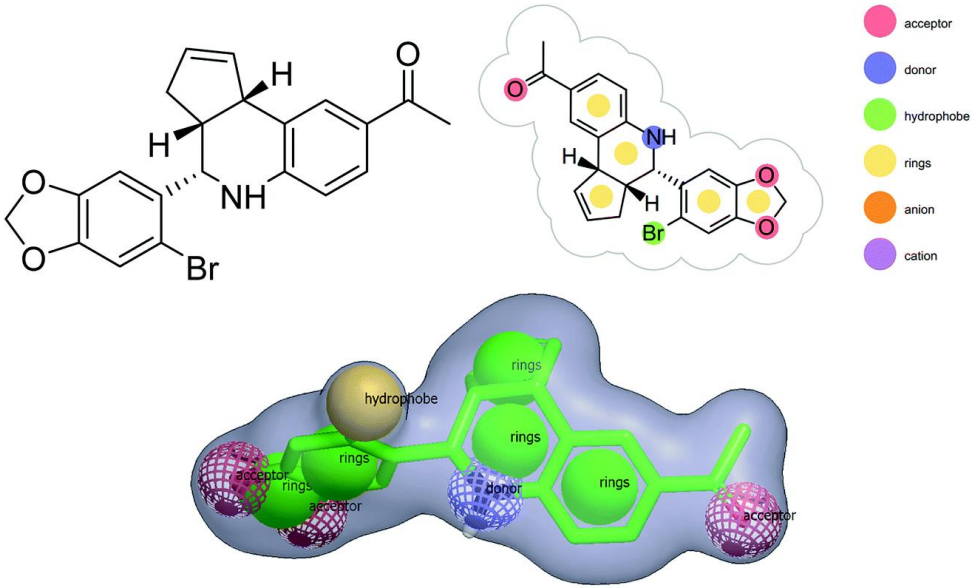
- Presence or absence of geometric features
  - Pairs of atoms at given distance range
  - Triplets of atoms and associated distance
  - Pharmacophore pairs and triplets (donors, acceptors, aromatic centres,....)
  - Valence angles, Torsion angles

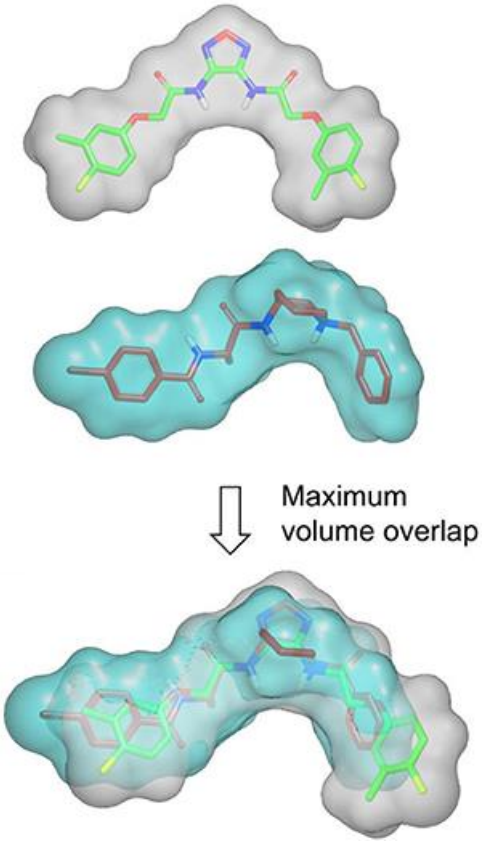


ROCS : Overlap of Field & Volume

- ROCS and Chemistry
  - 1 level : Shape only
  - 2 level : shape + color
  - 3 level : optchem







Data Fusion

- Consensus scoring method (~ docking study)

- Fusion of ranked lists generated from similarity search
  - Repeat with different reference active compounds, features (fingerprint), coefficients, …
  - Consider multiple bioactive conformation (3D similarity)
  - Sum the rank positions for a given structure to give an overall fused rank position
  - The fused rankings form the output from the search

- Consistency of search performance across a range of reference structures, types of fingerprint, biological activities etc.

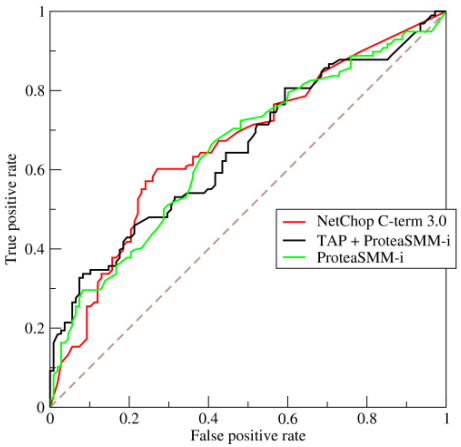
- Results are generally improved relative to using a single reference
- Best performance is achieved for diverse actives

Performance : ROC

- Enrichment factor : hit rate enhancement over random selection
  - EF = HR / HR_random
  - Top 1%, 2%, 5%

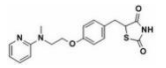
- ROC (Receive Operating Characteristic)
  - ROC Area = sensitivity / (1 – Specificity)
  - AUC[ROC] : TPR vs FPR

- Confusion matrix
  - Accuracy : *(TP+TN)/All*
  - Sensitivity, Recall, TPR : *TP/(TP+FN)*
  - Precision : *TP/(TP+FP)*
  - Specificity : *TN/(FP+TN)*
  - FPR : *FP / (FP + TN)*

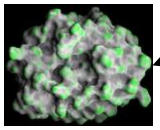


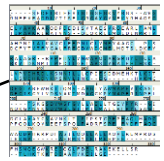
		Prediction	
		yes	no
Actual	Yes	TP	FN
	no	FP	TN

Virtual Screening



SMILES
Fingerprint
Library filtering





3D Structure of Target

Homology modeling

Binding site detection

Ligand-Based Methods

Structure-Based Methods

Similarity search

Descriptors
Fingerprint

Ligand superposition

Protein-ligand docking

Similarity Searching

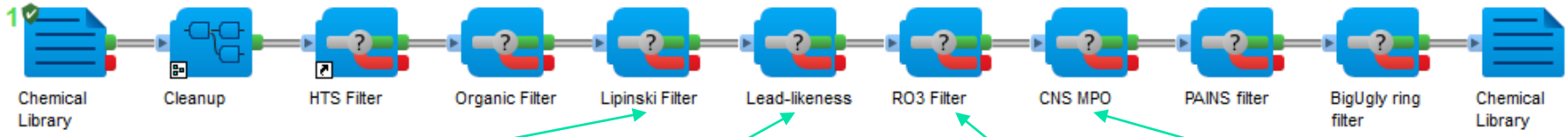
QSAR / Machine Learning

Pharmacophore Mapping

Protein-Ligand Docking

Library Filtering : Drug-likeness

- No simple rules in lead-like, drug-likeness and drugs
- Lipinski's RO5 is not for "drug likeness", but for early HTS hit analysis
  - RO3 (7.8% of drugs), lead-like (58.9%), RO5 (75.7~91.8%)
- Library filtering rules vary depending on targets
- Choosing the smaller lead may of benefit



Lipinski's RO5
- HBD <= 5
- HBA <= 10
- MW <= 500
- LogP <= 5.0
- RotBonds <= 10

Leadlikeness
- num_rings < 5
- fused arom_rings < 4
- HBD < 5
- halides < 5
- CF3 < 3

RO3 for Lead
- HBD <= 3
- HBA <= 3
- MW <= 300
- cLogP <= 3
- RotBonds <= 3
- PSA < 60

CNS-MPO
- MW < 470
- TPSA <= 90
- HBD <= 2

Library Filtering : PAINS

- PAINS (Pan Assay Interference Scaffold)

- Public tools
  - ChemAGG (http://admet.scbdd.com/ChemAGG/index)
  - Github : https://github.com/iwatobipen/rdkit_pains

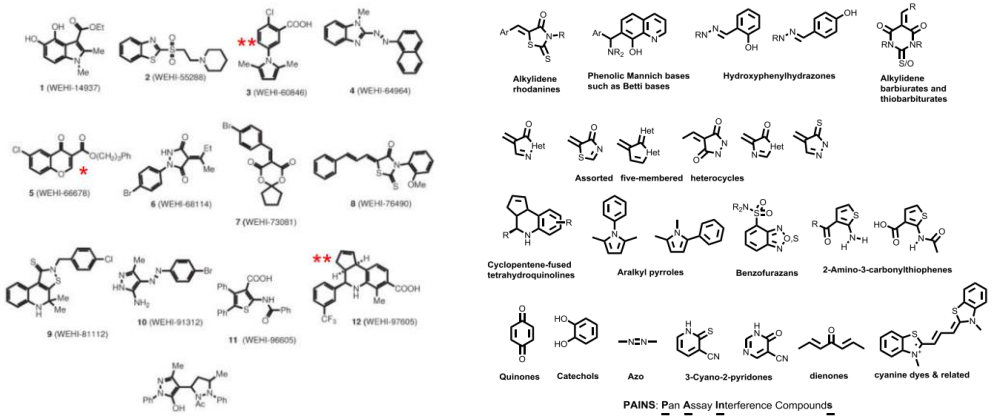


Table 4. Summary of Aggregation Rules

basic rules	log *D* is higher than 6.
	The number of aromatic carbons attached to a hydrogen atom is higher than 14.
	The number of hydroxyl groups is higher than 3.
	The number of sulfur atoms attached to more than three atoms is higher than 2.
additional rules	CATS-RR10 > 0.03

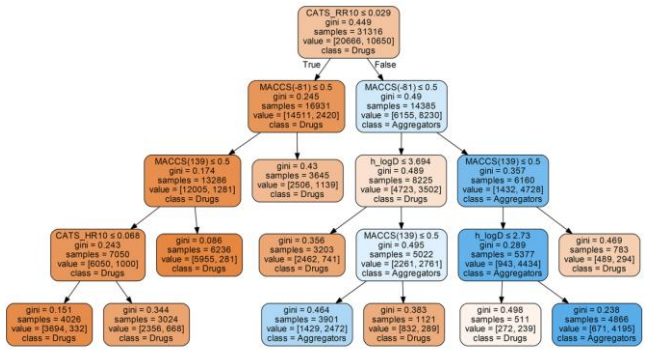
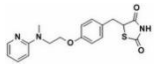


Figure 4. First four levels of the decision tree model.

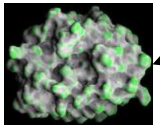
J. Chem. Inf. Model. 2019, 59, 9, 3714–3726

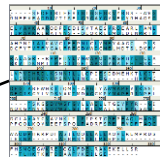
Virtual Screening



SMILES
Fingerprint
Library filtering

3D Structure of Target





Homology modeling

Binding site detection

Ligand-Based Methods

Structure-Based Methods

Similarity search

Descriptors
Fingerprint

Ligand superposition

Protein-ligand docking

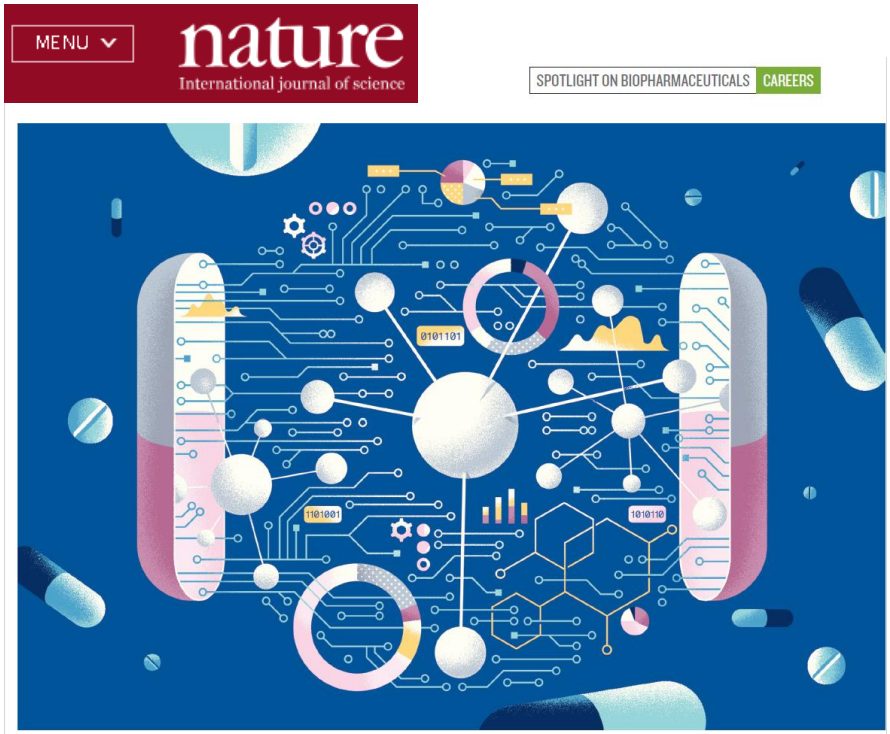
Similarity Searching

QSAR / Machine Learning

Pharmacophore Mapping

Protein-Ligand Docking

AI-based Drug Discovery (Global)



nature
International journal of science

MENU

SPOTLIGHT ON BIOPHARMACEUTICALS CAREERS

Computer-calculated compounds

Researchers are deploying artificial intelligence to discover drugs.

BY NIC FLEMING

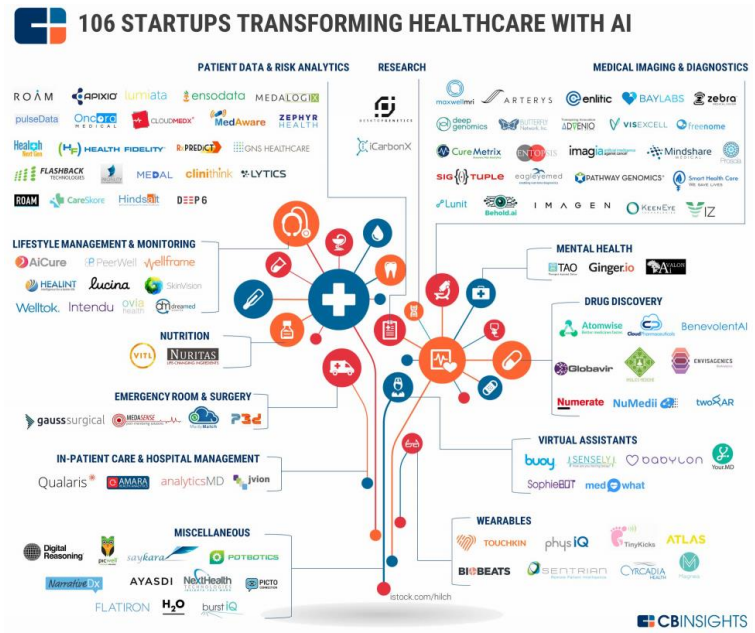
An enormous figure looms over scientists searching for new drugs: the estimated US$2.6-billion price tag of developing a treatment. A lot of that effectively goes down the drain, because it includes money spent on the nine out of ten candidate therapies that fail somewhere between phase I trials and regulatory approval. Few people in the field doubt the need to do things differently.

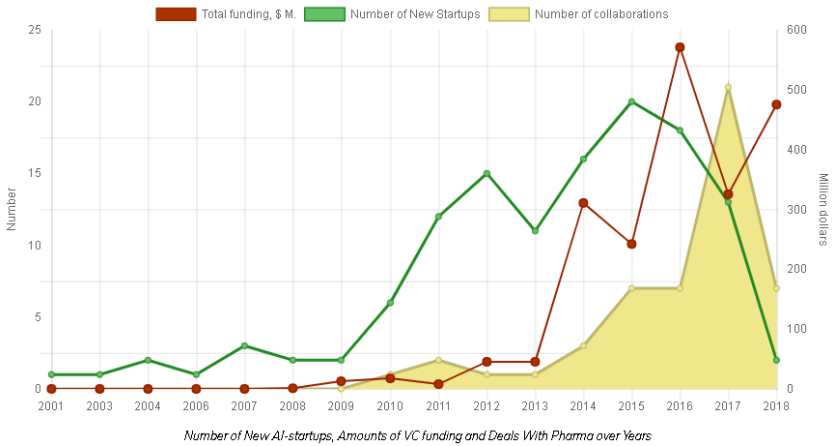
Leading biopharmaceutical companies believe a solution is at hand. Pfizer is using IBM Watson, a system that uses machine learning, to power its search for immuno-oncology drugs. Sanofi has signed a deal to use UK start-up Exscientia's artificial-intelligence (AI) platform to hunt for metabolic-disease therapies, and Roche subsidiary Genentech is using an AI system from GNS Healthcare in Cambridge, Massachusetts, to help drive the multinational company's search for cancer treatments. Most sizeable biopharma players have similar collaborations or internal programmes.

If the proponents of these techniques are right, AI and machine learning will usher in

ILLUSTRATION BY MICHELE MARCONI

"AI IS GOING TO LEAD TO THE FULL **UNDERSTANDING** OF HUMAN BIOLOGY AND GIVE US THE MEANS TO **FULLY ADDRESS** HUMAN DISEASE."





Number of New AI-startups, Amounts of VC funding and Deals With Pharma over Years

https://www.biopharmatrend.com/m/charts/

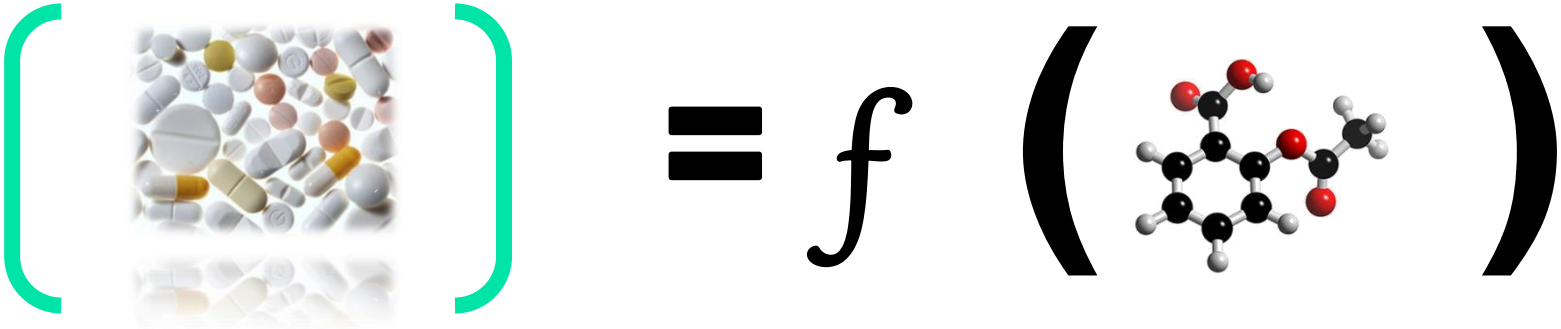
AI <Chemical> Drug Discovery Startups

Company	Research Science	Budget	
AtomWise	structure-based discovery using DL; AtomNet/CNN	Merck, Abbvie	13
Cloud Pharmaceuticals	structure-based discovery - QM/MD; computational molecular modeling based		4
BenevolentAI	drug repositioning, text mining using NLP	Janssen	74
Numerate		Takeda	15
Numedi	bio big data, pathway analysis, drug repositioning		9
twoXAR	drug repositioning; gene and protein expression analysis	Santen	15
In Silico Medicine	pathway analysis; druGAN for ligand design	GSK, Servier, Merck,Takeda	43
ExScientia	automated drug design system using AI, BigData; ChEMBL/PP/Bayesign/ECFP6	Sanofi($300M), Sumitomo, GSK($42M)	22
ATOM	Accelerating Therapeutics for Opportunities; GSK, FNLCR, UCSF, LLNL, Numerate		

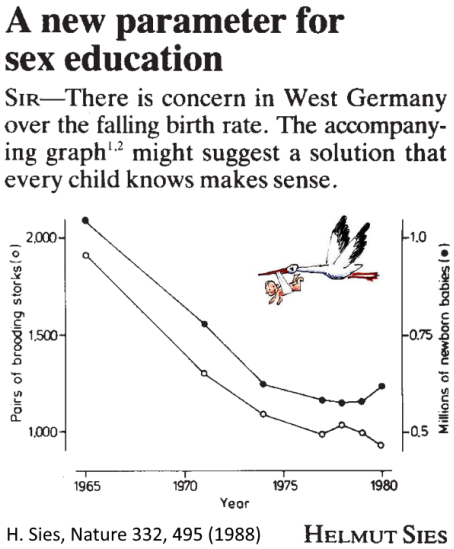
https://medium.com/the-ai-lab/artificial-intelligence-in-drug-discovery-is-overhyped-examples-from-astrazeneca-harvard-315d69a7f863
http://blogs.sciencemag.org/pipeline/archives/2018/01/10/objections-to-some-drug-discovery-ai

LBDD : QSAR

- Quantitative Structure–Activity Relationship (**구조-활성의 정량적 상관관계**)
- 화합물의 활성과 구조적 특징 사이의 상관관계를 찾아 예측모델을 만들고, 신규 화합물의 활성을 예측



인과관계 vs 상관관계

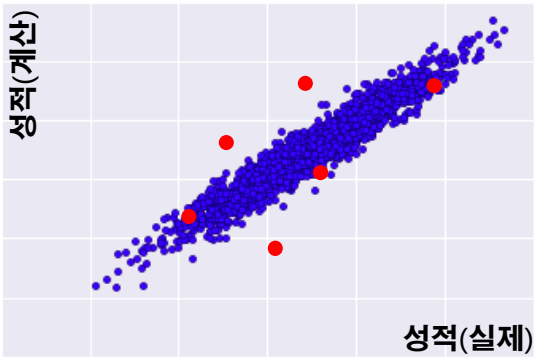


A new parameter for sex education

SIR—There is concern in West Germany over the falling birth rate. The accompanying graph[1,2] might suggest a solution that every child knows makes sense.

H. Sies, Nature 332, 495 (1988) HELMUT SIES

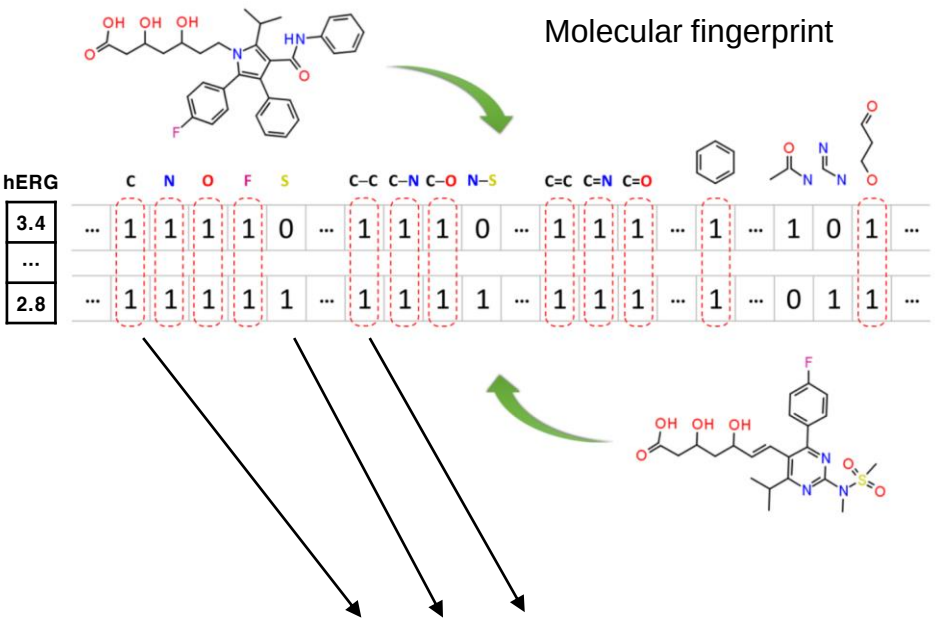
Features, Descriptors, Big Data

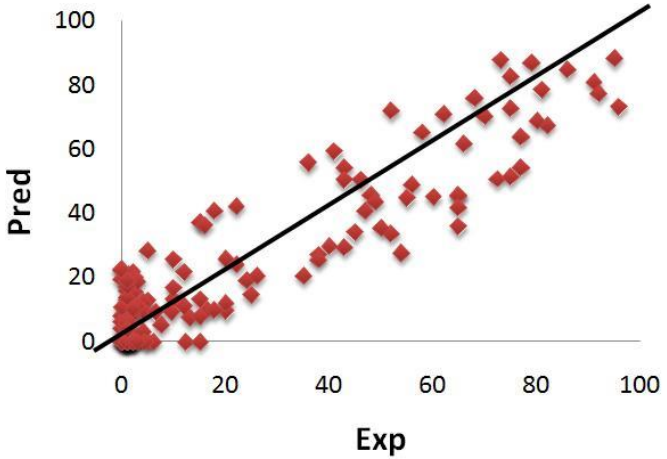


이름	성적	키	몸무게	독서	레고	게임	안경	…
	78	132	38	58	11	24	1	…
	83	124	32	79	14	15	0	…
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	…
	69	145	33	43	5	38	1	…
중요도		0.01	0.05	0.74	0.48	-0.2	0.04	…
	?	132	38	58	11		1	…

성적
74
86
⋮
63

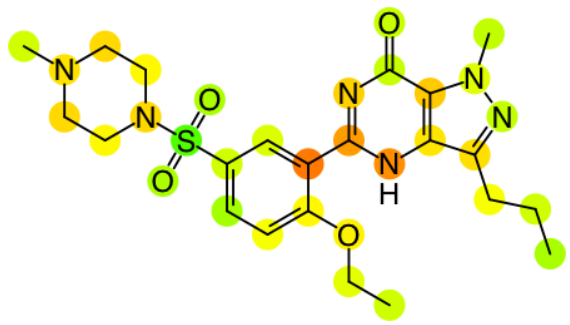
Molecular Descriptor, Fingerprint, Similarity, QSAR, AI





	Bio	P_1	P_2	...	LogP	MW	...	P_n	Y
Cpd 1	0.7	1	0		3.5	250.4			0.74
Cpd 2	3.2	1	1		4.1	340.8			4.18
…									
Cpd m	6.1								5.7

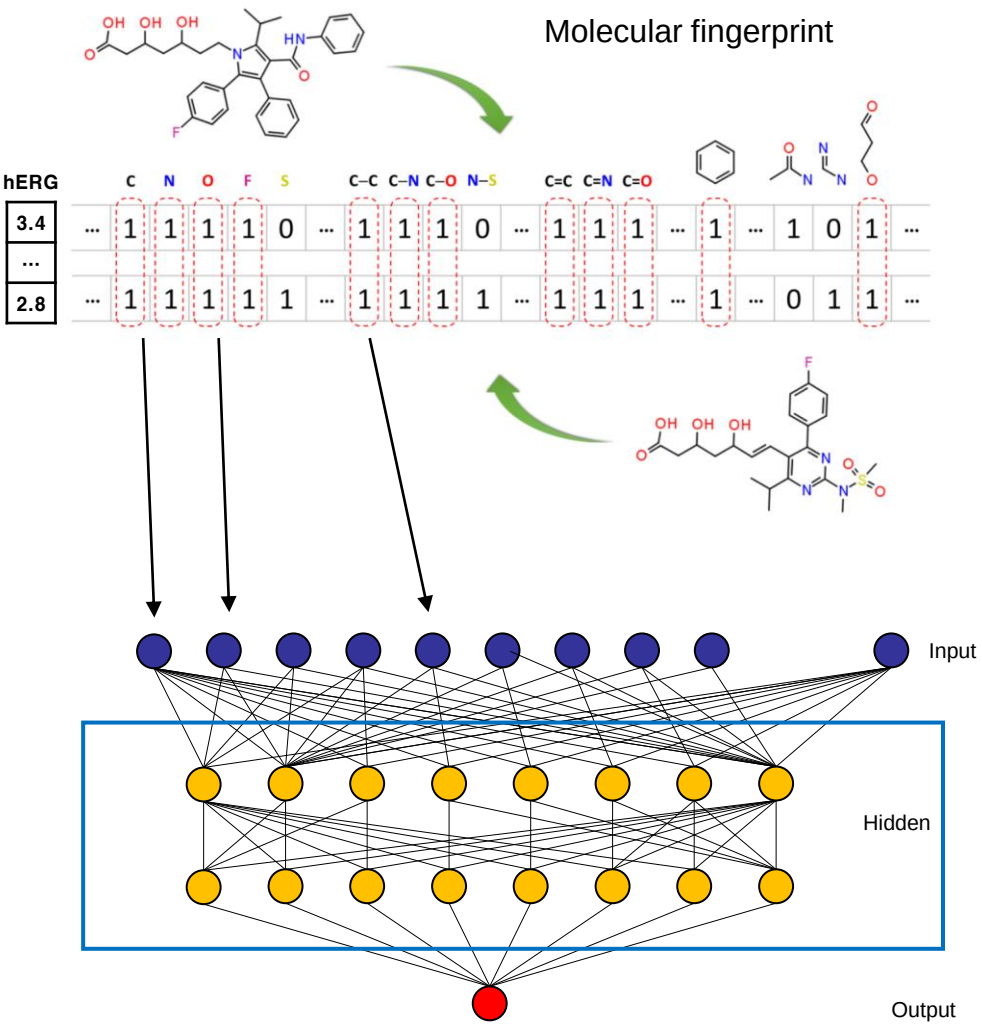
0.123	-0.312	...	0.071	0.003	...	

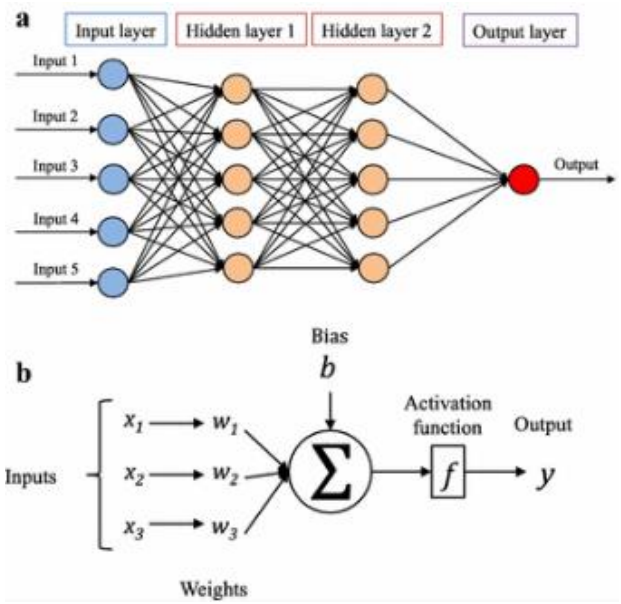


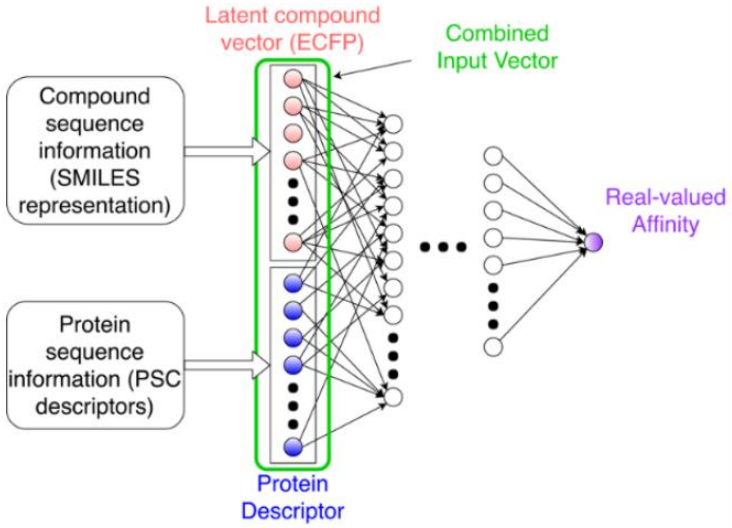
Biological activity = C_0+(C_1*P_1)+…+(C_n*P_n)

PLS (1988)
SVM (2000)
Bayesian (2002)

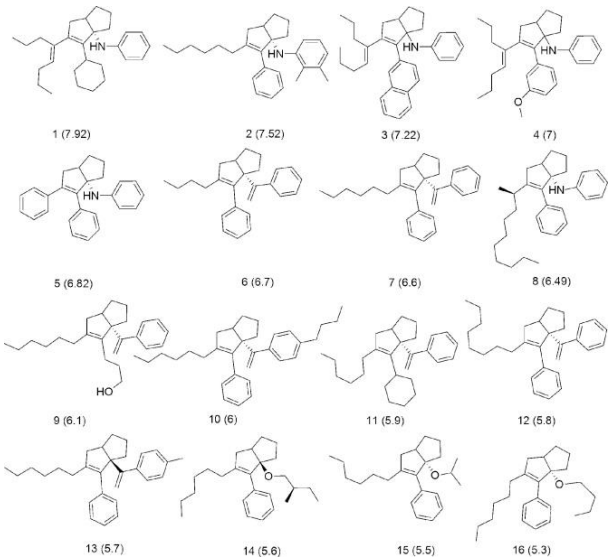
DNN/AI-based Drug Discovery

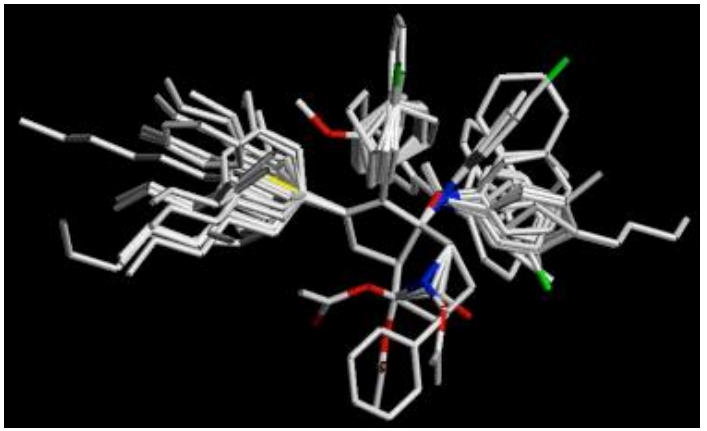




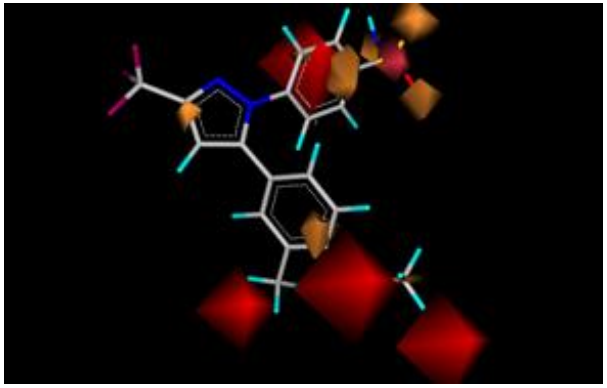


3D-QSAR : CoMFA

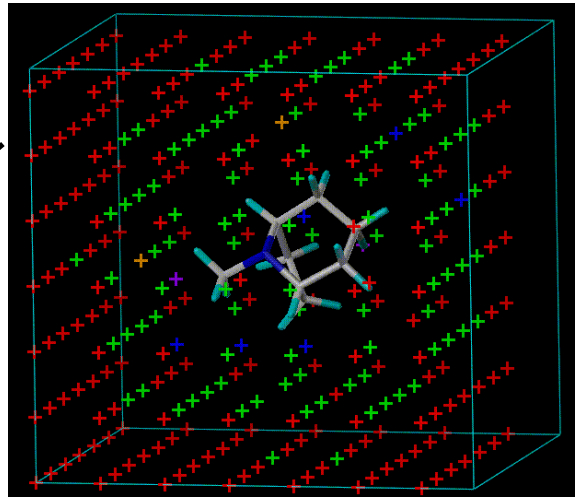




Liver receptor homolog-1 (LRH-1) (nuclear receptors) agonists



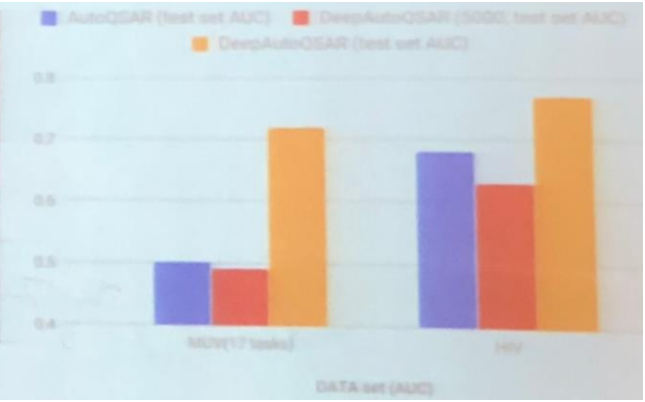
	Bio	P_1	P_2	...	P_n
Cpd 1	0.7	1	1		
Cpd 2	3.2	1	0		
…					
Cpd m	6.1				



Community: EuroQSAR-2018

- Performance enhancement of DL is marginal

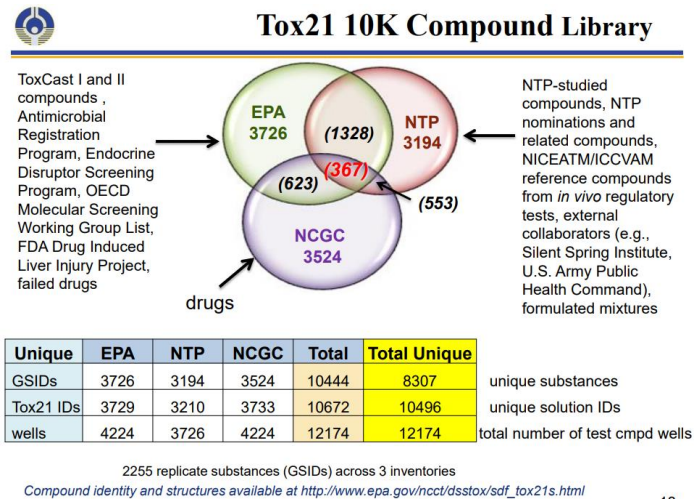
Data set	AutoQSAR (5000 training AUC)	AutoQSAR/DC (5000 training AUC)	AutoQSAR/DC (90% training AUC)	Data size (number of tasks)
MUV	0.50	0.49	0.72	~14700(17)
HIV	0.68	0.63	0.77	40426



- MD / Free Energy Perturbation

Tox21 Data Challenge : DeepTox2018

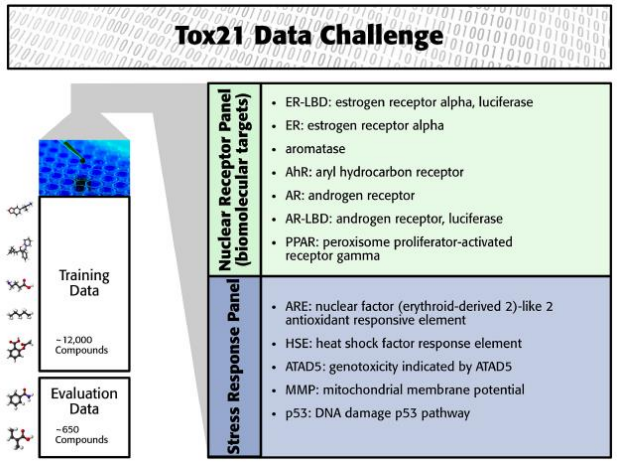
- Data set :
  - 12,707 environmental chemicals and drugs with 12 different toxic effects; 11,764 (training)/296 (leaderboard)/647 compounds (test)
  - 801 dense features : MW, solubility, surface are, …
  - 272,776 sparse features : fingerprints : ECFP10, DFS6, DFS8, …



Unique	EPA	NTP	NCGC	Total	Total Unique	
GSIDs	3726	3194	3524	10444	8307	unique substances
Tox21 IDs	3729	3210	3733	10672	10496	unique solution IDs
wells	4224	3726	4224	12174	12174	total number of test cmpd wells

2255 replicate substances (GSIDs) across 3 inventories

Compound identity and structures available at http://www.epa.gov/ncct/dsstox/sdf_tox21s.html



	AVG	NR	SR	AhR	AR	AR-LBD	ARE	Aromatase	ATAD5	ER	ER-LBD	HSE	MMP	p53	PPAR.g
DNN	0.837	0.827	0.851	0.923	0.778	0.825	0.829	0.804	0.775	0.791	0.811	0.863	0.930	0.860	0.856
SVM	0.832	0.819	0.849	0.919	0.822	0.748	0.818	0.819	0.781	0.799	0.798	0.848	0.946	0.854	0.827
RF	0.820	0.805	0.840	0.917	0.776	0.812	0.810	0.806	0.786	0.770	0.746	0.826	0.945	0.835	0.805
ElNet	0.803	0.787	0.826	0.897	0.788	0.692	0.778	0.763	0.768	0.765	0.805	0.844	0.924	0.818	0.799

Target Prediction by ML

Cite this: *Chem. Sci.*, 2018, **9**, 5441

Large-scale comparison of machine learning methods for drug target prediction on ChEMBL†

Andreas Mayr, ‡[a] Günter Klambauer, ‡[a] Thomas Unterthiner, ‡[a] Marvin Steijaert,[b] Jörg K. Wegner,[c] Hugo Ceulemans,[c] Djork-Arné Clevert[d] and Sepp Hochreiter[a]

Deep learning is currently the most successful machine learning technique in a wide range of application areas and has recently been applied successfully in drug discovery research to predict potential drug targets and to screen for active molecules. However, due to (1) the lack of large-scale studies, (2) the compound series bias that is characteristic of drug discovery datasets and (3) the hyperparameter selection bias that comes with the high number of potential deep learning architectures, it remains unclear whether deep learning can indeed outperform existing computational methods in drug discovery tasks. We therefore assessed the performance of several deep learning methods on a large-scale drug discovery dataset and compared the results with those of other machine learning and target prediction methods. To avoid potential biases from hyperparameter selection or compound series, we used a nested cluster-cross-validation strategy. We found (1) that deep learning methods significantly outperform all competing methods and (2) that the predictive performance of deep learning is in many cases comparable to that of tests performed in wet labs (*i.e.*, *in vitro* assays).

Received 10th January 2018
Accepted 16th May 2018

DOI: 10.1039/c8sc00148k

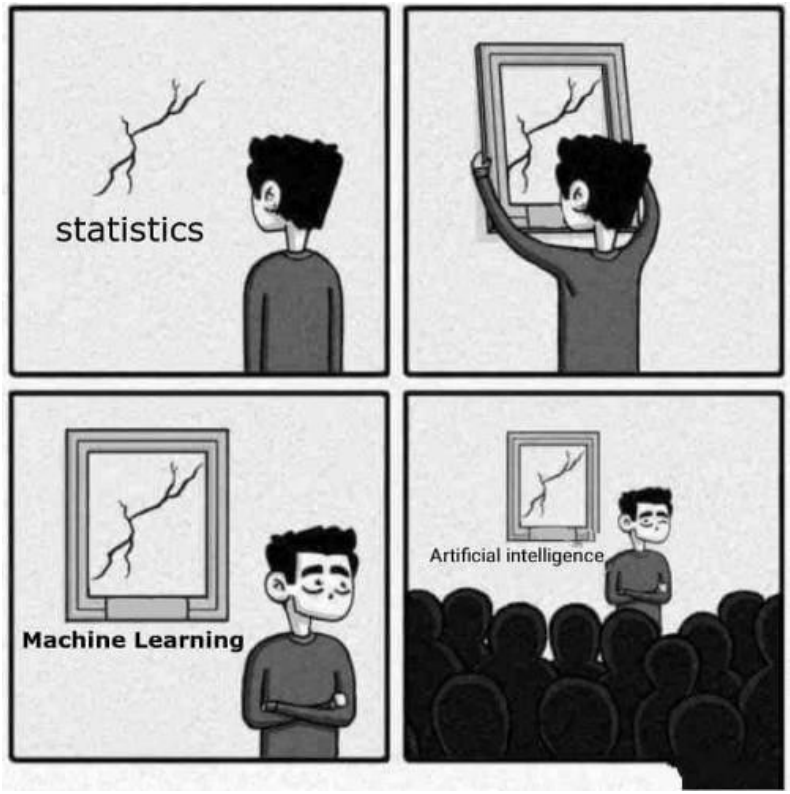
rsc.li/chemical-science

Table 1 Performance comparison of target prediction methods. The table gives the means and standard deviations of assay-AUC values for the compared algorithms and feature categories or input types. Overall, FNNs (second column) performed best. They significantly ($\alpha = 0.01$) outperformed all other considered methods. The methods GC and Weave work directly on graph representations of compounds and SmilesLSTM uses the SMILES representations of compounds

	FNN	SVM	RF	KNN	NB	SEA	GC	Weave	SmilesLSTM
StaticF	0.687 ± 0.131	0.668 ± 0.128	0.665 ± 0.125	0.624 ± 0.120					
SemiF	0.743 ± 0.124	0.704 ± 0.128	0.701 ± 0.119	0.660 ± 0.119	0.630 ± 0.109				
ECFP6	0.724 ± 0.125	0.715 ± 0.127	0.679 ± 0.128	0.669 ± 0.121	0.661 ± 0.119	0.593 ± 0.096			
DFS8	0.707 ± 0.129	0.693 ± 0.128	0.689 ± 0.120	0.648 ± 0.120	0.637 ± 0.112				
ECFP6 + ToxF	0.731 ± 0.126	0.722 ± 0.126	0.711 ± 0.131	0.675 ± 0.122	0.668 ± 0.118				
Graph							0.692 ± 0.125	0.673 ± 0.127	
SMILES									0.698 ± 0.130

Remarks on Machine Learning

- Lot's of lot's of hype
- It all depends on your ~~data~~ feature
- Good at interpolation, bad at extrapolation



Advantages and Drawbacks of VS

Advantages	Limitations
Time and cost reduction of screening process of millions of small molecules, compared to HTS	Many VS tools are applicable and successful to specific case studies (based on the training set) and not in general cases.
There is no need for physically existing compounds to perform the screening process, unlike HTS.	Compounds being identified by HTS are usually more bioactive than compounds identified by VS.
Different approaches of VS have been created for lead discovery depending each time on the availability of experimental information (SBVS Ligand-Based VS, Fragment-Based VS,etc.)	Weakness in perfect inclusion of receptor structural flexibility and of water in docking computations due to computational-cost and high complexity of its modeling
Several successful examples of identifying low nM leads that show the intended biological activity	Very potent leads (i.e. low nM) are rarely identified through VS.
A large number of docking programs and scoring functions	Scoring is still challenging in predicting accurately the correct binding pose and ranking of the compounds due to the difficulty in parameterizing the complexity of the ligand-receptor binding interactions and the approximations in calculating desolvation and entropic terms.
VS can use as input a desirable target structure complexed with a specific ligand even if there are no experimental data, through molecular modeling.	Predicted protein structures from homology modeling and predicted protein-ligand complexes may result to increased rates of false positive/negative results.

Free Software for Virtual Screening

Program	Functionality	Link
PROPKA	Determination of protonation states	http://propka.ki.ku.dk/
H++	Determination of protonation states	http://biophysics.cs.vt.edu
SPORES	Determination of protonation states	http://www.tcd.uni-konstanz.de/research/spores.php
Maestro	Protein Preparation	http://www.schrodinger.com/downloadcenter/
PDB2PQR	Protein Preparation	http://www.poissonboltzmann.org/pdb2pqr/
WebPDB	Protein Preparation	http://reccr.chem.rpi.edu/Software/WebPDB/WebPDB-index.html
JAWS	Prediction of water molecules in binding site	http://www.julienmichel.net/lab/
SiteHound-web	Binding Site Identification	http://scbx.mssm.edu/sitehound/sitehound-web/Input.html
FTMap	Binding Site Identification	http://ftmap.bu.edu/param/
Fpocket	Binding Site Identification	http://fpocket.sourceforge.net/
Mdpocket	Binding Site Identification	http://fpocket.sourceforge.net/
QsiteFinder	Binding Site Identification	http://www.modelling.leeds.ac.uk/qsitefinder/
MED-SUMO	Binding Site Identification	http://www.medit-pharma.com/index.php?page=med-sumo
MDMix	Binding Site Identification	http://sourceforge.net/projects/mdmix/
FTFlex	Binding Site Identification	http://ftflex.bu.edu/
CLEVER	Library Design	http://datam.i2r.a-star.edu.sg/clever/
e-LEA3D	Scaffold Hopping, Library Design, VS	http://chemoinfo.ipmc.cnrs.fr/
FAF-Drugs2	Compound Filtering	http://bioserv.rpbs.univ-paris-diderot.fr/FAF-Drugs/
ChemBioServer	Compound Filtering, Post-processing of docking results	http://bioserver-3.bioacademy.gr/Bioserver/ChemBioServer/
DISI	Ligand Preparation	http://wiki.uoft.bkslab.org/index.php/Preparing_the_ligand
MAPS	Ligand Preparation	http://scienomics.com/products/molecular-modeling-platform
Grinter et al.	Ligand and Protein Preparation	
Autodock	Docking, Protein and Ligand Preparation	http://autodock.scripps.edu/
Drugster	Docking, Protein and Ligand Preparation	http://www.bioacademy.gr/bioinformatics/drugster/Home.html
Dock	Docking, Protein and Ligand Preparation	http://dock.compbio.ucsf.edu/
SLIDe	Docking	http://www.bmb.msu.edu/~kuhn/software/slide/
ROSETTA_DOCK	Docking	http://graylab.jhu.edu/docking/rosetta/
CovalentDock	Covalent docking	http://docking.sce.ntu.edu.sg/
Drug-Design with PyMOL	Docking, MD, QSAR	http://people.pharmacy.purdue.edu/~mlill/software/pymol_plugins

http://bip.weizmann.ac.il/toolbox/structure/binding.htm

Open Source Modeling Tools

List of CADD software, databases and web services

분류	Package
Database	ZINC, ChEMBL, Bingo, PDBBind, Expasy, UniProt, PubChem, eMolecules, ChemSpider, SMPDB, …
Chemical Structure Representations	ChemDraw, MarvinSketch, ACD/ChemSketch, Jmol, Pymol, InChI, OpenBabel, …
Molecular Dynamics	CHARMM, GROMACS, Amber, Desmond, …
Homology Modeling	Modeller, I-TASSER, SIWSS-MODEL, …
Binding Site Prediction	MED-SuMo, TRAPP, sc-PDB, Pocketom, 3DLigandSite, …
Docking	Autodock, DOCK, GOLD, SwissDock, …
Ligand Design	
.	.
.	.
.	.

Virtual Screening : My Best Practices

- Not computational screening, but Visual Screening (Human factor !)
- Use computation filters as early as possible, to reduce human labor
- Performance of different methods varies on different target, datasets
- Combining different approaches can lead to improved results
- Increased complexity in descriptors and method does not necessarily lead to greater success