Basics of protein-ligand docking

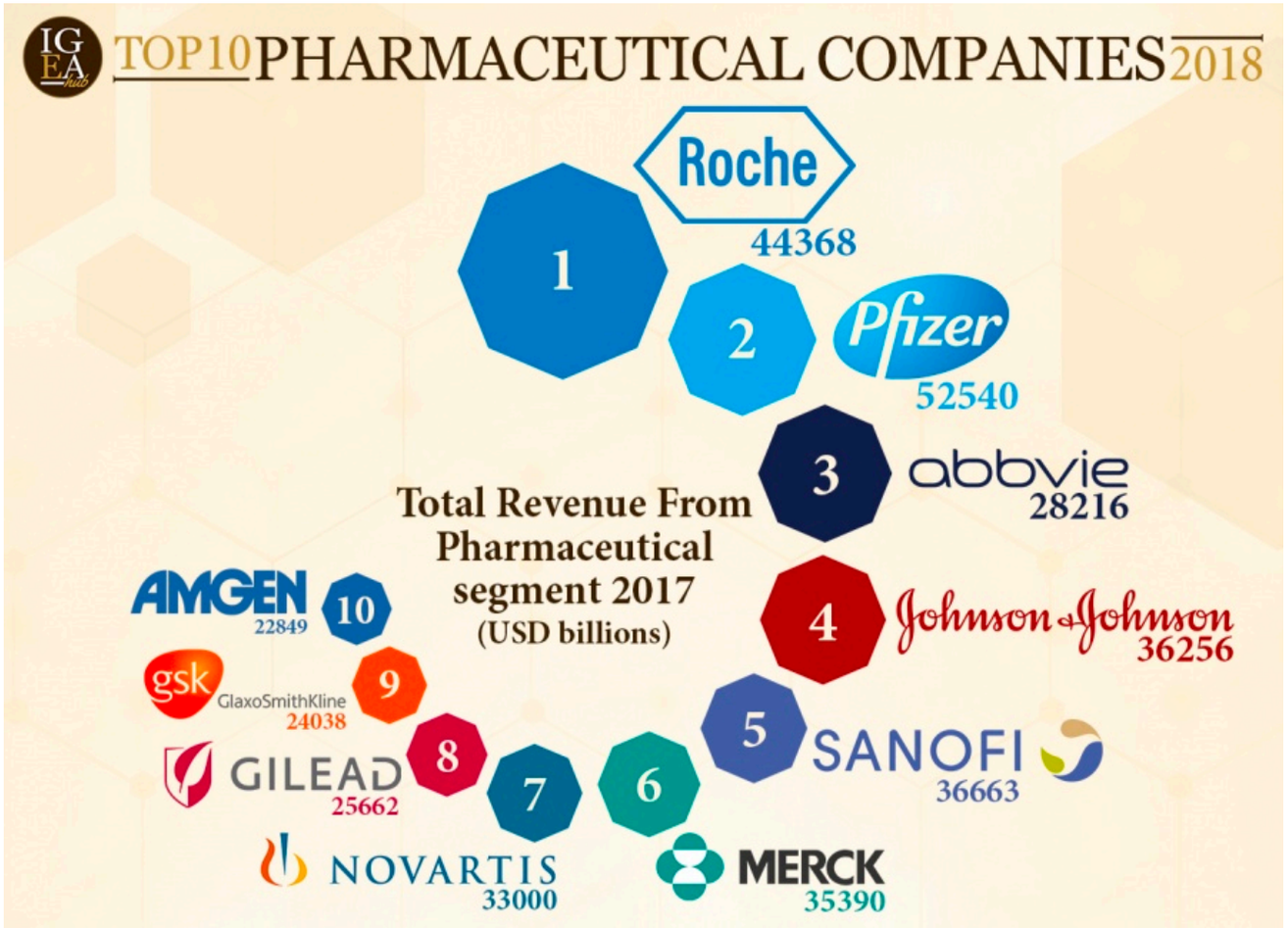
AI-assisted drug discovery

College of Pharmacy, Seoul National University
Prof. Juyong Lee

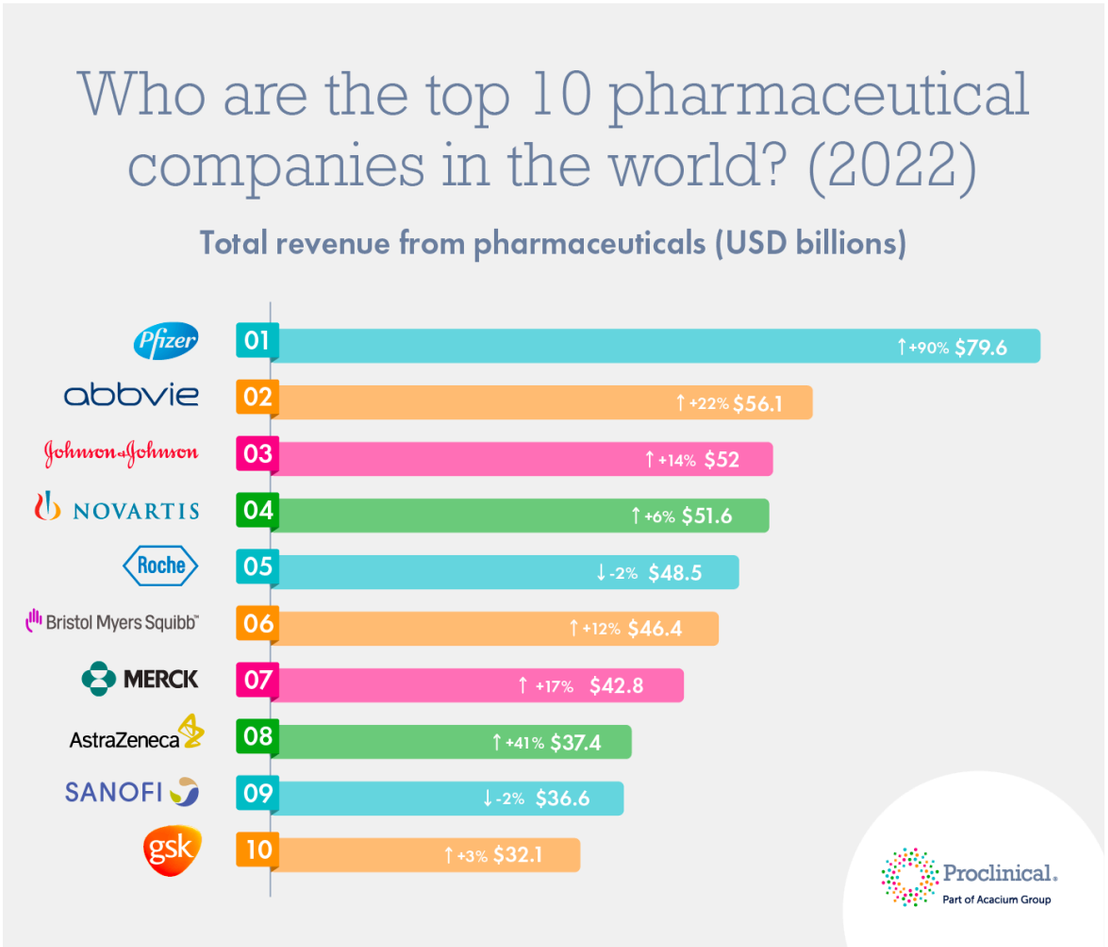
Slido: 2698620

https://app.sli.do/event/58MDz45pkXgC3SXJLVc6fW

Pharmaceutical industry can be our next economic boost



https://intelligence-pharma.com/2018/06/17/top-10-pharmaceutical-companies-2018-2/



https://www.proclinical.com/blogs/2022-6/who-are-the-top-10-pharma-companies-in-the-world-2022

- The top10 pharmaceutical companies made about 1.4 trillion dollars in 2021
- However, Korean companies occupy less than 1% in global market

Artificial Intelligence in Precision Medicine

비만 치료제 위고비 – Wegovy (semaglutide)

뉴스홈 | 최신기사

'살 빼는 주사제 열풍'…美서 수요가 공급 앞질러

송고시간 | 2023-11-03 10:57

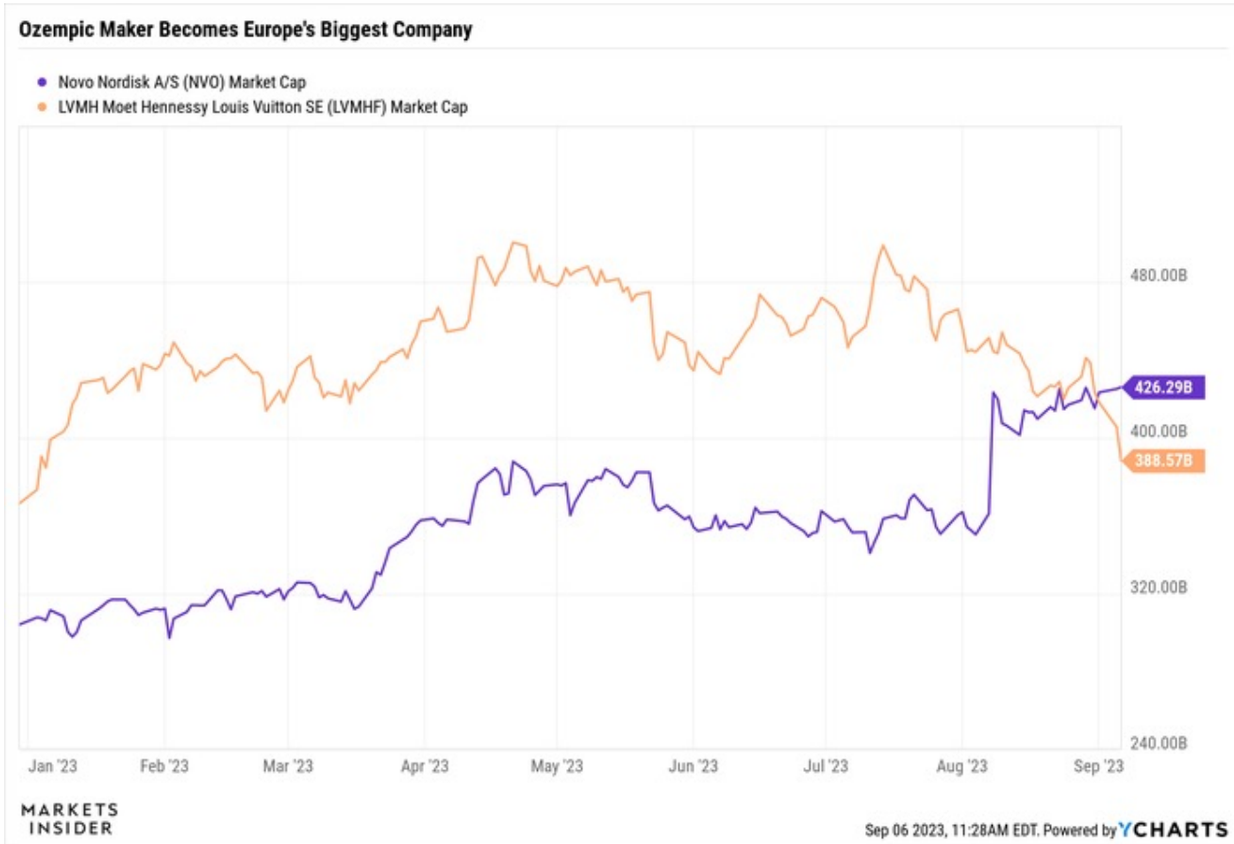
임상수 기자
기자 페이지

노보 노디스크의 비만치료제 위고비

[로이터 연합뉴스 자료사진. 재판매 및 DB 금지]

덴마크의 노보 노디스크(Novo Nordisk)에 의해서 개발

노보 노디스크는 현재 유럽에서 가장 큰 회사가 되었음



2024년 3월 기준

삼성 전자 시가 총액: ~4000억 달러

노보 노디스크 시가 총액: ~5700억 달러

덴마크 GDP: ~3983억 달러

Artificial Intelligence in Precision Medicine

국내 신약 개발 사례: 레이저티닙

2023.05.16(화) 19:00
Virtual MASTERCLASS Web Symposium

MedicalTimes

전체 정책 병·의원 제약·바이오 의료기기·AI 학술 오피니언 메타TV

제약·바이오 > 국내사

유한양행, 얀센에 1조 4천억 규모 기술 이전 계약

최선 기자 | 발행날짜: 2018-11-05 09:29:31

가

비소세포폐암 치료제 라이선스 및 공동개발 계약 체결…레이저티닙 개발 제조 독점 권리 부여

[메디칼타임즈=최선 기자] 유한양행이 얀센에 1조 4000억원 규모의 항암 치료제 기술 이전 계약을 체결했다.

5일 유한양행은 얀센 바이오텍(얀센)과 비소세포폐암 치료를 위한 임상단계 신약인 레이저티닙(Lazertinib)의 라이선스 및 공동개발 계약 체결을 발표했다.

유한양행

유한양행은 본 계약에 따라 계약금 미화 5000만 달러을 지급받고, 개발 및 상업화까지 단계별 마일스톤 기술료로 최대 미화 12억 500만 달러, 그리고 상업화에 따른 매출 규모에 따라 두 자릿수의 경상기술료를 지급받게 된다.

이에 대한 대가로 얀센은 한국을 제외한 전세계에서 레이저티닙에 대한 개발, 제조 및 상업화에 대한 독점적 권리를 가지며, 국내에서 개발 및 상업화 권리는 유한양행이 유지하게 된다.

UPDATED. 2024-03-28 13:41 (목)

MEDICAL Observer

홈 > 제약 > 제약바이오

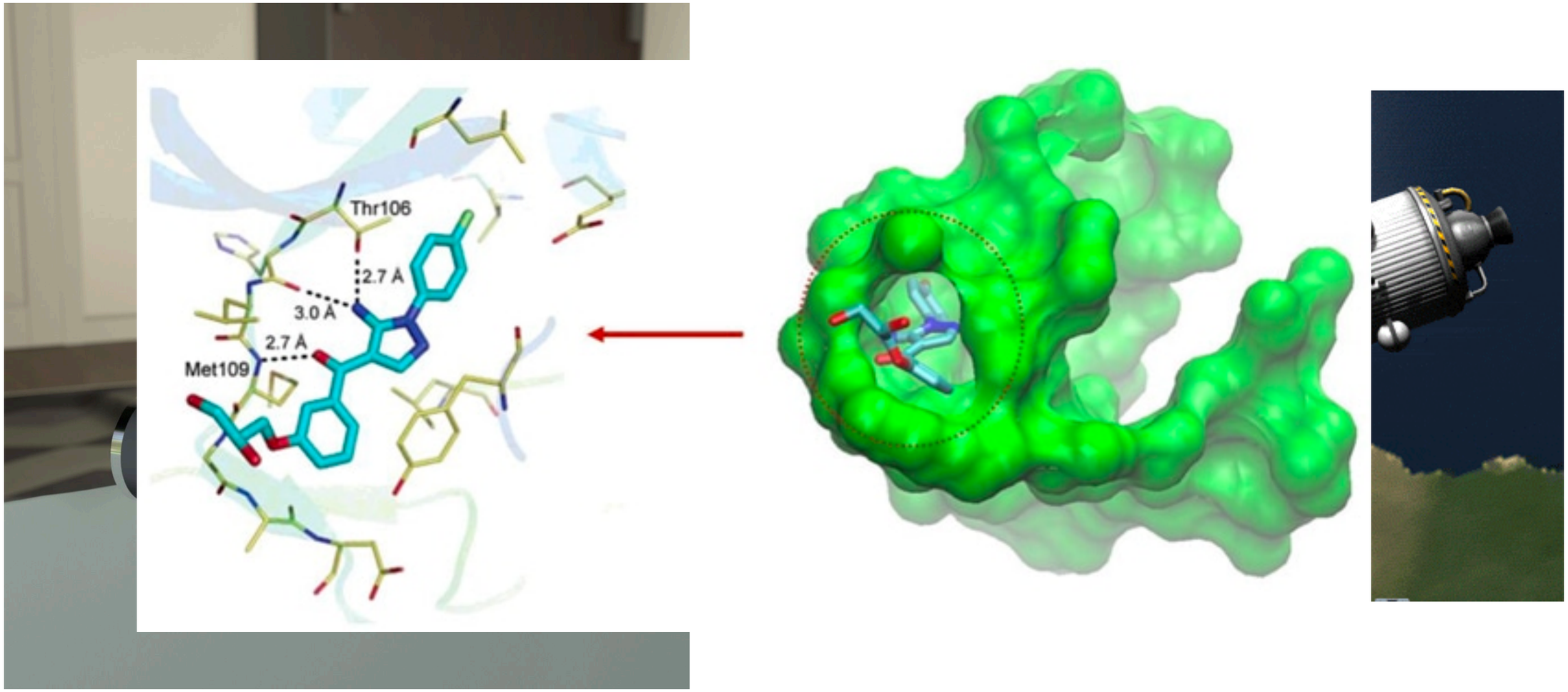
1차 치료 급여된 타그리소·렉라자 올해 어디까지 성장될까?

신형주 기자 | 입력 2024.02.19 06:18 | 수정 2024.02.19 09:46 | 댓글 0

렉라자, 올해 연 매출 1280억원…순매출 830억원 예상
타그리소, 지난해 890억원 매출 기록…타그리소 포트폴리오 강화

[메디칼업저버 신형주 기자] 지난 1월부터 비소세포폐암(NSCLC) 1차 치료제로 건강보험 적용을 받는 아스트라제네카의 타그리소(성분명 오시머티닙)와 유한양행 렉라자(레이저티닙)가 올해 어디까지 성장될지 제약업계의 이목이 쏠리고 있다.

What is docking?



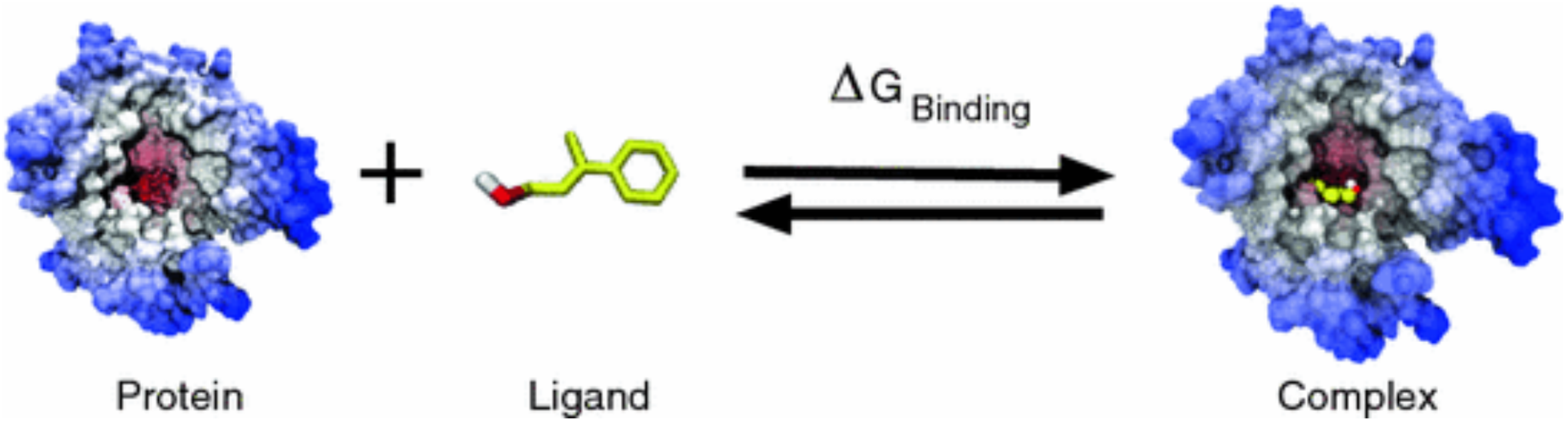
- Docking attempts to find the “**best**” matching between two molecules

Definition of protein-ligand docking

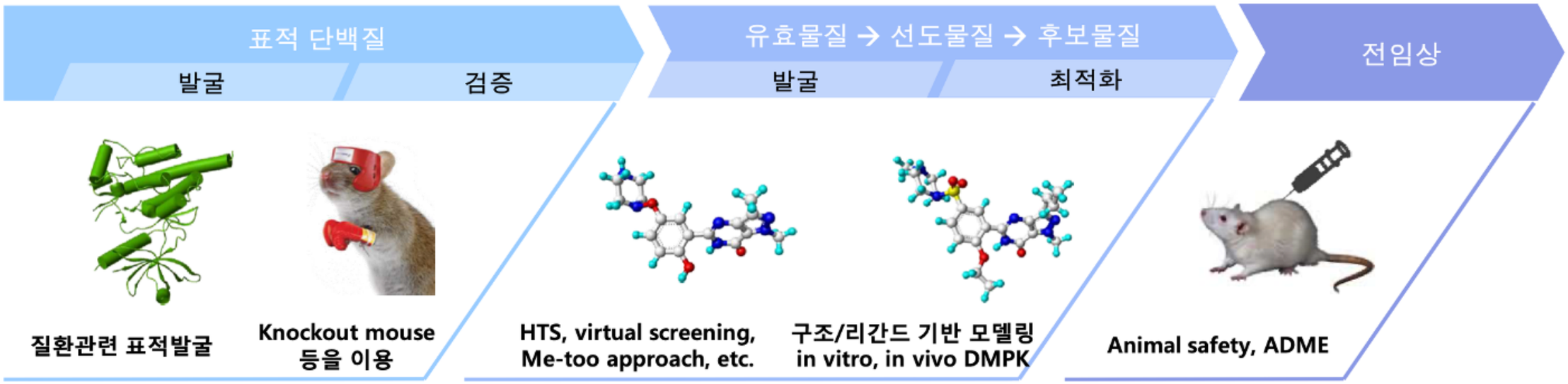
- Protein–ligand docking is a molecular modeling technique.

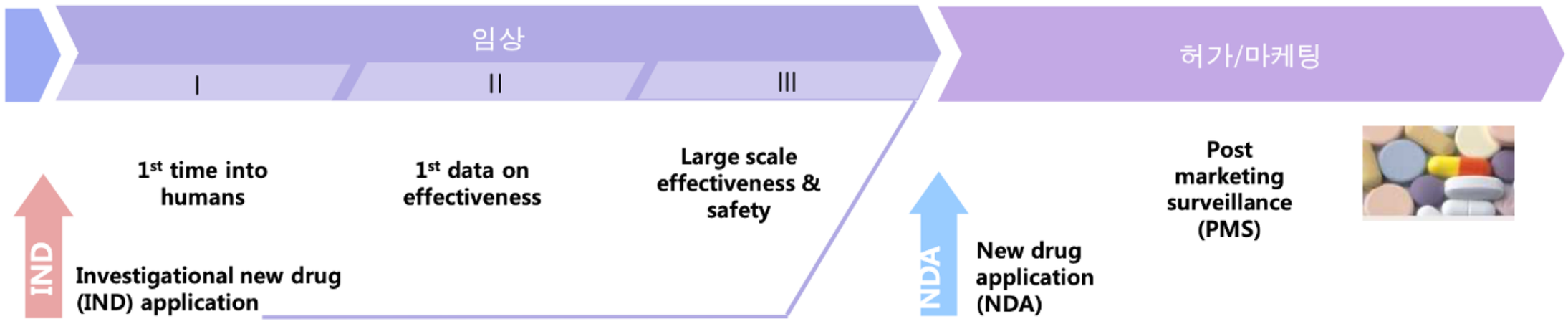
- The goal of protein–ligand docking is to **predict the position** and **orientation (pose) of a ligand** (a small molecule) when it is bound to a protein receptor or enzyme.

- Finding **the global Gibbs free energy minimum conformation**



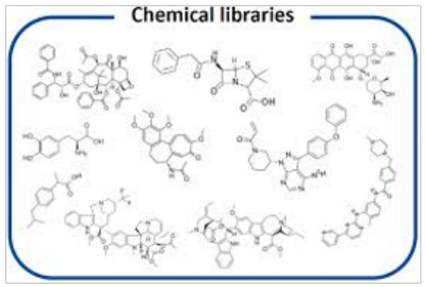
Drug-discovery process



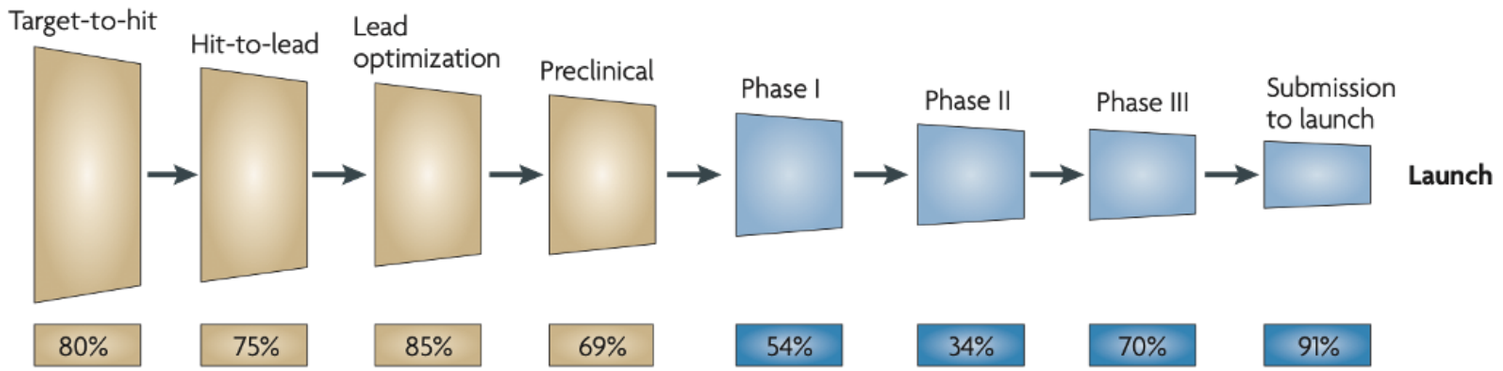


Drug discovery process is expensive

유기합성, 조합화학
(수천~수만개)



p(TS)



	80%	75%	85%	69%	54%	34%	70%	91%	
***WIP* needed for 1 launch**	24.3	19.4	14.6	12.4	8.6	4.6	1.6	1.1	1
Cost per WIP per Phase	$1	$2.5	$10	$5	$15	$40	$150	$40	
Cycle time (years)	1.0	1.5	2.0	1.0	1.5	2.5	2.5	1.5	
Cost per launch (out of pocket)	**$24**	**$49**	**$146**	**$62**	**$128**	**$185**	**$235**	**$44**	**$873**
% Total cost per NME	3%	6%	17%	7%	15%	21%	27%	5%	
Cost of capital	11%								
Cost per launch (capitalized)	**$94**	**$166**	**$414**	**$150**	**$273**	**$319**	**$314**	**$48**	**$1,778**

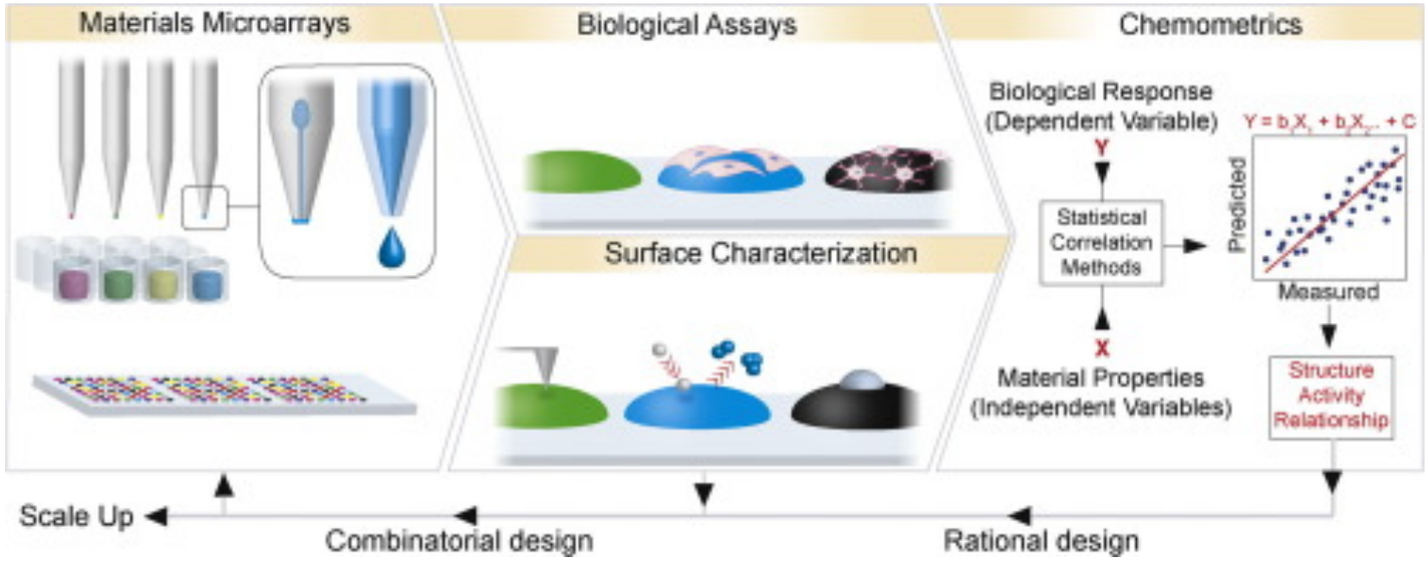
Discovery Development

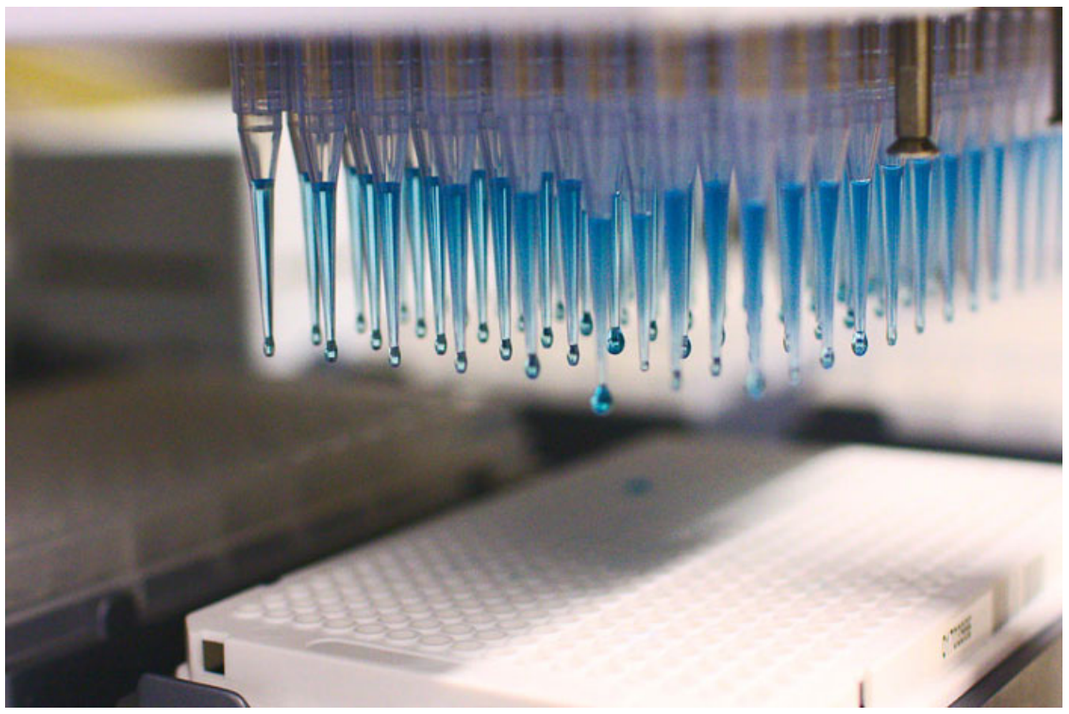
p(TS): 각 단계의 성공률
WIP(work in process) needed for 1 launch: 연구중인 분자의 개수
Cost per WIP per Phase: 각 단계별로 한 분자를 연구하는데 들어가는 비용
Cycle time: 각 단계에서 걸리는 시간
Cost per launch (out of pocket): 순수하게 각 단계에서 필요한 비용
Cost per launch (capitalized): 투자 비용의 이자등을 고려한 전체 투자비용

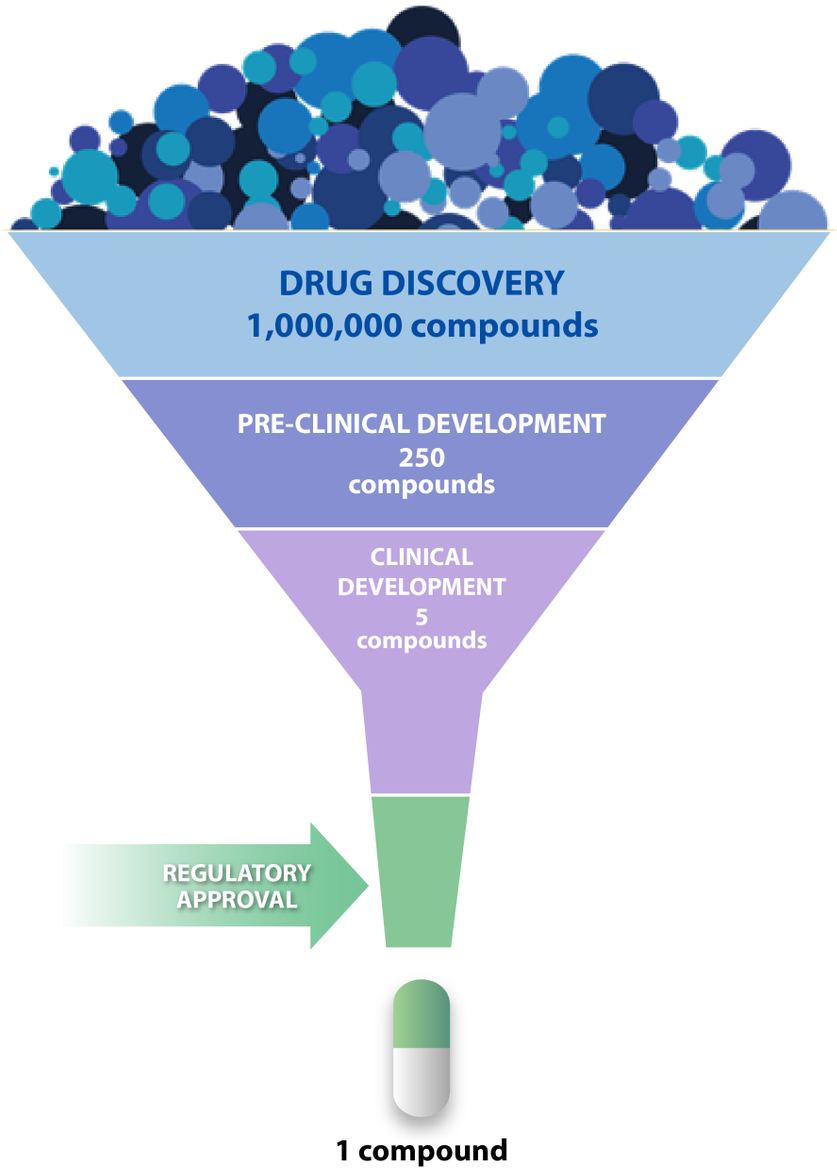
하나의 신약을 개발하기 위해서 대략 **13년** 정도의 기간과 약 **$9억 (1조 2천억)** 정도의 개발 비용이 필요!

Paul, S. M. et al. How to improve R&D productivity: the pharmaceutical industry's grand challenge. Nat. Rev. Drug Discov. 9, 203–14 (2010).

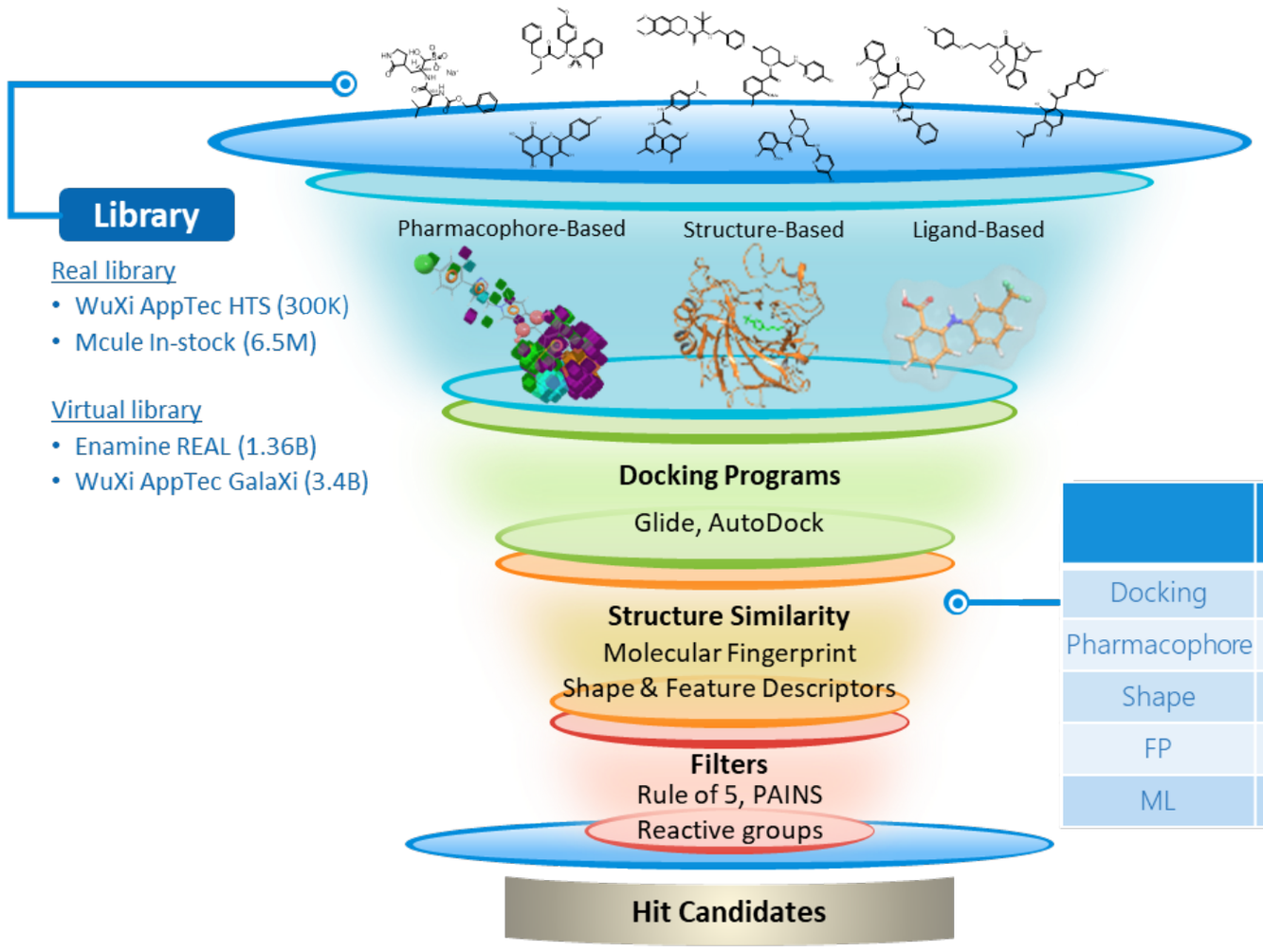
High-throughput screening: the first stage of drug discovery







Protein-ligand docking is important for virtual screening

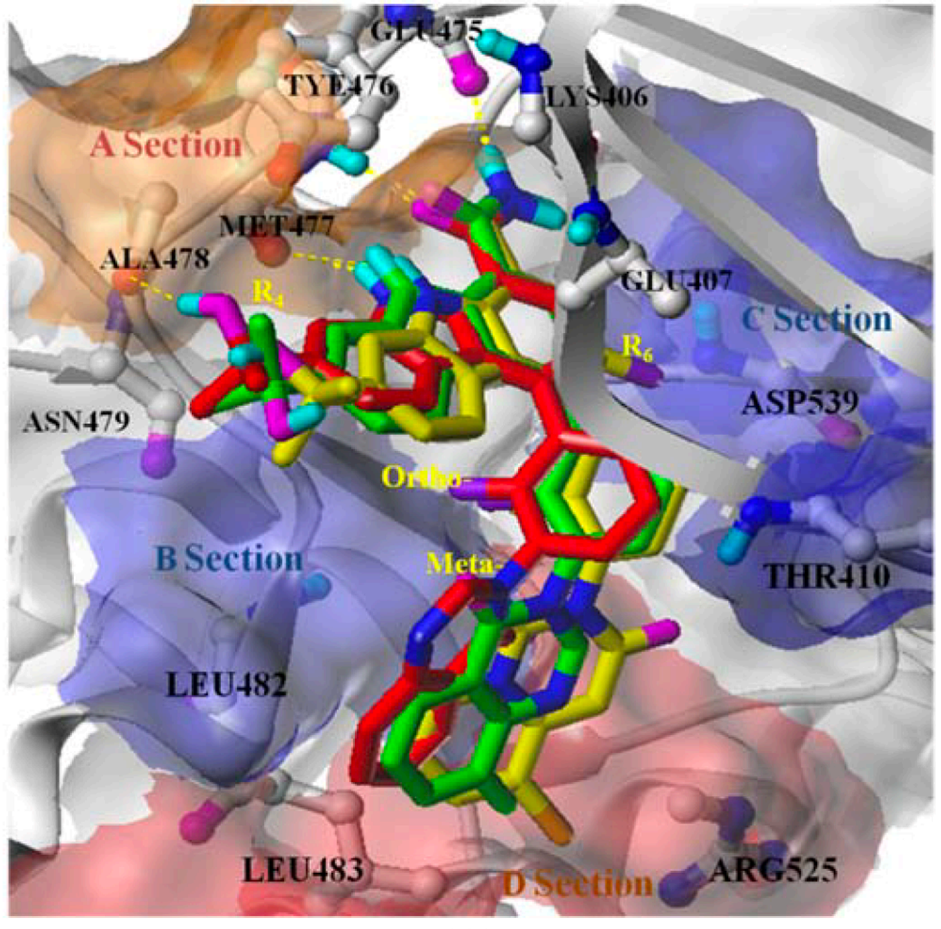


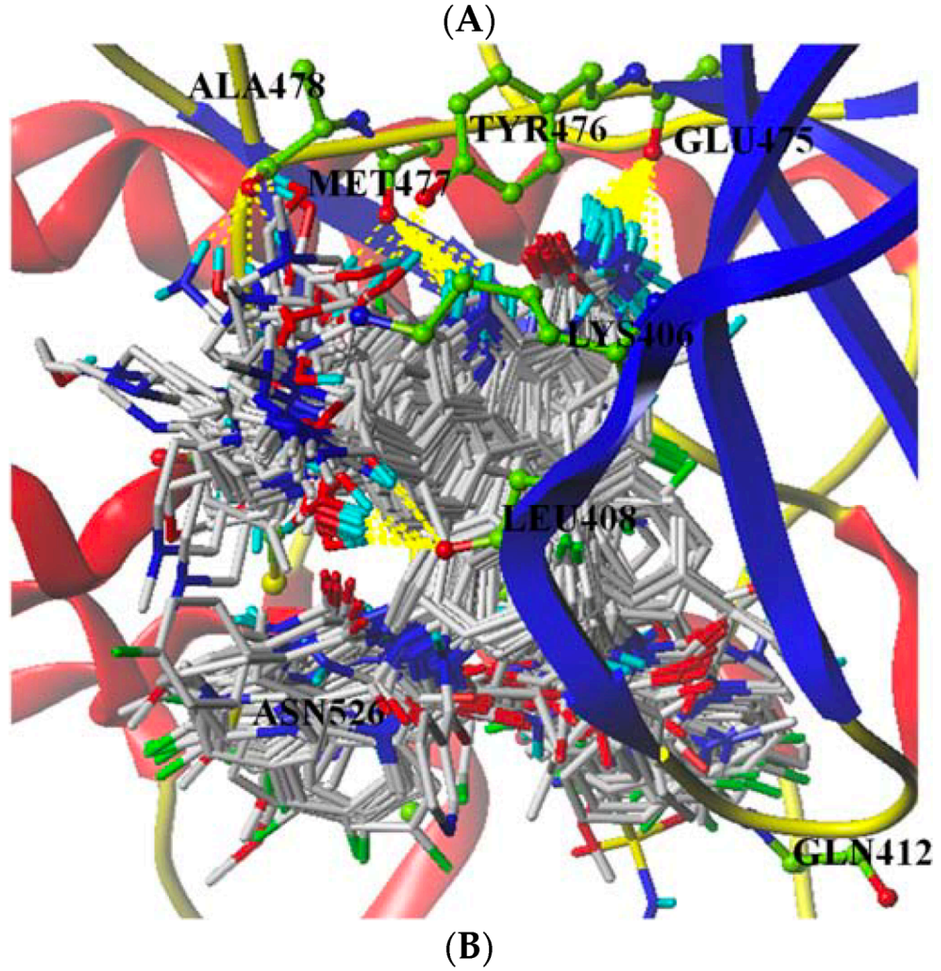
	Chemical Space	Time consuming
Docking	Millions	Days
Pharmacophore	Millions	Days
Shape	Billions	Hours
FP	Billions	Hours
ML	Billions	Hours

- Efficiently finding drug candidates from huge chemical DBs
  - Optimizing new drug candidates with higher affinity

The challenge of protein-ligand docking : pose sampling

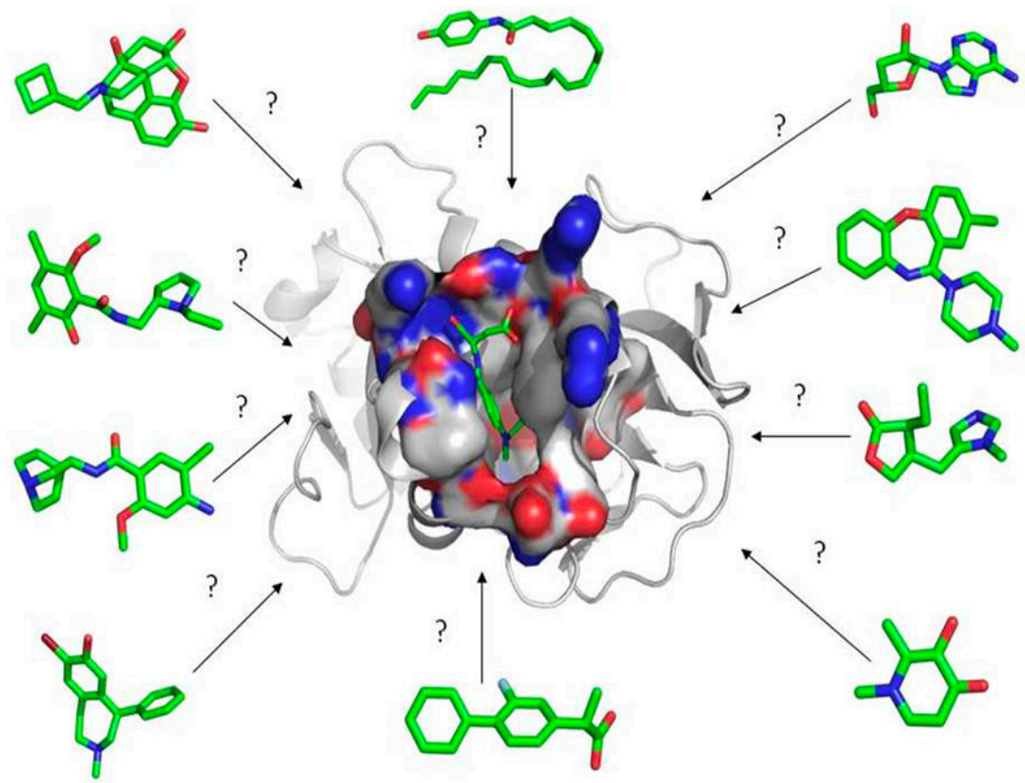
- Sampling of conformational space of ligands
  - What is the true pose of bound ligand?
- Ligands are **flexible**
- It can adopt multiple different binding poses
- Ligand binding sites are unclear





The challenge of protein-ligand docking - scoring

- Scoring of protein-ligand interaction
  - **Predicting the binding affinity** (binding free energy)
  - Free energy change of
    - [P] + [L]⟶[P-L]
    - $\Delta G_{\text{bind}} = -kT \ln \frac{[P][L]}{[PL]}$
  - True binders should have high affinities: < -7 kcal/mol
  - False binders should have low affinities: > -7 kcal/mol



Which ones will bind?
Which ones will not bind?
Which ligand binds most strongly?

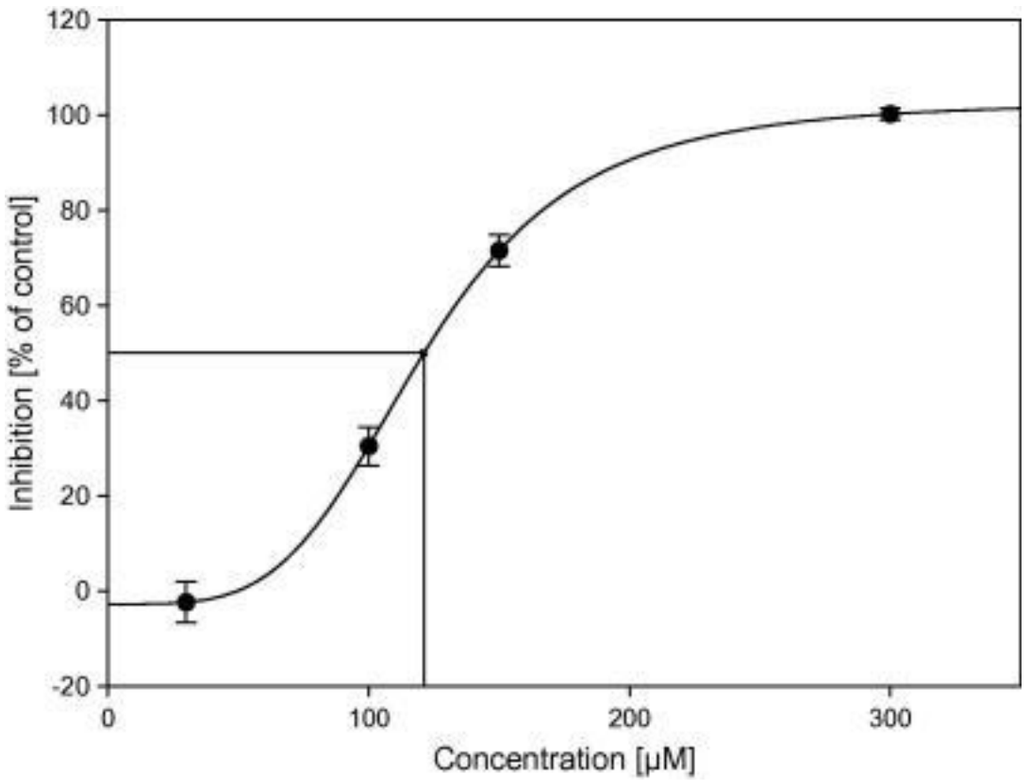
The challenge of protein-ligand docking - scoring

- Protein-ligand binding affinities are generally measured by half maximal inhibitory concentration IC50.

- Lower indicates higher affinity

  - Hit molecules ~ μM

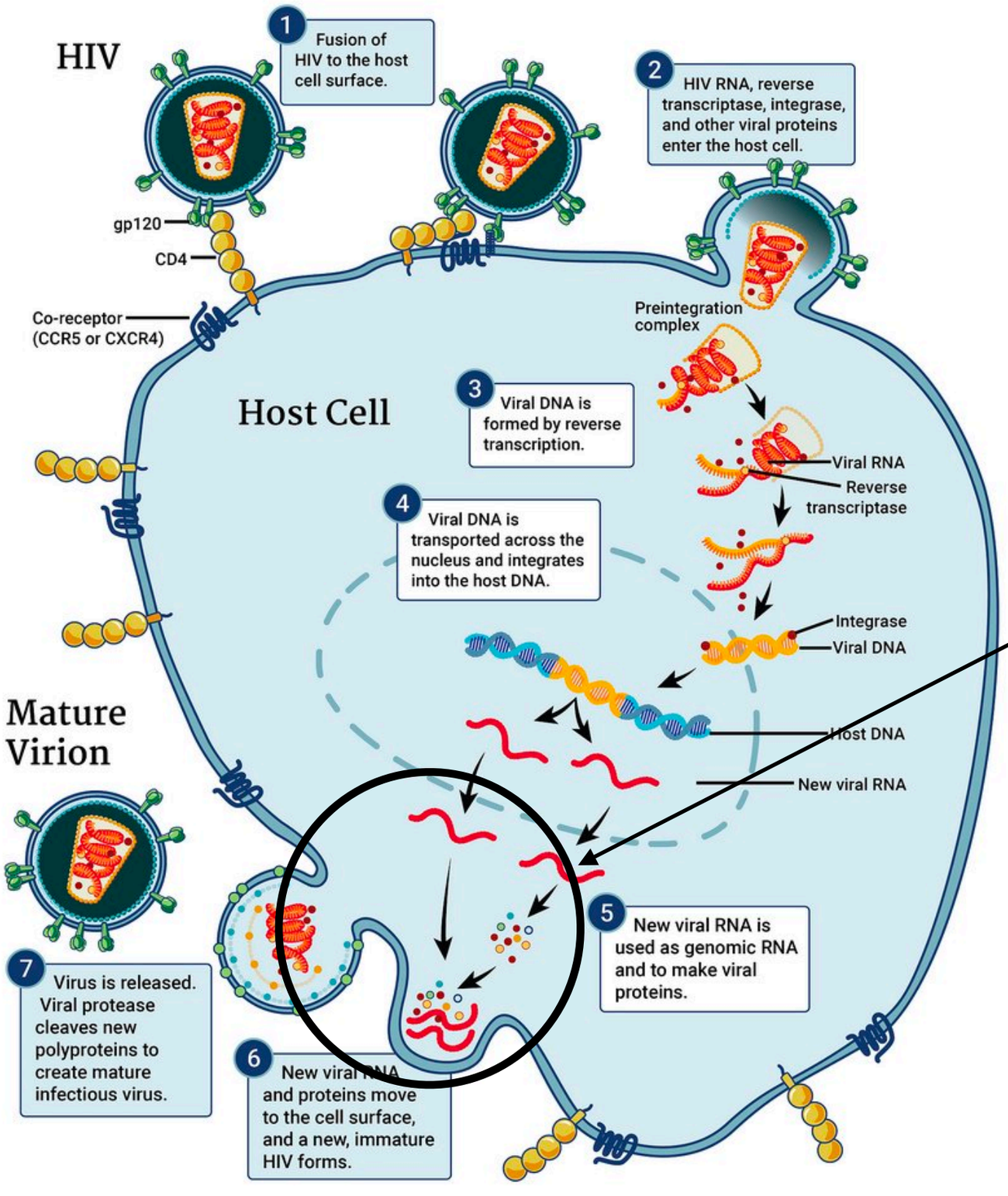
  - Lead molecules ~ nM



- With some assumptions

- $$\Delta G_{\text{bind}}(\text{IC50}) = -kT \ln \frac{[P][L]}{[PL]} = -kT \ln[L_{\text{IC50}}]$$

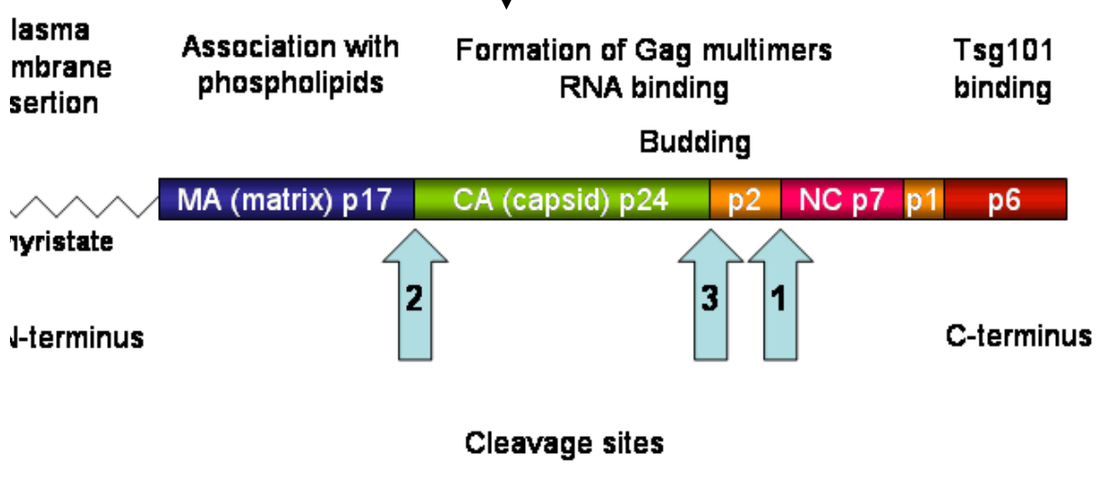
Example of protein-ligand docking - HIV protease



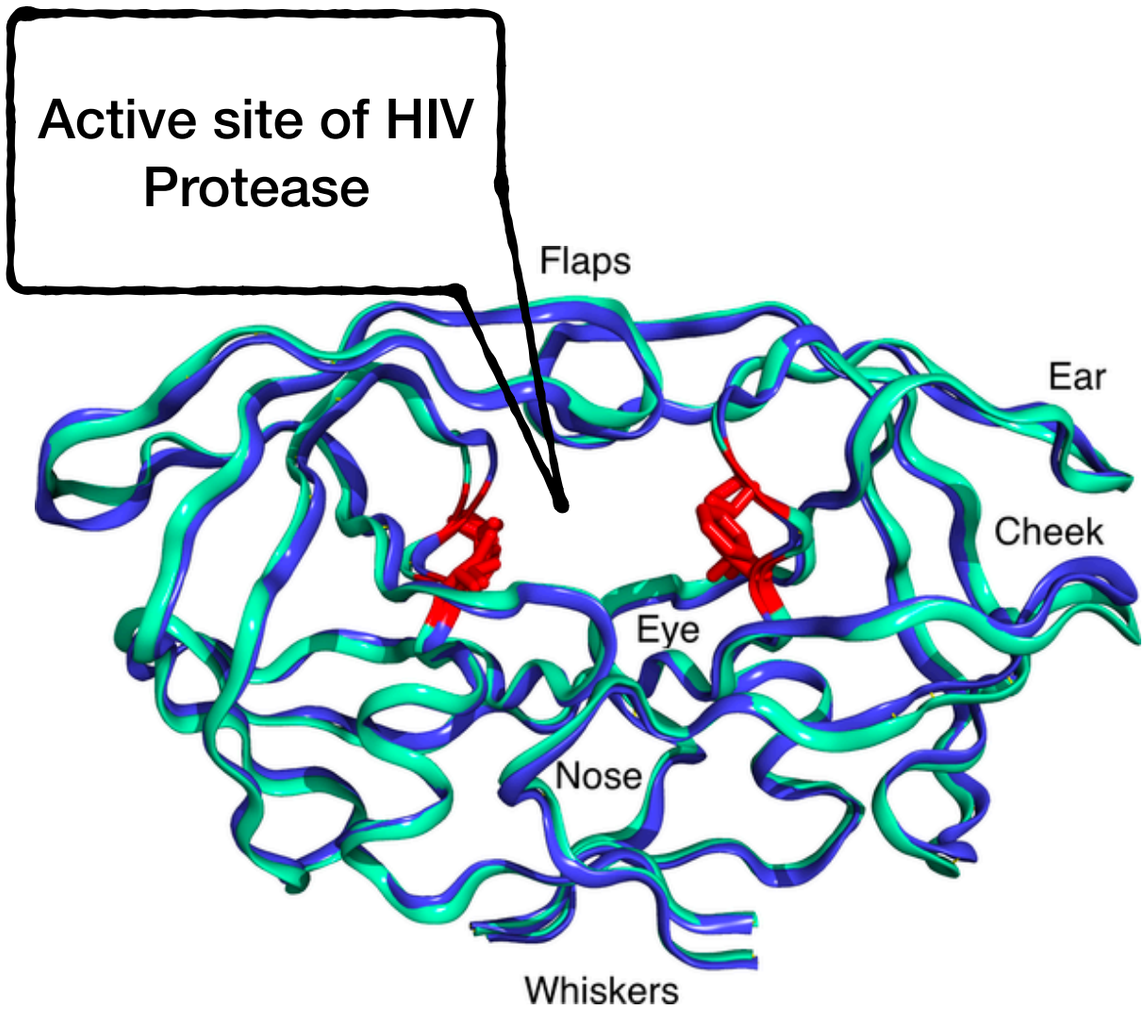
1. When HIV virus is infected, it fuses with a host cell
2. It transcripts its DNA using host cell's ribosome
3. Creates a long protein chain including multiple proteins

A long protein-chain must be cut to form separate proteins

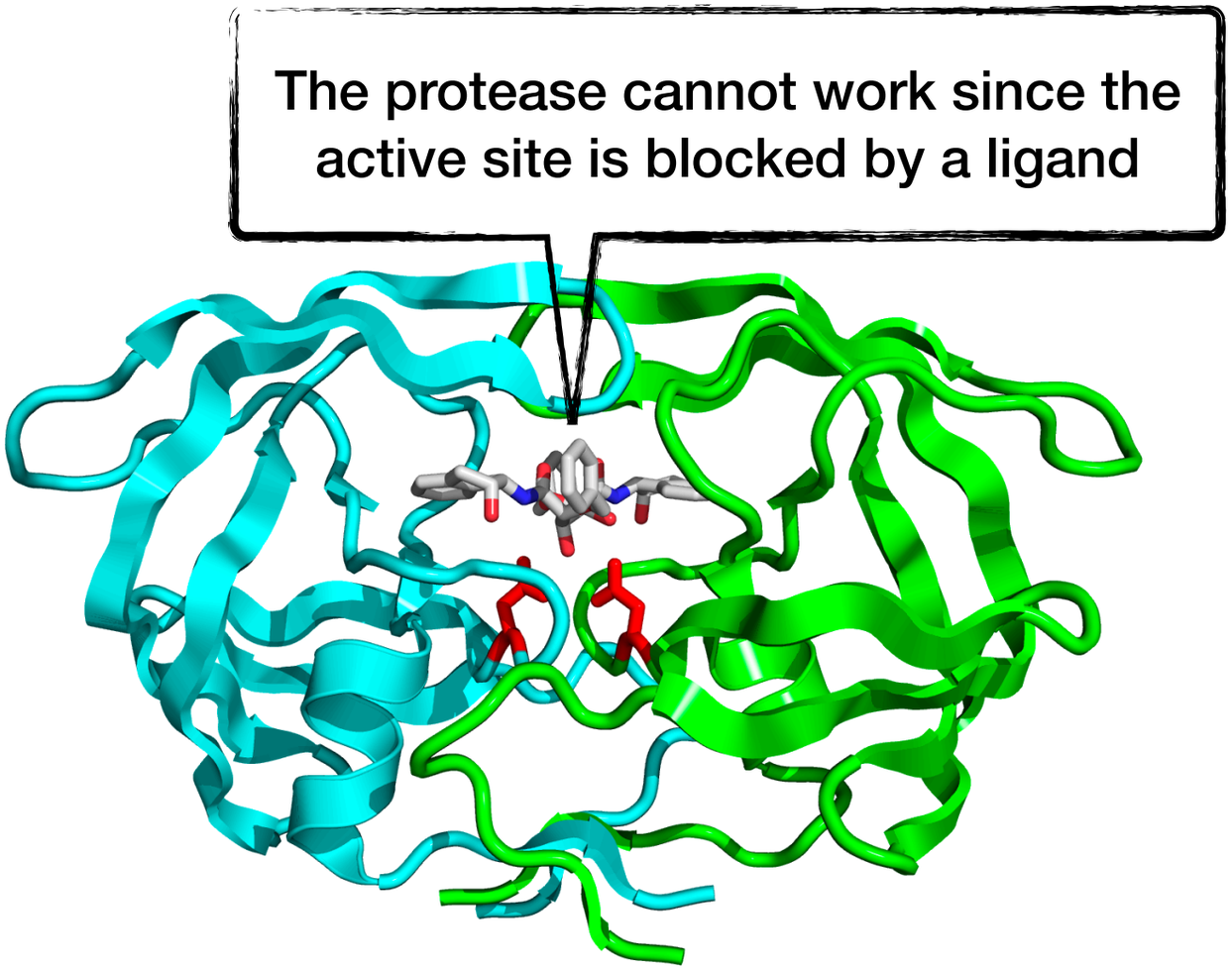
HIV protease is an enzyme that cuts the chain.



The structure of HIV-protease



The structure of HIV-protease
PDB ID: 1D4S

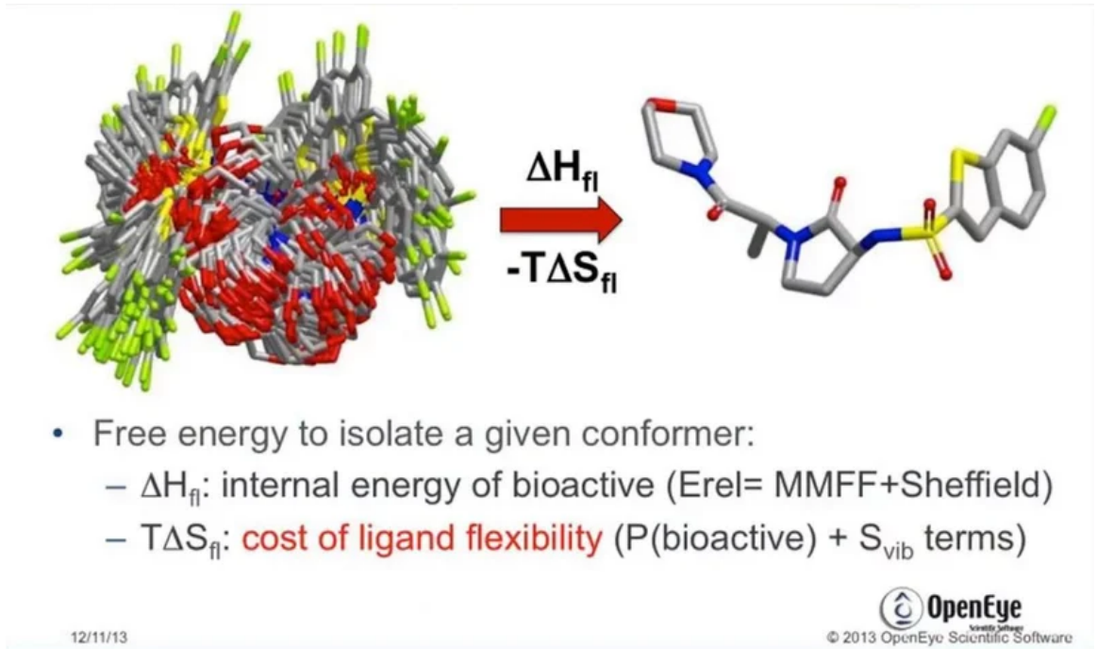


The structure of HIV-protease docked with a ligand
PDB ID: 1EBY

- A site where reactions occur or ligands bind is called an **active site** or **a binding site**

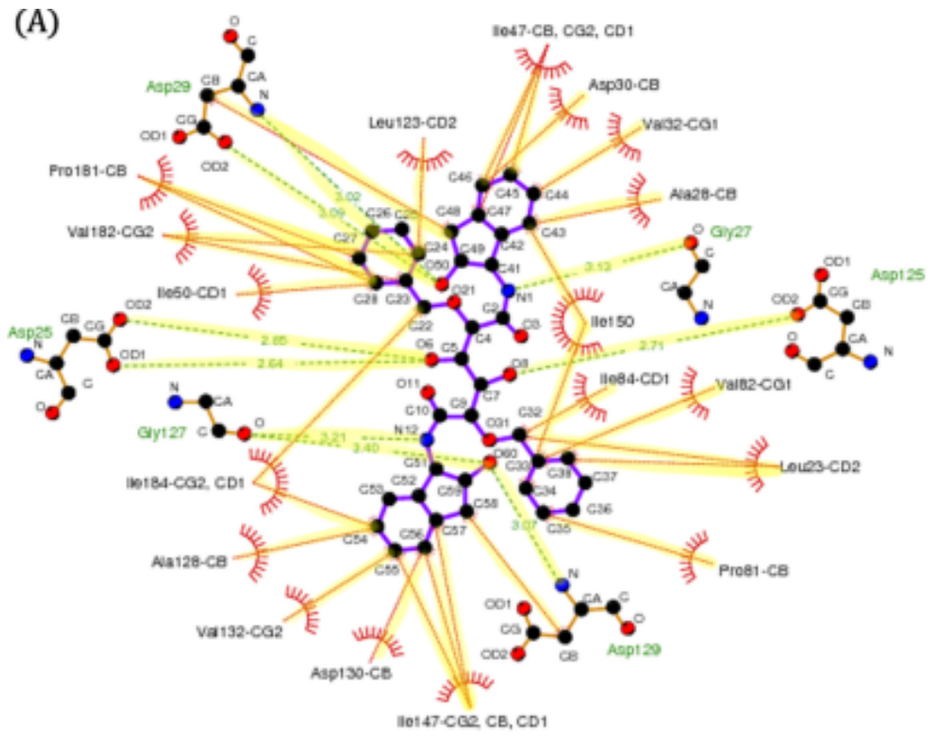
Ligands have diverse conformation

- **Sampling Problem - many possible structures**
  - The structures of proteins and ligands are flexible
  - The structures of proteins may change if ligands bind
  - Ligands (small molecules) have multiple conformations



Estimating exact binding affinity is difficult

- **Scoring Problem - accurate energy function**
  - It is **hard** to calculate an accurate **binding free energy** **(binding affinity, 결합 친화도)**
  - Proteins and ligands consist of many atoms
    - Solving the Schrödinger equation of a protein takes forever
  - **Entropy** is hard to calculate since it is related to the number of possible states



$$\Delta G_{\text{bind}} = \Delta H_{\text{bind}} - T\Delta S_{\text{bind}}$$

$$\Delta G = (V_{bound}^{L-L} - V_{unbound}^{L-L}) + (V_{bound}^{P-P} - V_{unbound}^{P-P}) + (V_{bound}^{P-L} - V_{unbound}^{P-L} + \Delta S_{conf})$$

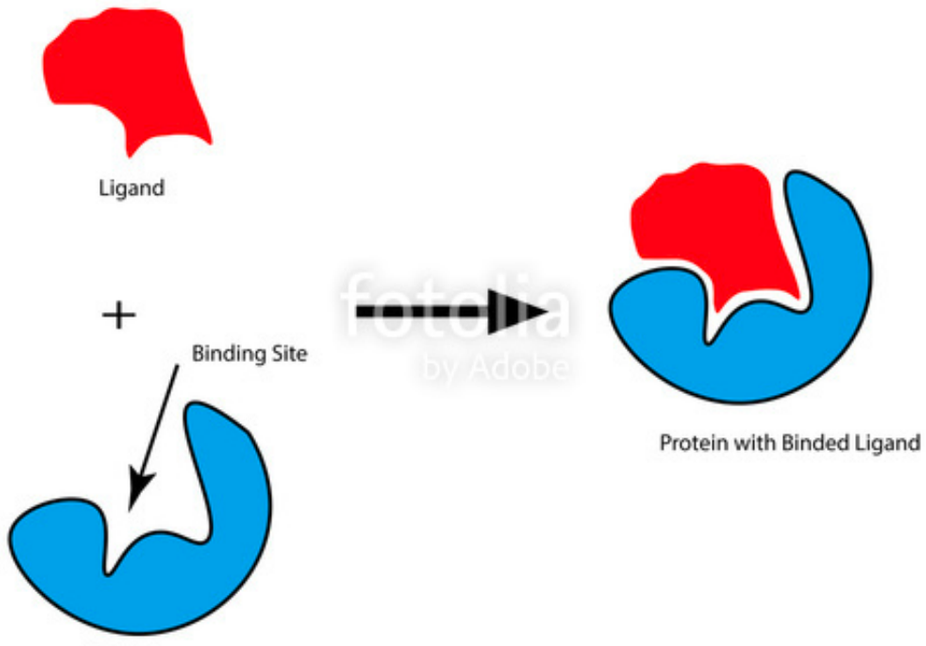
Basic assumption of protein-ligand docking

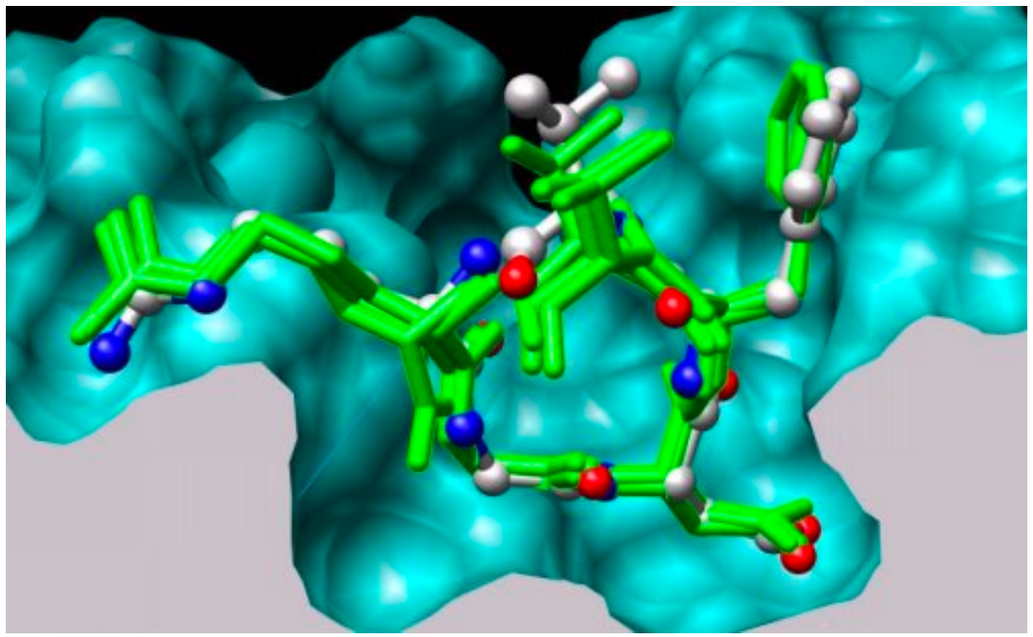
- Ligands generally bind to **structurally complementary sites**

- Why?

- For many protein-ligand complexes, van der Waals interactions are important.

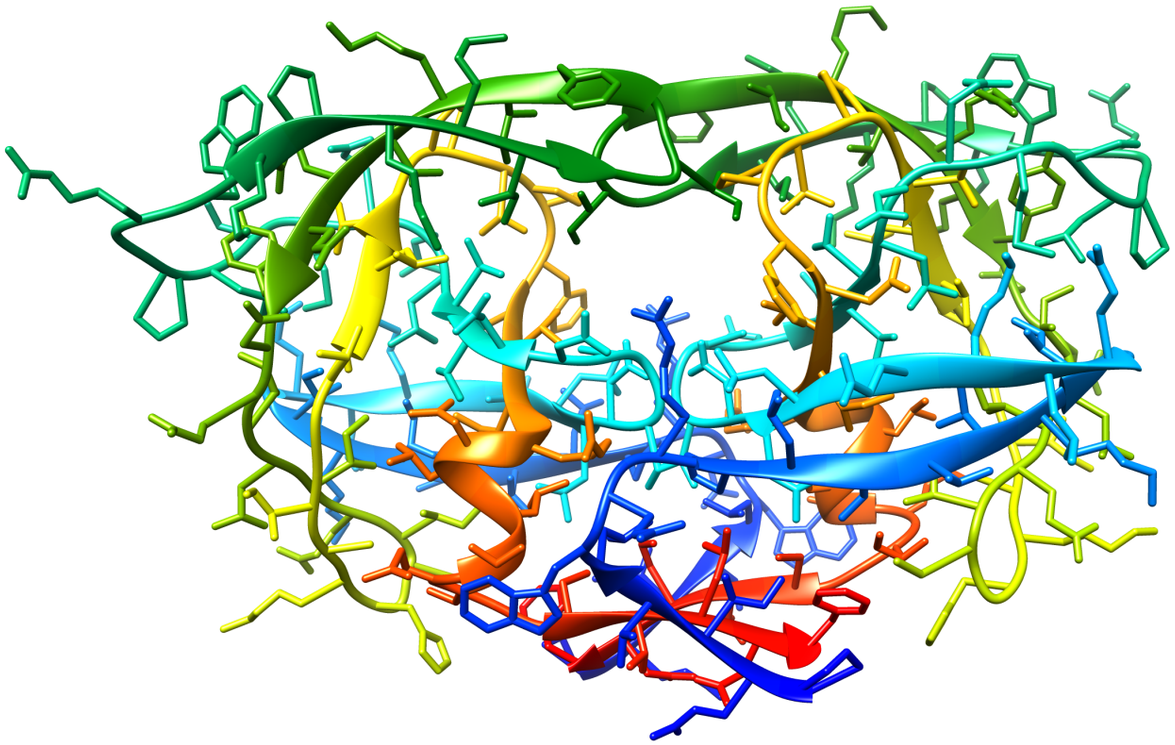
- vdW interactions become **larger** as more and **more atoms interact**



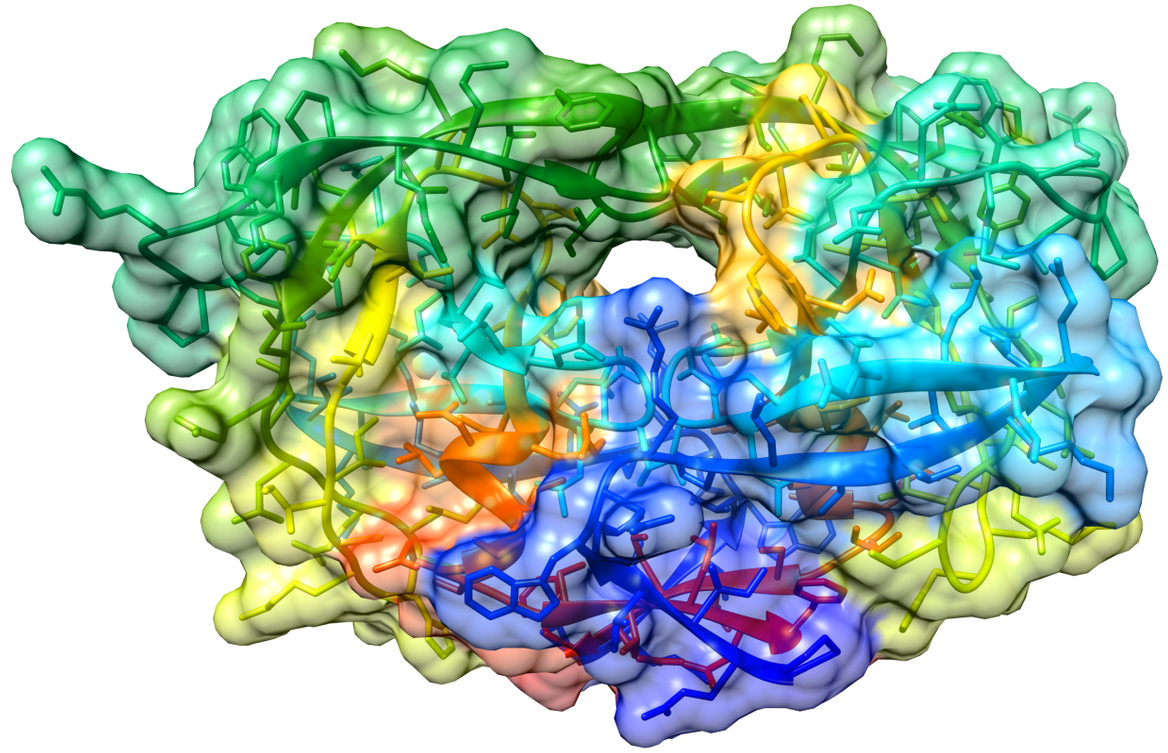


Structural complementary

How to define binding sites and structural complementary?



Ribbon & Atom representation of HIV Protease

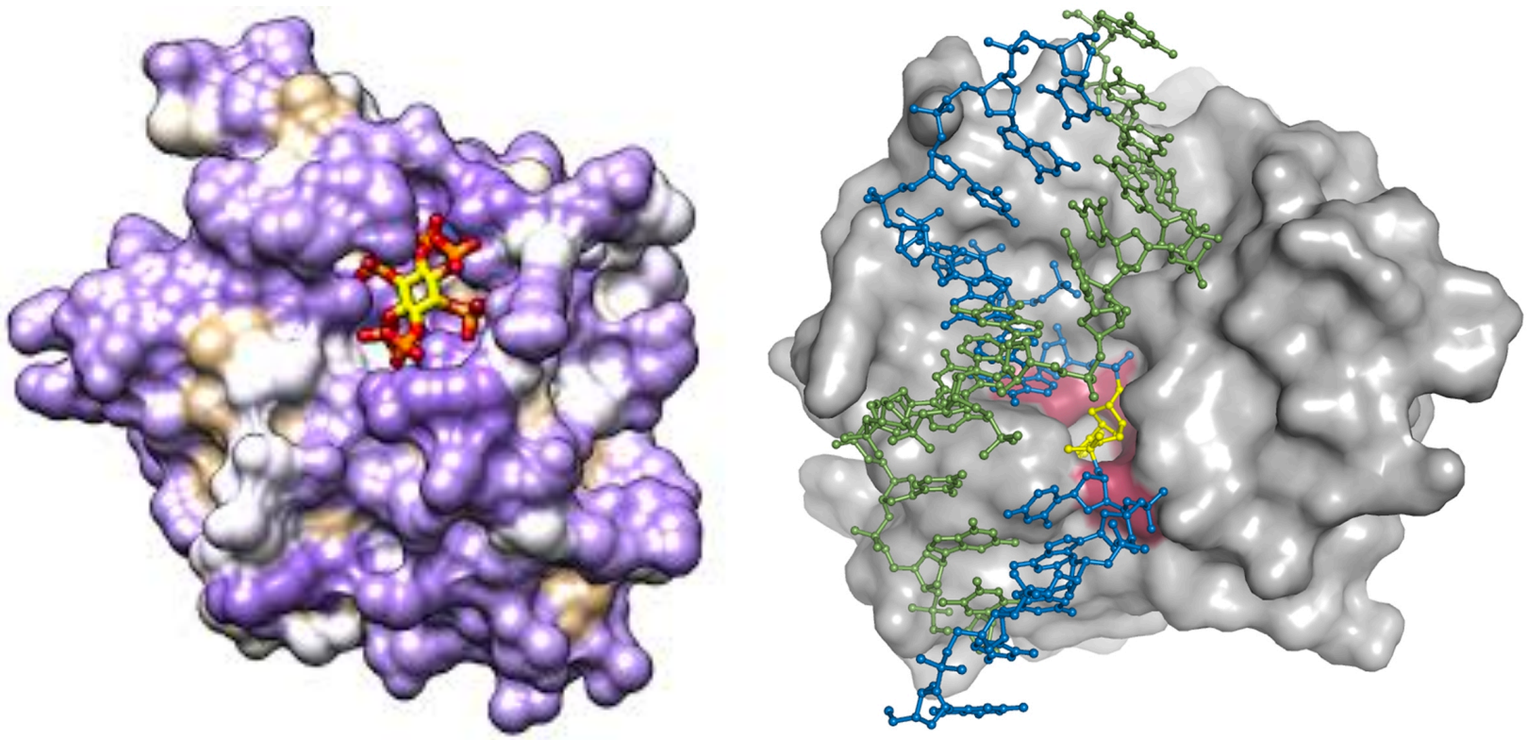


Surface of HIV protease

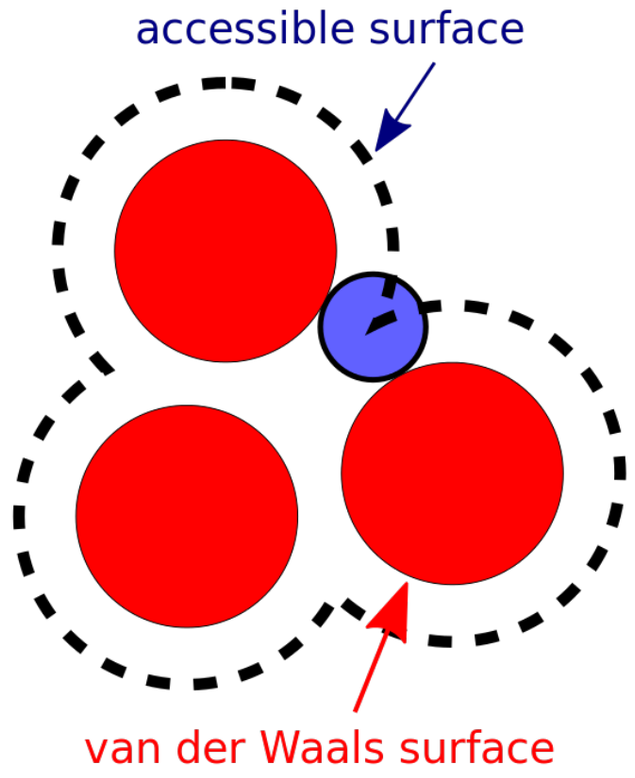
- To identify structural complementary, the surface of protein should be considered

- Remember that each atom has van der Waals radius

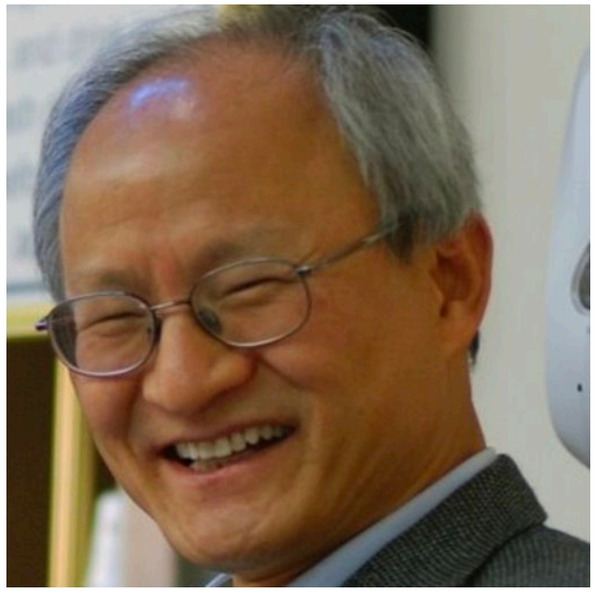
How to generate protein surface?



A ligand and DNA bound on protein surface

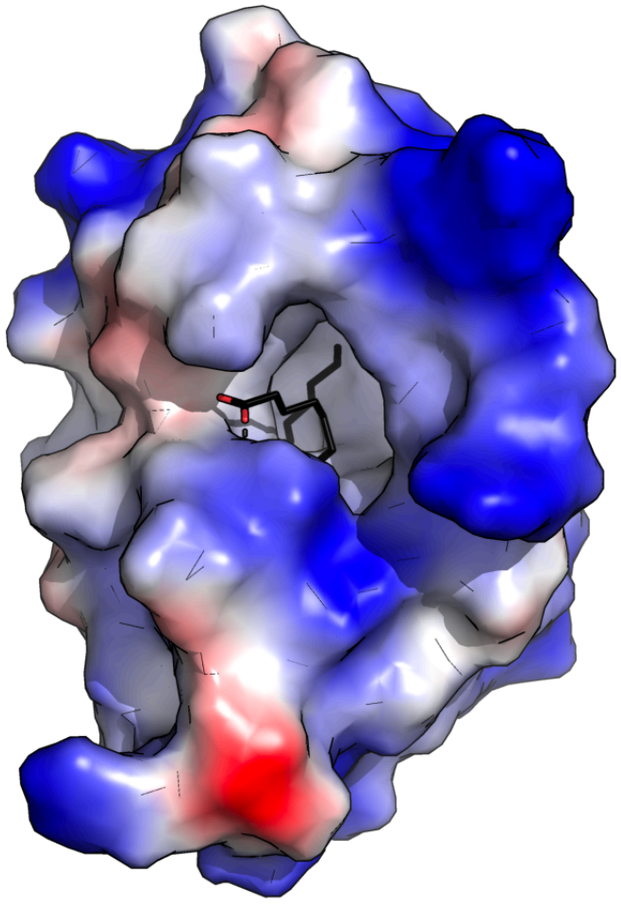


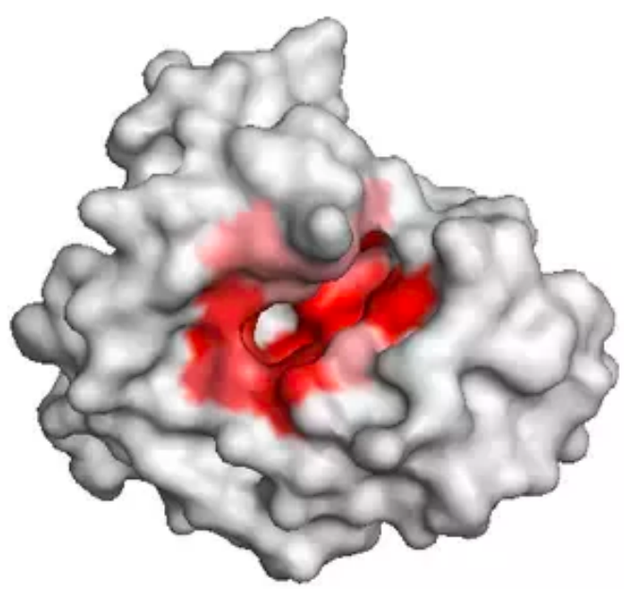
- Accessible surface area (**ASA**)
- Also known as Lee and Richardson surface
- Roll a probe on a protein surface

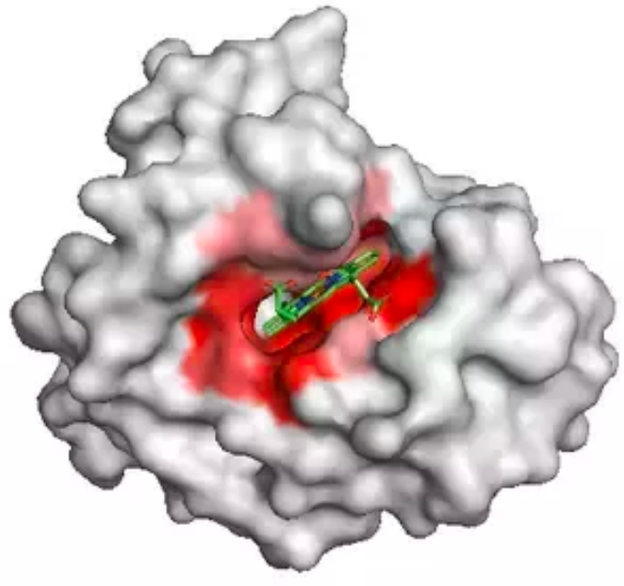


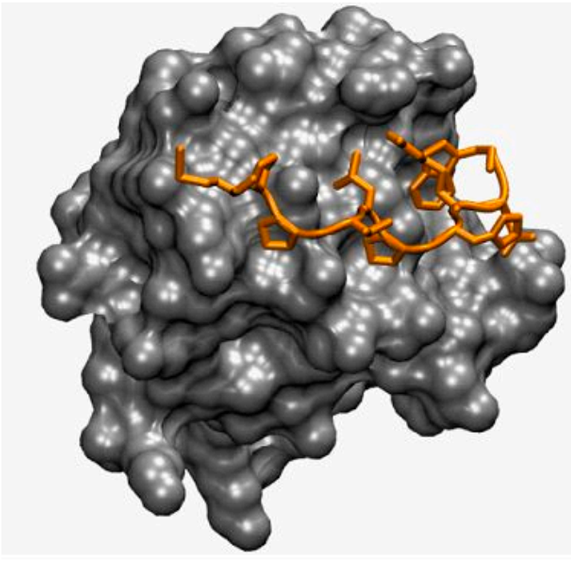
Byoung-Kuk Lee
NIH

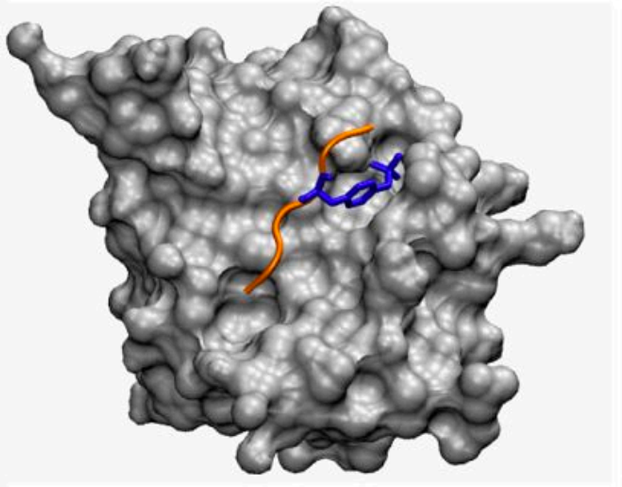
Ligands bind to pockets









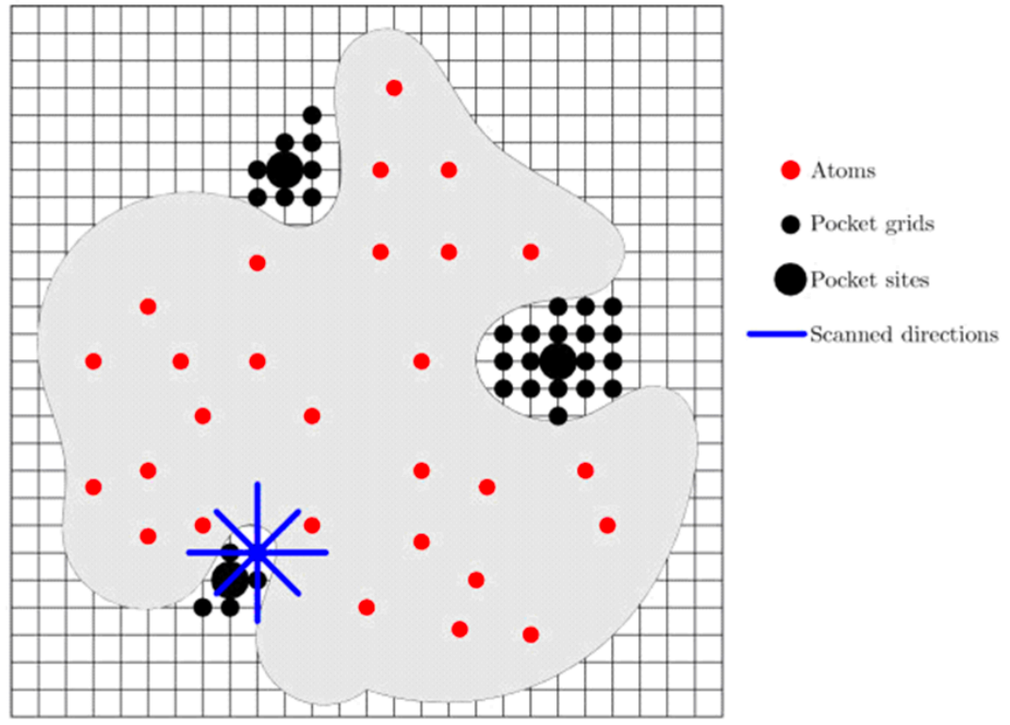


- Examples of ligand binding sites

- Ligands generally bind to **concave (오목한)** sites of a protein surface

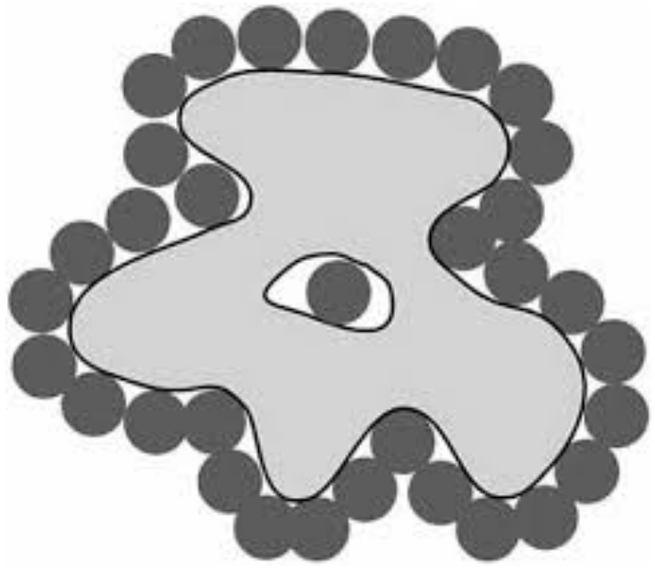
How to find pockets?

- From the geometry of protein surface, pockets can be detected



- If a probe is inside concave pockets, they have more interactions with neighboring atoms

- If a probe is at a convex surface, they have less interactions with protein atoms



Questions?

Slido: 2698620

https://app.sli.do/event/58MDz45pkXgC3SXJLVc6fW

Conventional protein-ligand docking

Example of Autodock

Scoring functions

- Energy functions
- To predict binding free energy of protein-ligand complex
  - **Nonbonded interactions are important**
    - **Electrostatic**
    - **Van der Waals**
    - **Solvation**
- Two types of scoring functions
  - **Physics-based** (물리 법칙에 기반을 둔) function
  - **Empirical** (경험적인) functions
- Every program has different scoring function

Empirical scoring functions

- **Fit parameters to experimental data**
  - Assume a function based on intuition and fit parameters
- **An example of empirical scoring function (Autodock)**
  - **http://autodock.scripps.edu**

$$\Delta G = \Delta H_{vdW} \quad W_{vdW} \sum_{i,j} \left(\frac{A_{ij}}{s(r_{ij})^{12}} - \frac{B_{ij}}{s(r_{ij})^{6}} \right)$$ **Van der Waals**

$$+ \Delta H_{\text{hbond}} \quad W_{hb} \sum_{i,j} E(t) \left(\frac{C_{ij}}{s(r_{ij})^{12}} - \frac{D_{ij}}{s(r_{ij})^{10}} \right)$$ **Hydrogen bond**

$$+ \Delta H_{\text{elec}} \quad W_{el} \sum_{i,j} \left(\frac{q_i q_j}{\varepsilon(r_{ij}) r_{ij}} \right)$$ **Electrostatics**

$$+ \Delta G_{\text{desolv}} \quad W_{\text{desolv}} \sum_{i,j} (S_i V_j + S_j V_i) e^{(-r_{ij}^2 / 2\sigma^2)}$$ **Desolvation term**

$$+ \Delta S_{\text{tor}} \quad W N_{\text{tor}}$$ (1) **Torsion angle term**

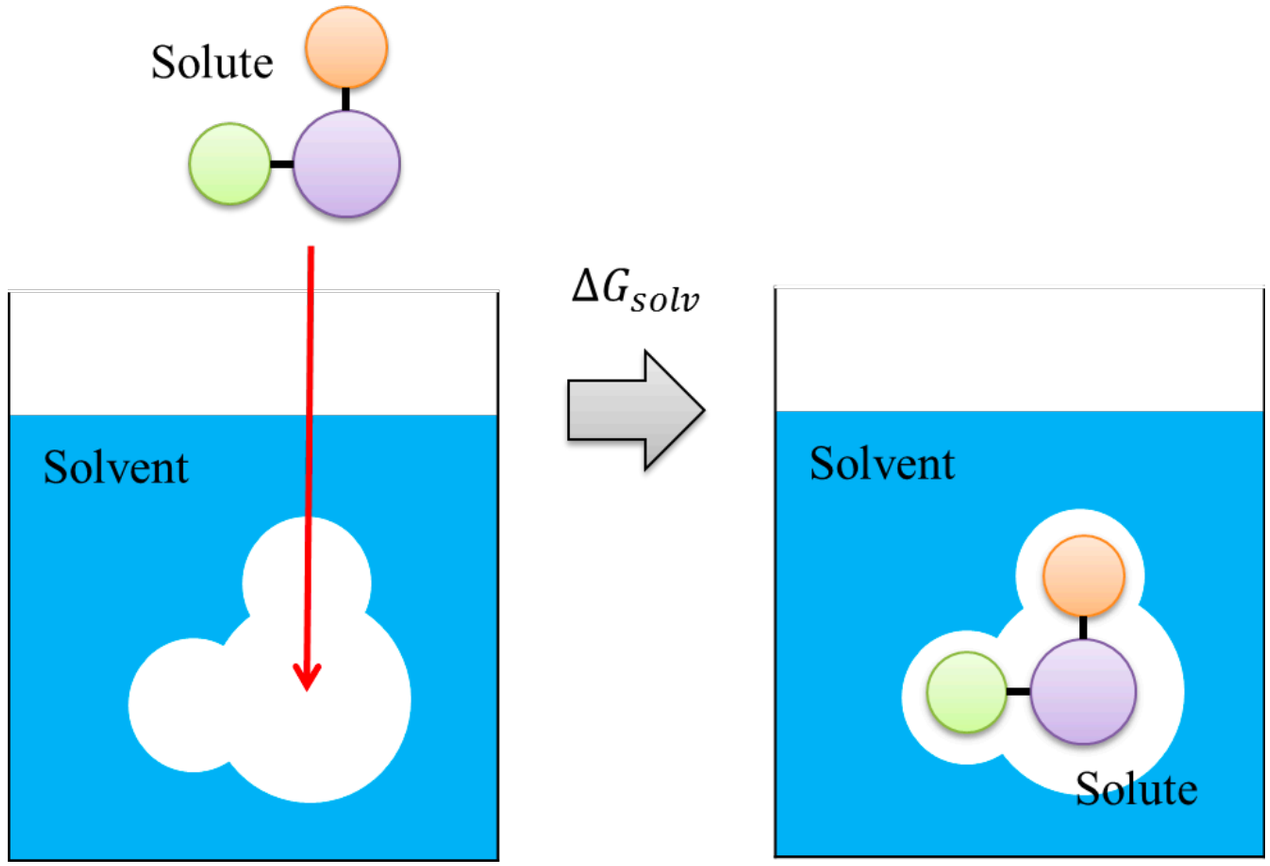
Two most important non-bonded interactions

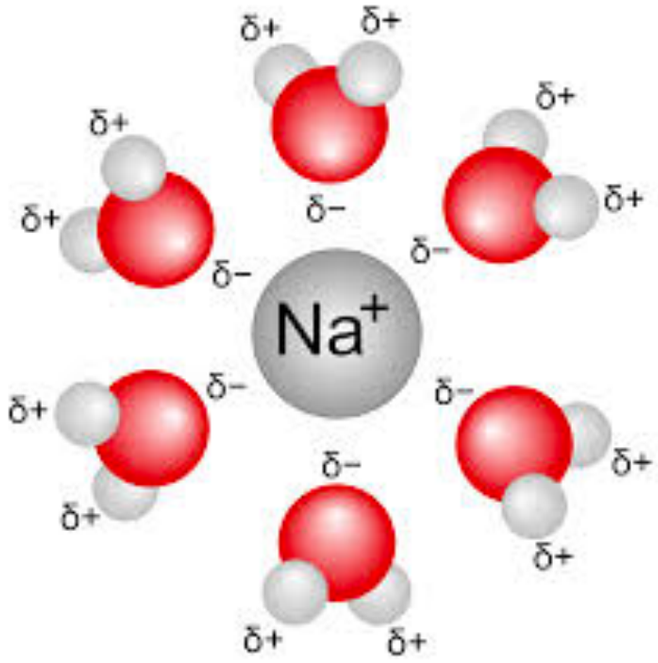
$$E_{nonbond} = \sum_{i}^{lig} \sum_{j}^{prot} \left[\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} + c \frac{q_i q_j}{r_{ij}} \right]$$

van der Waals | Column

- Based on Column's law and van der Waals interaction
- Parameters are obtained from quantum mechanical calculations, not from experimental data

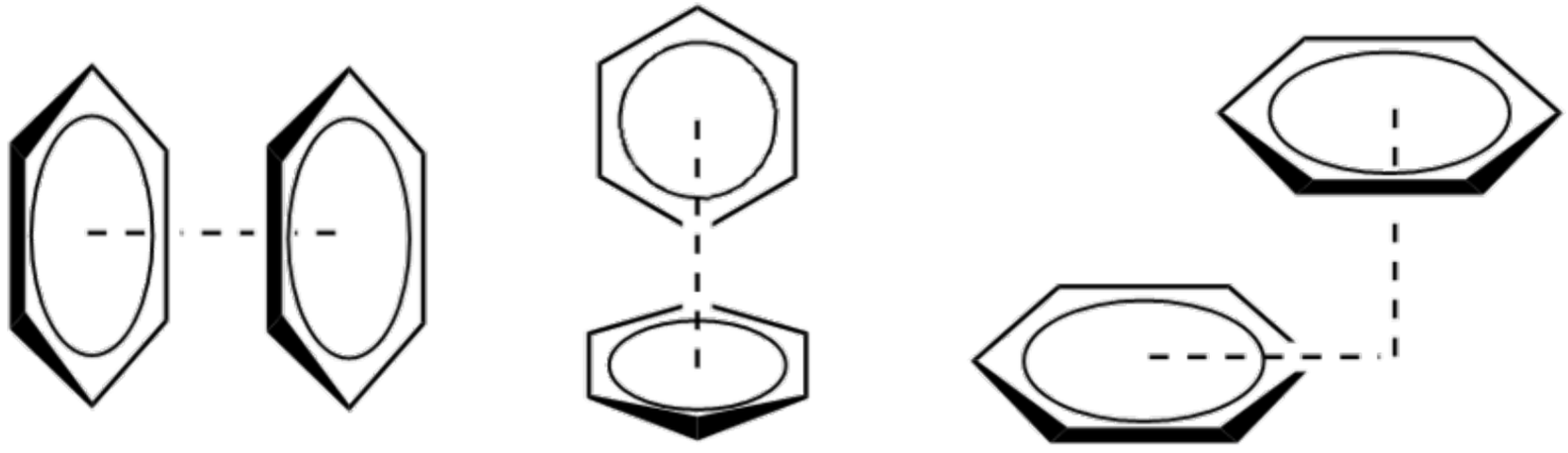
(De)Solvation energy





- Energy required to transfer a molecule from vacuum to solvent
- Can be decomposed into three terms
  - $\Delta G_{solv} = \Delta G_{elec} + \Delta G_{vdw} + \Delta G_{cavity}$
  - Elec: electrostatic interaction between solute & solvent
  - Vdw: van der Waals energy
  - cavity: energy to form a cavity inside solvent

π–π stacking interaction

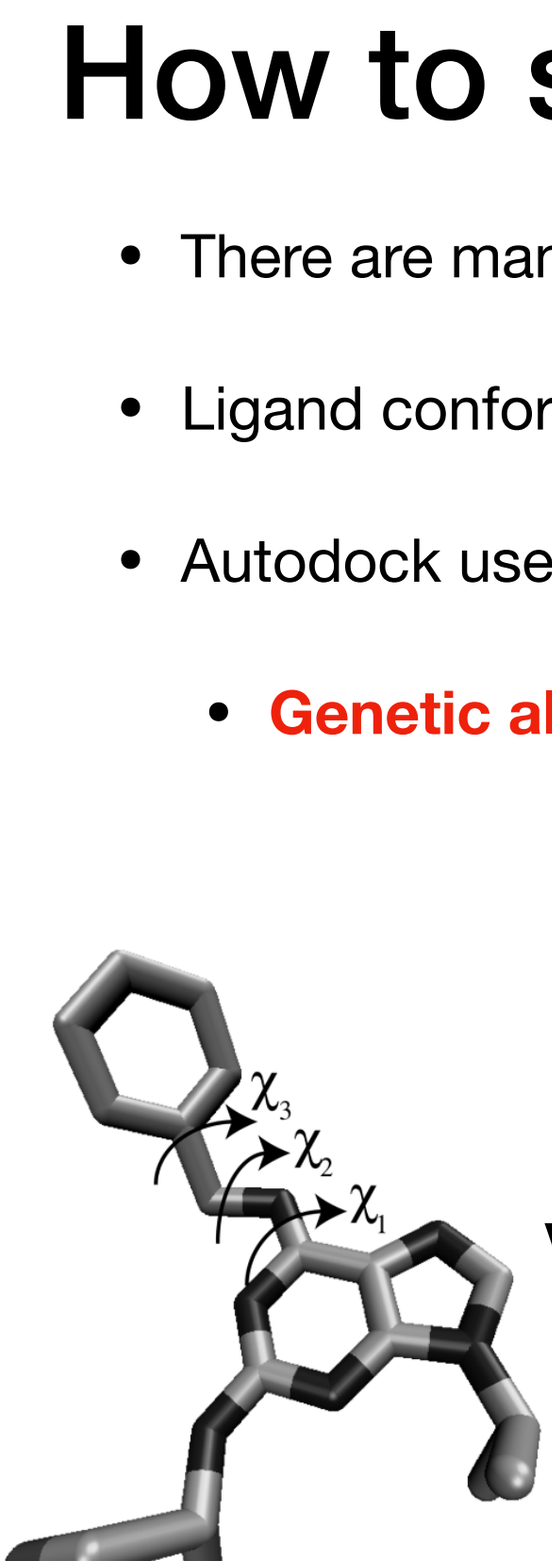


Sandwich T-shaped Parallel-displaced

- Aromatic rings tend to stack together closely
- Stacking interactions are generally in a range of 2~3 kcal/mol
- Due to dispersion and quantum effects
- Mostly represented by the van der Waals interaction term

How to sample ligand conformations?

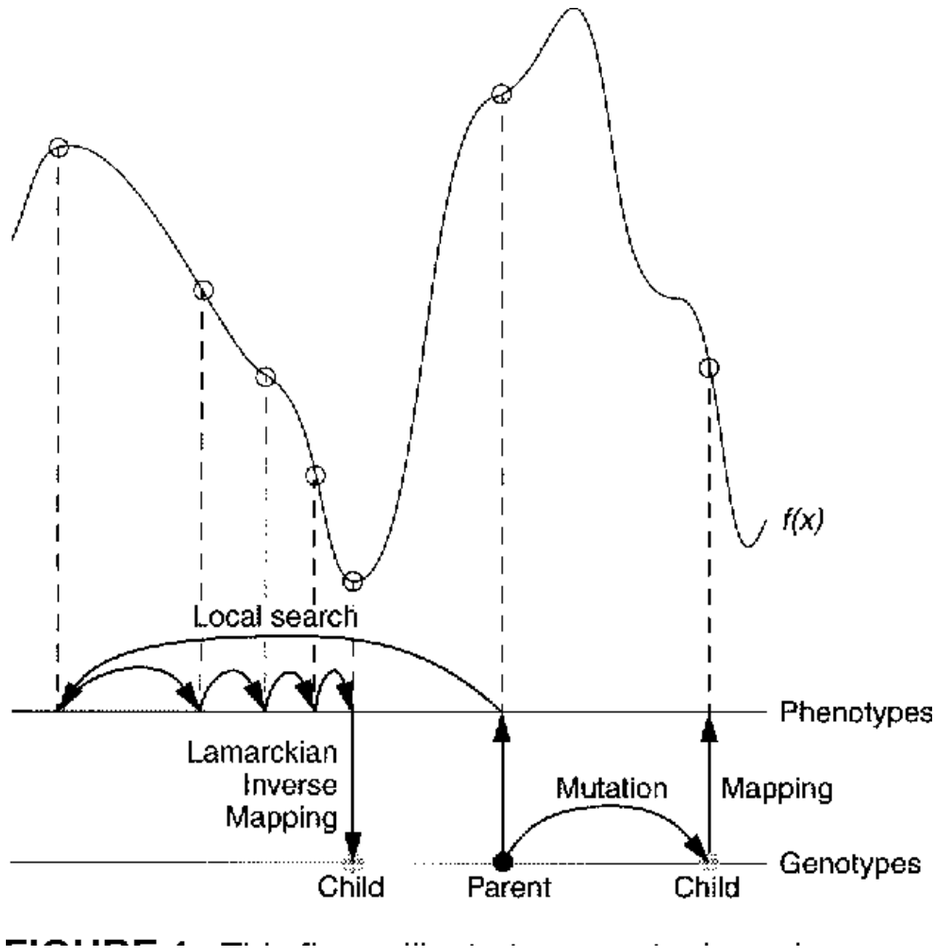
- There are many different algorithms to consider ligand flexibility
- Ligand conformations are represented as a set of rotatable torsion angles
- Autodock uses **Lamarkian genetic algorithm**
  - **Genetic algorithms + local Monte Carlo minimization**



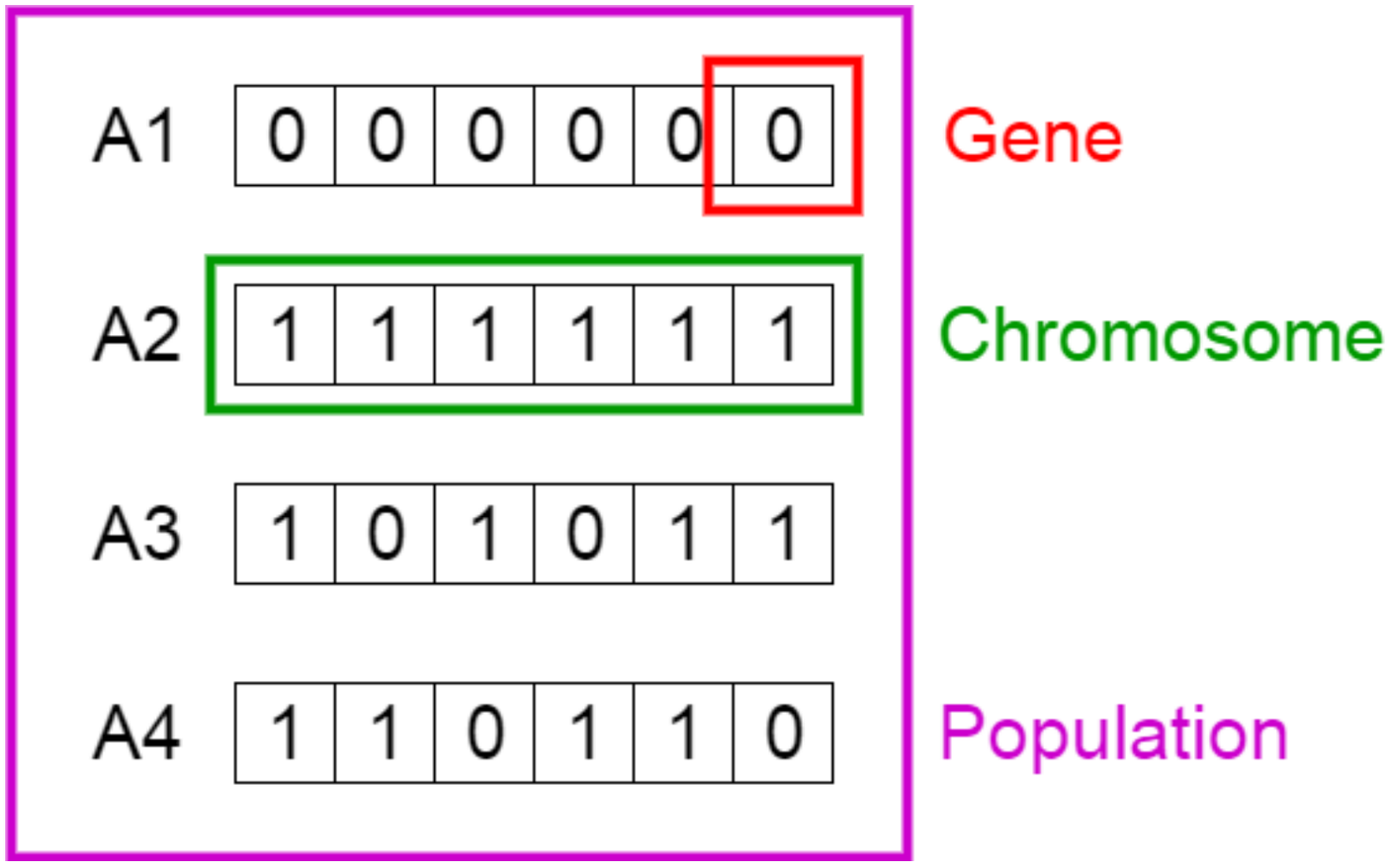
vectorize

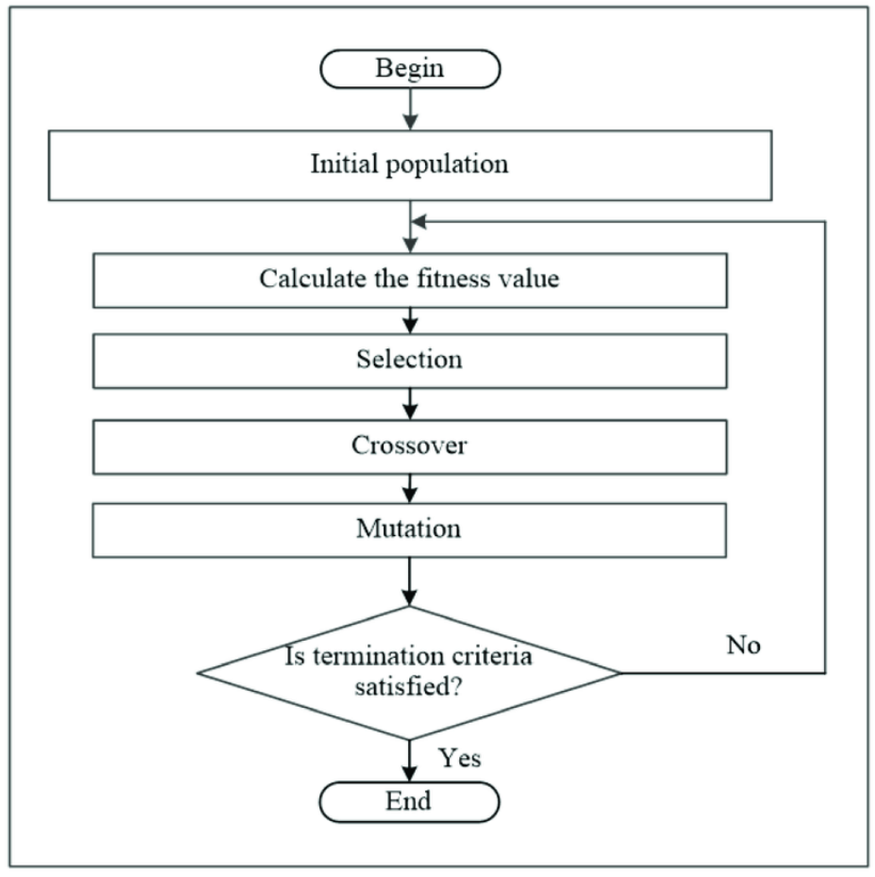
$$(\chi_1, \chi_2, \chi_3, \ldots, \chi_n)$$

Global optimization



Genetic algorithm





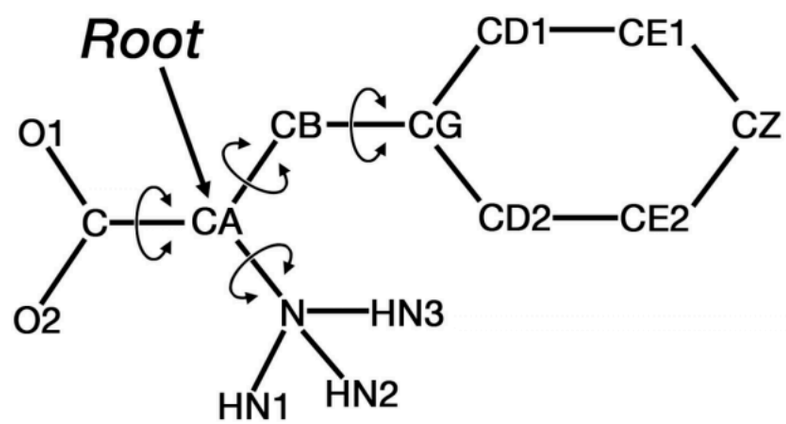
- Inspired by evolution, genes (features) are mixed and mutated
- Better solutions are kept
- Iterate until no better solution is found

PDBQT file

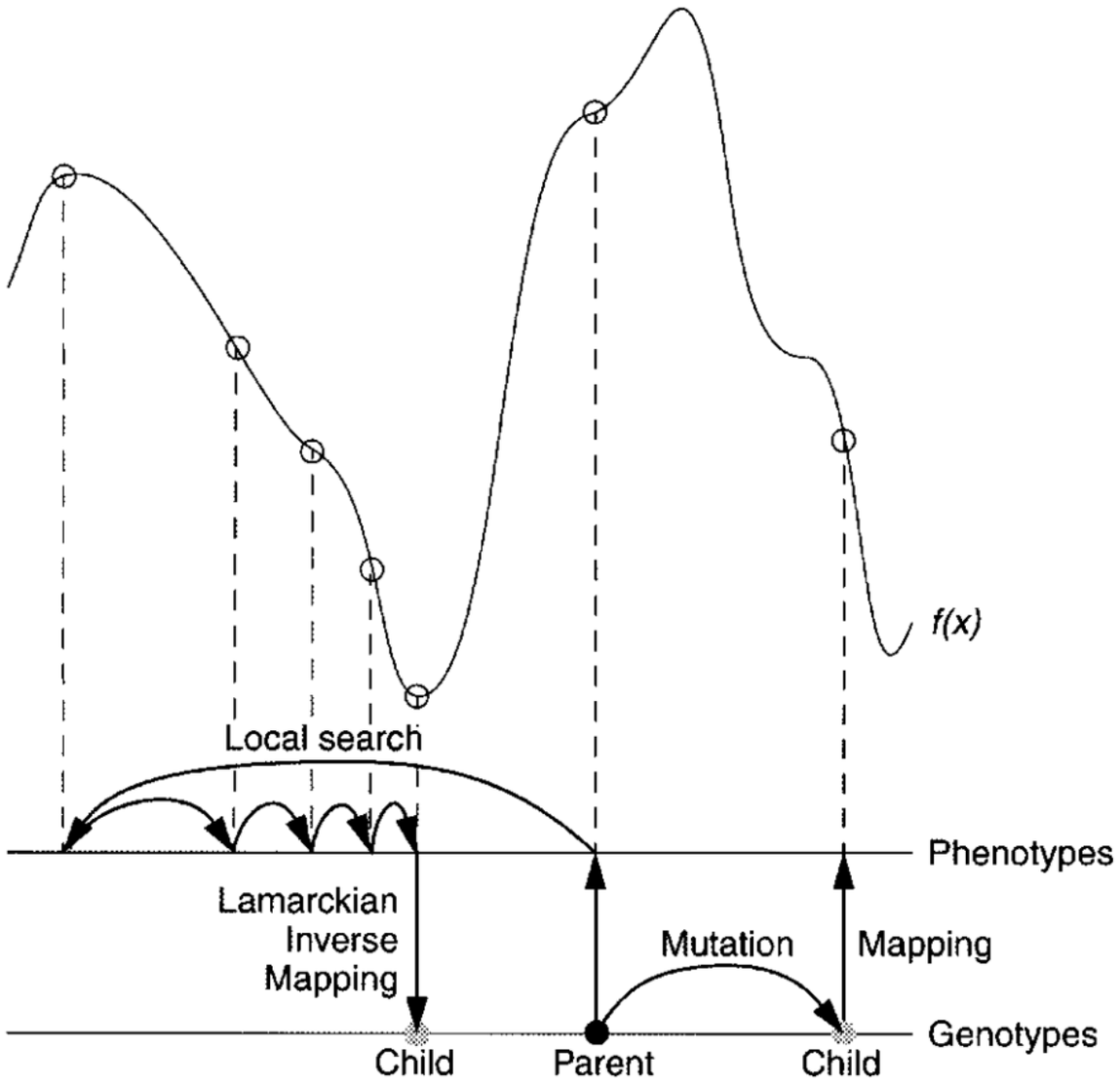
- PDBQT file
  - Rigid fragments are connected via branch
- Torsion angles of rotatable bonds are sampled

Sample PDBQT file

```
REMARK  4 active torsions:
REMARK  status: ('A' for Active; 'I' for Inactive)
REMARK    1  A    between atoms: N_1  and  CA_5
REMARK    2  A    between atoms: CA_5  and  CB_6
REMARK    3  A    between atoms: CA_5  and  C_13
REMARK    4  A    between atoms: CB_6  and  CG_7
ROOT
ATOM      1  CA  PHE A   1      25.412  19.595  12.578  1.00 12.96     0.287 C
ENDROOT
BRANCH   1   2
ATOM      2  N   PHE A   1      25.225  18.394  13.381  1.00 13.04    -0.065 N
ATOM      3  HN3 PHE A   1      25.856  17.643  13.100  1.00  0.00     0.275 HD
ATOM      4  HN2 PHE A   1      25.558  18.517  14.337  1.00  0.00     0.275 HD
ATOM      5  HN1 PHE A   1      24.247  18.105  13.350  1.00  0.00     0.275 HD
ENDBRANCH   1   2
BRANCH   1   6
ATOM      6  CB  PHE A   1      26.873  20.027  12.625  1.00 12.45     0.082 C
BRANCH   6   7
ATOM      7  CG  PHE A   1      27.286  20.629  13.923  1.00 12.96    -0.056 A
ATOM      8  CD2 PHE A   1      27.470  22.001  14.050  1.00 12.47     0.007 A
ATOM      9  CE2 PHE A   1      27.877  22.571  15.265  1.00 13.98     0.001 A
ATOM     10  CZ  PHE A   1      28.108  21.754  16.360  1.00 13.84     0.000 A
ATOM     11  CE1 PHE A   1      27.919  20.380  16.242  1.00 13.77     0.001 A
ATOM     12  CD1 PHE A   1      27.525  19.821  15.027  1.00 11.32     0.007 A
ENDBRANCH   6   7
ENDBRANCH   1   6
BRANCH   1  13
ATOM     13  C   PHE A   1      25.015  19.417  11.141  1.00 13.31     0.204 C
ATOM     14  O2  PHE A   1      24.659  20.534  10.507  1.00 12.12    -0.646 OA
ATOM     15  O1  PHE A   1      25.024  18.283  10.608  1.00 13.49    -0.646 OA
ENDBRANCH   1  13
TORSDOF 4
```



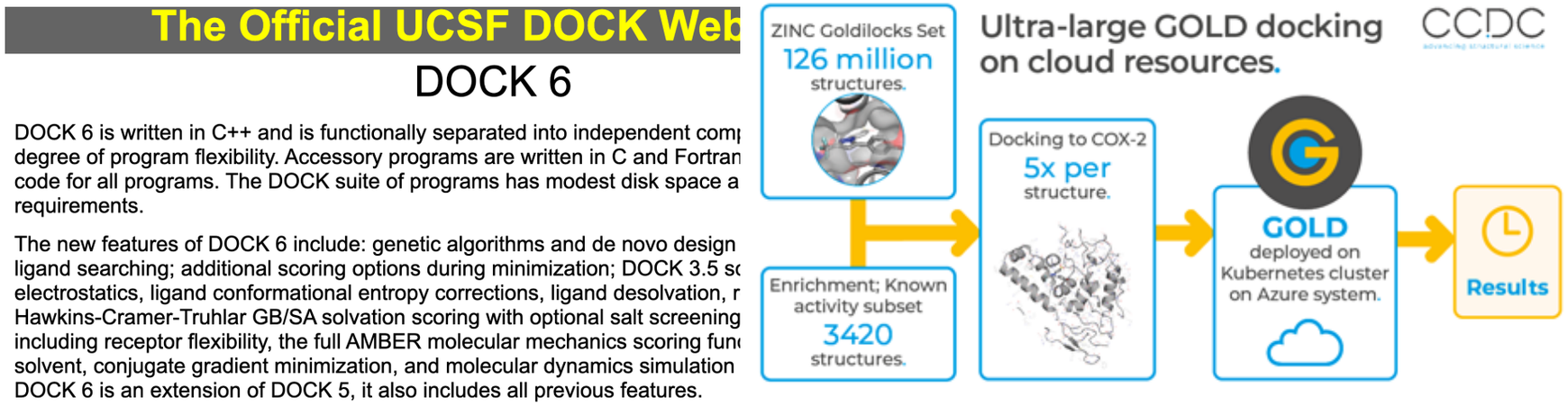
Lamarckian genetic algorithm

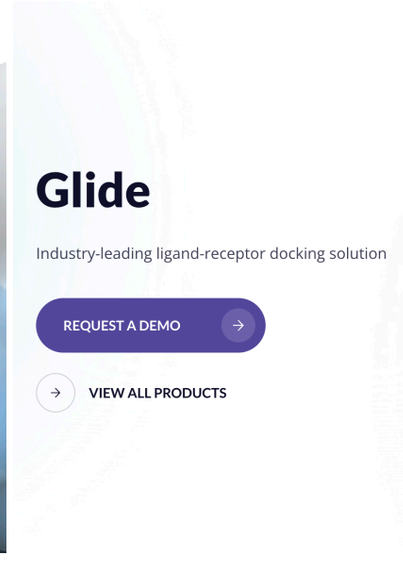


- A molecular conformation is represented as a set of rotatable torsion angles
- Torsion angles are mixed to generate new conformations
- After generating new conformations, local energy minimizations are performed

Famous conventional docking programs

- GOLD
- Glide
- UCSF Dock
- Flare
- More…





Benchmarking docking performance

- PDBbind is the most widely used data and benchmark set in protein-ligand docking

- Docking performances are generally tested with three categories

  - Scoring power

  - Docking power

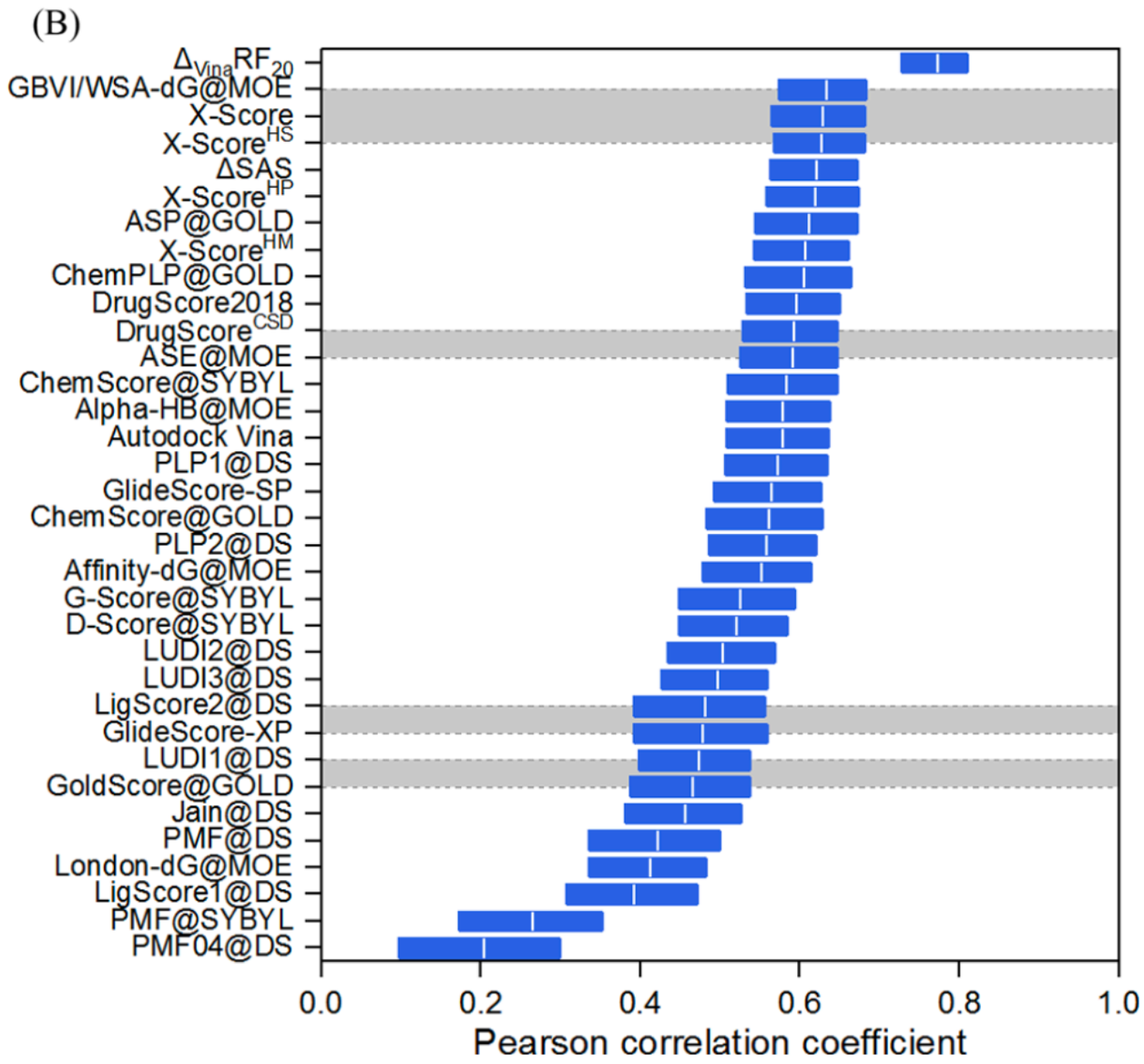
  - Screening power

http://www.pdbbind.org.cn/index.php

Scoring power

- How much experimental and predicted binding affinities are correlated?

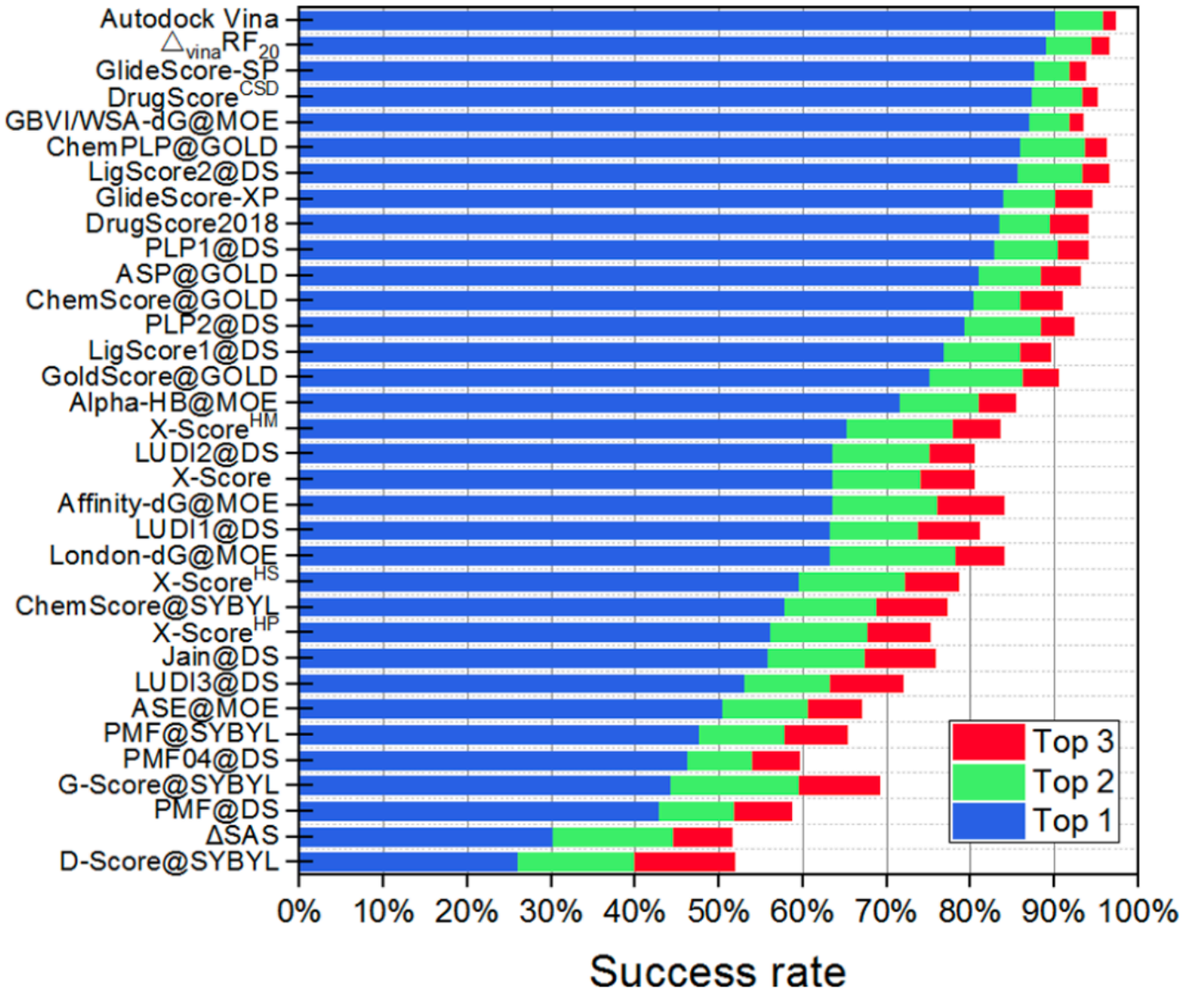
Re-docking to crystal structure



Journal of Chemical Information and Modeling 2019 59 (2), 895-913

Docking power

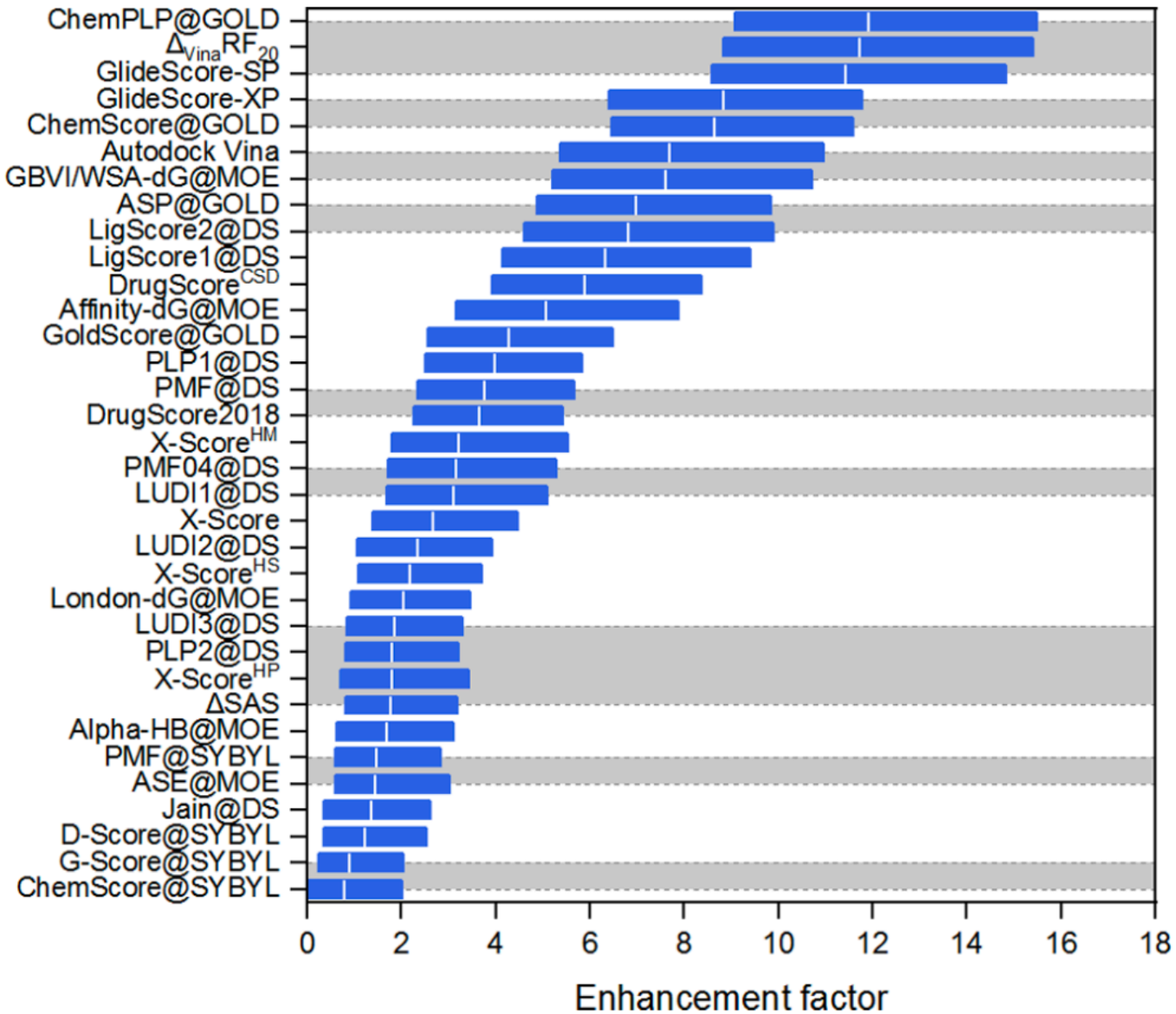
- How accurately native poses are predicted (RMSD < 2Å)?



Journal of Chemical Information and Modeling 2019 59 (2), 895-913

Screening power

- How accurately a scoring function can discriminate true and false binders?
- Enrichment factor (EF)
- $$EF_{\alpha\%} = \frac{N_{\alpha\%}^{\text{true}}}{\alpha\,\%\,N_{total}^{\text{true}}}$$

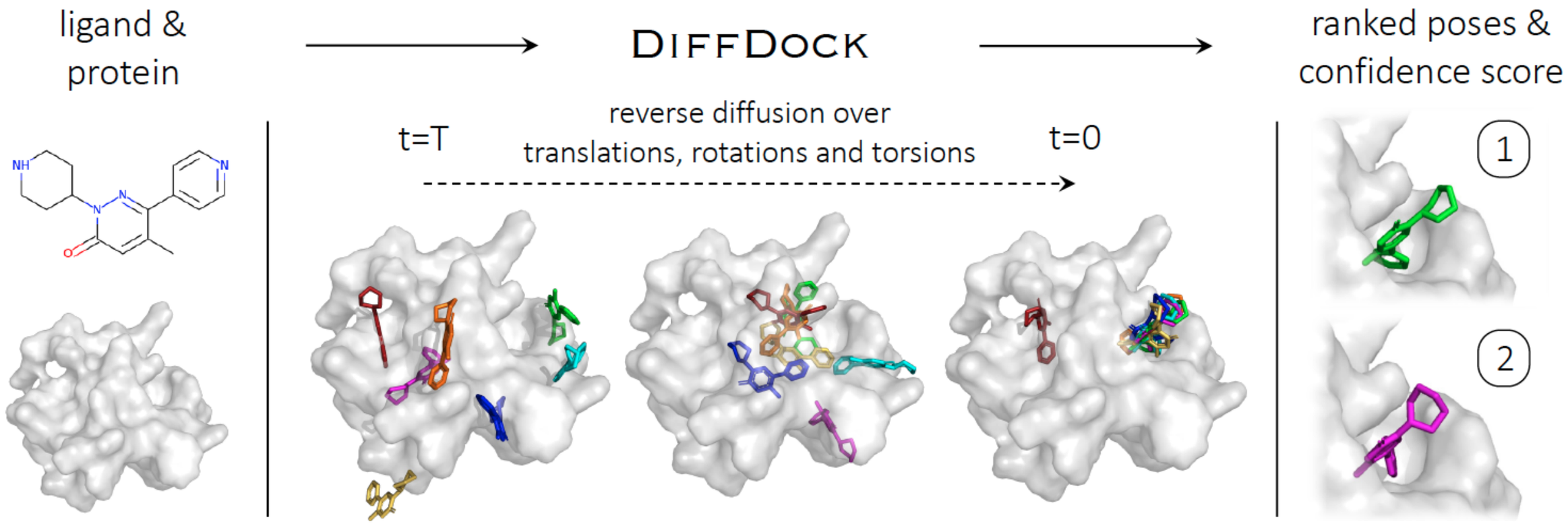


Limitations of docking programs

- **High false positive rates**
  - Docking programs bind true and false binders to proteins
  - In reality, false binders should not bind to proteins
- Protein flexibility is still hard to be considered
  - Accuracy of flexible docking is still limited
- Predicting accurate binding affinity is still limited
  - 1.3 kcal/mol error ~ 100 fold difference in Kd or IC50 values

Deep-learning assisted docking

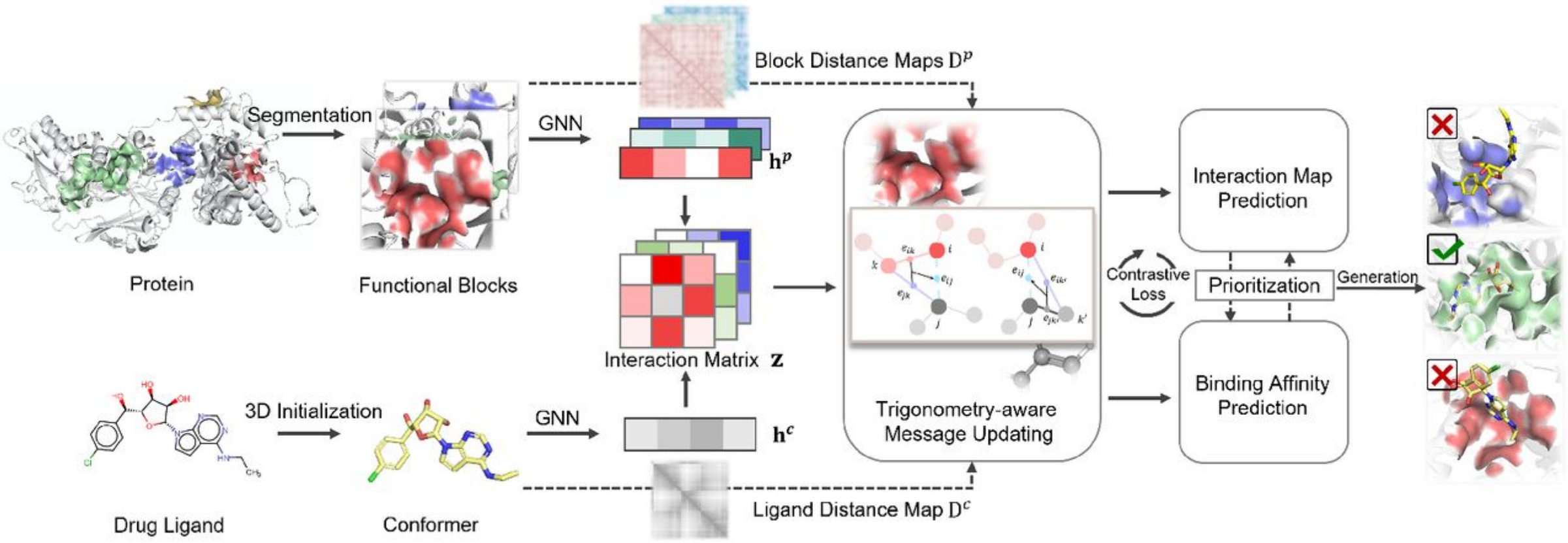
Diffusion-based ligand pose prediction



https://doi.org/10.48550/arXiv.2210.01776

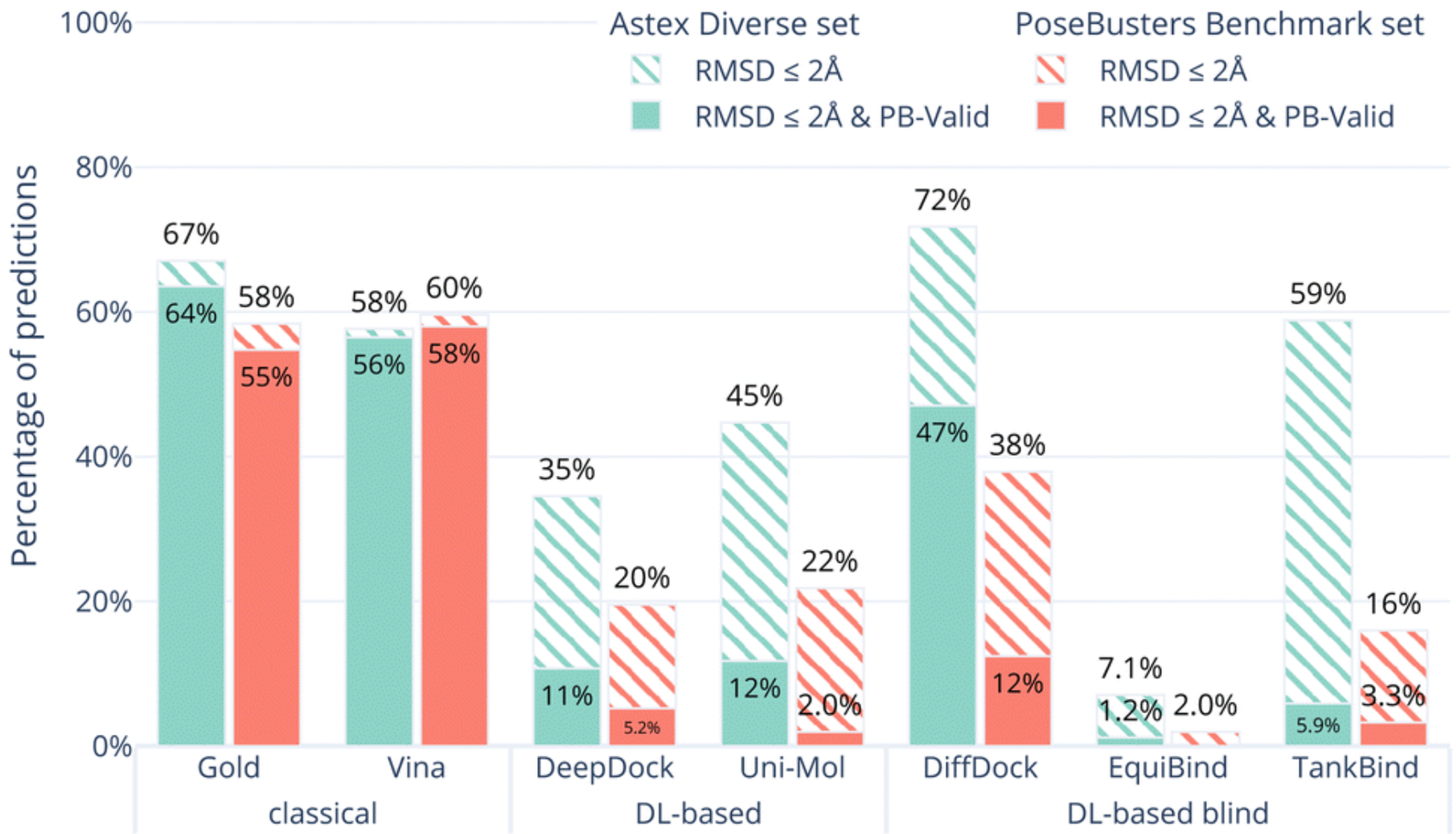
- Deep-learning-based protein-ligand pose prediction
- Advantage: faster by avoiding testing multiple conformations
- Disadvantage: unphysical chemical conformations

Another diffusion-based model: TankBind



- Accept protein and ligand shapes and merge them via interaction matrix

How accurate docking programs in pose prediction?



- Benchmark results using the PoseBuster benchmark set.
- Deep-learning-based methods are not necessarily better than traditional models

AI/DL assisted scoring functions

- To enhance the accuracy of scoring and screening, many AI/DL based scoring functions have been developed

- Mostly, re-scoring a given protein-ligand complex pose

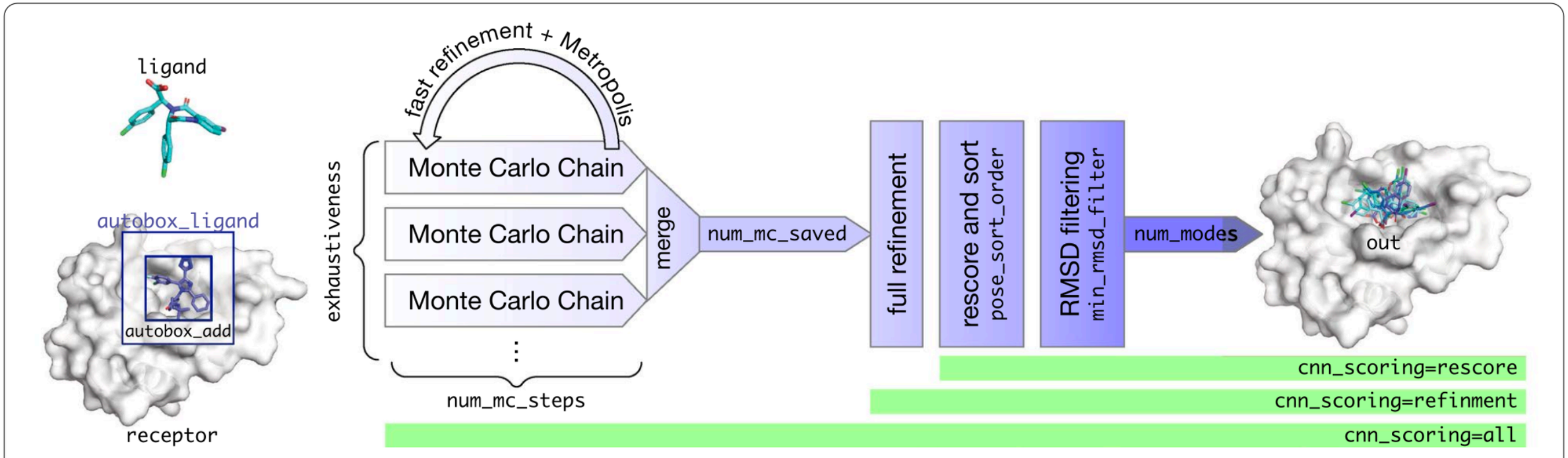
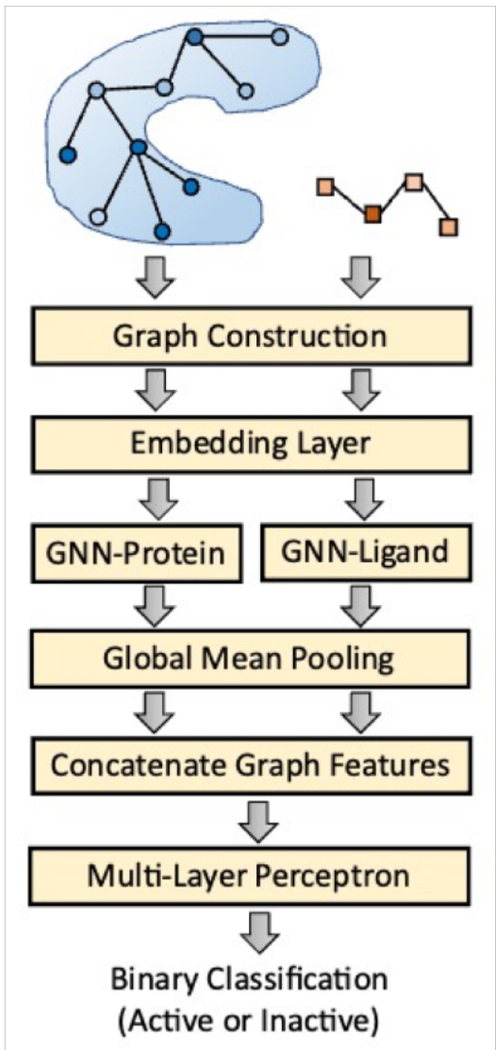
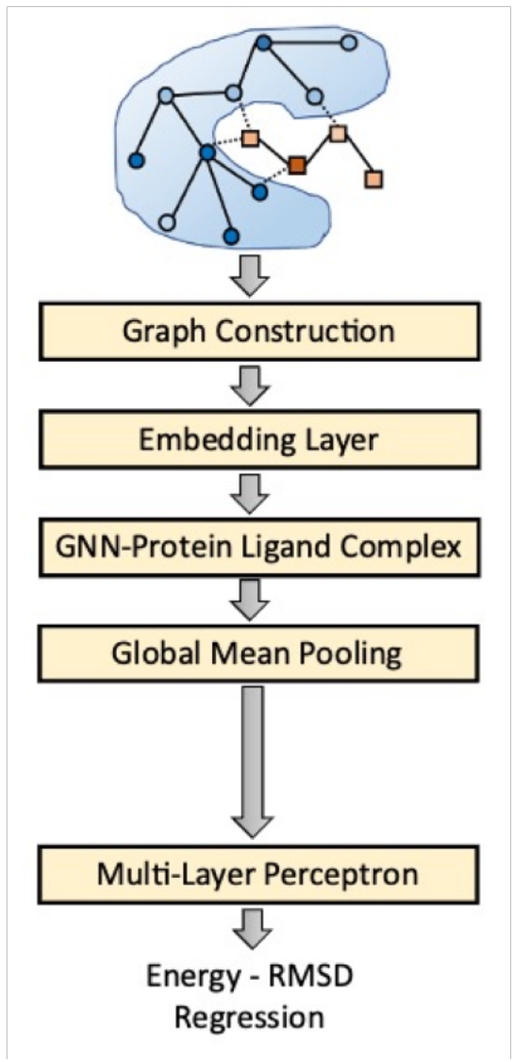


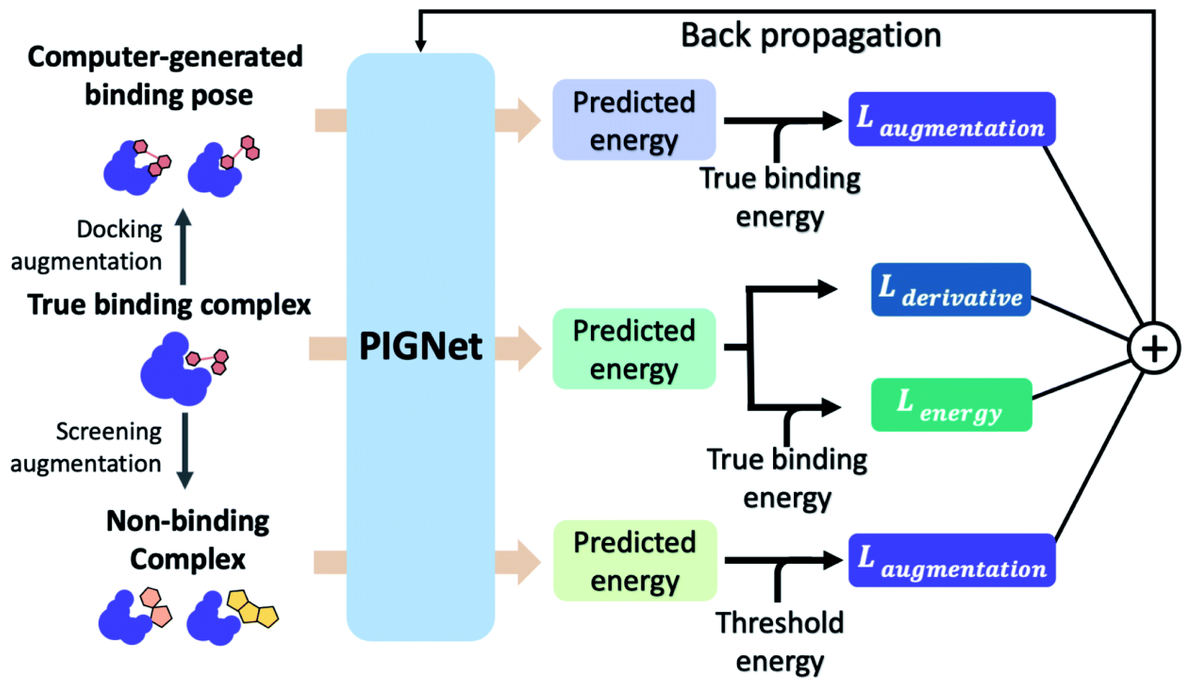
Fig. 1 The GNINA sampling and scoring algorithm shown with relevant `commandline parameters` and the scope of CNN scoring

GNINA, *J Cheminform* 13, 43 (2021)

Deep-learning-based rescoring models





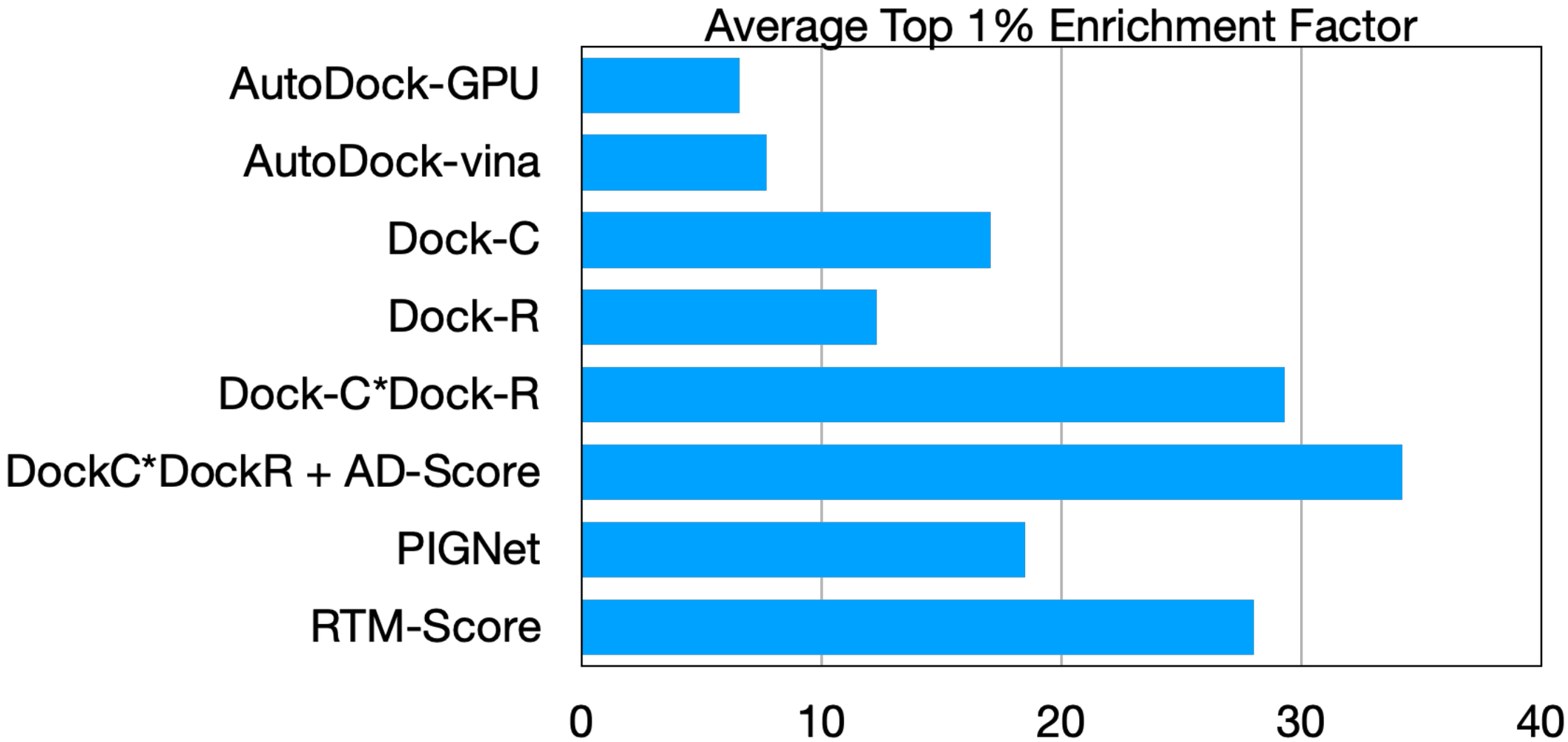


Chem. Sci., 2022, **13**, 3661-3673

AK-score, manuscript in revision

PigNet, developed by Prof. W. Kim's group at KAIST

How accurate are docking programs? in virtual screening



- Deep-learning assisted rescoring is actually improving protein-ligand docking quality

- Average EFs using CASF-2016 set is 10~30

Conclusion

- Protein-ligand docking problem has three challenges
  - Global combinatorial optimization problem
  - Correct docking pose prediction
  - Correct scoring prediction
  - Only native ligands with correct poses should have high affinity
  - Native ligand with wrong poses or non-binders should have low affinity