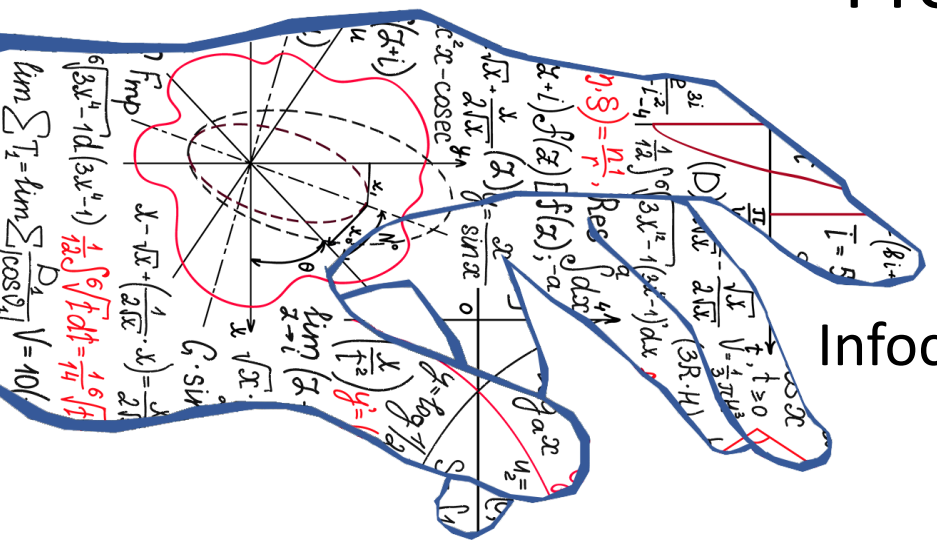
ITMO UNIVERSITY

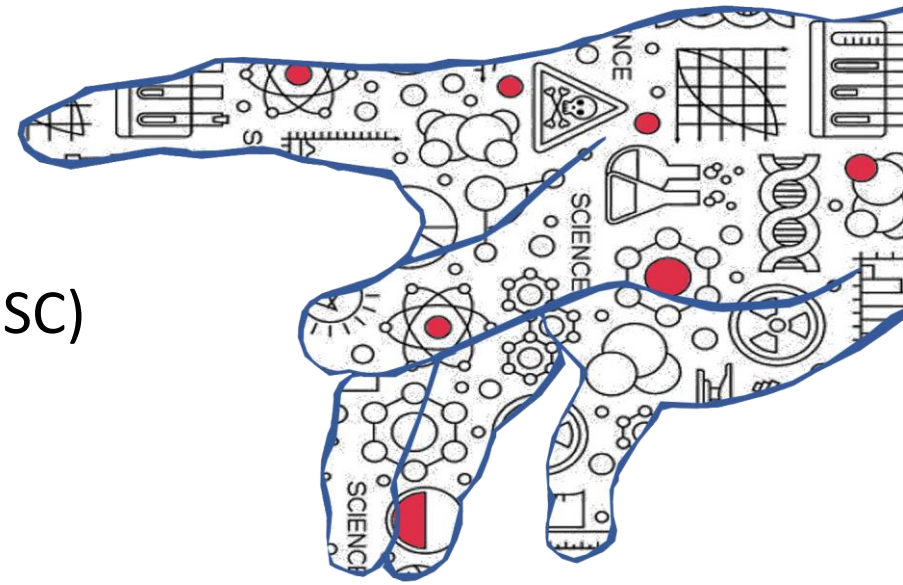
INFOCHEMISTRY SCIENTIFIC CENTER

Cheminformatics: Molecular Docking

Prof. Dr. Sergey Shityakov



Infochemistry Scientific Center (ISC)
ITMO University
Saint-Petersburg, 2021

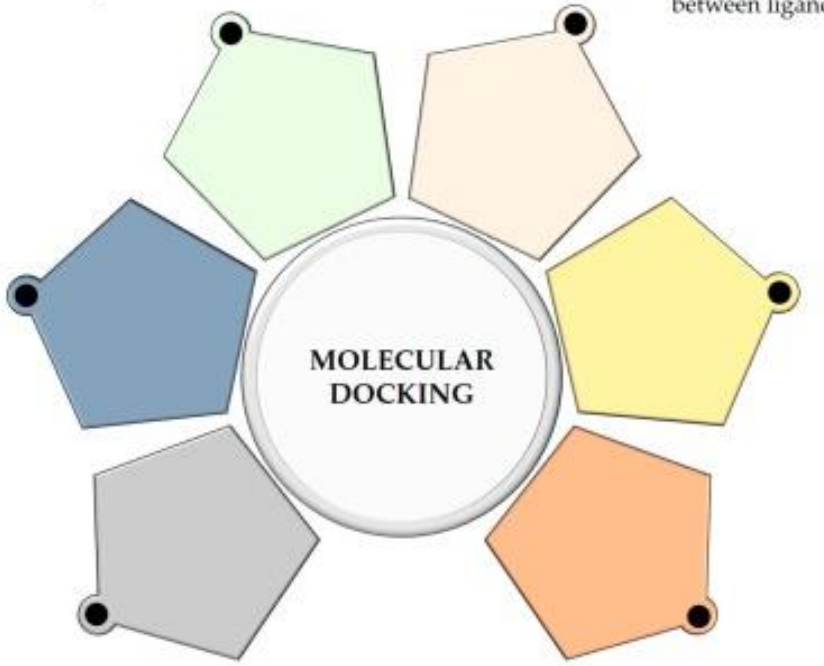


Molecular docking

Target fishing and profiling
Prediction of targets for compounds on the basis of ligand-receptor complementarity

Prediction of Adverse Drug Reactions
Prediction and rationalization of drug off-target activities based on the complementarity between ligands and targets

Virtual Screening
Identification of compounds modulating disease-related targets and their optimization



Polypharmacology
Identification and optimization of compounds that simultaneously modulate a set of targets involved in the same disease

Ligand-Target binding rationalization
Identification of the structural determinants necessary for the efficient ligand-receptor binding

Drug Repositioning
Identification of novel therapeutic-relevant targets for already marketed drugs, and known chemical and natural entities

Molecular docking

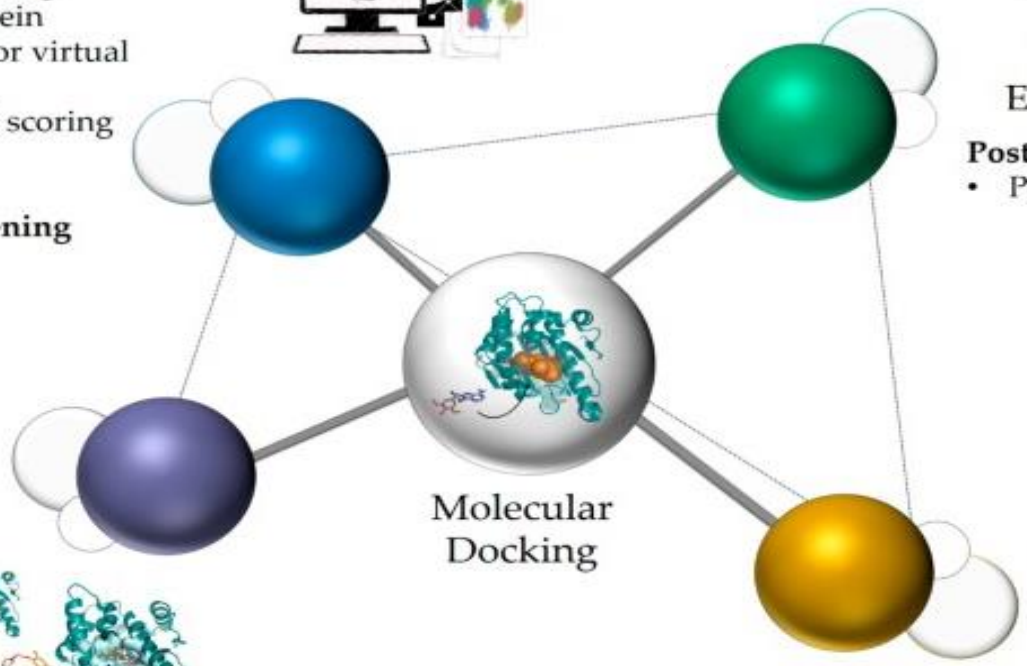
Artificial intelligence (AI) and statistical methods

Pre-docking screening
- Selection of protein conformations for virtual screening
- Improvement of scoring functions

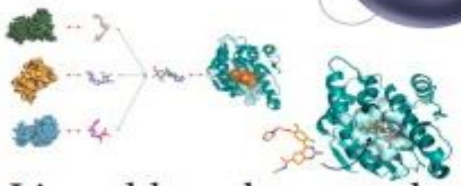
Post-docking screening
- Pose rescoring

Binding Free Energy methods

Post-docking screening
- Pose rescoring



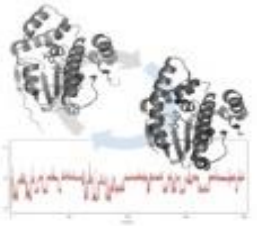
Molecular Docking



Ligand-based approaches

Pre-docking screening
- Selection of protein conformations for virtual screening

Post-docking screening
- Pose selection
- Pose rescoring



Molecular Dynamics (MD)

Pre-docking screening
- Identification of representative conformations for virtual screening

Post-docking screening
- Pose refinement
- Ligand-target complex stability assessment

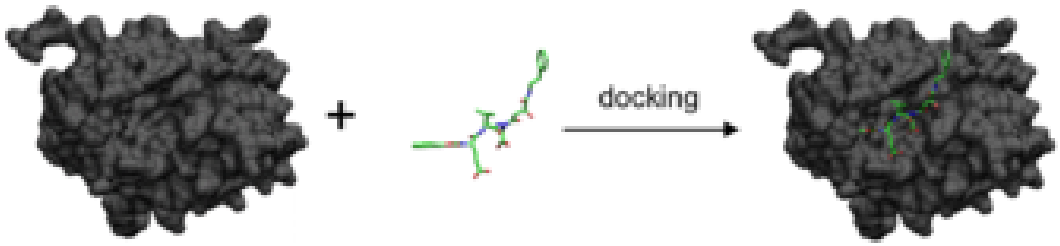
Molecular docking

- Molecular docking technique is a computational determination of binding affinity between the protein (receptor) and drug-like molecule (ligand).

- Molecular docking attempts to find the “best” matching between two molecules based on the appropriate scoring function.

- The molecular docking methods usually use a rigid receptor molecule and flexible ligand.

- Various genetic algorithms are widely implemented in molecular docking methods. These algorithms describe ligand conformational changes by mimicking the process of natural evolution (mutation, selection, and crossover).

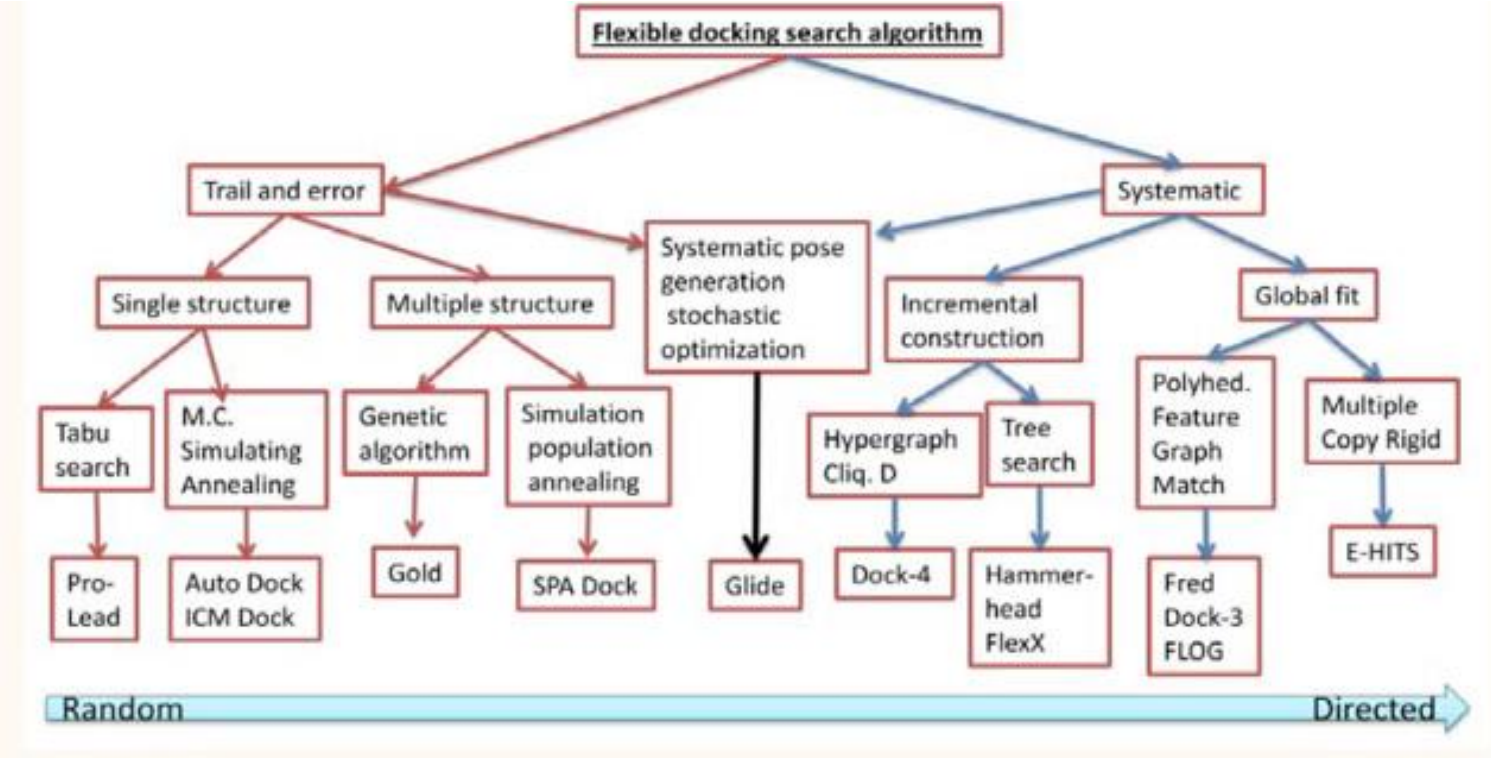


Molecular docking software

- AutoDock and AutoDockVina (USA, Scripps Research Institute and Olson Laboratory, 1990, 2010)
- GOLD (UK, Center for Research Computing, 1995)
- Glide (USA, Schrödinger, Inc., 2004)
- DOCK (USA, University of California and San Francisco (UCSF), 1998)
- PLANTS (Germany, University of Konstanz, 2004)
- GEMDOCK (Taiwan, National Chiao Tung University 2004)
- HEX (UK, University of Glasgow, 2008)

Algorithms

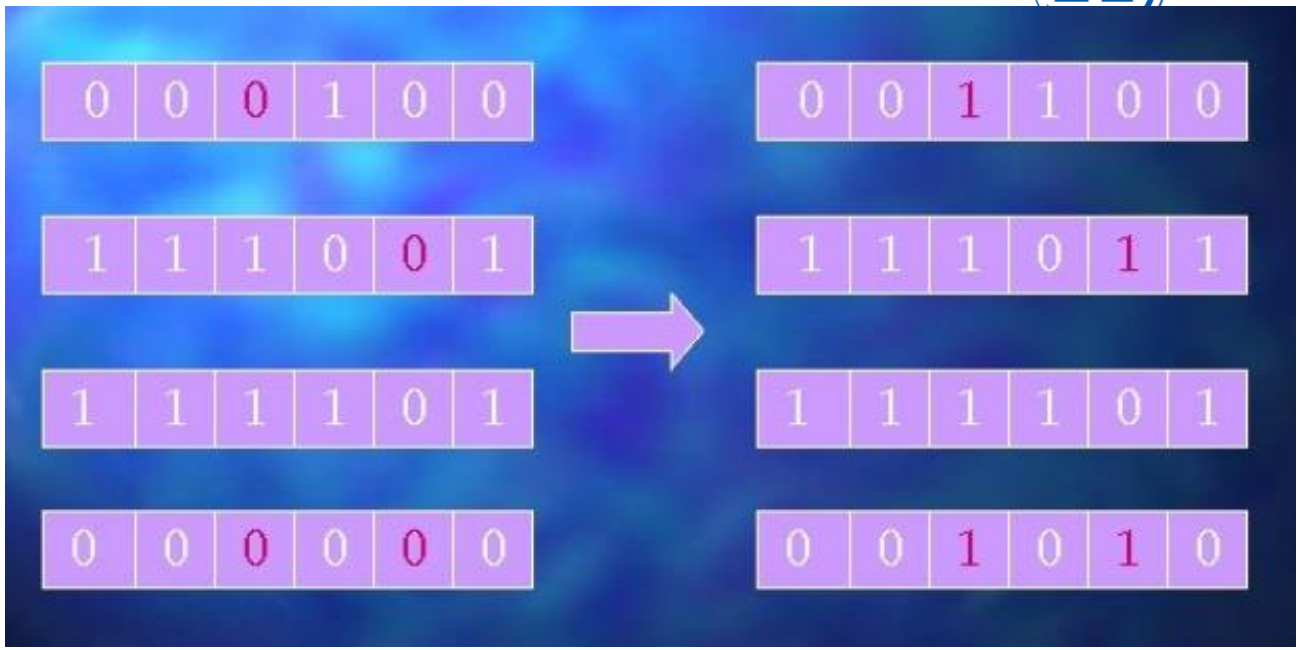
- Genetic algorithm
- Monte-Carlo algorithm
- Simulated annealing
- Induced Fit
- Exhaustive search
- Voronoi tessellation
- Hidden Markov Models (?)



Genetic algorithm

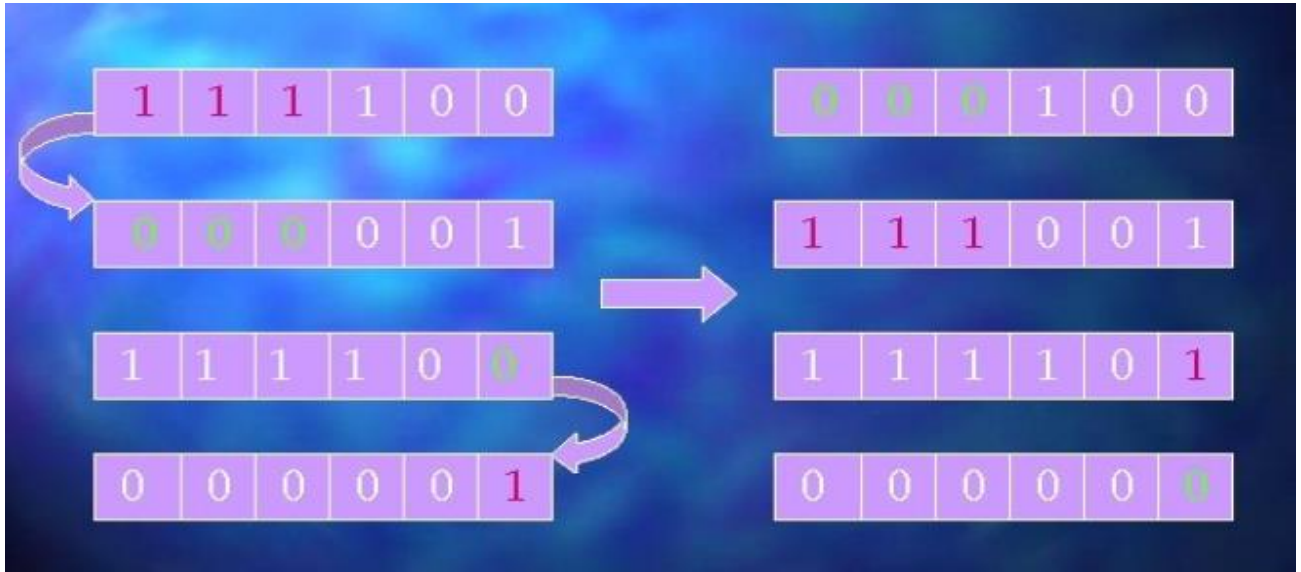
- Initial population of binary creatures having 6 "genes"
- Each gene has two different alleles, either a 0 or a 1
- Three operators: crossover, mutation and selection

1	1	1	1	0	0	
0	0	0	0	0	1	
1	0	0	0	0	1	
0	0	0	0	0	0	



- Selection based on a fitness function f(x)
- This operator chooses those individuals with the lowest values
- Those with higher values chosen with a very low probability

						Score
1	1	1	1	0	0	20
0	0	0	0	0	1	13
1	0	0	0	0	1	48
0	0	0	0	0	0	52



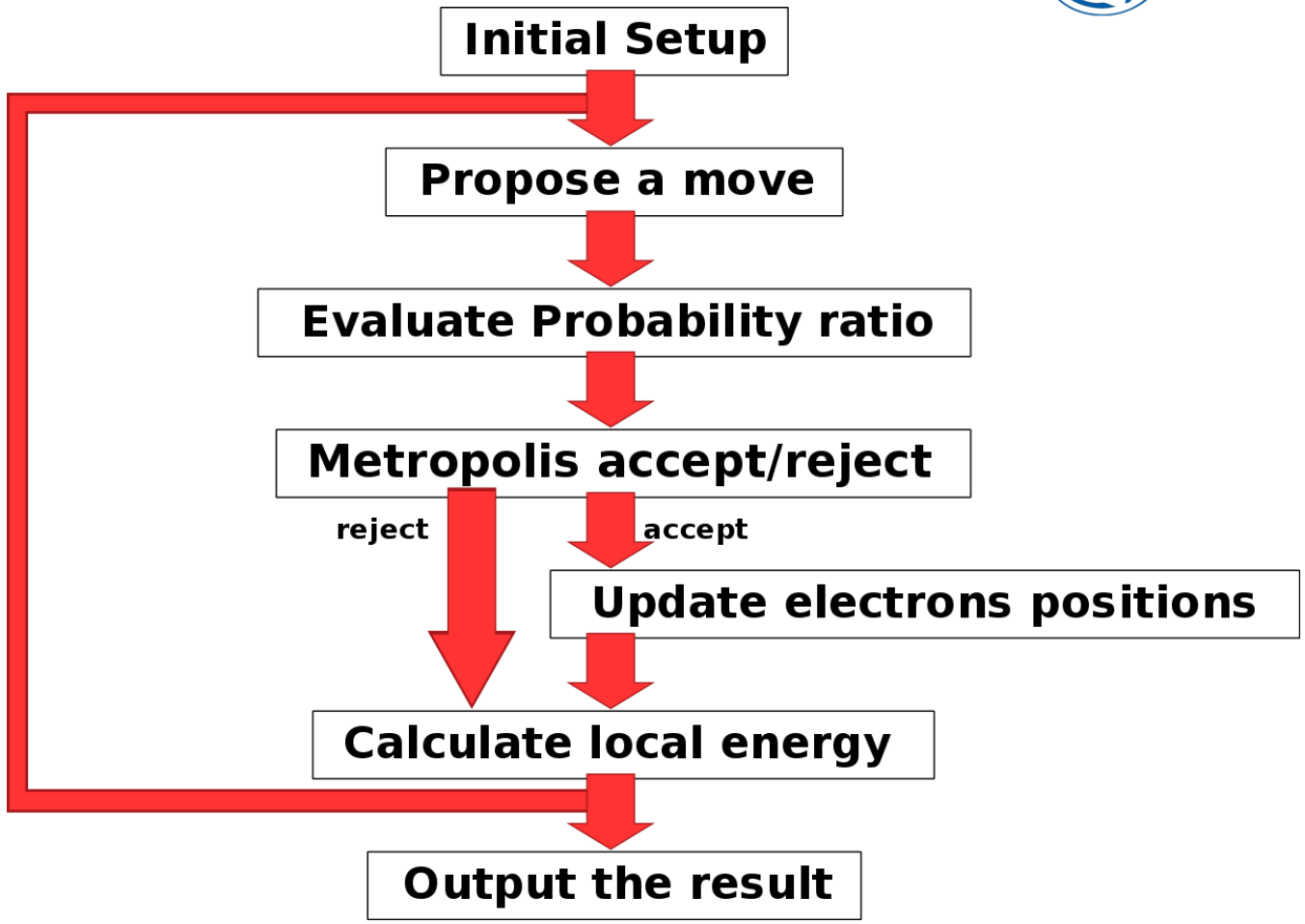
Monte-Carlo

- Start with configuration A (energy E_A)
- Make random move to configuration B (energy E_B)
- Accept move when:

  $E_B < E_A$ or if

  $E_B > E_A$ except with probability P:

$$P = \exp(-[E_A - E_B]/kT)$$



$$E[f(X)] \approx \frac{1}{N}\sum_{i=1}^{N} f(x_i)$$

Where:

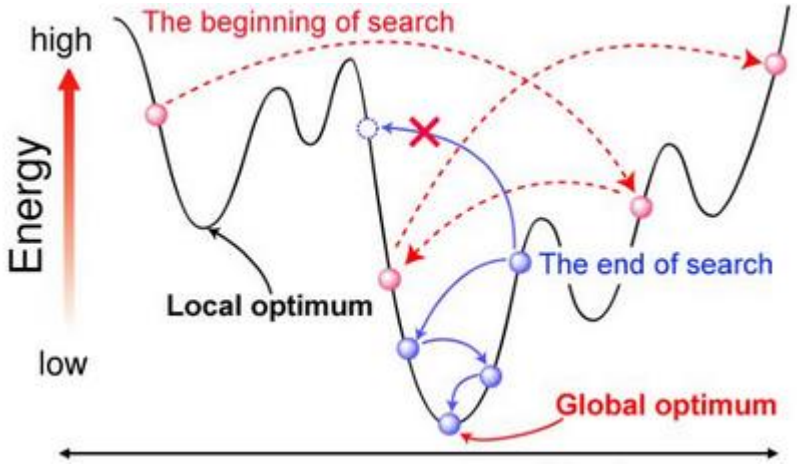
- $E[f(X)]$ is the expected value of the function $f(X)$.
- N is the number of samples drawn from the probability distribution of X.
- x_i are random samples drawn from the probability distribution of X.

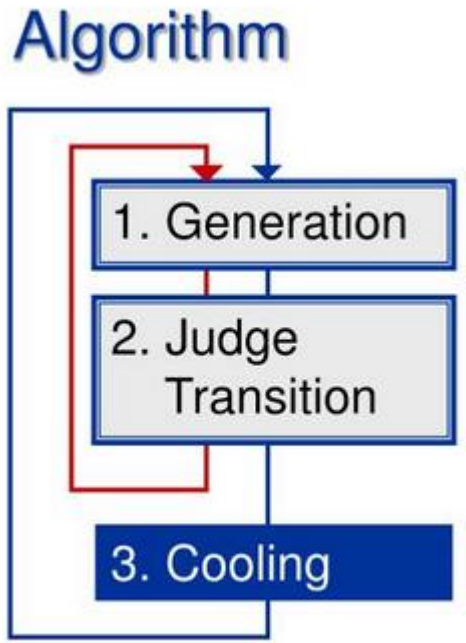
Algorithms

- Start a random run (position , orientation, conformation)
- Make random state changes, accepting up-hill moves with probability dictated by “temperature”
- Reduce temperature after each move
- Terminate when temperature gets very small

$$\text{acceptance} \quad \text{Exp}\left(\frac{-\Delta E}{\text{Temperature}}\right)$$

$$(\Delta E = E_{next} - E_{now})$$





Scoring functions

4 classes of scoring functions:

1) **Force field-based function** (FF). FF estimates the affinities by summing intermolecular vdW and electrostatic interactions of the receptor and ligand.

2) **Knowledge-based function** as statistical potentials. They based on statistical observations of intermolecular close contacts in large 3D databases.

3) **Empirical function.** These functions are based on counting the number of various types of interactions between the docked molecules, such as SASA, hydrophobic/hydrophilic contacts, h-bonds, etc.

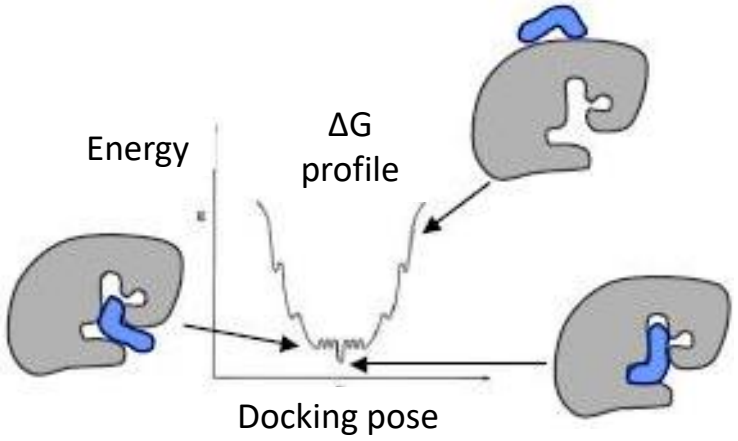
4) **Hybrid functions** use combination of functions (FF-based + empirical and so on). They are usually represented as Gibbs free energy of binding (ΔG) to estimate binding affinity (Ki).

$$\Delta G = RT(\ln Ki)$$

R – gas constant

T – room temperature

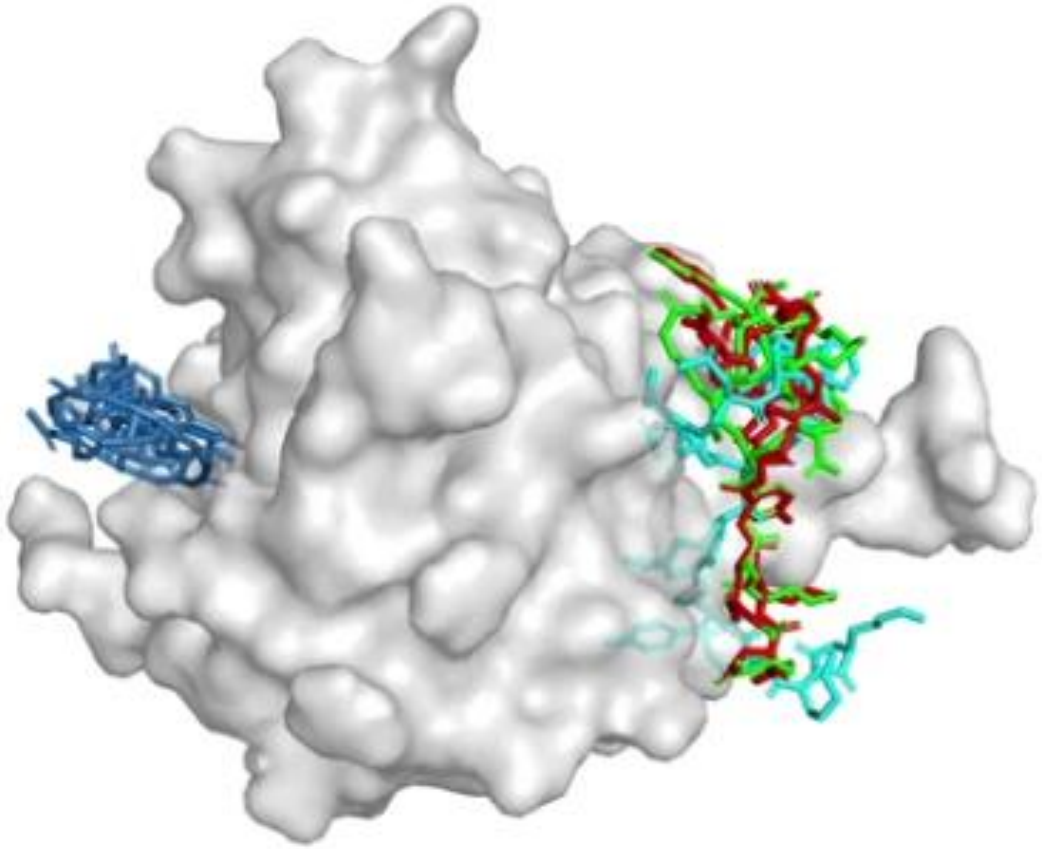
Ki – inhibition constant



DiffDock

Regression vs Generation for Docking

Epistemic uncertainty causes distortions and steric clashes

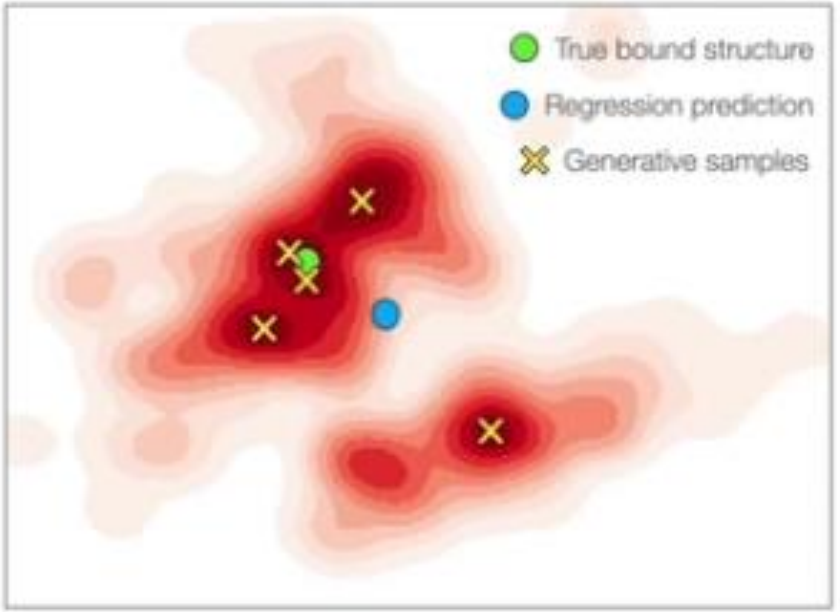


Crystal Structure
EquiBind (regression)
TANKBind (regression)
DiffDock top-1

Regression vs Generation for Docking

Explaining the issue with existing deep learning models

- Docking has significant aleatoric and epistemic uncertainty
- Methods will exhibit uncertainty about correct pose between multiple alternatives (multimodal d.)
- Regression methods to minimize squared error predict (weighted) mean
- Generative model will populate all/most modes and we'll then predict the MAP among these



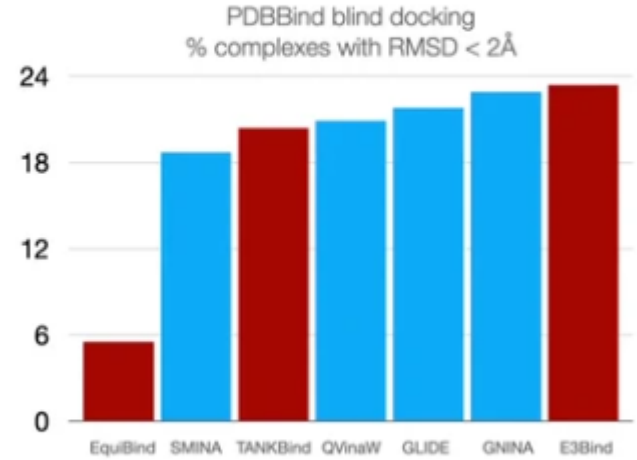
Model's posterior over the space of docked structures

DiffDock

Existing Deep Learning methods

Fail to outperform search-based methods

- EquiBind and TANKBind are regression-based deep learning methods
- Significant speedup over search-based methods
- No improvements in accuracy
- Recently, also E3Bind, closely following AlphaFold's framework did not show any boost



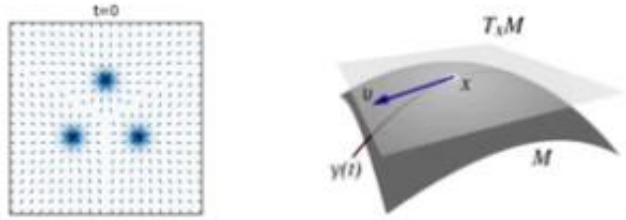
EquiBind Stärk et al. 2022

TANKBind Lu et al. 2022

E3Bind Zhang et al. 2022

Product Space Diffusion

Diffusion generative modeling works on manifolds [de Bortoli et al, '22]
…provided the score model predicts in the **tangent space**



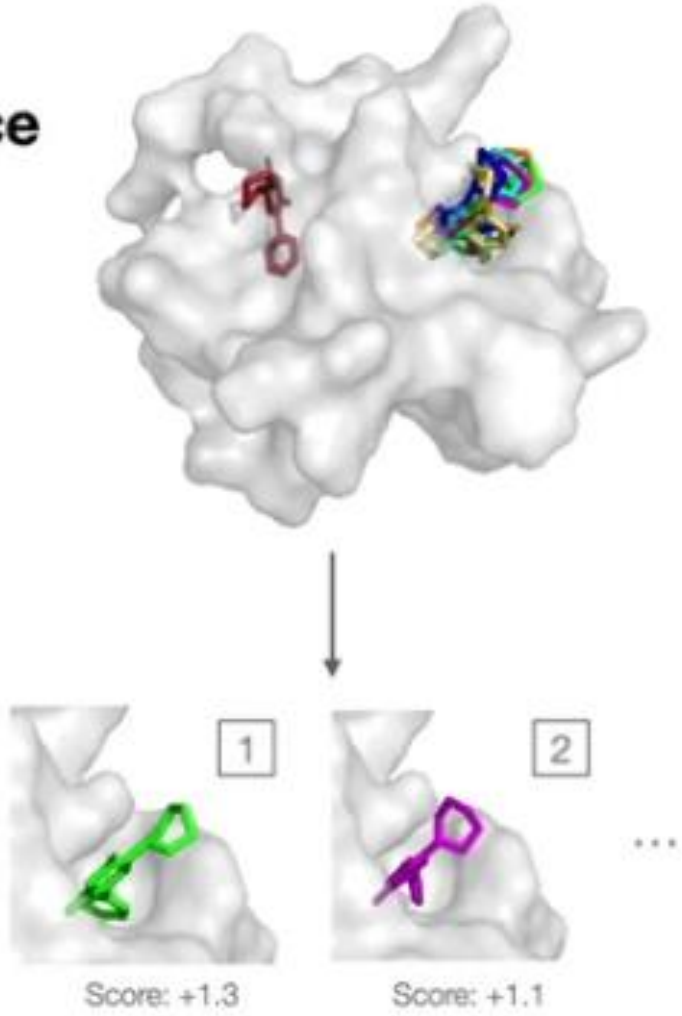
…and that we can:

1. sample the heat kernel for arbitrary t
2. compute its score
3. sample from the stationary distribution (t = T)

Confidence Model

Ranking and evaluating samples confidence

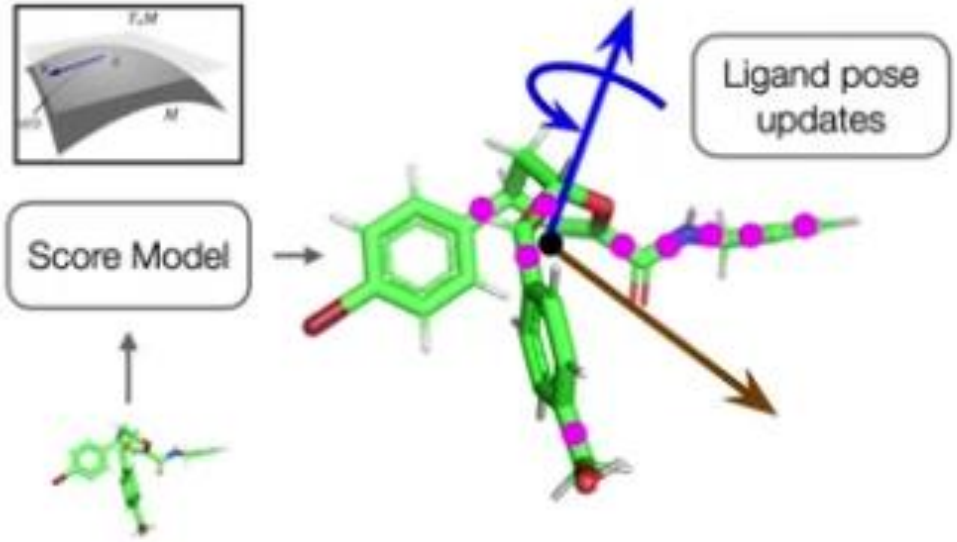
- The generative model can sample an arbitrary number of poses, but researchers are interested in one or a fixed number of them
- Confidence predictions are very useful for downstream tasks
- We train a confidence model to take in the generative model samples and return a score for each sample
- Samples are ranked by score and the score of the best is used as overall confidence score



DiffDock

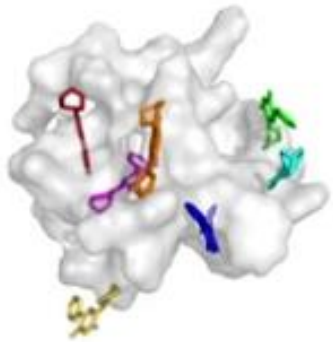
Score Model

Space	$\mathbb{R}^3$ (position)	SO(3) (orientations)	$\mathbb{T}^m$ (torsion angles)
Tangent space	$\mathbb{R}^3$ (translation vec.)	$\mathbb{R}^3$ (rotation vectors)	$\mathbb{R}^m$ (torsion updates)
SE(3) symmetry	Equivariant	Equivariant	Invariant



Workflow

1. Embed with RDKit
2. **Sample N random poses**
3. Simulate reverse diffusion
4. Rank and select top M poses



Space	$\mathbb{R}^3$ (position)	SO(3) (orientations)	$\mathbb{T}^m$ (torsion angles)
Stationary distribution	Normal	Uniform	Uniform

DiffDock

Standard benchmark PDBBind

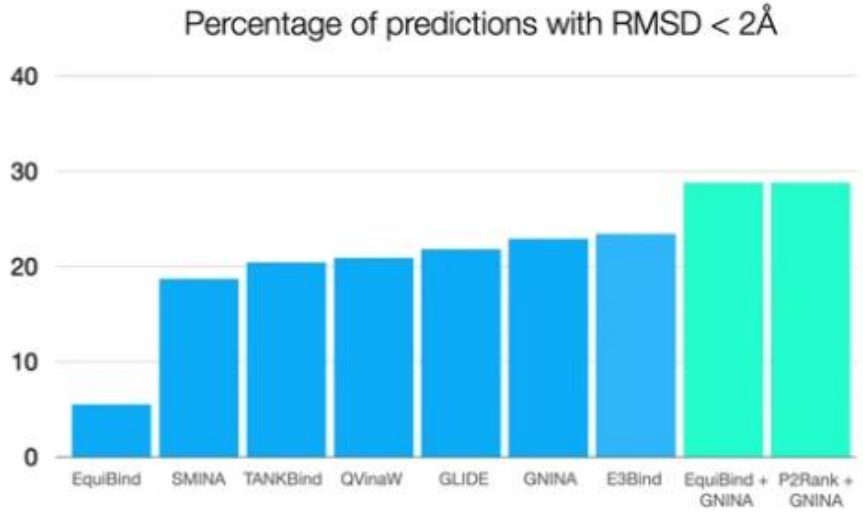
19k experimentally determined structures of small molecules + proteins

Baselines: search-based and deep learning

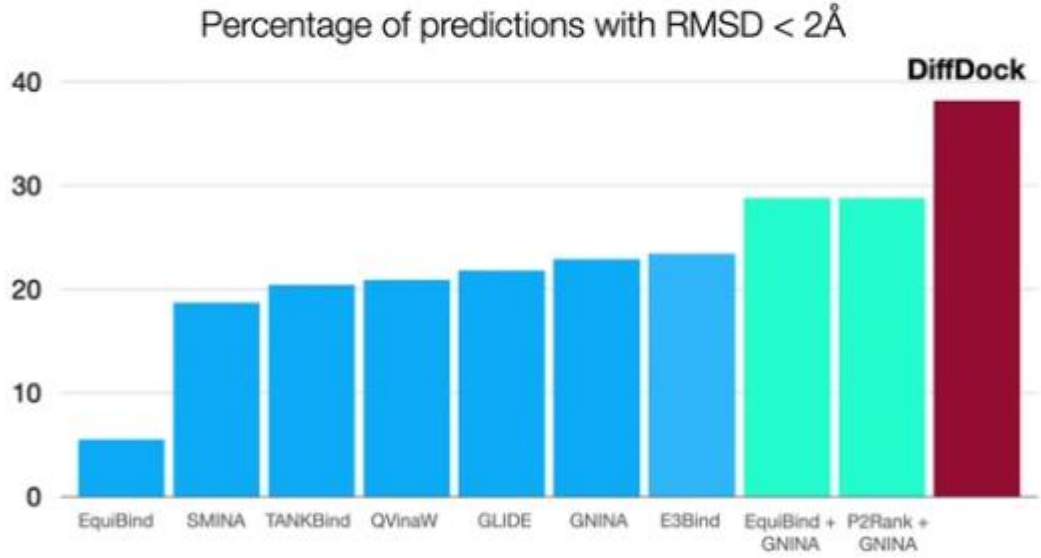
GNINA	SMINA	QuickVina-W	GLIDE	EquiBind	TankBind
McNutt et al. 2021	Koes et al. 2013	Hassan et al. 2017	Schrödinger, Release 2021-4	Stärk et al. 2022	Lu et al. 2022

DiffDock

Prediction correctness

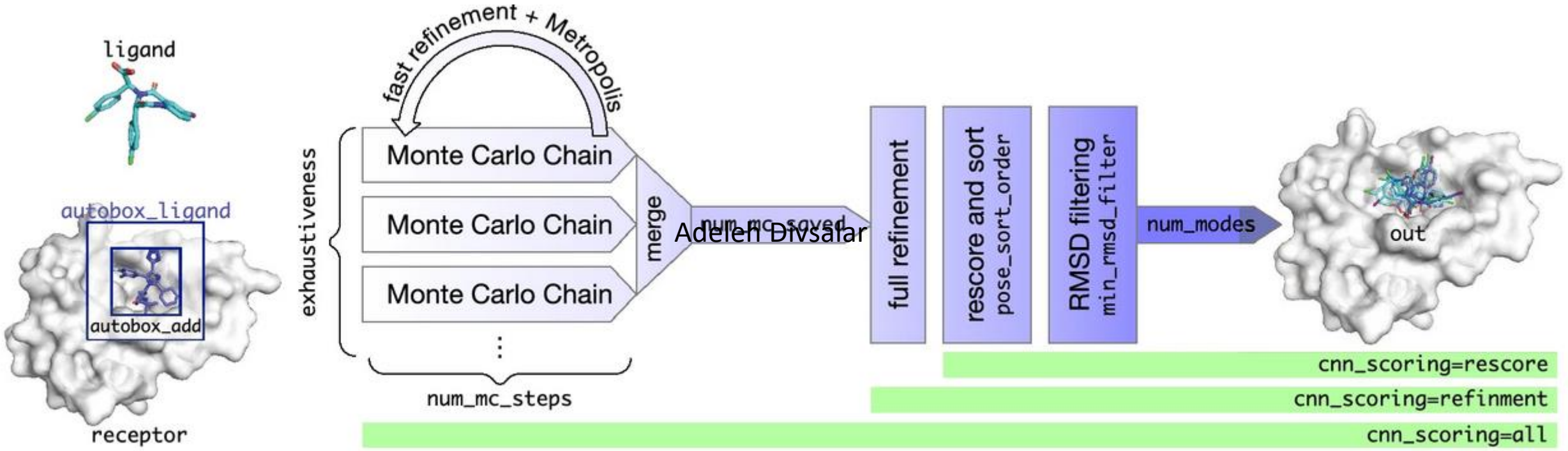


Prediction correctness



Outperform search-based, deep learning, and pocket prediction + search-based methods

GNINA Molecular docking



Adelen Divsalar

Design of drug-like molecules

Molecular docking of HIV protease inhibitors (indinavir)

1) Indinavir molecule is the ligand

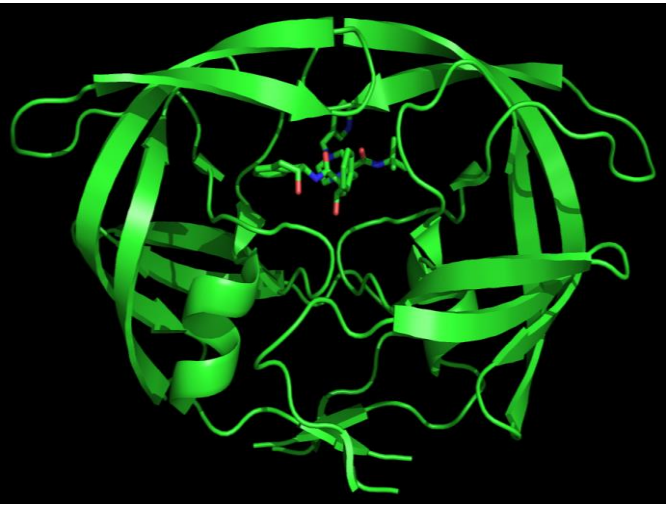
2) HIV-1 holo-structure (ligand-bound state) is the target receptor (X-ray model ID: 1HPV):

a) C2-symmetric (2-fold) homodymeric enzyme

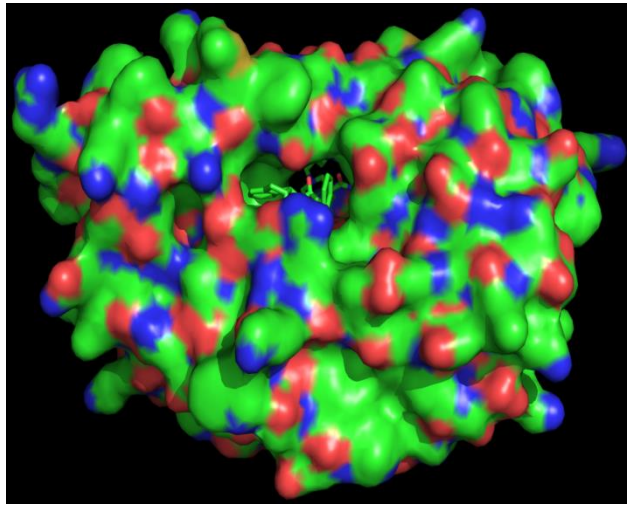
b) Aspartic acid residues are its active site (Asp25, Asp29, Asp30)

c) Catalytic triad (Asp-Thr-Gly sequence)

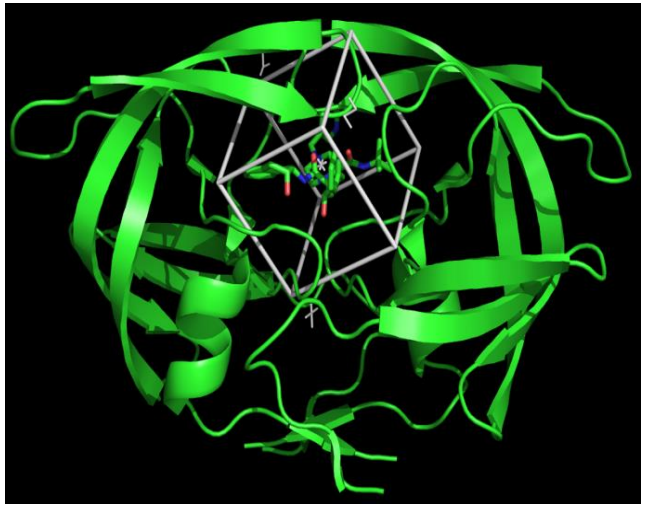
Cartoon



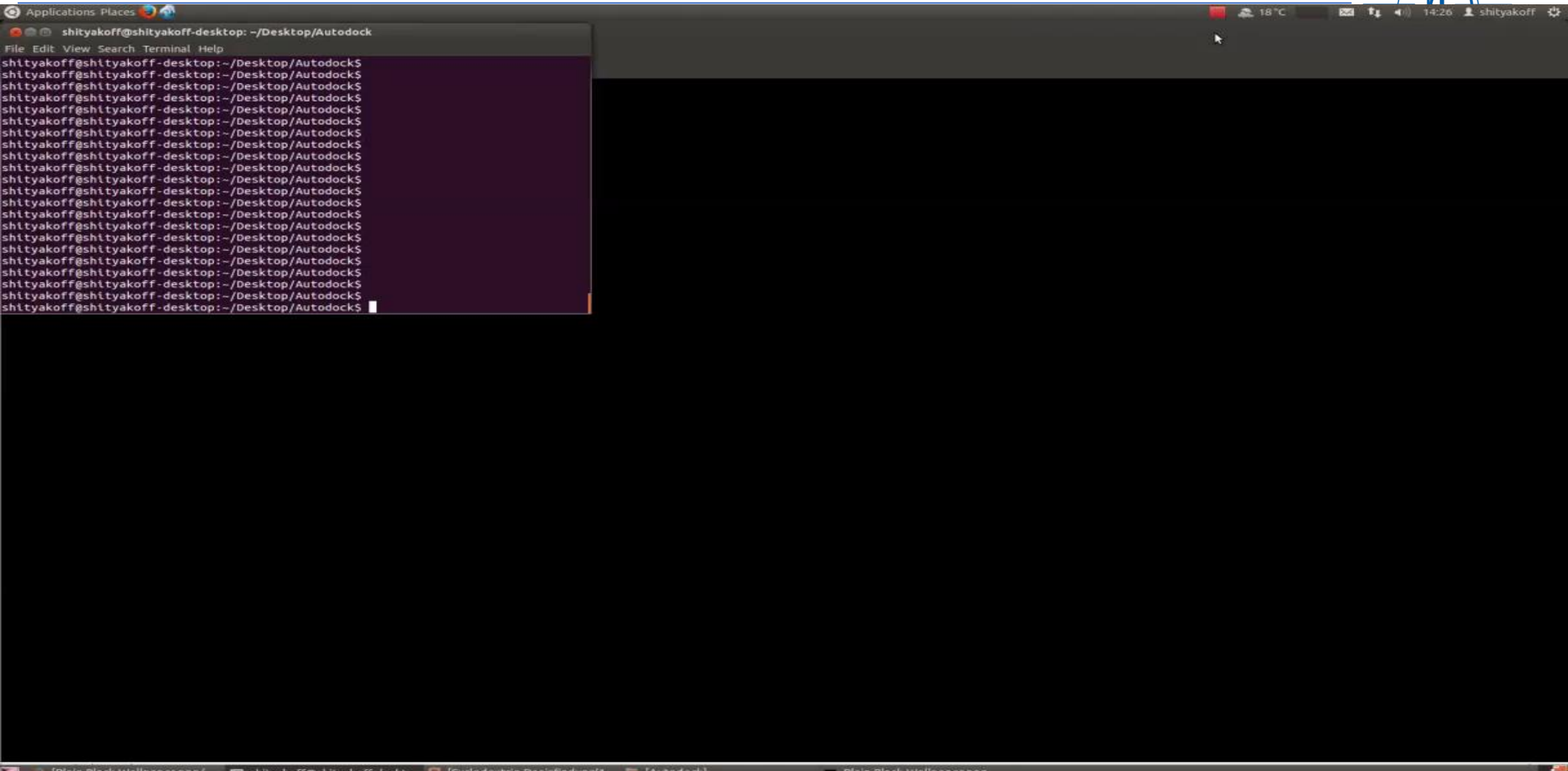
Molecular Surface



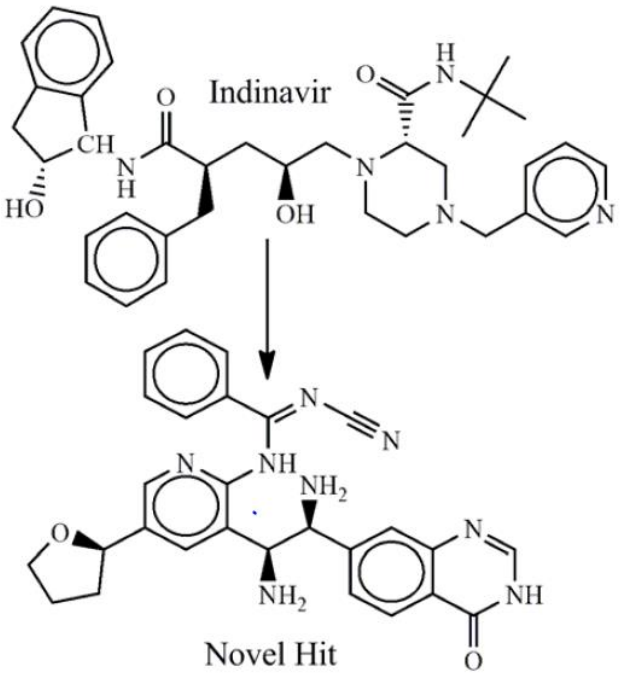
Docking grid

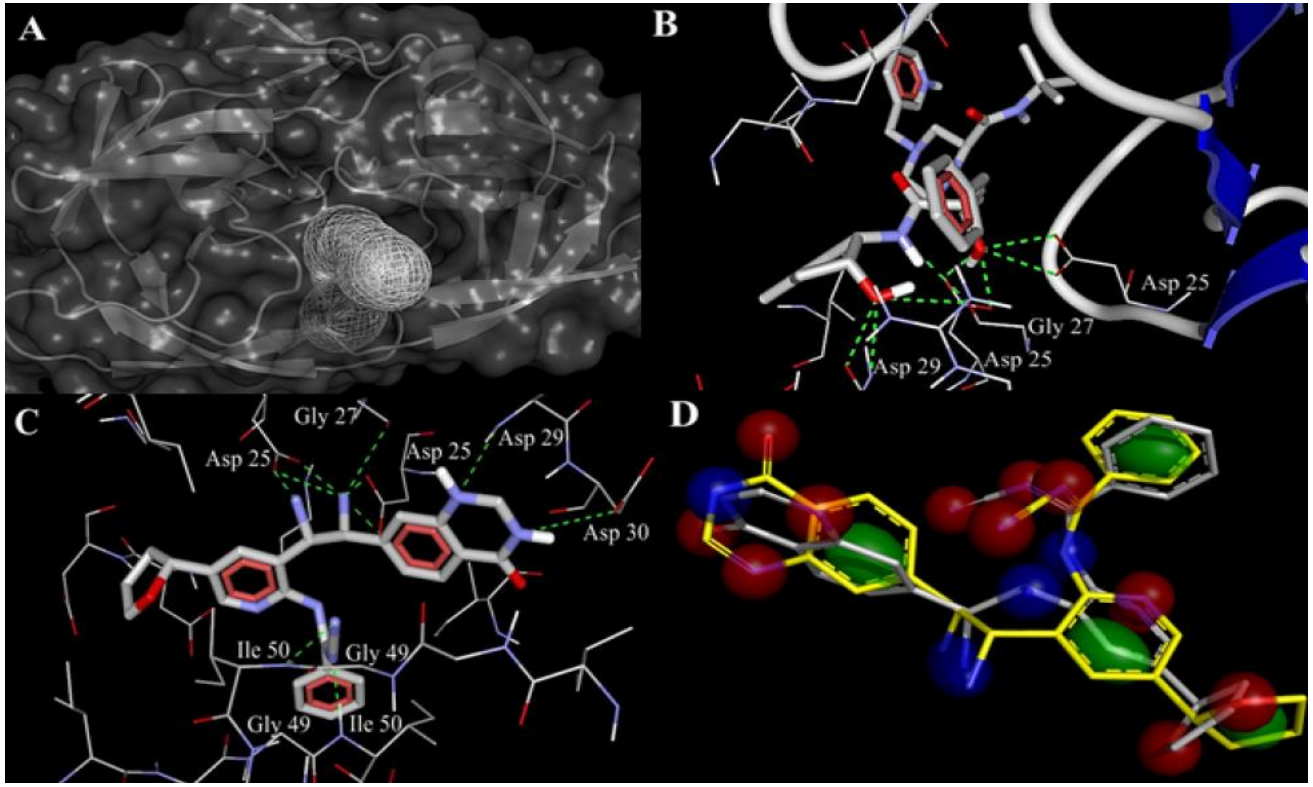


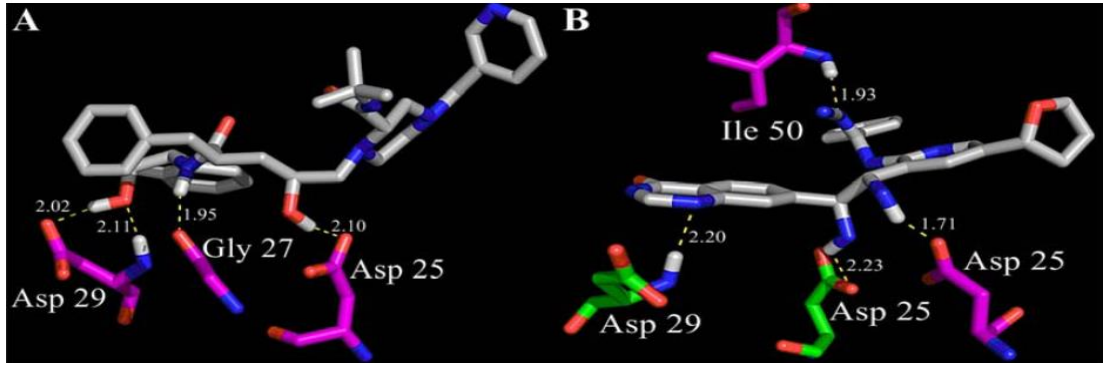
Molecular docking protocol



HIV-1 protease docking results



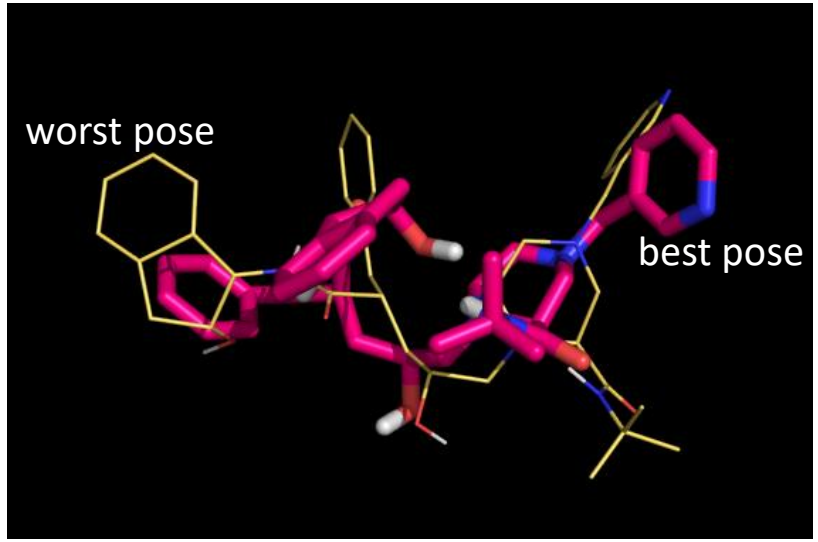


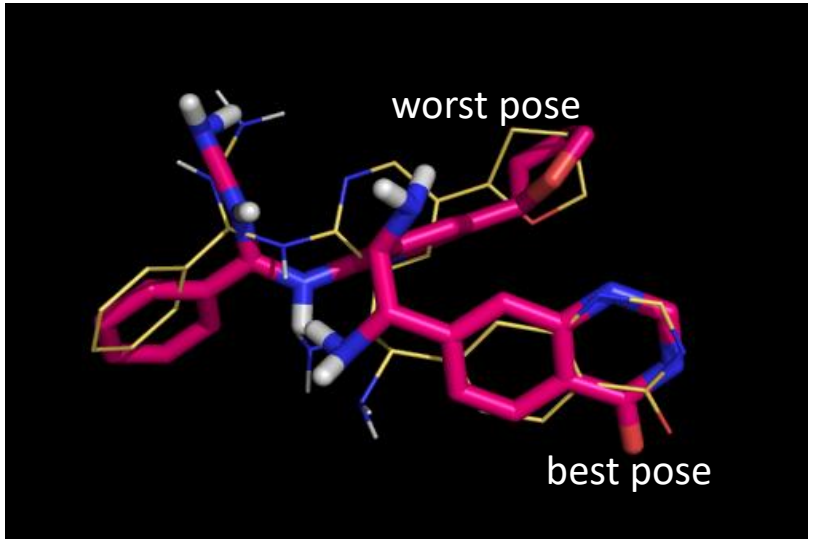


Molecular docking results

Rank	Ligand	Pose #	Score
1	ind	1	-10.15
2	ind	3	-8.93
3	ind	2	-8.93
4	ind	4	-8.9
5	ind	5	-8.68
6	ind	6	-8.2
7	ind	7	-7.81
8	ind	8	-7.75
9	ind	9	-6.72
10	ind	10	-6.08

Rank	Ligand	Pose #	
1	novelhit	1	-9.75
2	novelhit	2	-8.8
3	novelhit	3	-8.74
4	novelhit	4	-8.13
5	novelhit	5	-8.11
6	novelhit	6	-7.83
7	novelhit	7	-7.32
8	novelhit	8	-7.29
9	novelhit	9	-6.89
10	novelhit	10	-6.34



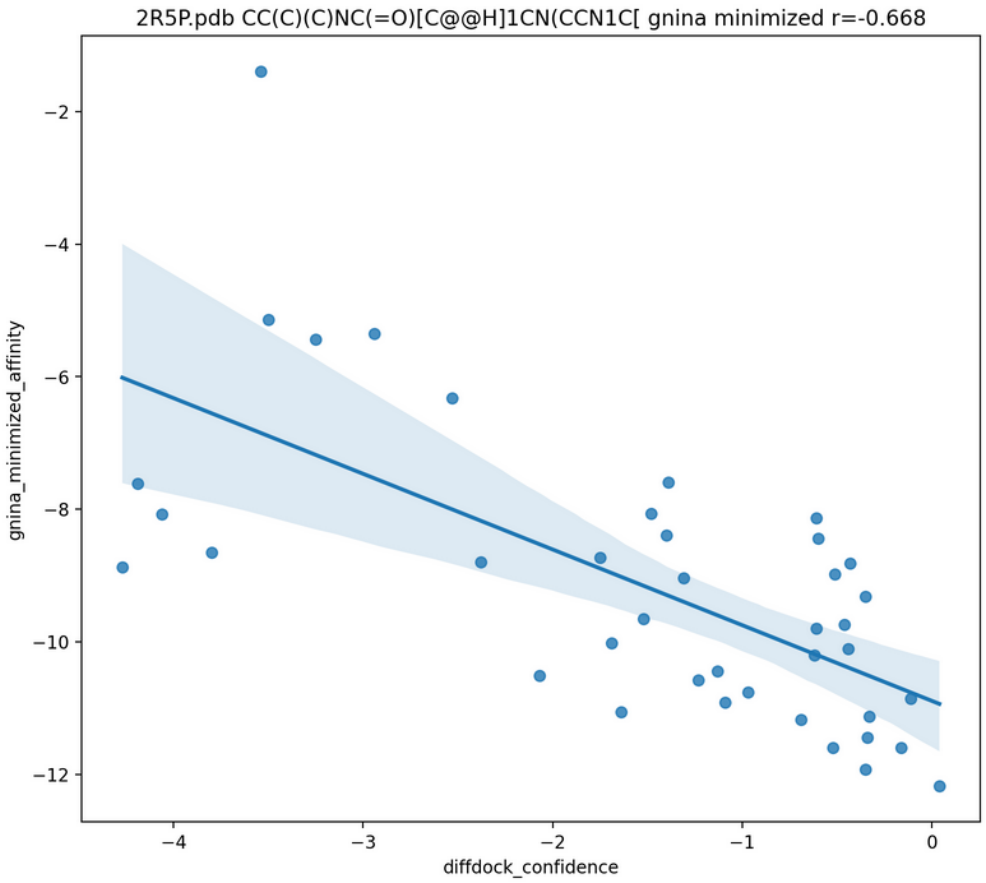


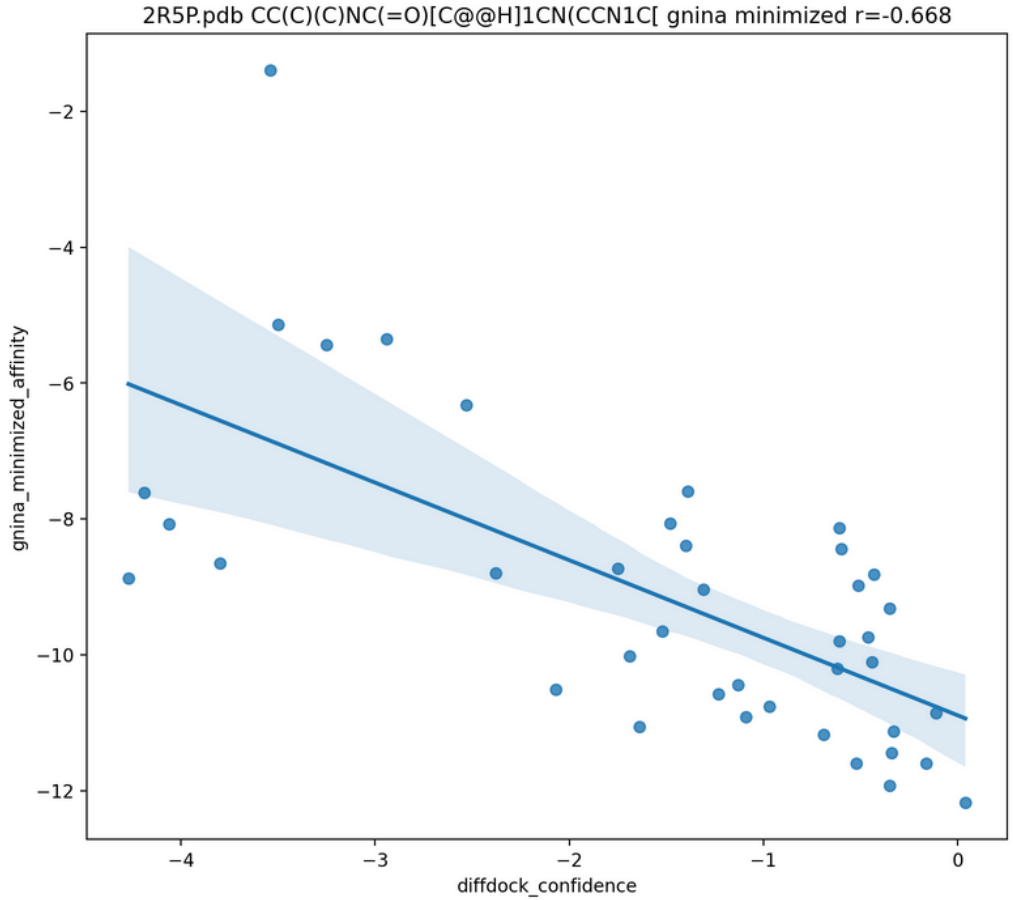
ΔG (indinavir) = – 10.15 kcal/mol ΔG (novel hit) = – 9.75 kcal/mol

Ki (indinavir) = 34.11 nM Ki (novel hit) = 67.17 nM

Ki (exp) = 0.052 – 398 nM

Molecular docking results (indinavir)

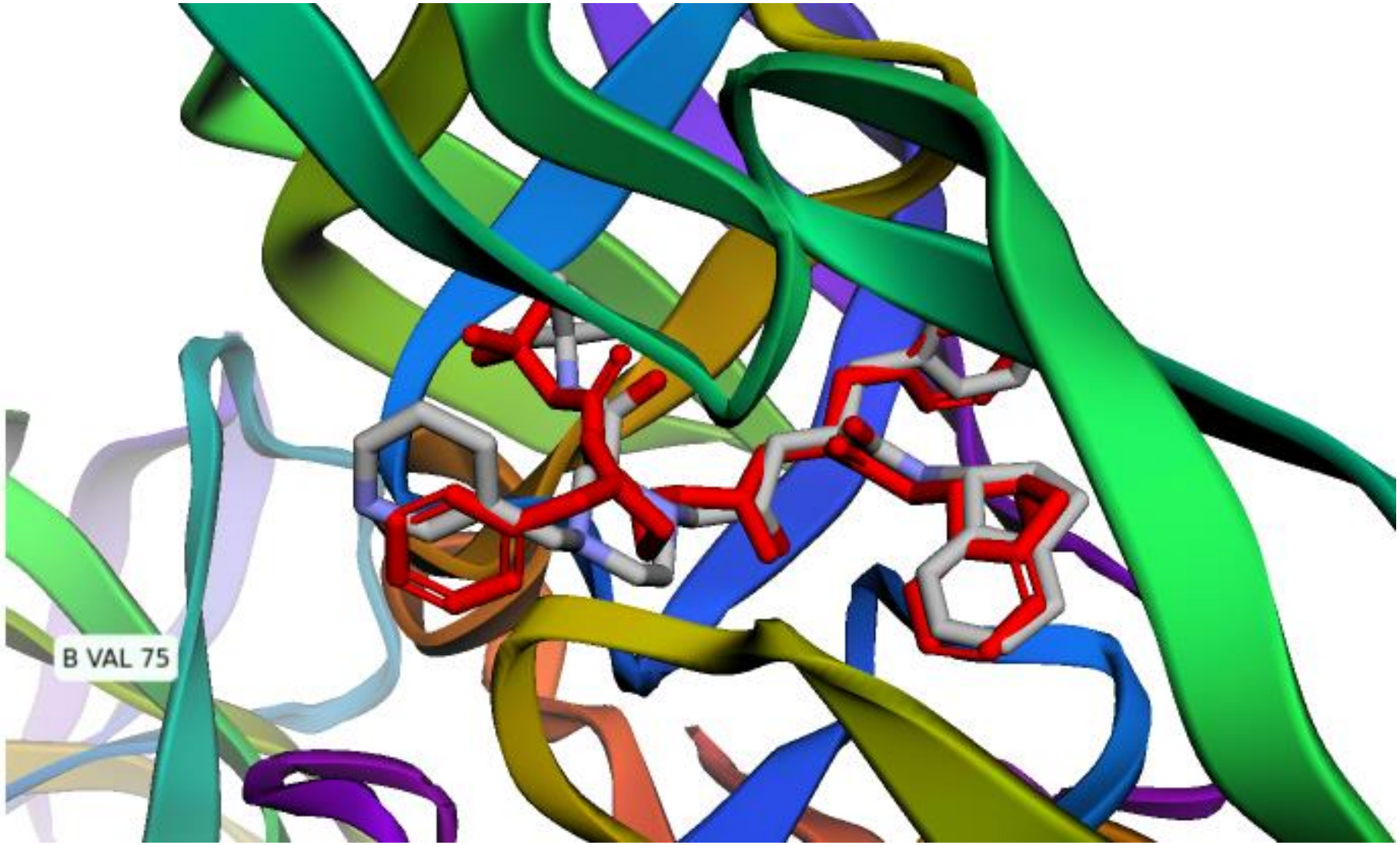




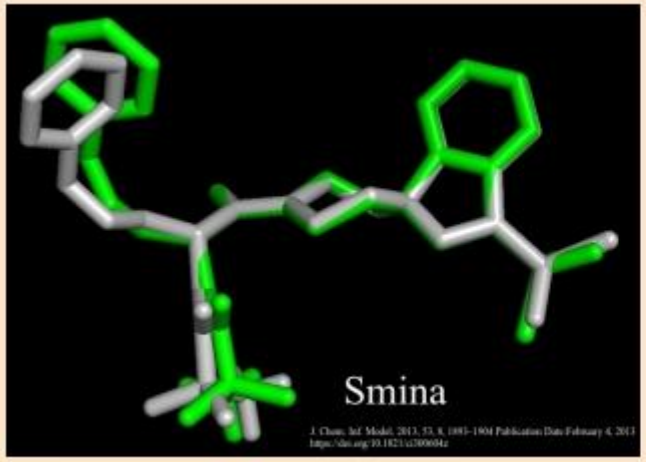
	pdb_file	smiles	diffdock_confidence	gnina_scored_affinity	gnina_minimized_affinity
27	2R5P.pdb	CC(C)(C)NC(=O)[C@@H]1CN(CCN1C[C@H](C[C@@H](CC2...	0.04	-8.90635	-12.18005
34	2R5P.pdb	CC(C)(C)NC(=O)[C@@H]1CN(CCN1C[C@H](C[C@@H](CC2...	-0.11	-5.75277	-10.85702
16	2R5P.pdb	CC(C)(C)NC(=O)[C@@H]1CN(CCN1C[C@H](C[C@@H](CC2...	-0.16	-9.22007	-11.60135

time: 267 ms (started: 2024-02-20 10:58:01 +00:00)

HIV-1 protease docking results



Ghrelin receptor docking results



Tool	RMSD
Gnina	0,953
AD Vina	0,349
Smina	**0,259**

Binding energy

Ligand / Tool	Gnina	AutoDock Vina	Smina
Ibutamoren	-11,66	-11,3	-11,5
Agrelax	-10,99	-11,0	-10,9
YIL 781	-10,52	-9,9	-10,0
D-Lys3-GHSR-6	-10,33	-9,1	-9,4

Experimental data

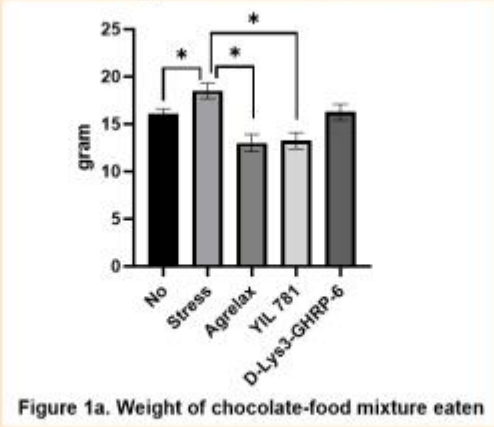
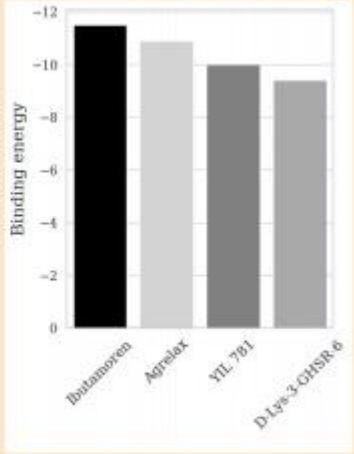
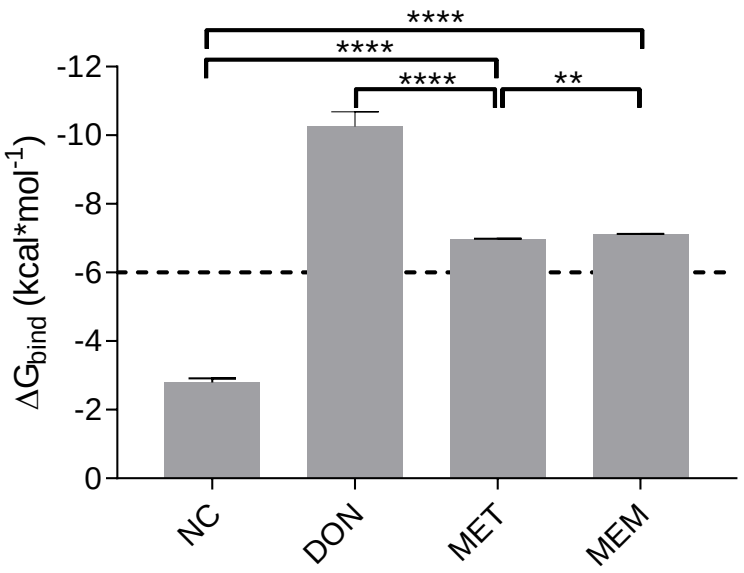


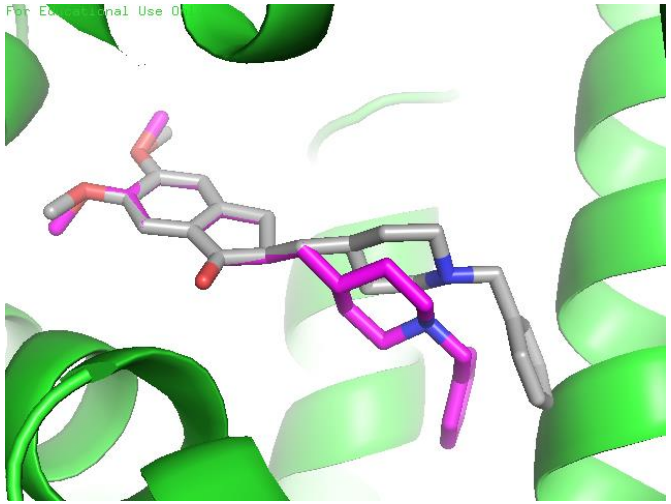
Figure 1a. Weight of chocolate-food mixture eaten

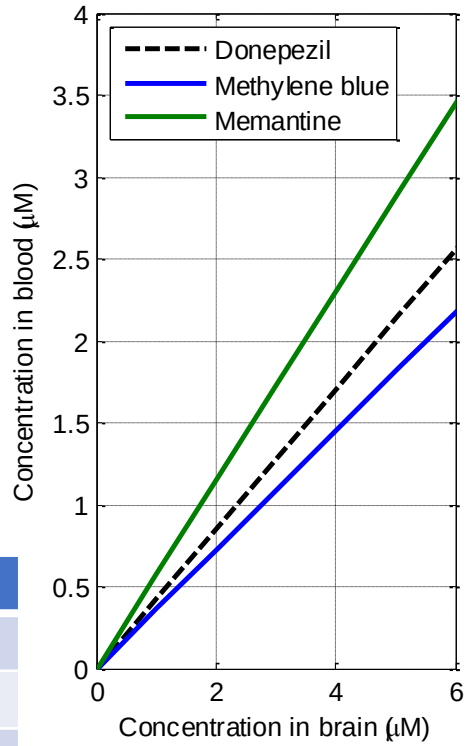
Molecular docking

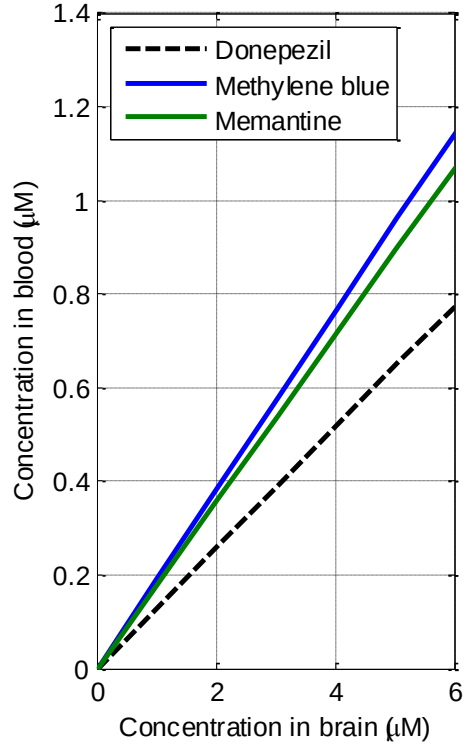


BBB-ChT docking result









$$logBB_{active} = \frac{logBB_{passive} * pK_i}{n}$$

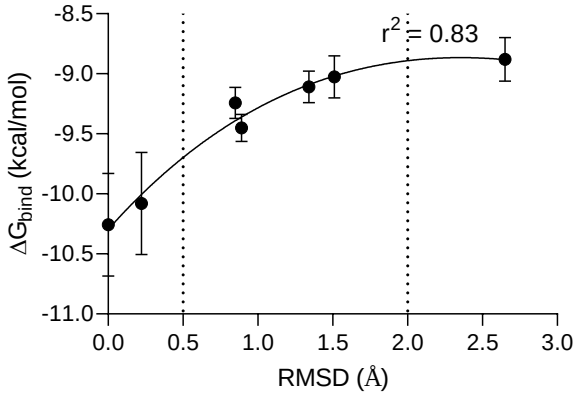
$$BSA = \frac{(ASA_{host} + ASA_{guest}) - ASA_{complex}}{2}$$

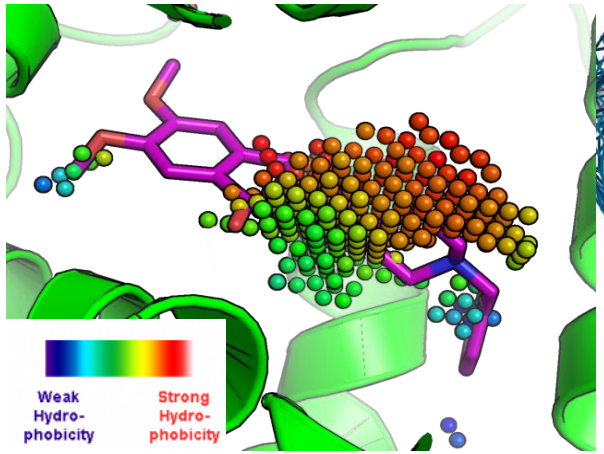
Complex	$logBB_{passive}$	$logBB_{active}$
DON	0.37	0.89
MET	0.44	0.72
MEM	0.24	0.4

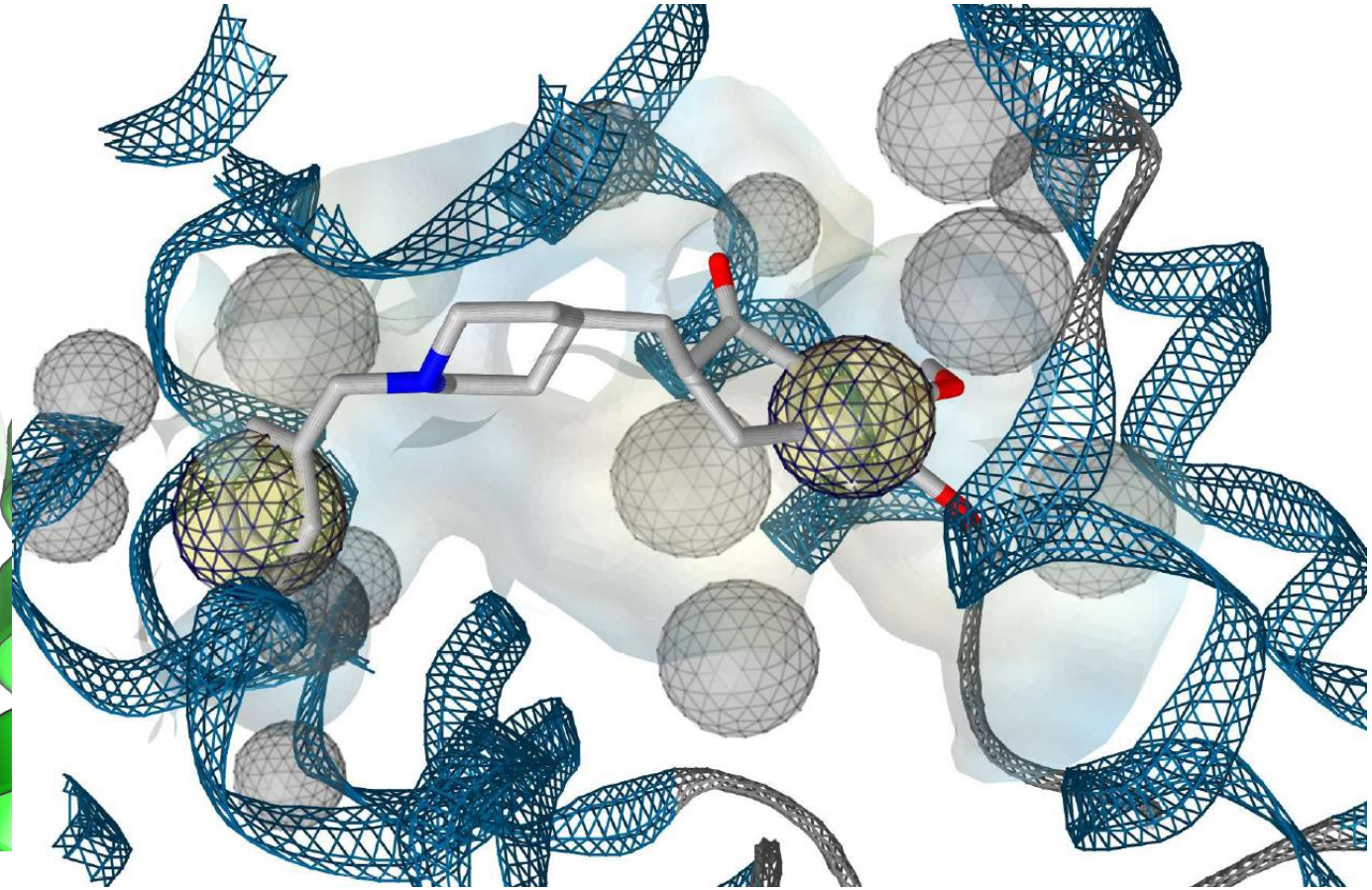
Complex	ΔG_{bind} (kcal*mol^{-1})	K_i (μM)	pK_i	BSA (Å^2)	ΔG_{GBSA} (kcal*mol^{-1})	ΔG_{PBSA} (kcal*mol^{-1})
DON	–10.26	0.03	7.52	461.99	-55.51	-10.75
MET	–6.97	7.45	5.13	223.1	-38.14	-4.49
MEM	–7.12	5.78	5.23	341.29	-33.73	-7.26

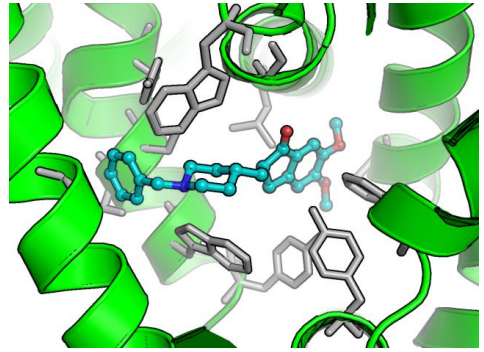
ITMOre than a UNIVERSITY

BBB-ChT docking result

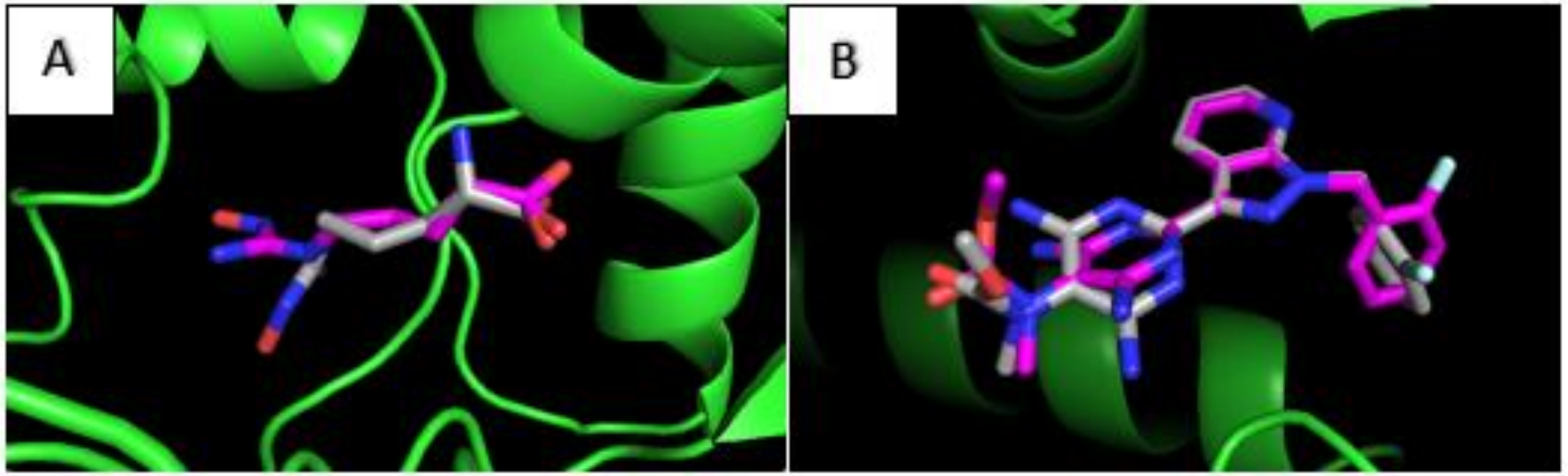


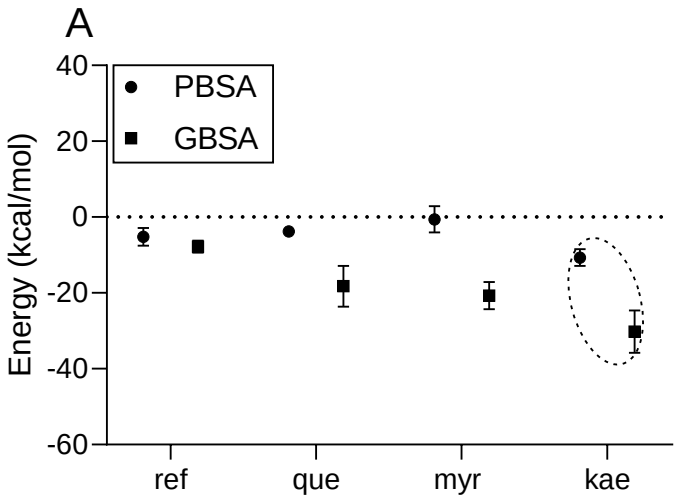


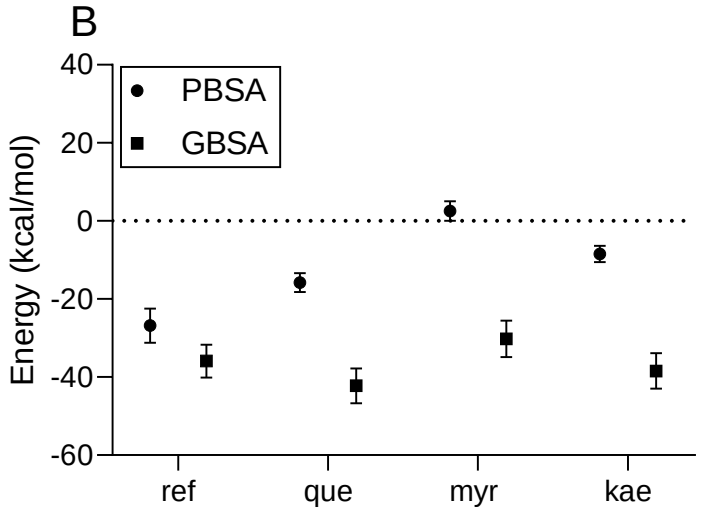


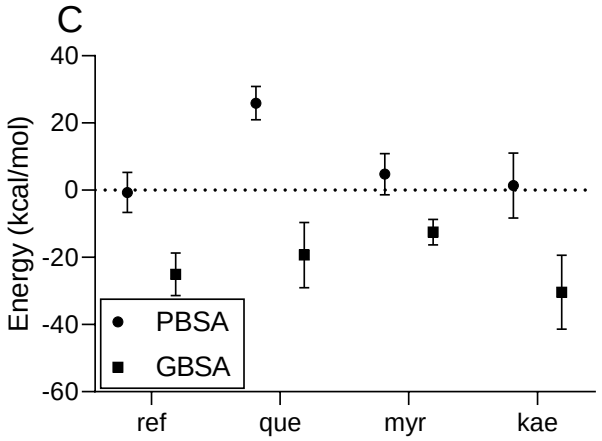


eNOS/NO/GC/SERCA pathway

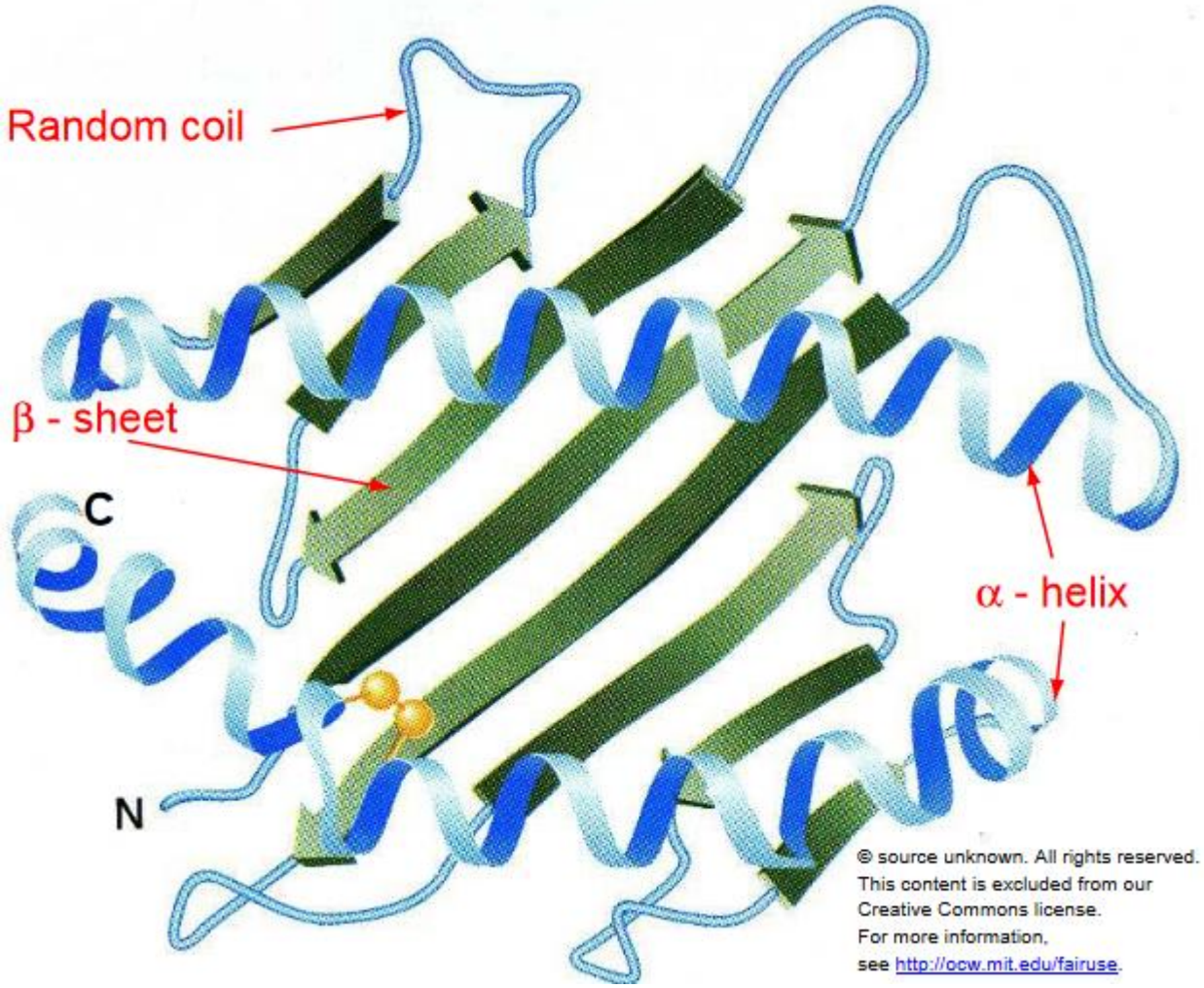




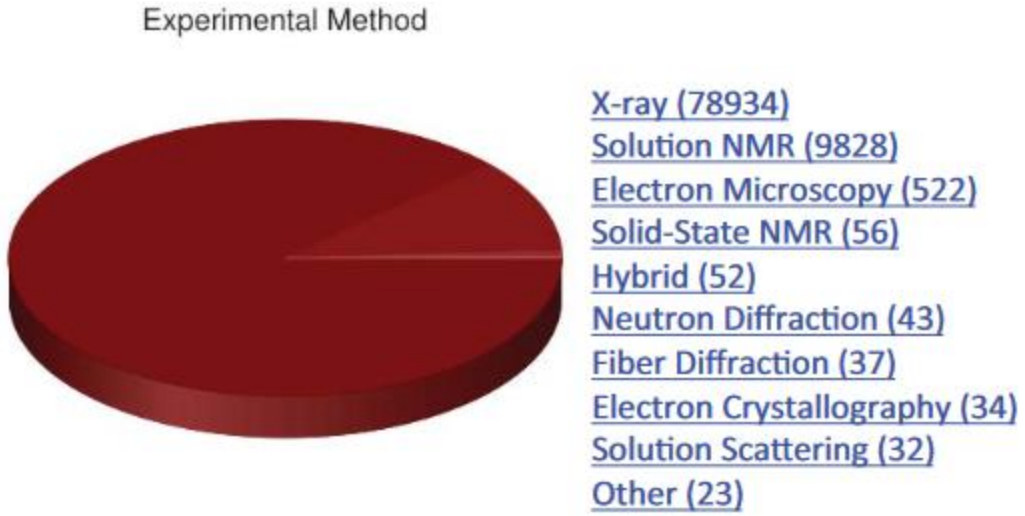




Protein topology



http://www.rcsb.org/pdb



Protein structure

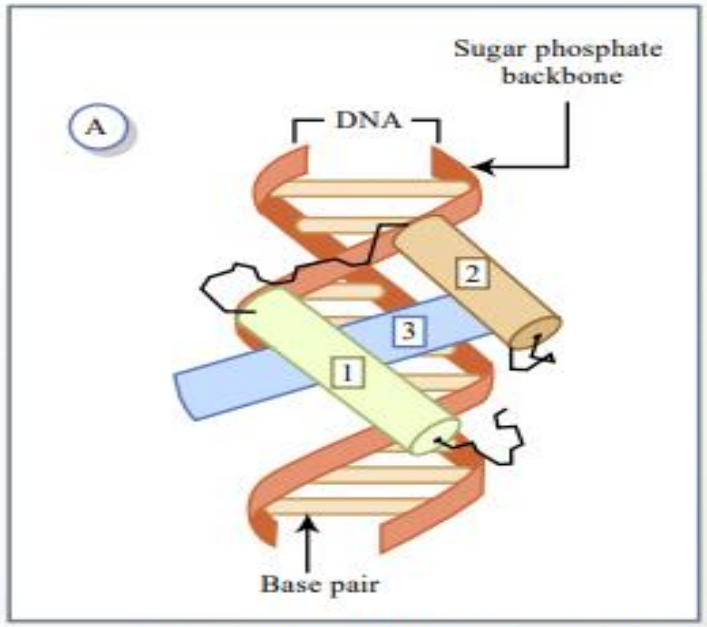


Image by MIT OpenCourseWare.

Helix-turn-helix

Common motif for DNA-binding proteins that often play a regulatory role as mRNA level transcription factors

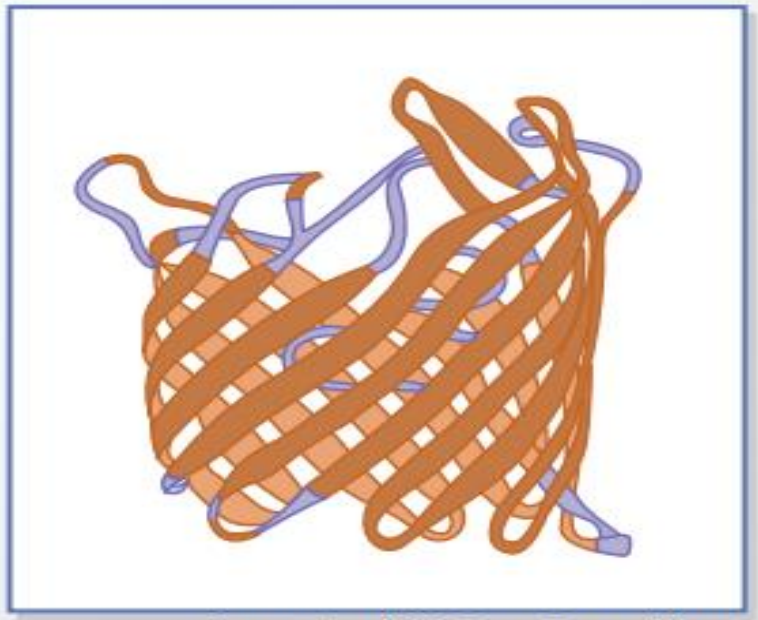


Image by MIT OpenCourseWare.

Beta-barrel

Some antiparallel b-sheet domains are better described as b-barrels rather than b-sandwiches, for example streptavadin and porin. Note that some structures are

intermediate between the extreme barrel and sandwich arrangements.

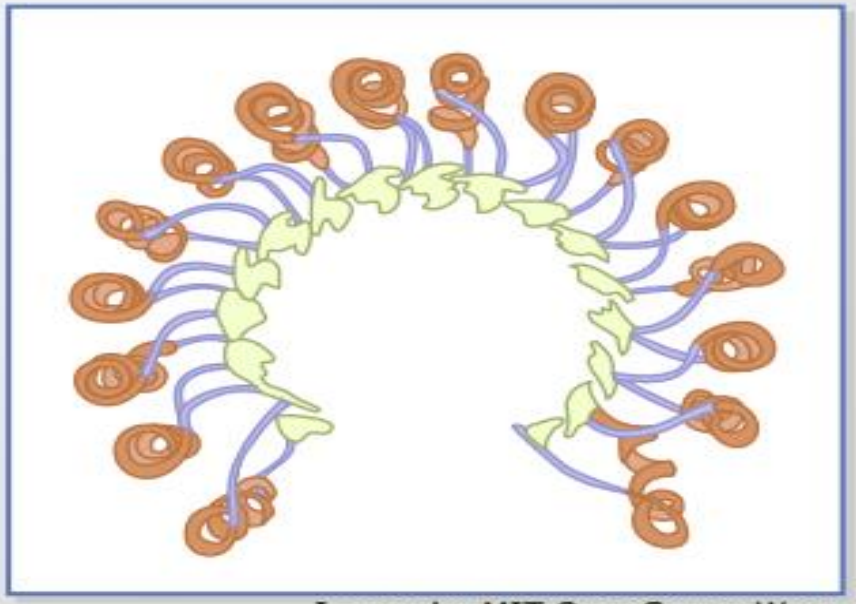


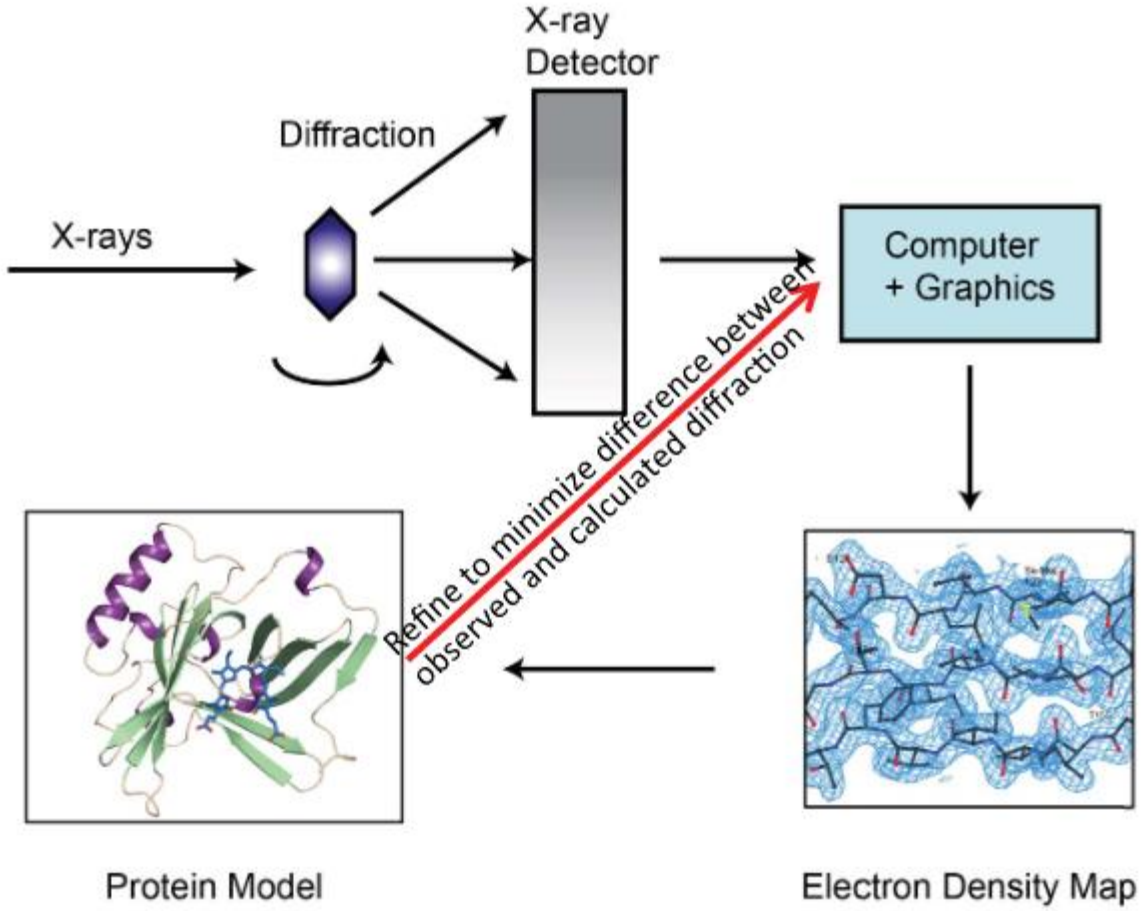
Image by MIT OpenCourseWare.

Alpha-beta horseshoe

this placental ribonuclease inhibitor is a cytosolic protein that binds extremely strongly to any ribonuclease that may leak into the cytosol. 17-stranded parallel b sheet curved into an open horseshoe shape, with 16 a-helices packed against the outer surface. It doesn't form a barrel although it looks as though it should. The strands are only very slightly slanted, being nearly parallel to the central `axis'.

Protein topology

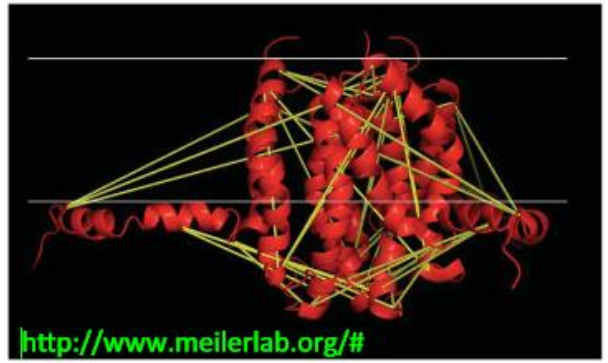
Overview of the X-ray Crystallographic Method



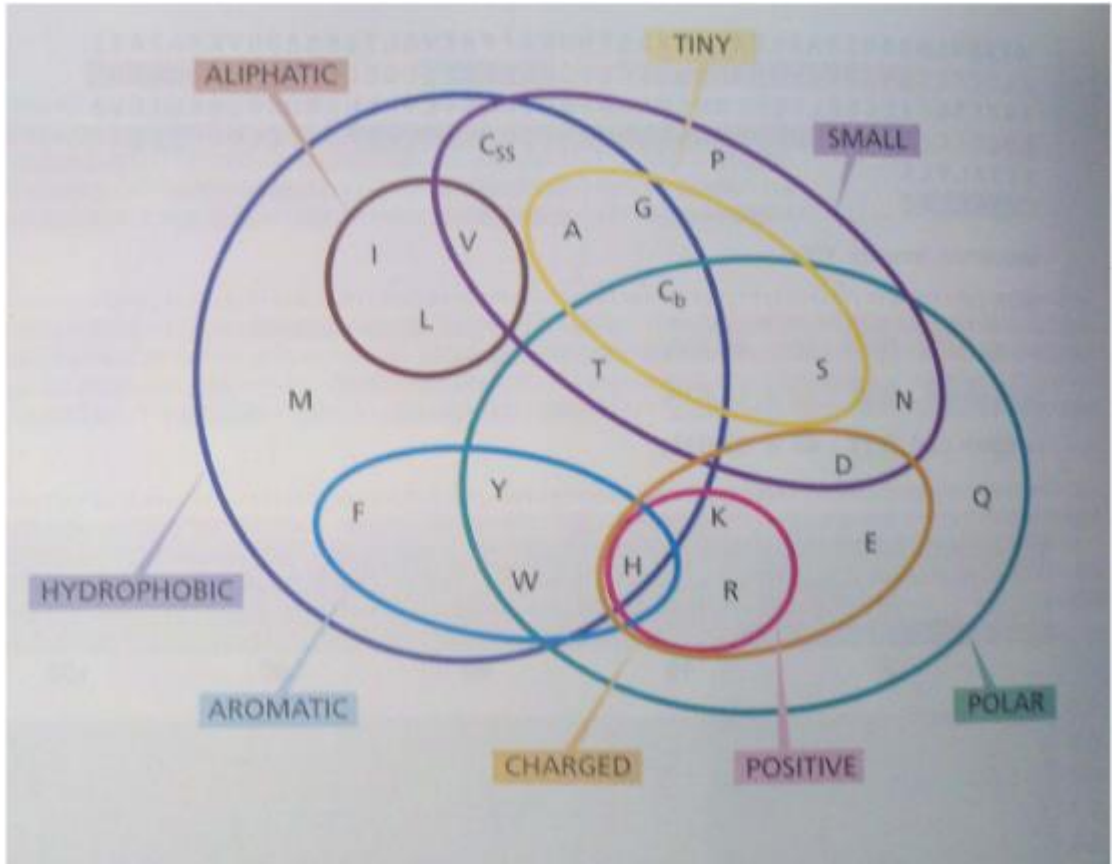
NMR

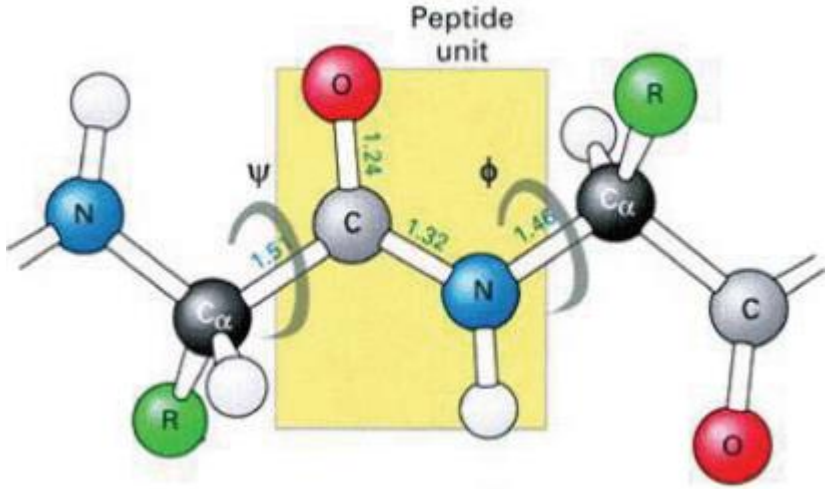
Courtesy of Kjaergaard on wikipedia.
Photograph in the public domain.

Courtesy of MartinSaunders on wikipedia.
Photograph in the public domain.

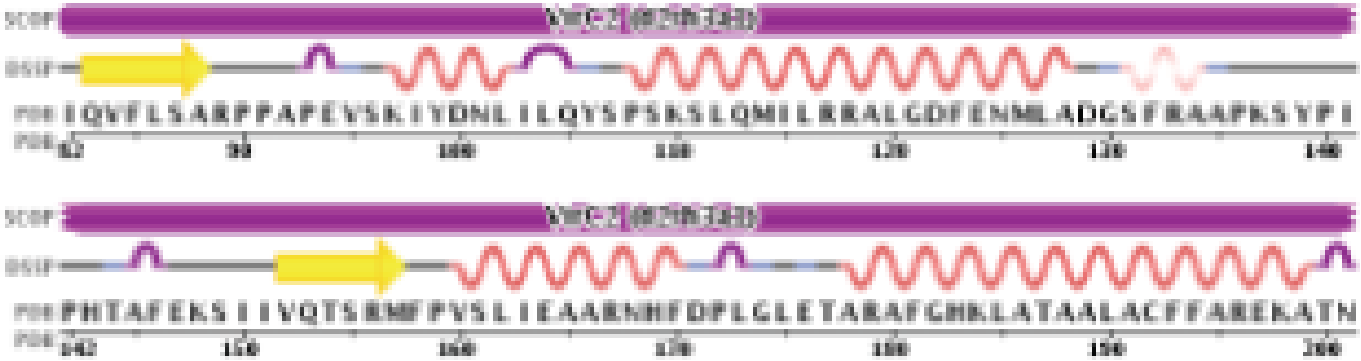


Protein topology





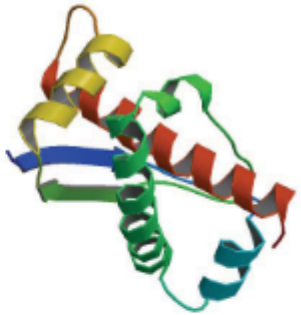
secondary structure

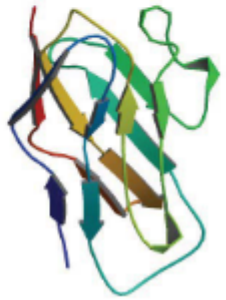


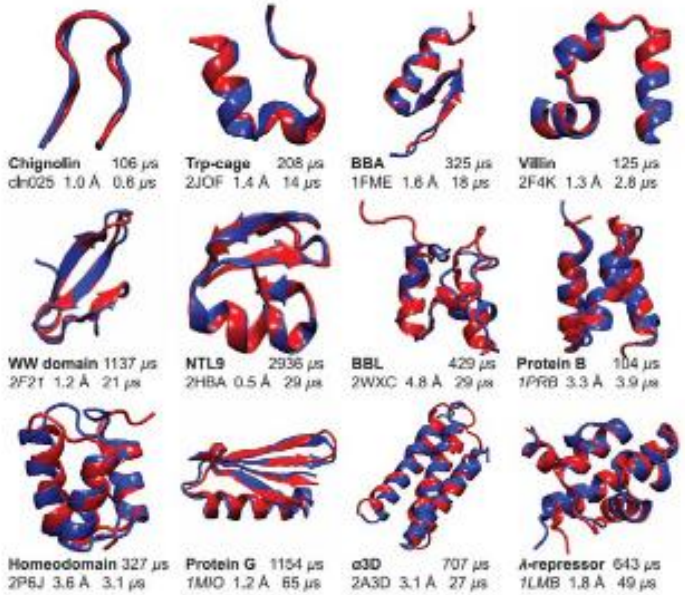
Protein topology

IQVFLSARPPAPEVSKIY
DNLILQYSPSKSLQMILR
RALGDFENMLADGSFR
AAPKSYPIPHTAFEKSIIV
QTSRMFPVSLIEAARN
HFDPLGLETARAFGHKL
ATAALACFFAREKATNS

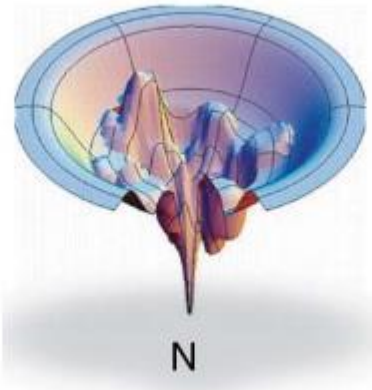
?





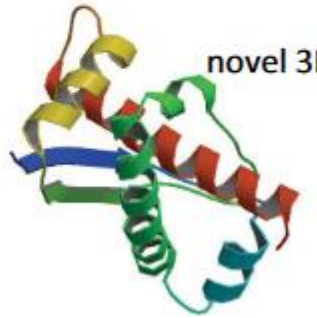


Energy



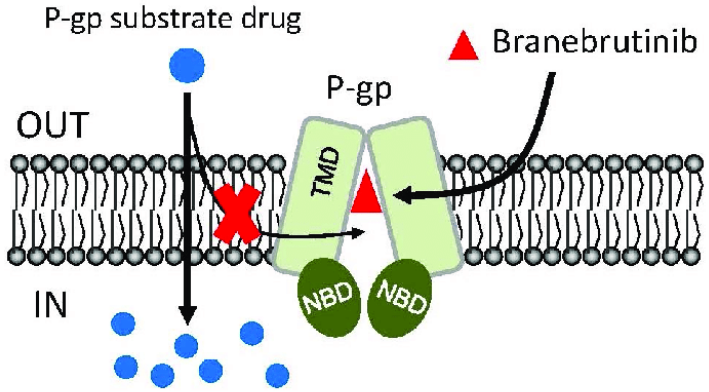
In principle, we don't even need a starting structure.

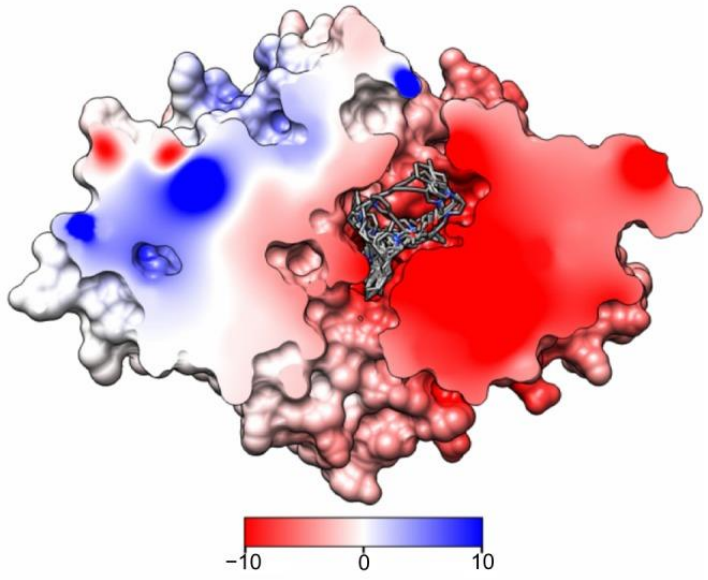
IQVFLSARPPAPEVSKIY
DNLILQYSPSKSLQMILR
RALGDFENMLADGSFR
AAPKSYPIPHTAFEKSIIV
QTSRMFPVSLIEAARN
HFDPLGLETARAFGHKL
ATAALACFFAREKATNS

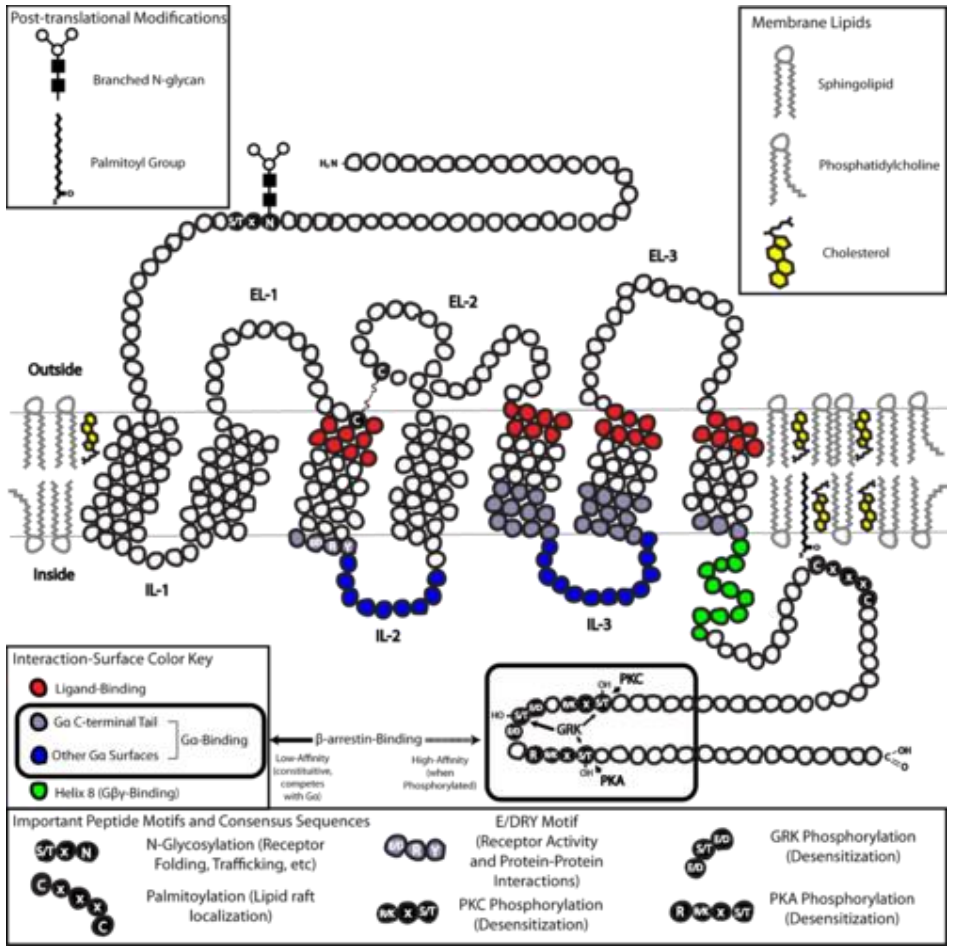


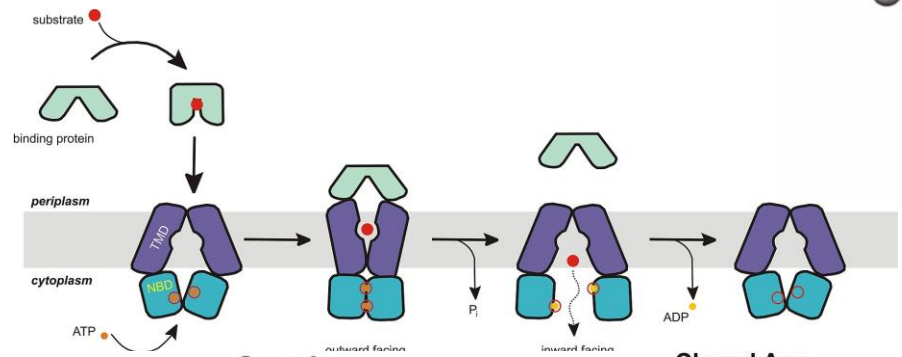
novel 3D structure

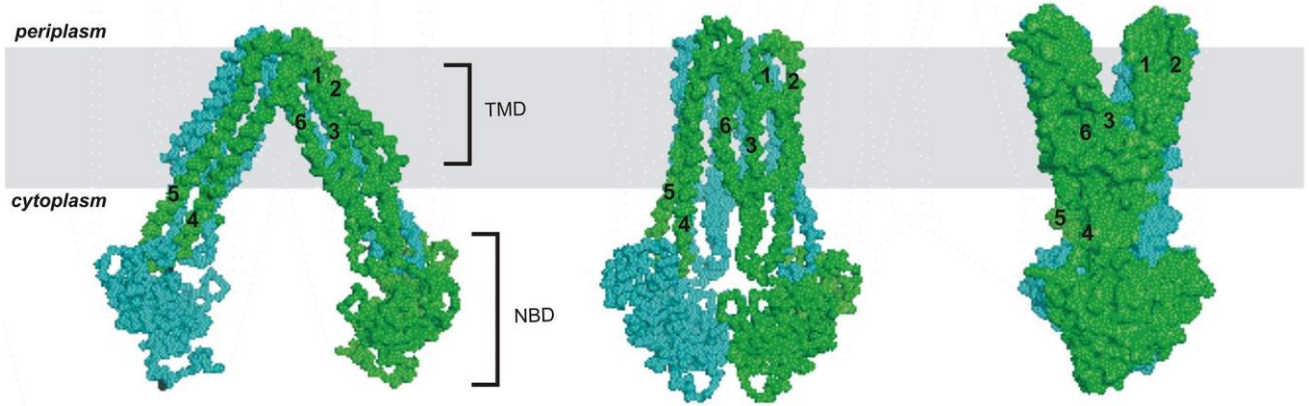
Protein domains and motifs











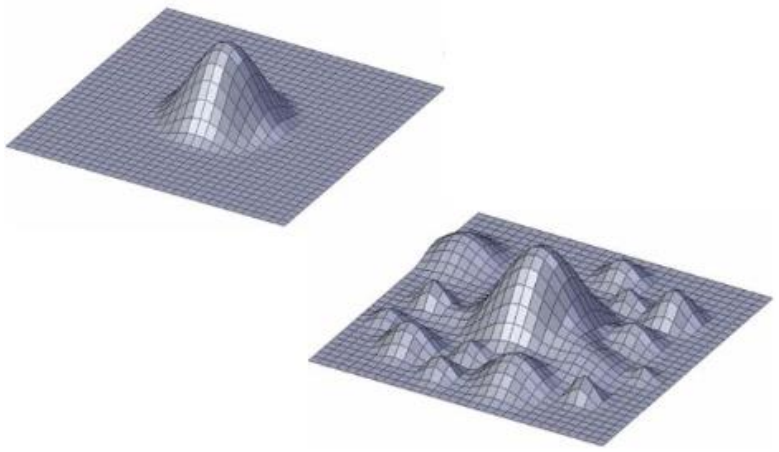
Protein motifs

What is a (biomolecular) sequence motif?

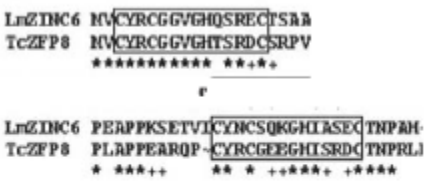
A pattern common to a set of DNA, RNA or protein sequences that share a common biological property, such as functioning as binding sites for a particular protein

Ways of representing motifs

- Consensus sequence
- Regular expression
- Weight matrix/PSPM/PSSM
- More complicated models



Examples of Protein Sequence Motifs

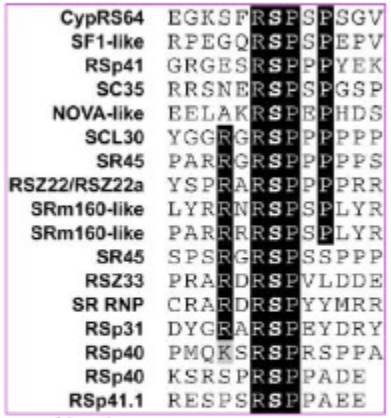


Source: Ericsson, A. C., L. C. Faria, et al. "TcZFP8, A Novel Member of the Trypanosoma Cruzi CCHC Zinc Finger Protein Family with Nuclear Localization." *Genetics and Molecular Research* 5, no. 3 (2006): 553-63.

$CX_2CX_4HX_4C$

Zinc finger (DNA binding)

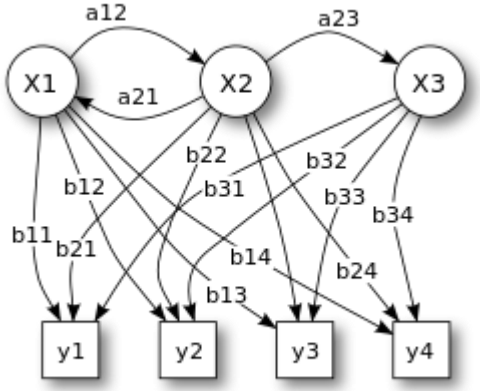
Ericsson et al. Genet. Mol. Res. 2006



Source: Benlloch, Van, Sergio de la Fuente, et al. "Phosphoproteomics Reveals Extensive in Vivo Phosphorylation of Arabidopsis Proteins Involved in RNA Metabolism." *Nucleic Acids Research* 34, no. 11 (2006): 3267-78.

Phosphorylation sites (*Arabidopsis* SRPK4)

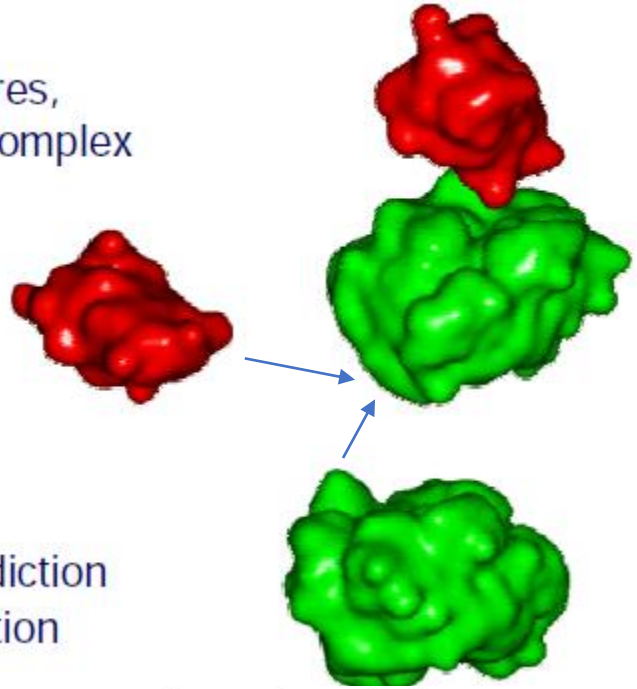
de la Fuente van Bentem et al. NAR 2006



Protein-protein interactions

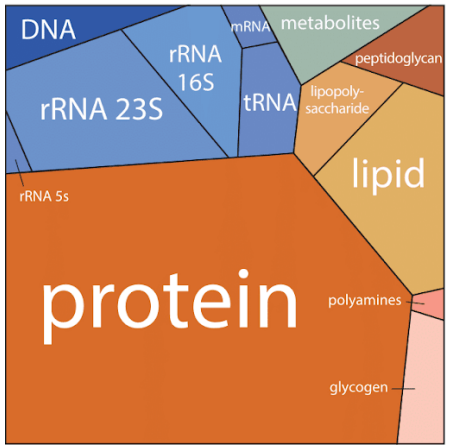
Goal:

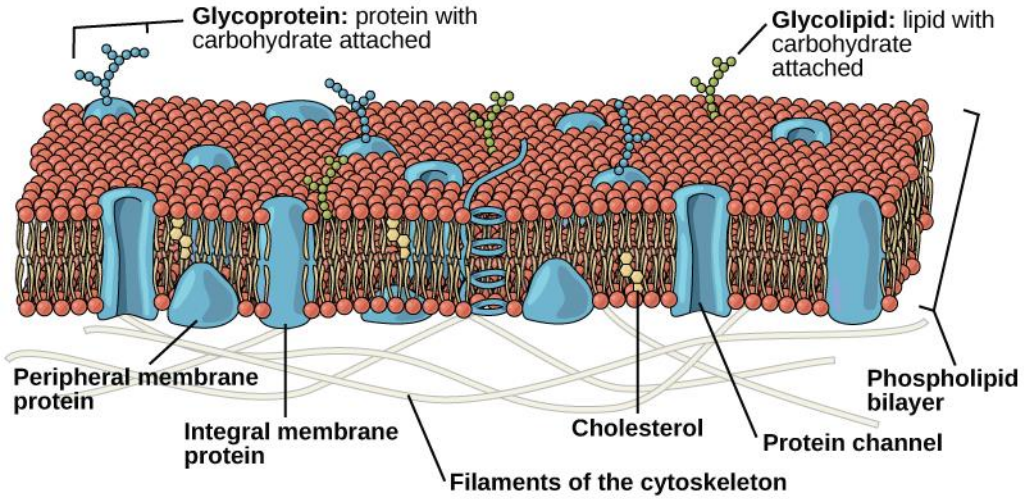
- Given two protein structures, predict how they form a complex



Applications:

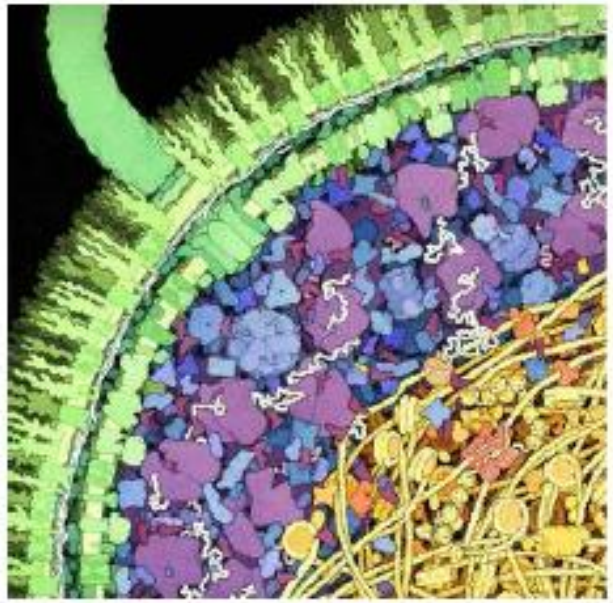
- Quaternary structure prediction
- Protein interaction prediction
- etc.





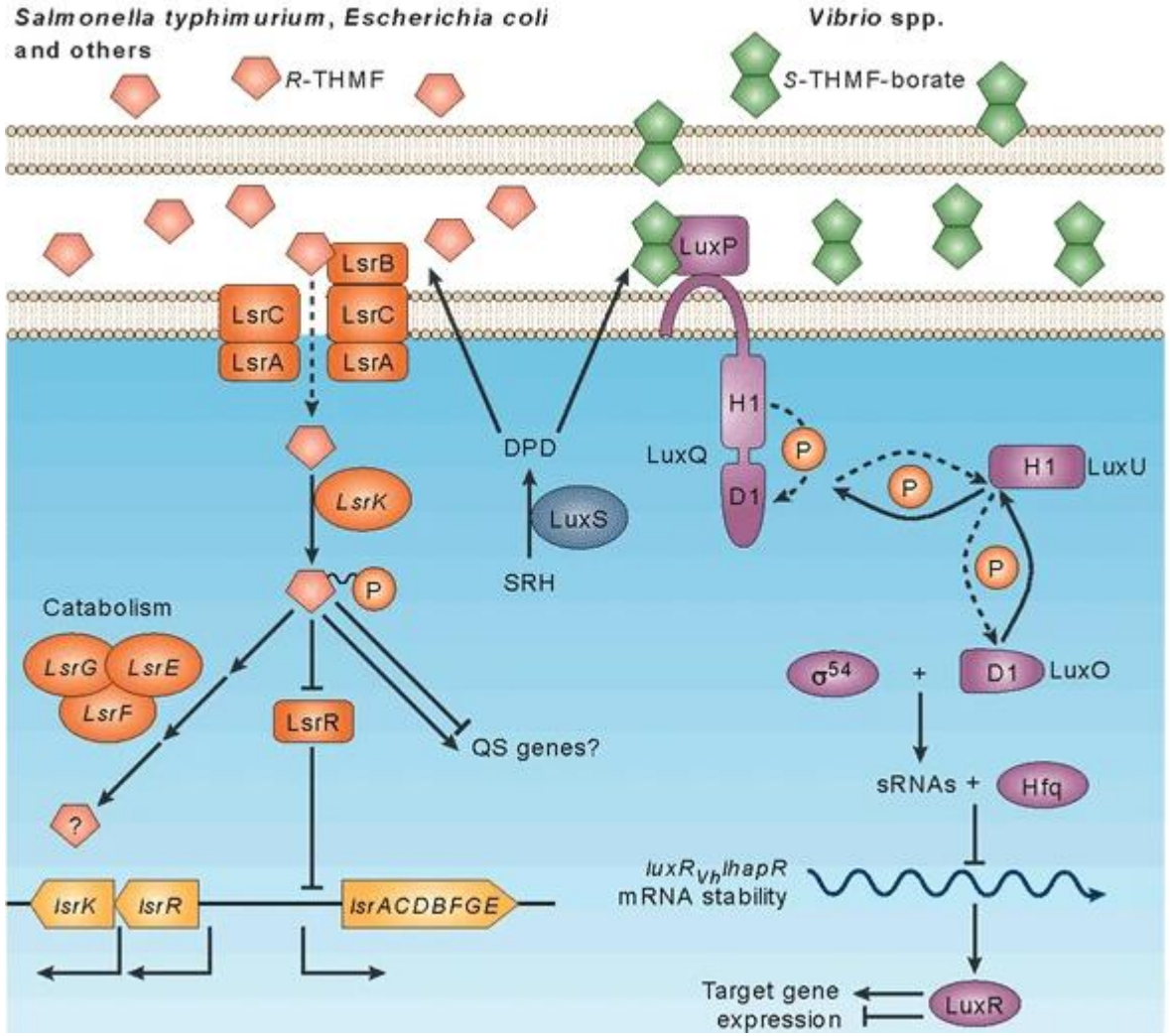
Proteins are densely packed inside cell

- 20-30% of total volume inside cell



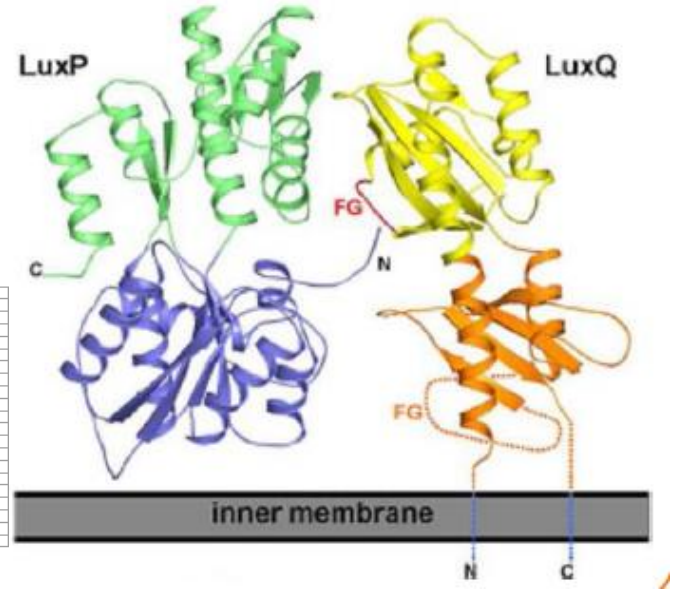
Representation of the approximate numbers, shapes and density of packing of macromolecules inside a cell of *Escherichia coli*. (Illustration by David S Goodsell)

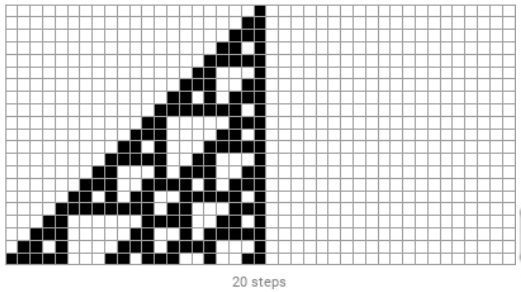
Protein-protein interaction quorum sensing

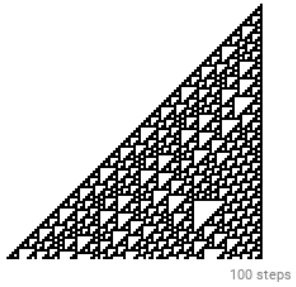


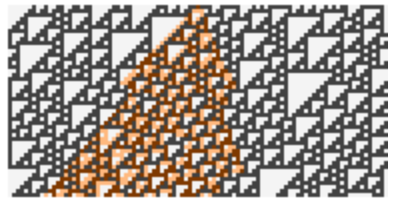
Many biological processes are controlled by protein-protein interactions

- Signal transduction
- Transport
- Cellular motion







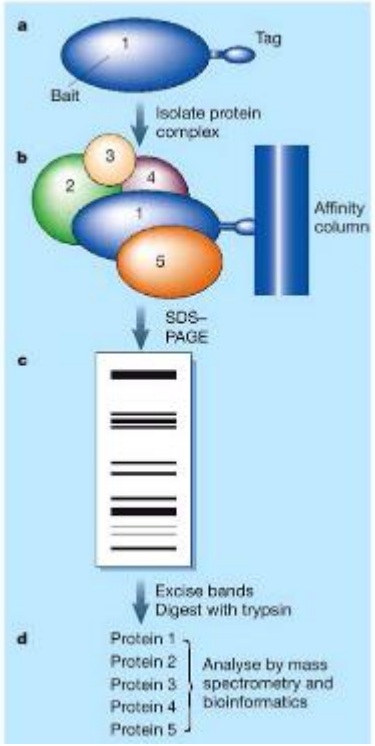


ITMOre than a UNIVERSITY

Protein-protein interaction

What are the likely false positives?

What are the likely false negatives?

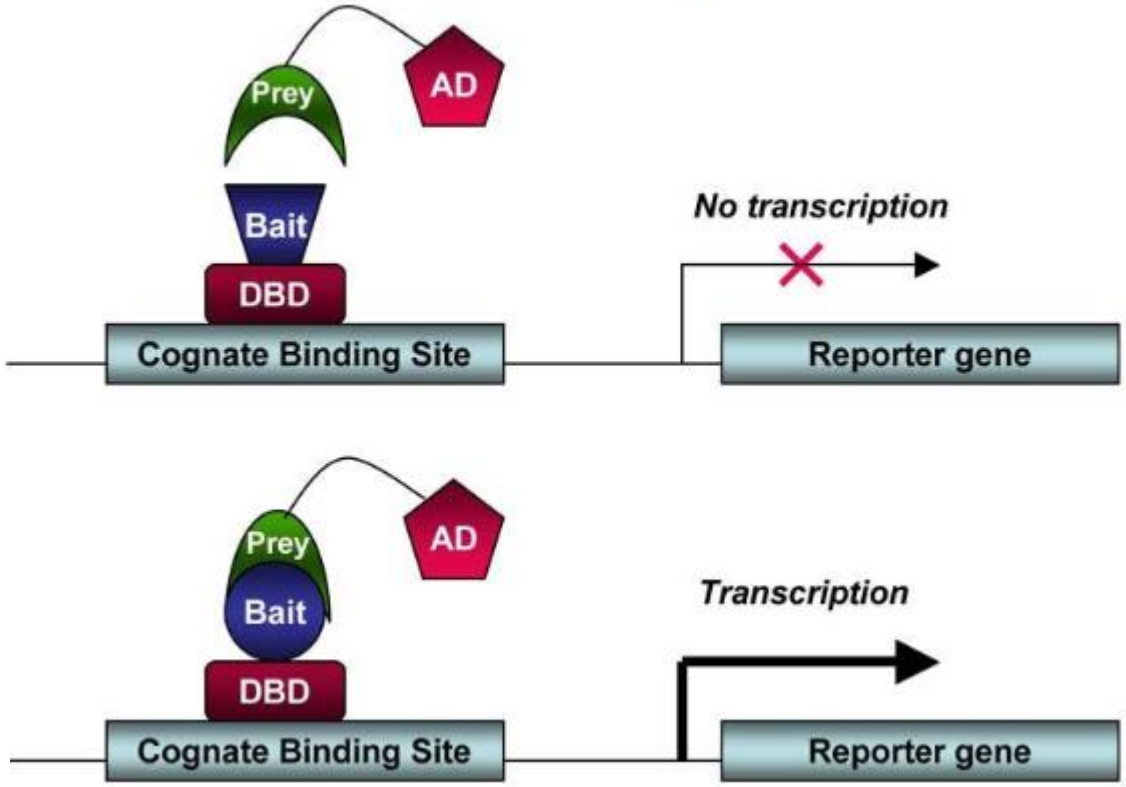


Gavin, A.-C. *et al. Nature* **415**, 141-147 (2002).

Ho, Y. *et al. Nature* **415**, 180-183 (2002).

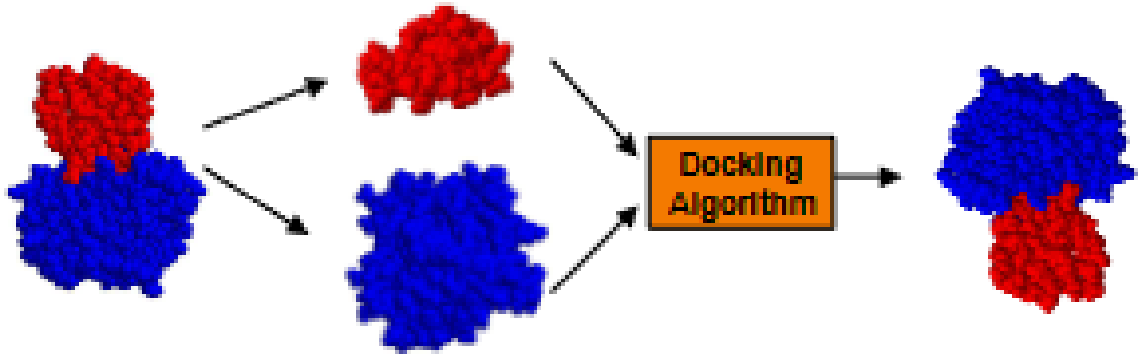
- **Does not require purification – will pick up more transient interactions.**
- **Biased against proteins that do not express well, or are incompatible with the nucleus**

Yeast two-hybrid

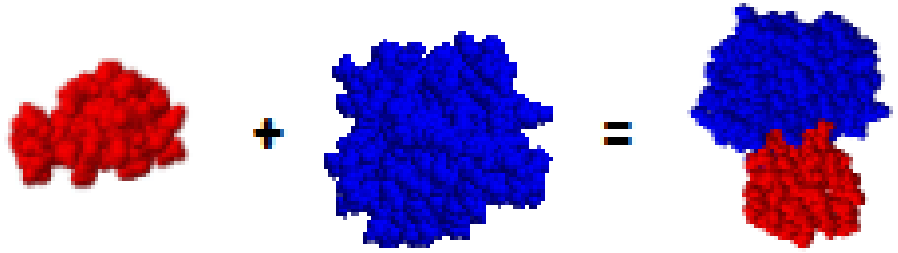


Protein-protein docking

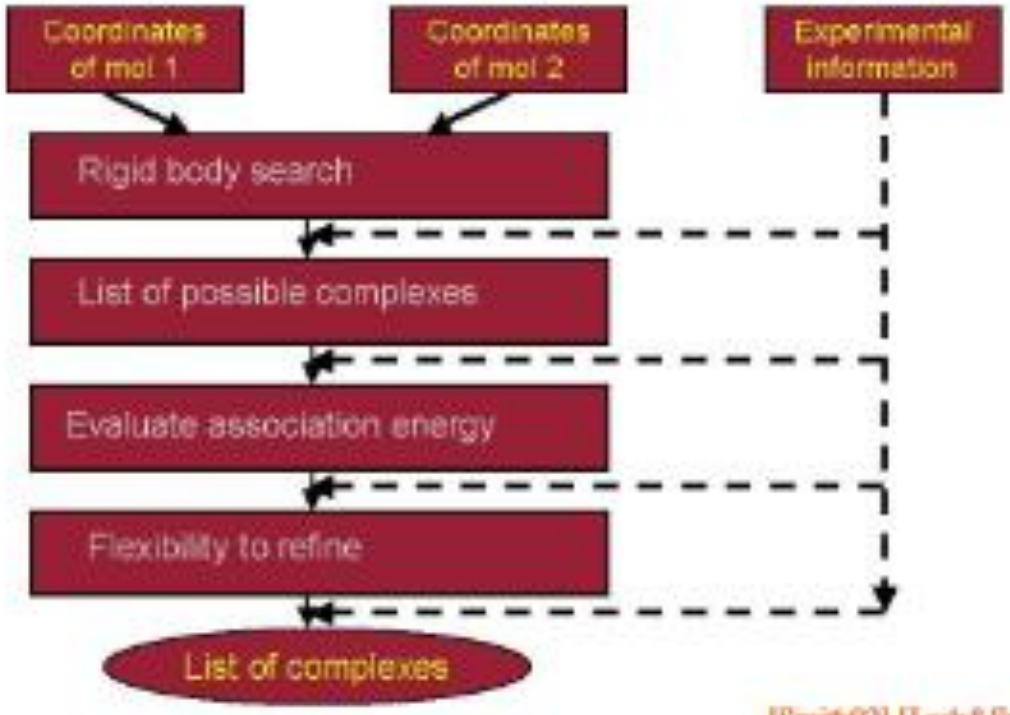
Bound docking:



Unbound docking:



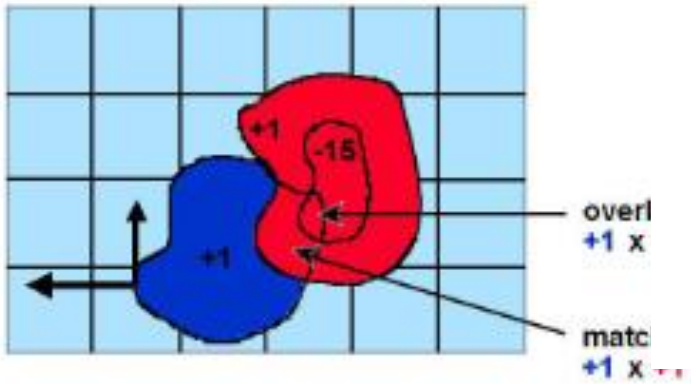
[Gidalevitz]



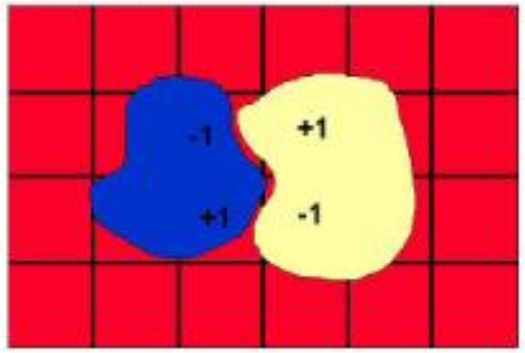
[Smith02] [Lesk&Sternberg]

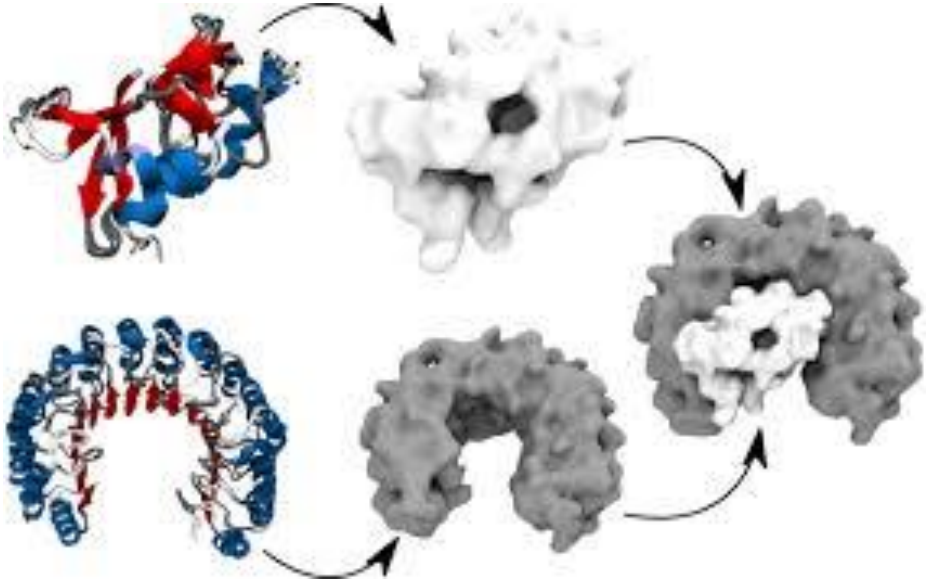
Protein-protein docking

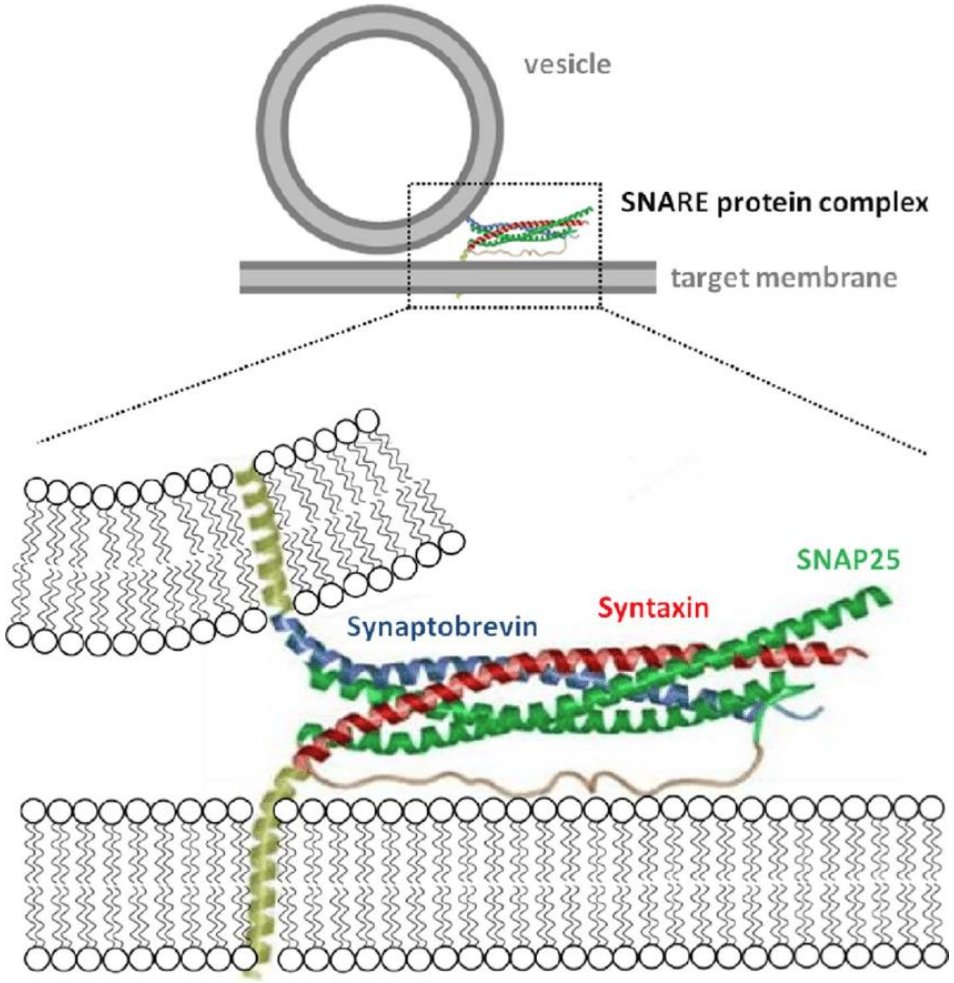
Shape complementarity:



Electrostatic complementarity:

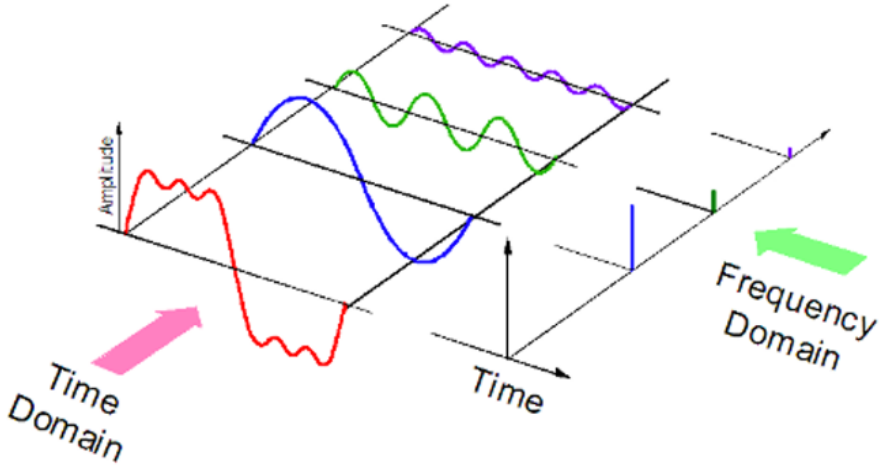






Protein-protein docking

Fast Fourier Transform

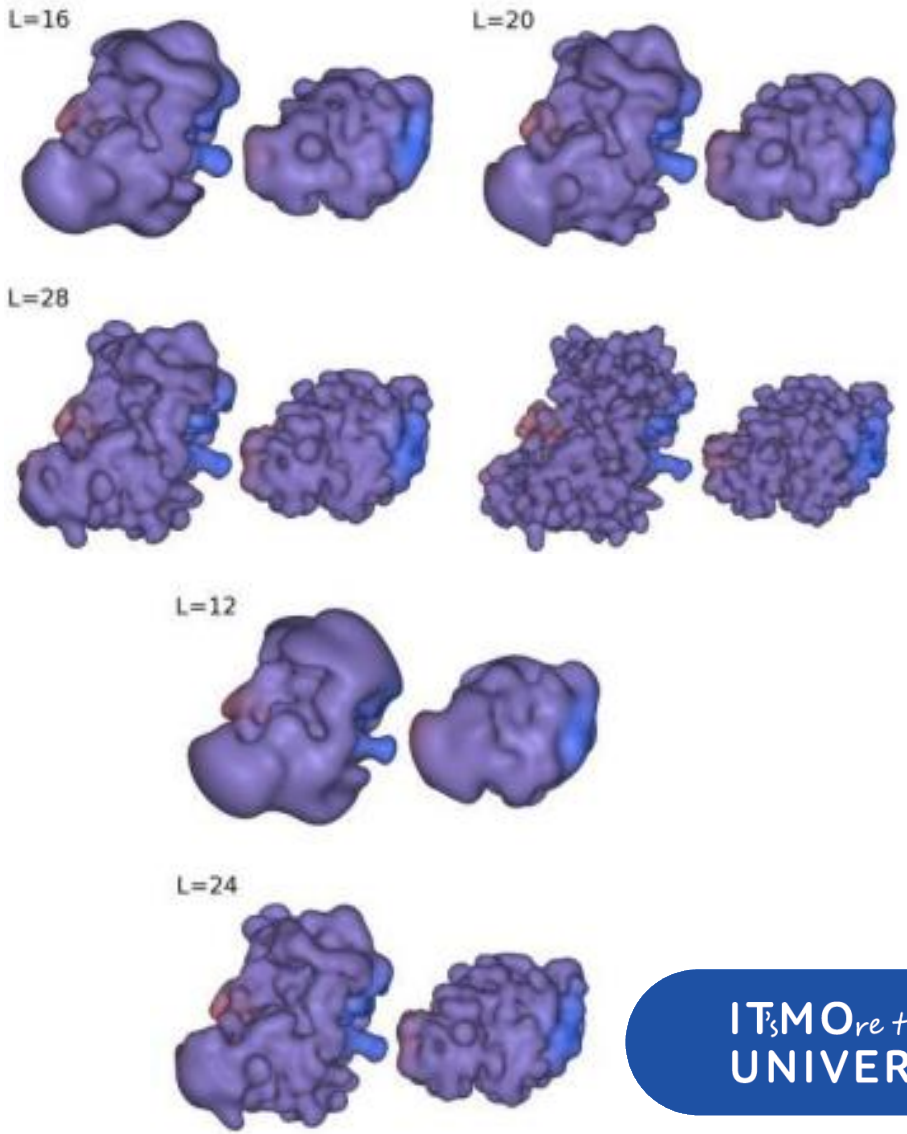


The spherical polar Fourier (SPF) algorithm involves mathematical operations to represent the molecular surfaces of ligand and receptor molecules in terms of spherical harmonics. The equation for the SPF algorithm can be expressed as follows:

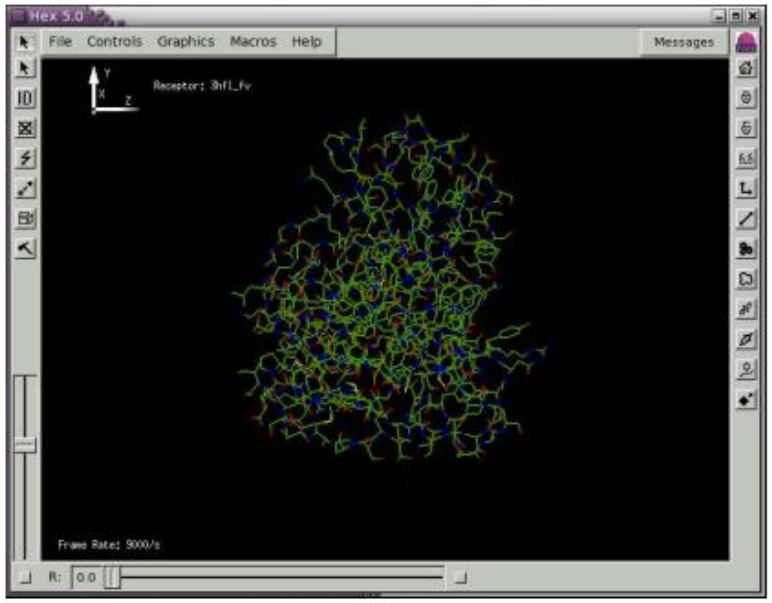
$$F(\theta, \phi) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} c_{lm} Y_{lm}(\theta, \phi)$$

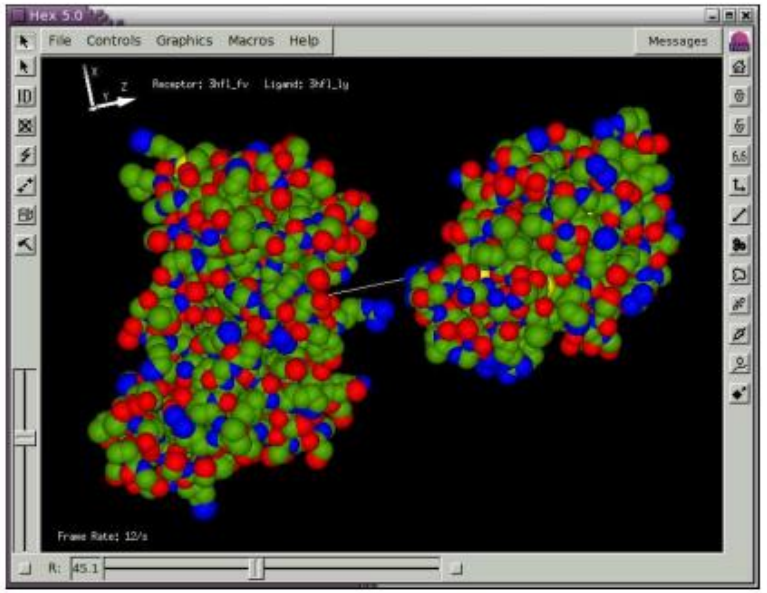
where:

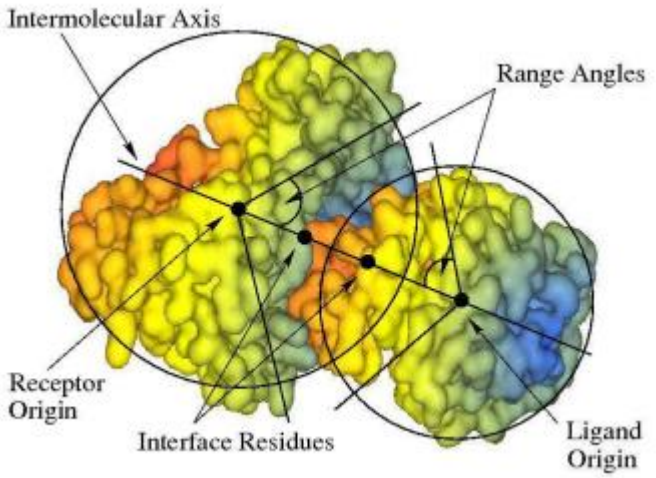
- $F(\theta, \phi)$ represents the molecular surface function in spherical polar coordinates (θ, ϕ),
- c_{lm} are the coefficients of the spherical harmonics,
- $Y_{lm}(\theta, \phi)$ denotes the spherical harmonics functions,
- l represents the degree of the spherical harmonics, and
- m represents the order of the spherical harmonics.

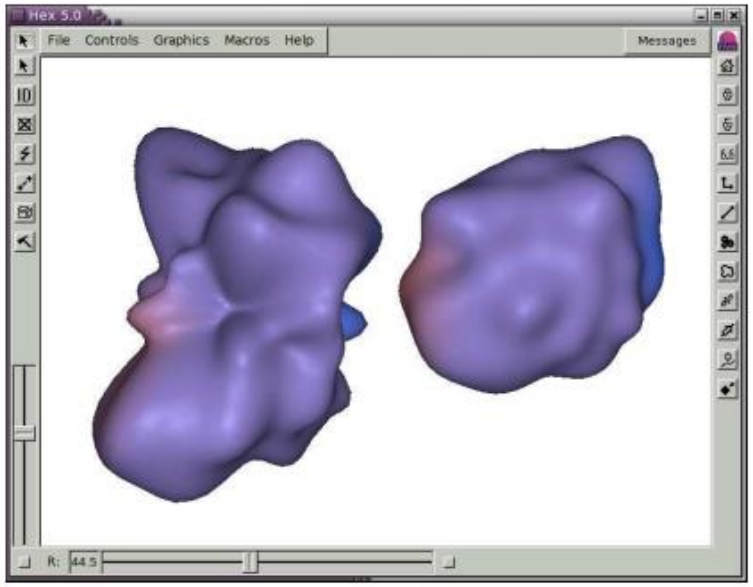


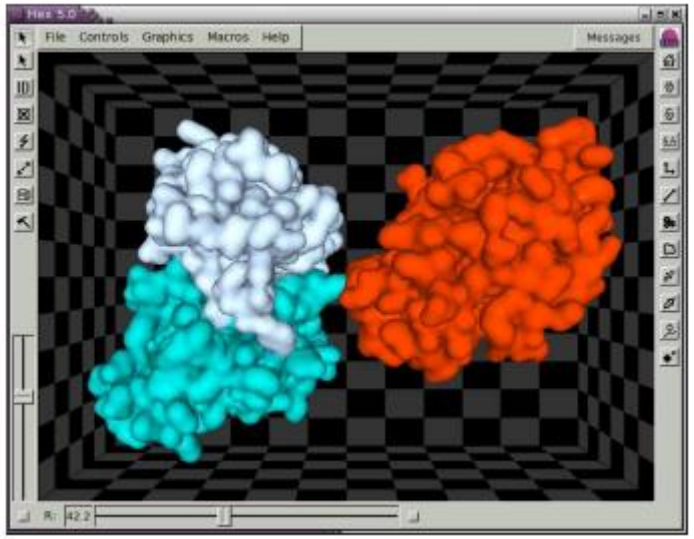
Protein-protein docking (HEX)

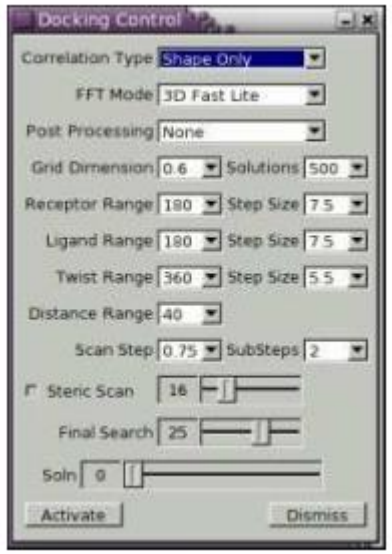




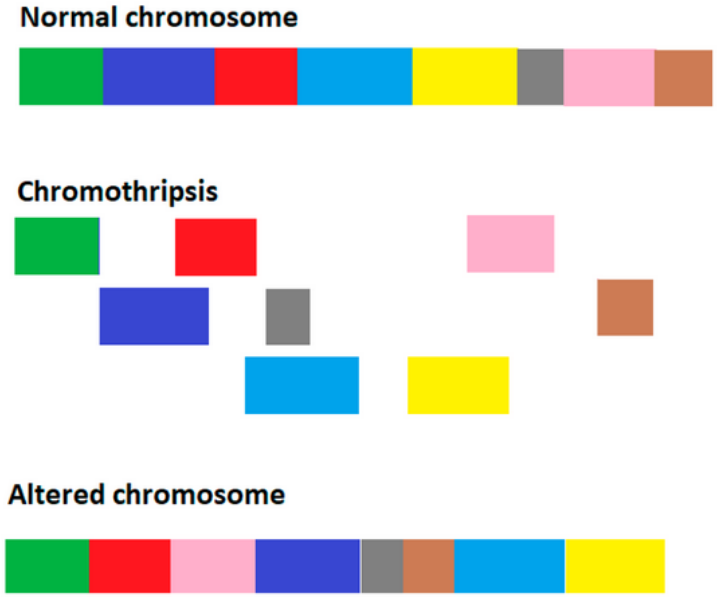


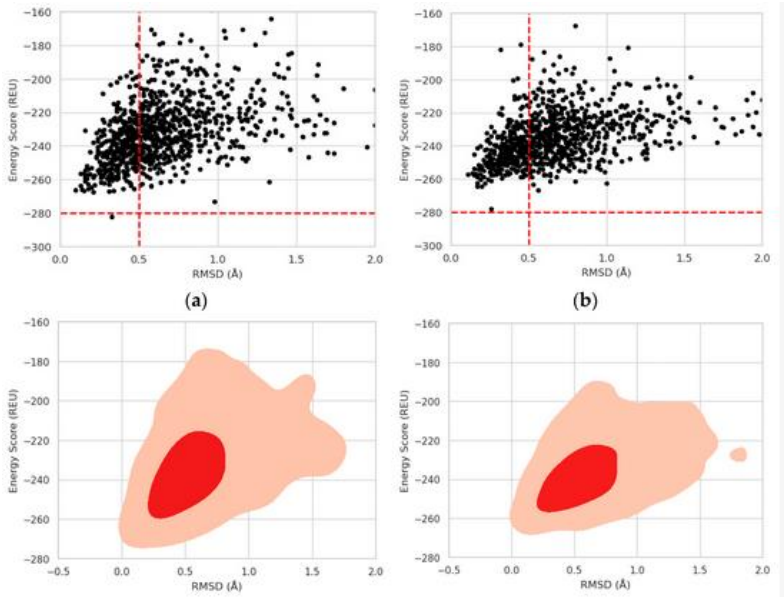


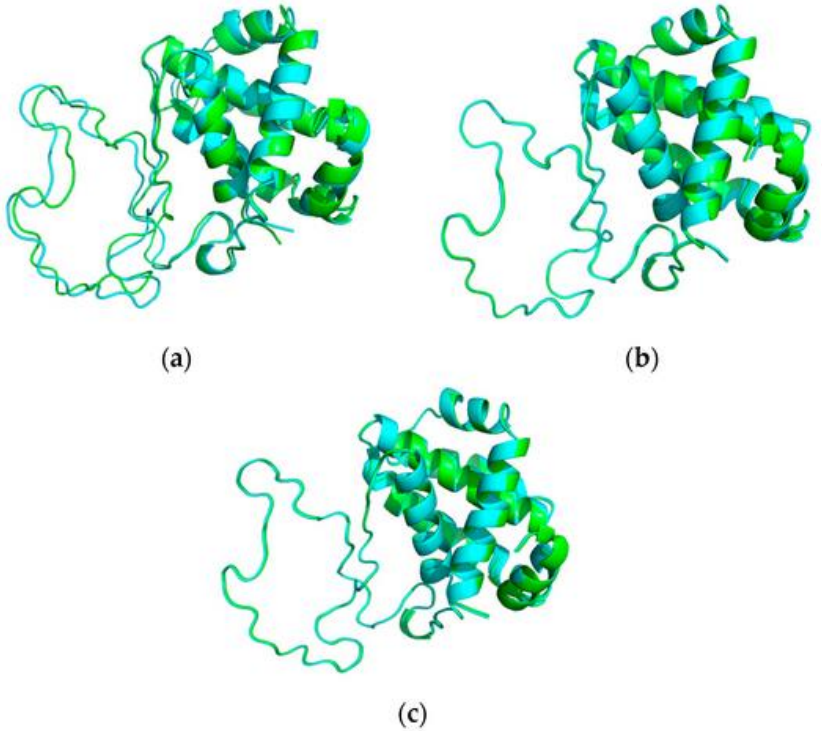


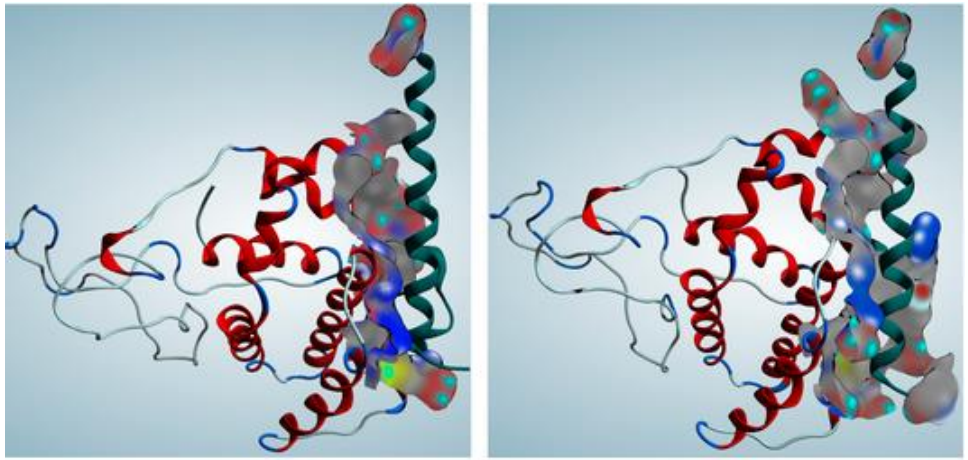


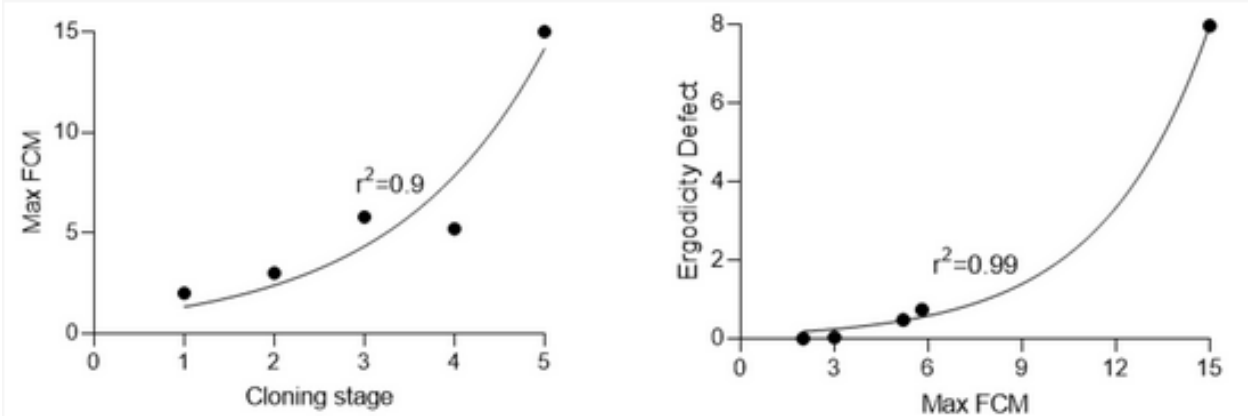
Molecular docking in synthetic biology











Shityakov et al., 2024

ITMOre than a UNIVERSITY

ITMO UNIVERSITY

INFOCHEMISTRY SCIENTIFIC CENTER

Thank you for your attention