Structural Biology - BIO315

Master SV - Spring Semester
Lecture 9

Matteo Dal Peraro
matteo.dalperaro@epfl.ch
AAB 048, phone: 31681

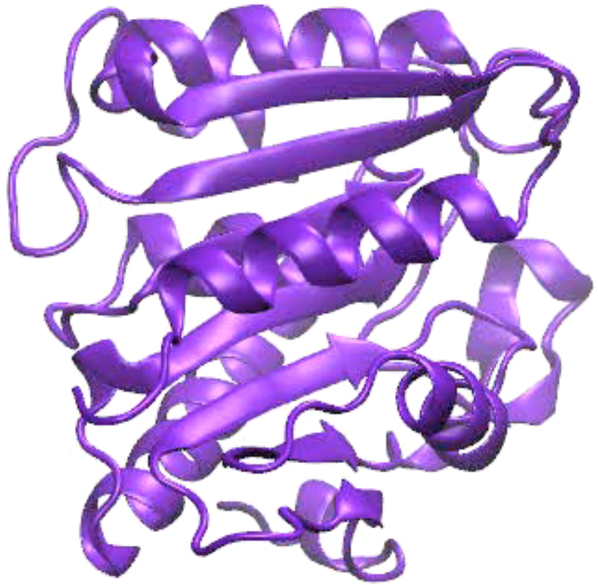
Outline of lecture 9:

- multiscale simulations
- integrative modeling
- structure-based drug design
- protein design

molecular modeling
and simulations

$$\{x_i(t), y_i(t), z_i(t)\}_{i=1,\ldots,N}$$

solvation
pH
post-translational modifications
interactions network
temperature effects (k_BT)
.....



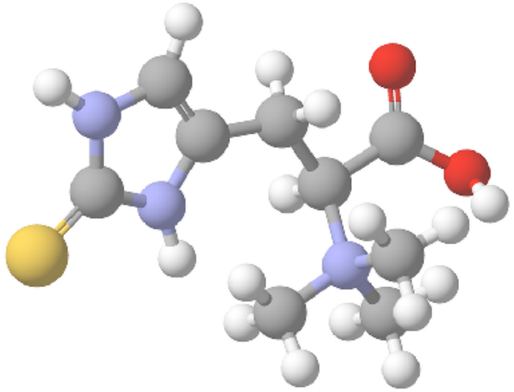
Current common MD engines

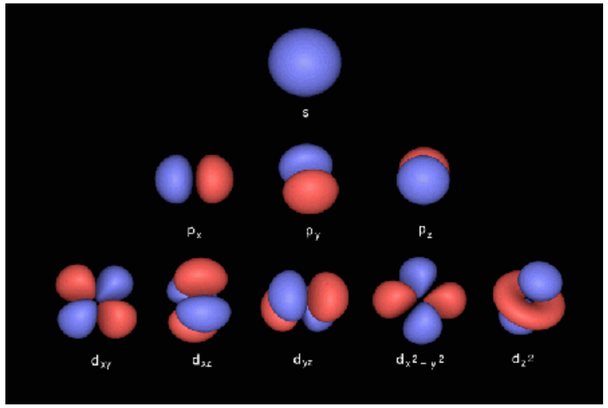
- **CHARMM**: Karplus Harvard, http://www.charmm.org/
- AMBER: Kollman UCSF, http://ambermd.org/
- GROMOS: van Gunsteren, ETHZ, www.igc.ethz.ch/GROMOS/index

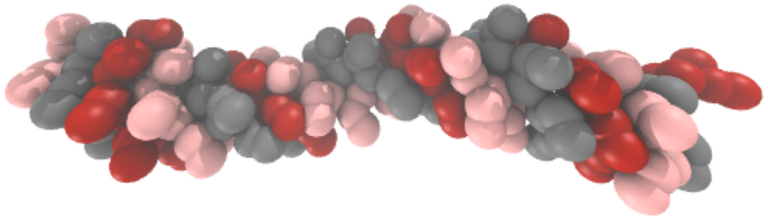
- DESMOND: Shaw, http://www.deshawresearch.com/
- GROMACS: http://www.gromacs.org
- LAMMPS: http://lammps.sandia.gov
- ACEMD: http://multiscalelab.org/acemd
- NAMD: http://www.ks.uiuc.edu/Research/namd/

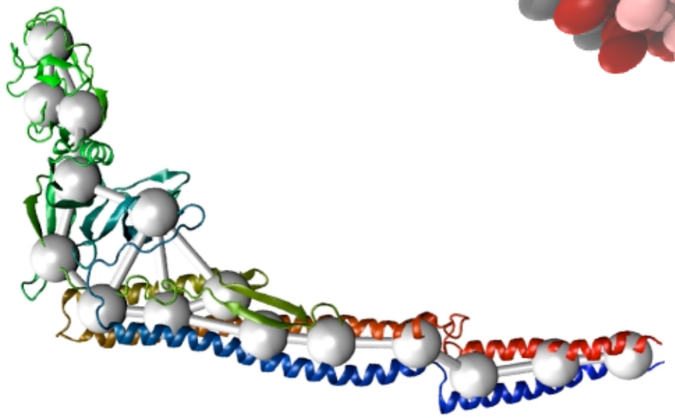
Multiscale resolution in modeling

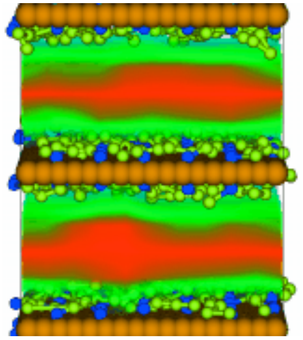
- electrons
- atoms
- amino-acids
- domains
- mesoscopic to continuum







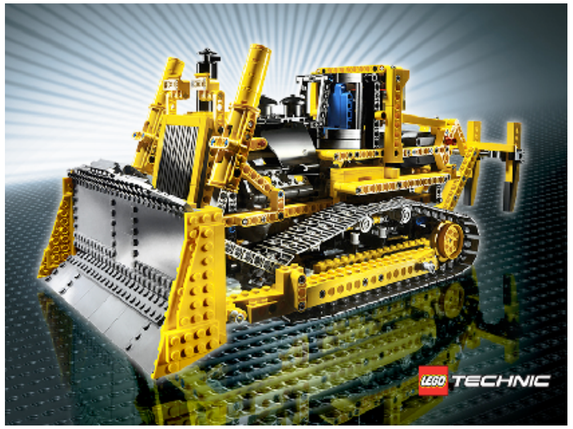




Building blocks

size/sampling

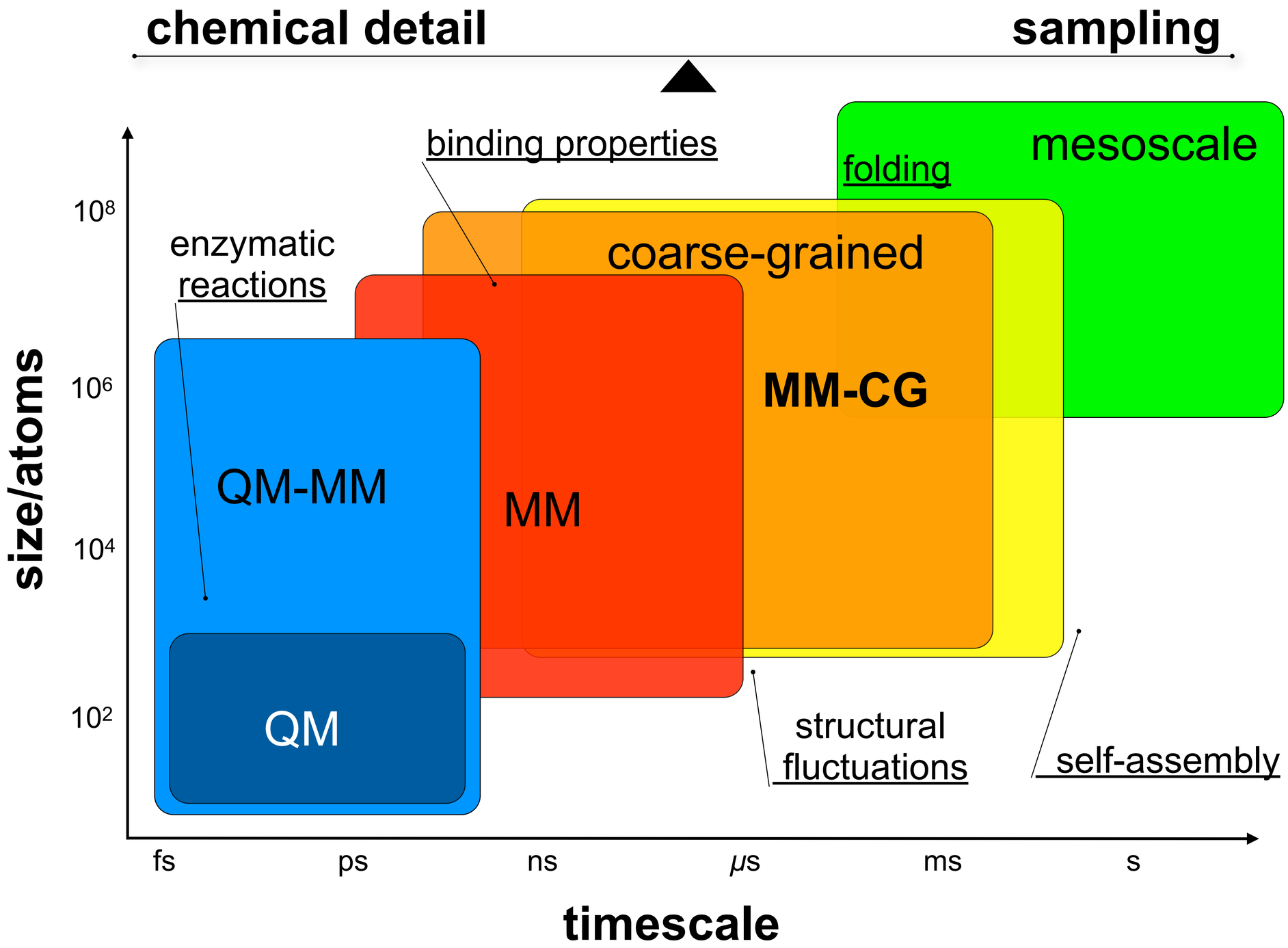
electrons



atoms

domains

accuracy^{-1}


chemical detail
sampling
binding properties
folding
mesoscale
enzymatic
reactions
coarse-grained
MM-CG
QM-MM
MM
QM
structural
fluctuations
self-assembly
size/atoms
10^8
10^6
10^4
10^2
fs
ps
ns
μs
ms
s
timescale

Speeding up timescales of Chemical Reactions

- **Enzymes** enhance the rate of chemical reactions by several orders of magnitude (e.g. arginine decarboxylase, alkaline phosphatase, staphylococcal nuclease **up to 10^{14} fold**)
- the transition rate depends on the activation barrier

$$\Gamma_{reactants \rightarrow products} \propto e^{-G_{barrier}/k_B T}$$

- and enzymes affect this, not the R and P states

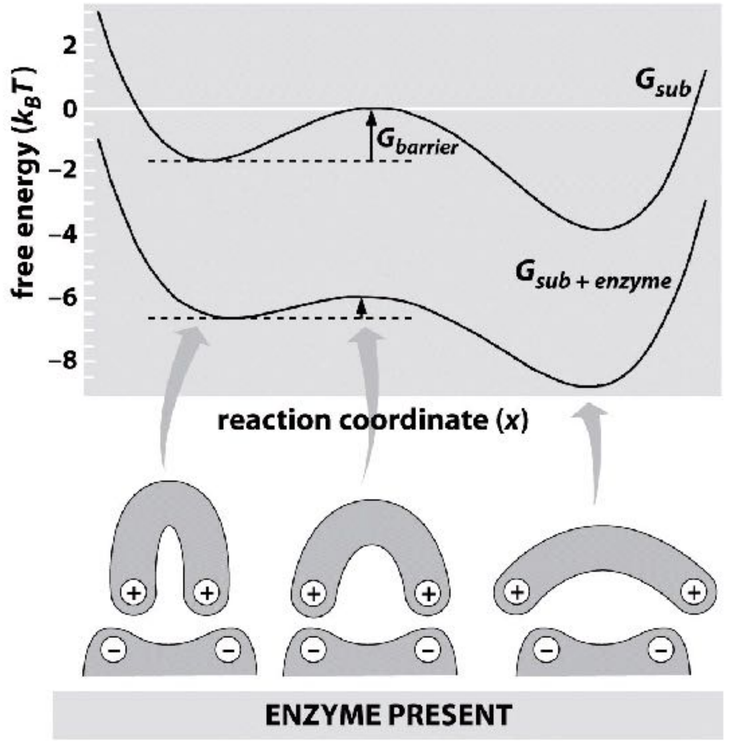
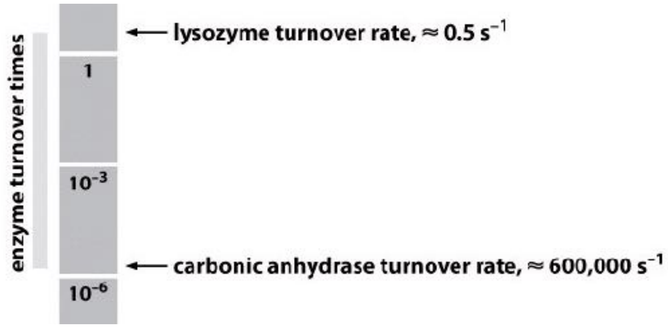
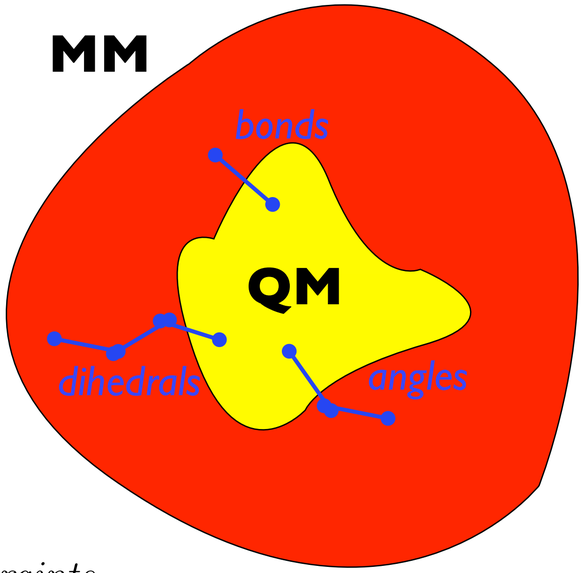


Figure 3.24b Physical Biology of the Cell (© Garland Science 2009)



Hybrid QM/MM molecular dynamics

$$H = H_{QM} + H_{MM} + \underbrace{H_{QM/MM}}_{\textit{coupling term}}$$



QM: First principles Density functional theory MD

$$\mathcal{L}_{\mathrm{CP}} = \underbrace{\sum_I \frac{1}{2} M_I \dot{\mathbf{R}}_I^2 + \sum_i \frac{1}{2}\mu_i \left\langle \dot{\psi}_i \middle| \dot{\psi}_i \right\rangle}_{\text{kinetic energy}} - \underbrace{\langle \Psi_0 | \mathcal{H}_{\mathrm{e}} | \Psi_0 \rangle}_{\text{potential energy}} + \underbrace{\textit{constraints}}_{\text{orthonormality}}$$

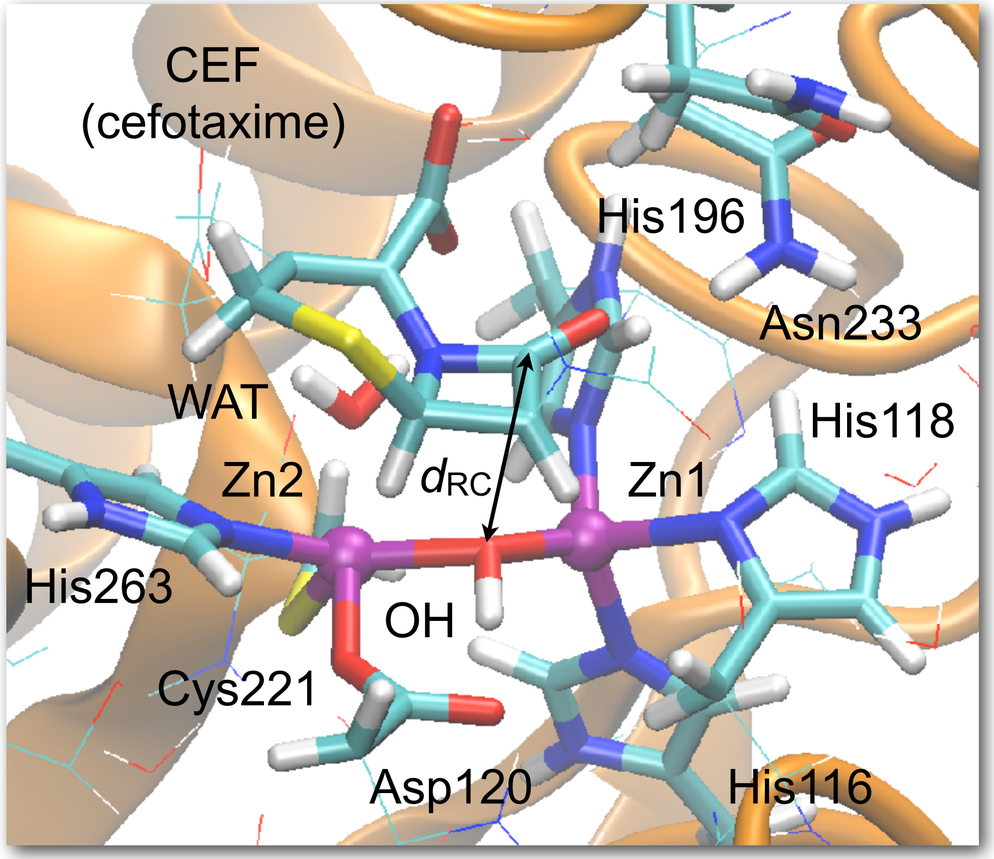
MM: Classical molecular dynamics (*e.g.* AMBER, Gromos force fields)

QM/MM: - boundary atom (*ad hoc* monovalent pseudopotential or H capping)

- hierarchical scheme to compute Coulomb interactions

Car, Parrinello, *PRL* 1985, Laio, Vandevondele, Rothlisberger *JCP* 2004, Dal Peraro *et al.*, *Curr. Opin. Struct. Biol.* 2007

CcrA MβL from *Bacteroides fragilis*



Thermodynamic integration along the reaction coordinate d_{RC}
DFT-BLYP, Martins-Troullier PPs, 70 Ry cutoff,
Nose' thermostat at 300 K,
2 reactions pathways for a total of ~150 ps trajectory

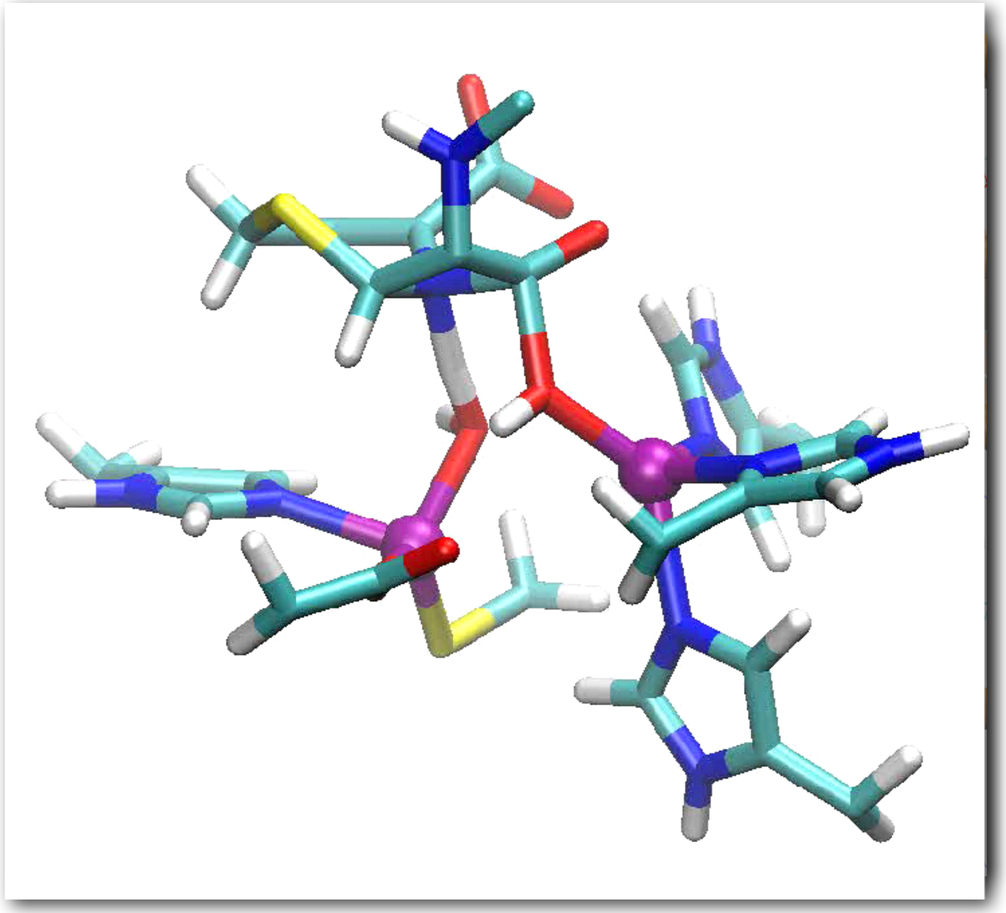
Reactant state

CcrA complexed with cefotaxime

- stable Michaelis complex
  OH-β-lactam distance=3.3(2)Å
  during 5 ns MD and 20ps QM/MM
- Zn2-bound WAT is the
  only water between the
  zinc center and CEF in 5Å

➡Classical force-field based MD is used as a tool to sample conformational space within the nanosecond timescale

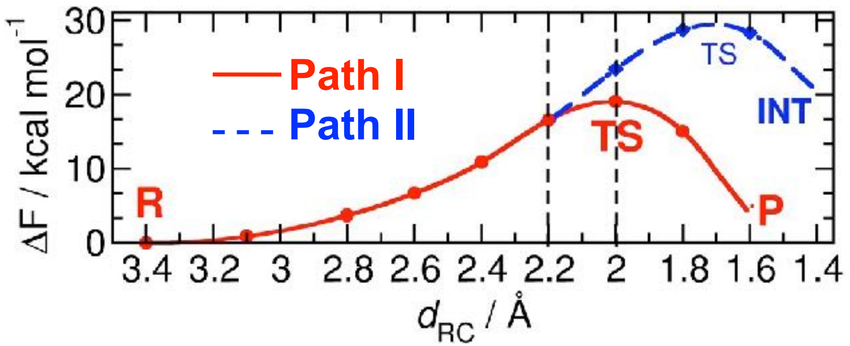
... from transition state to products



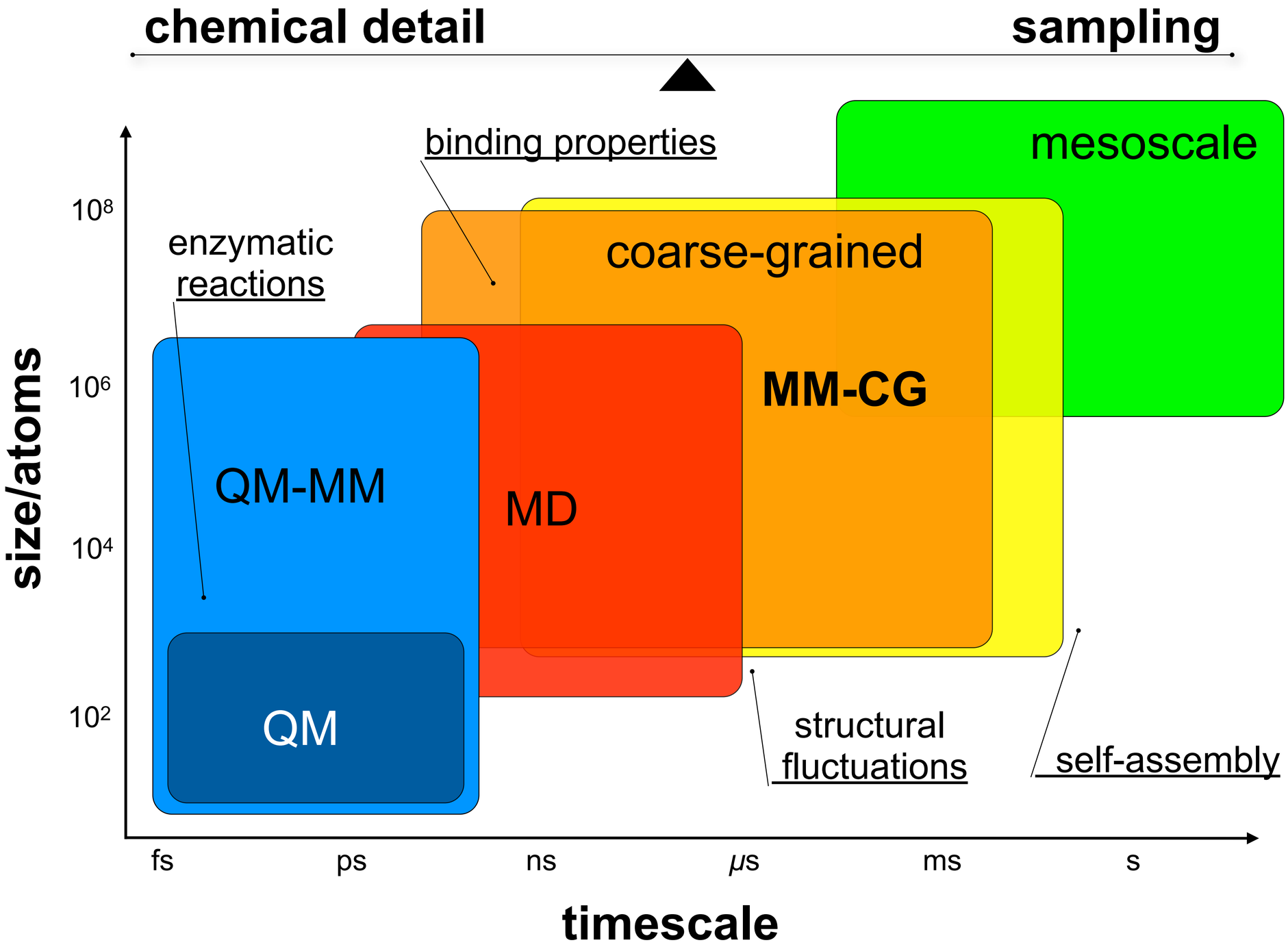
water-mediated single-step

- OH^- loses Zn2 coordination
- Zn1, Zn2 flexibility
- WAT protonates β-lactam N
- N-C β-lactam bond breaks
- WAT replaces OH^- as an hydroxide

- **ΔF = 18(2) kcal/mol** is in good agreement with experiments
- if Asn233 *does* H-bond β-lactam: formation of a high unfavorable intermediate (Path II)



Dal Peraro *et al.*, *JACS* 2007


chemical detail
sampling
binding properties
mesoscale
enzymatic
reactions
coarse-grained
MM-CG
QM-MM
MD
QM
structural
fluctuations
self-assembly
size/atoms
10^8
10^6
10^4
10^2
fs
ps
ns
μs
ms
s
timescale

Coarse-graining degrees of freedom

- **CG** is the process of consistently reduce the complexity of your problem integrating out degrees of freedom which can be in principle neglected for your system.

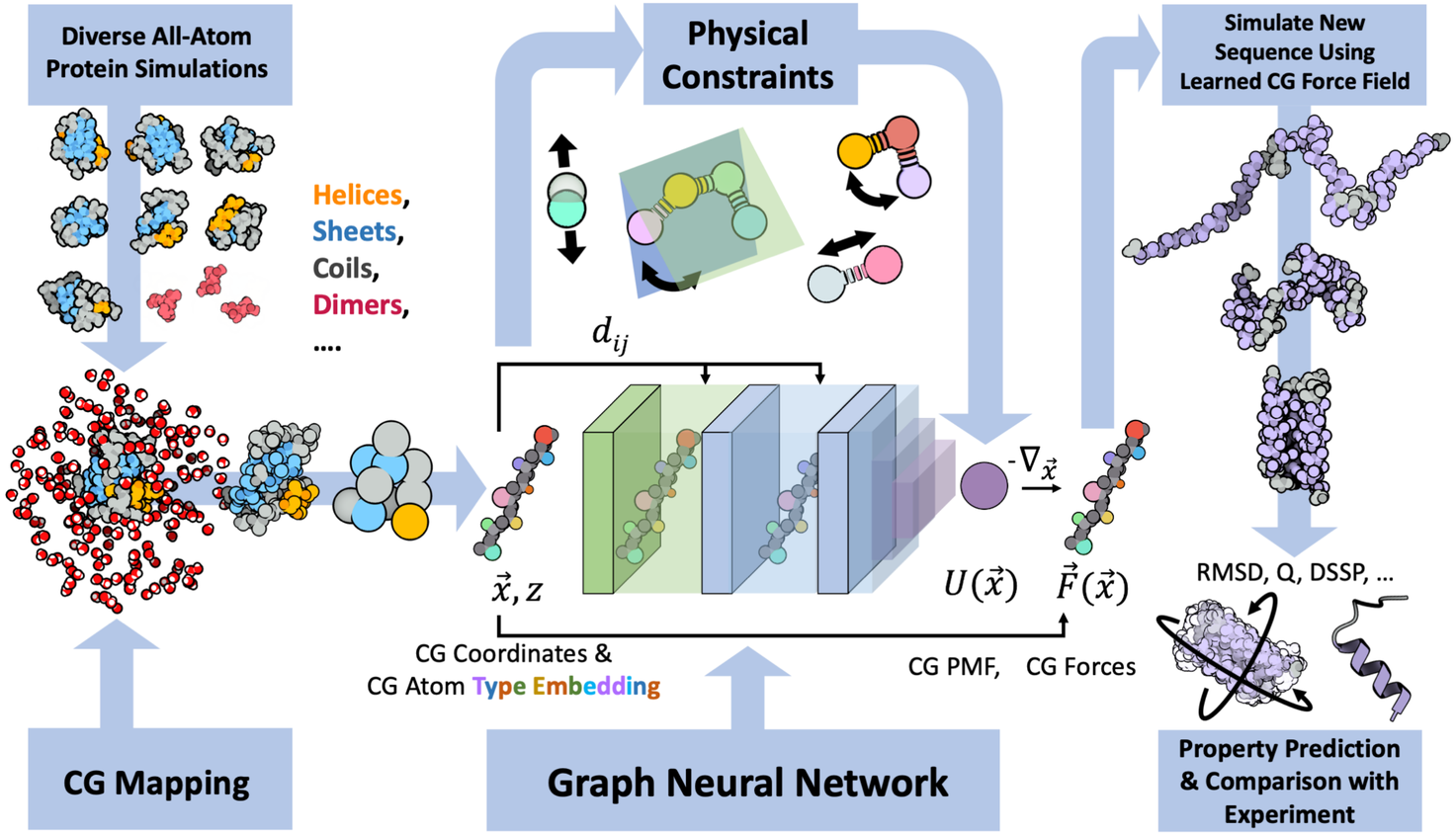
$$V_{QM} \rightarrow V_{MM} \rightarrow V_{CG-MM} \rightarrow V_{mesoscopic}$$

- the CG process implies a **simplification** of your potential that is not always rigorous and includes **approximations**

- what you obtain is an **effective** potentials which is parametrized to reproduce given properties

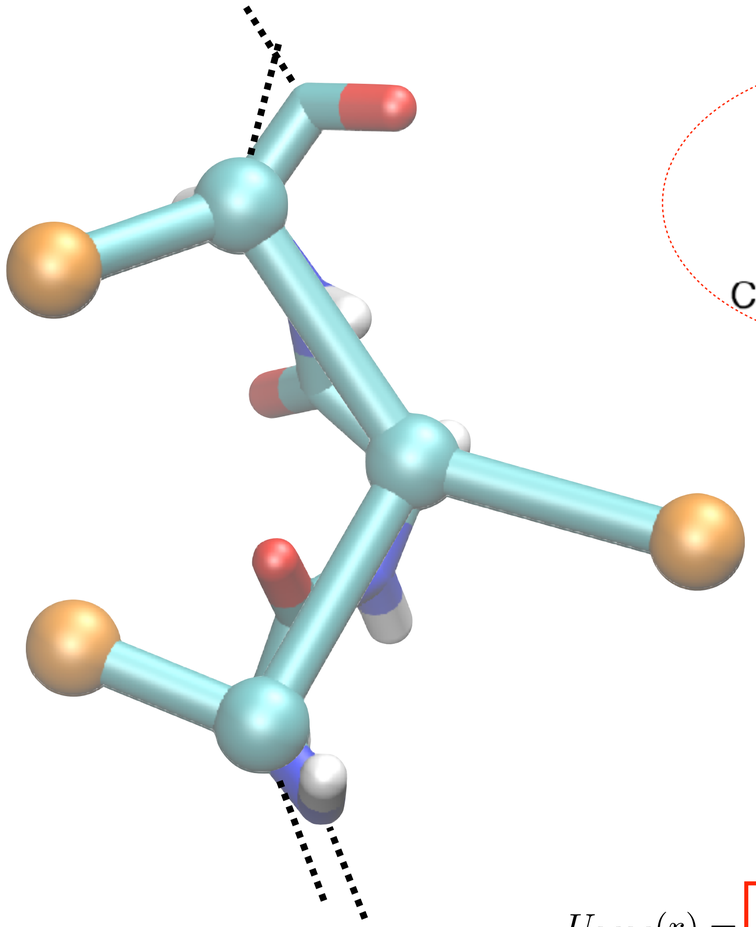
New directions

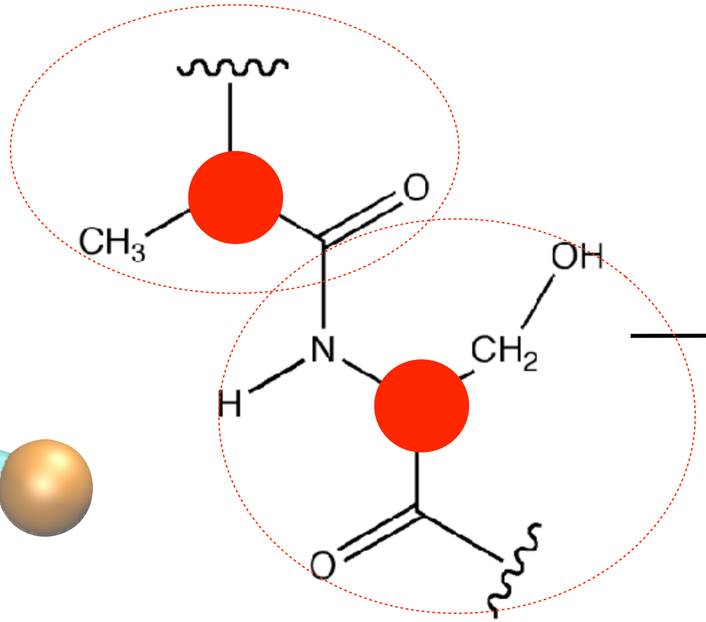
universal and computationally efficient machine-learned CG model for proteins

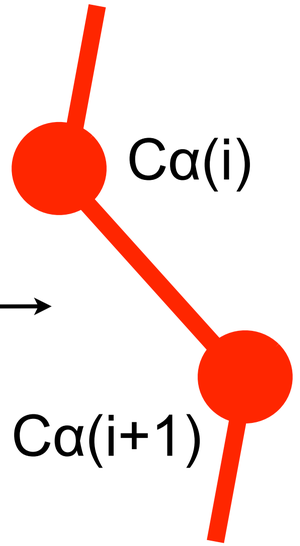


arXiv:2310.18278v1 [q-bio.BM] 27 Oct 2023

Coarse-graining degrees of freedom





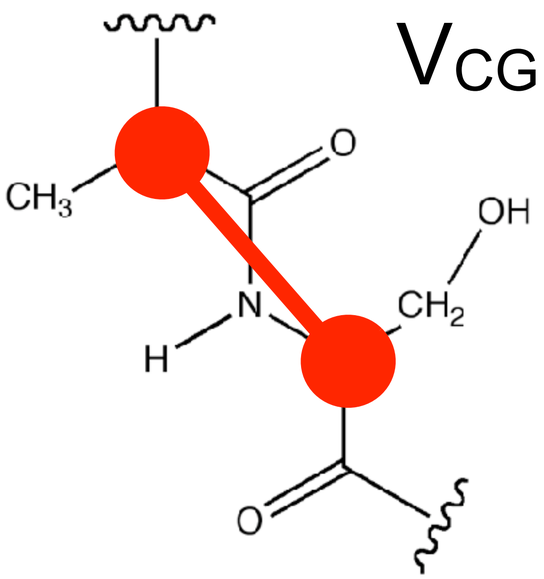


UMM **UCG**

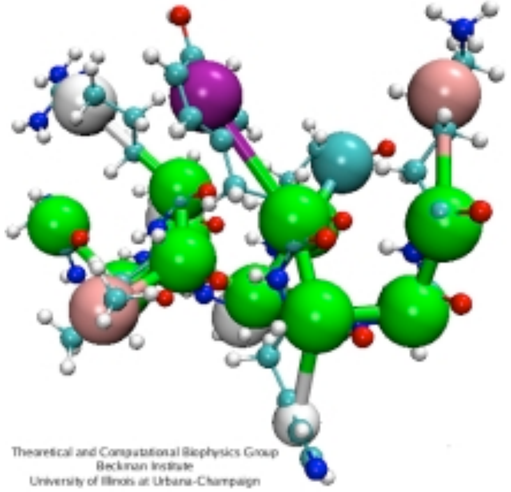
$$U_{MM}(r) = \sum_{bonds} \frac{k_b}{2}(r - r_0)^2 + \sum_{angles} \frac{k_\theta}{2}(\theta - \theta_0)^2 + \sum_{torsions,n} \frac{k_{\phi,n}}{2}[1 + cos(n\phi - \delta)] +$$

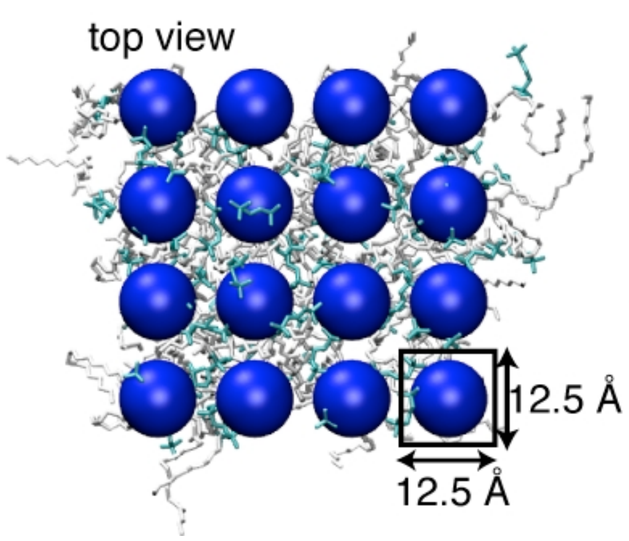
$$+ \sum_{i>j}^{N} \left(\frac{A}{r_{ij}^{12}} - \frac{C}{r_{ij}^{6}} \right) + \sum_{i>j}^{N} \frac{1}{4\pi\epsilon_0} \frac{q_i q_j}{r_{ij}}$$

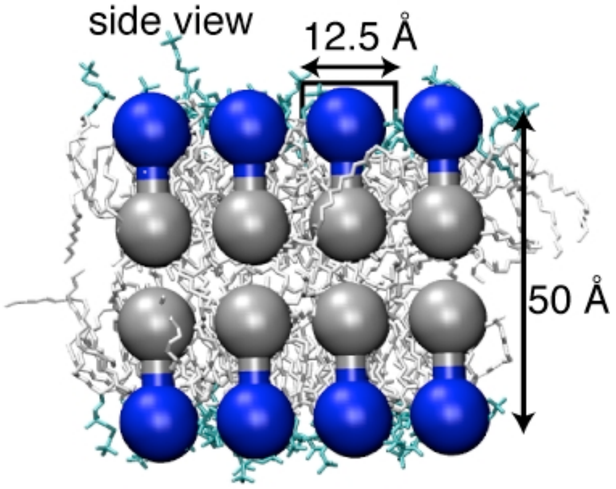
Coarse-grained force fields



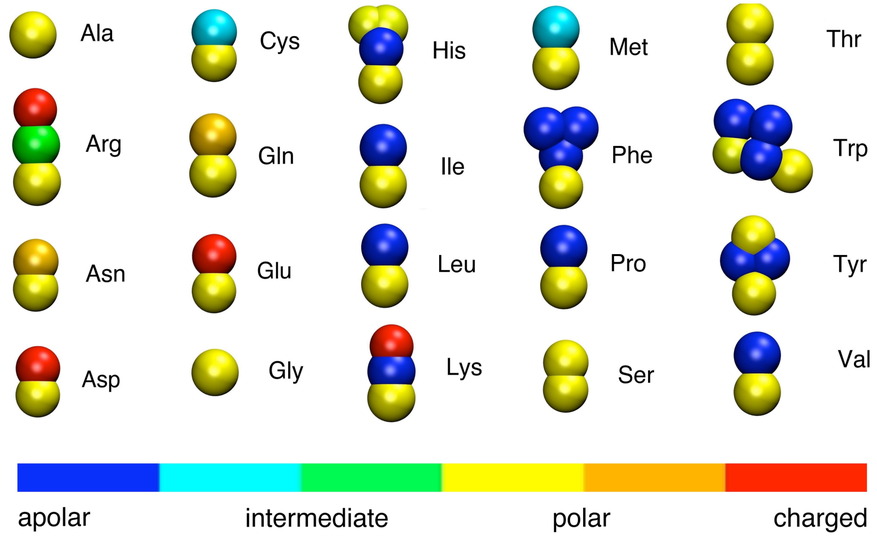
- CG FF models are not topologically biased on the native structure
- softer interactions allow for **longer** timestep in MD simulations
- sampling on the **millisecond** timescale
- accuracy can be a problem (e.g. **no explicit electrostatic** contribution)
- biases on the secondary structures





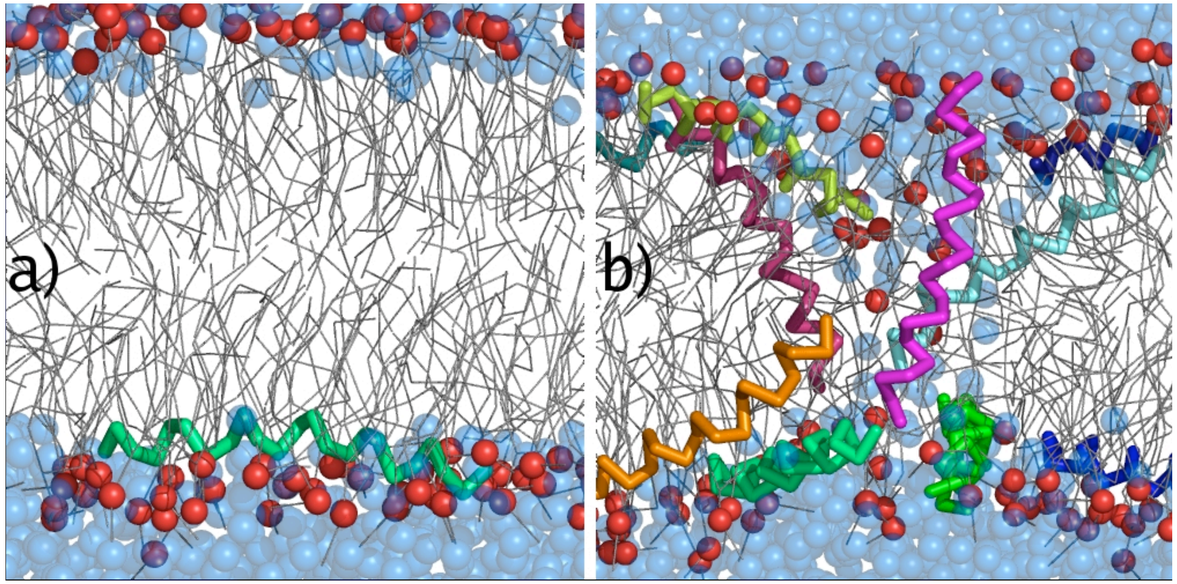


Coarse-grained MARTINI FF

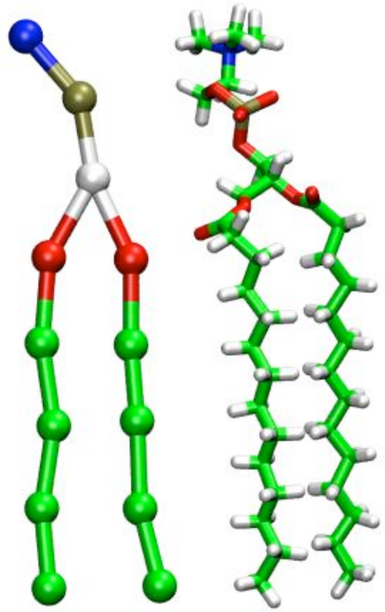


- MARTINI CG FF has functional form similar to MM FF
- 4-to-1 mapping from MM to CG
- very convenient for membranes and peptide-membrane interactions

Monticelli et al, JCTC 2008
Klein and coworkers



Magainin H2 in a DPPC bilayer, at low concentration (a) and high concentration



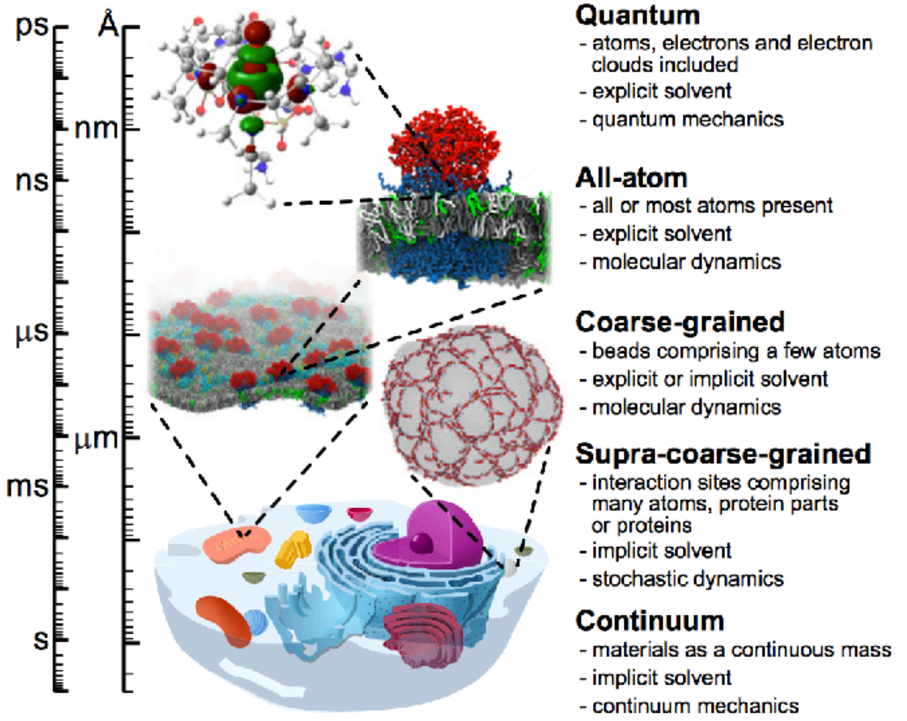
© 2016. Published by The Company of Biologists Ltd | Journal of Cell Science (2016) 129, 257-268 doi:10.1242/jcs.176040

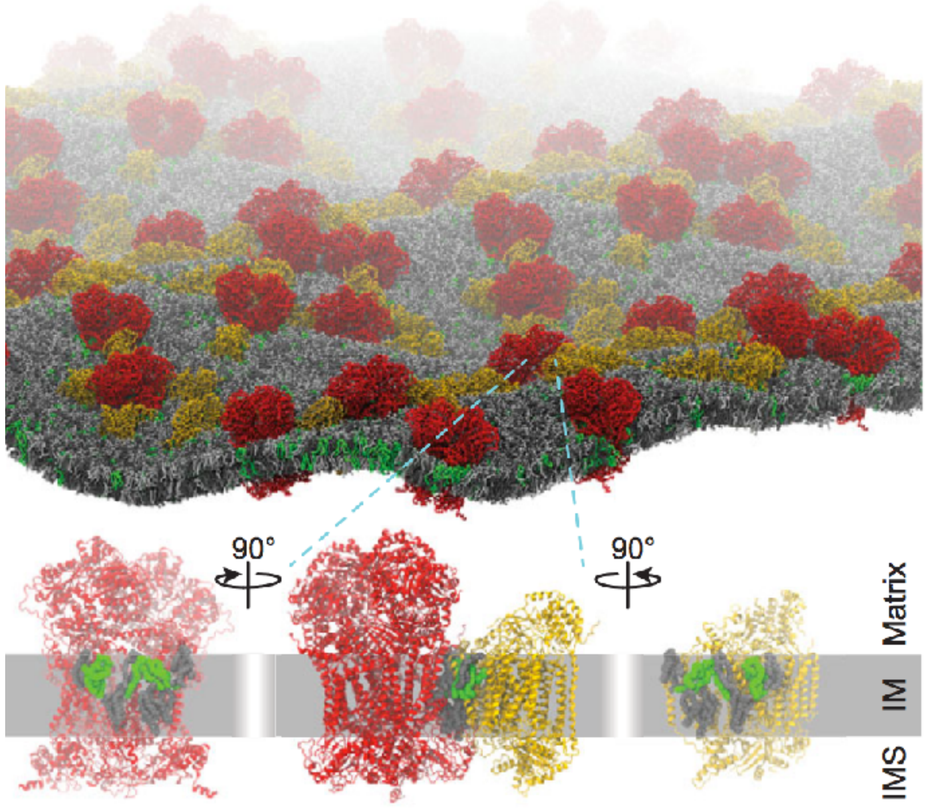
COMMENTARY

ARTICLE SERIES: IMAGING

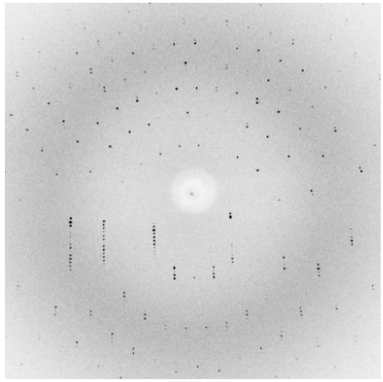
Computational 'microscopy' of cellular membranes

Helgi I. Ingólfsson, Clément Arnarez, Xavier Periole and Siewert J. Marrink*

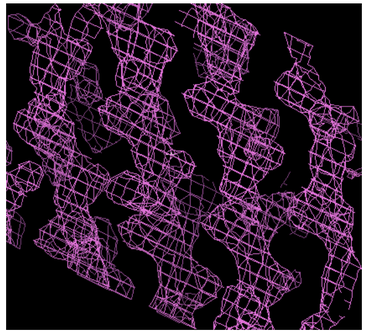


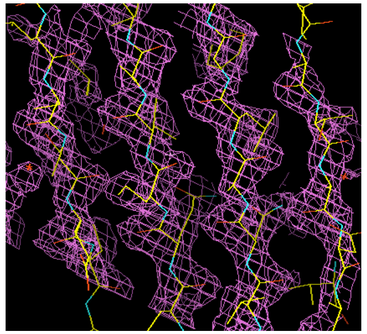


Structural biology methods are strongly based on theory and computation!



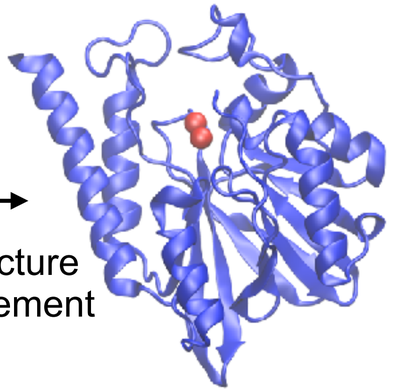
Fourier transform

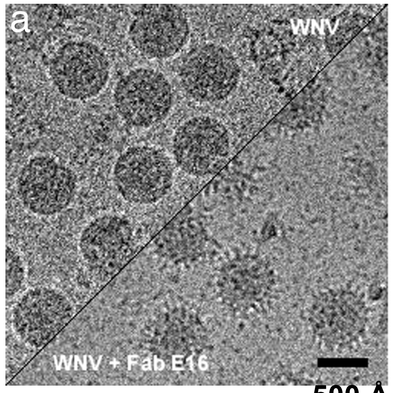


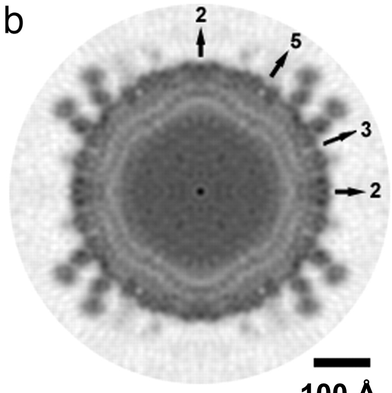


molecular replacement

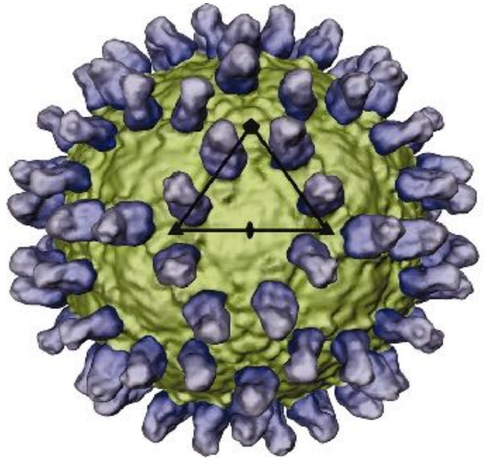
structure refinement







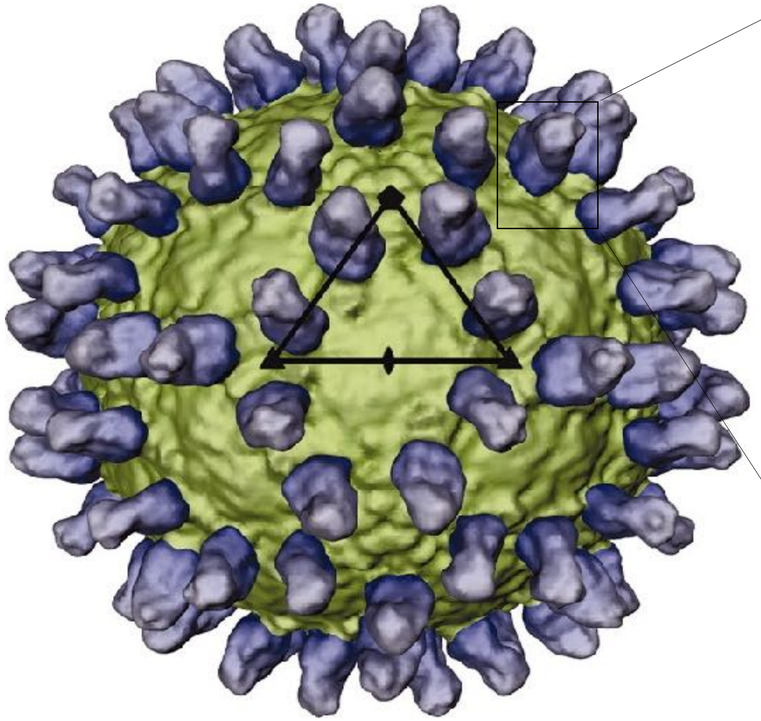
3D image reconstruction



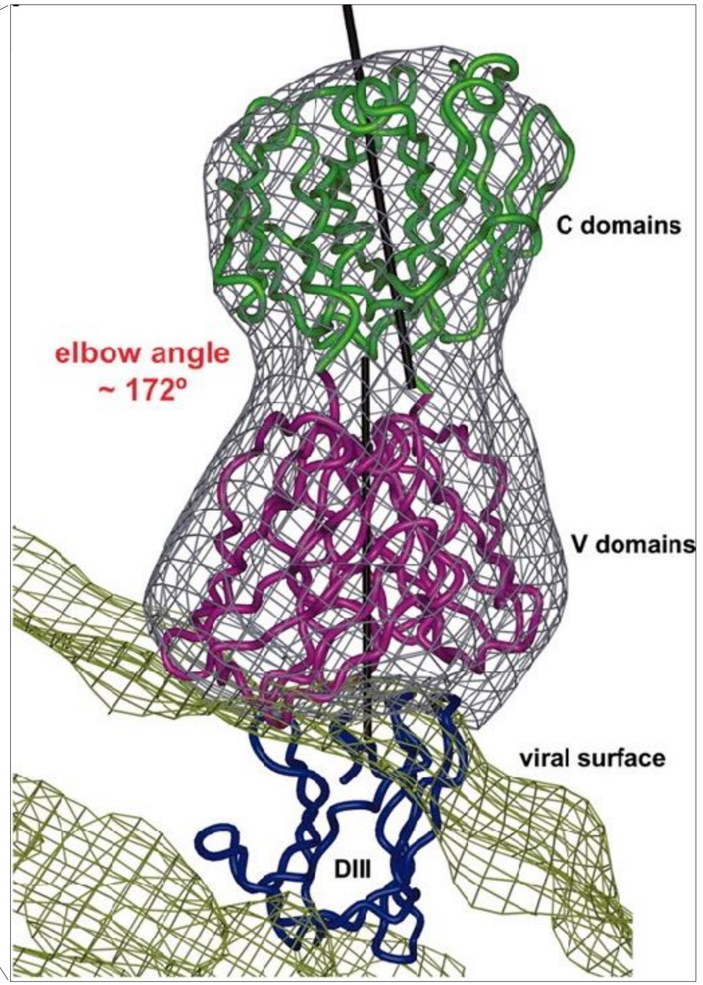
Kaufmann B et al. PNAS 2006;103:12400-12404

Combine cryo-EM and X-ray structures

surface rendering of the 3D image reconstruction of WNV (green) in complex with Fab E16 (blue) at 15-Å resolution



integrative modeling

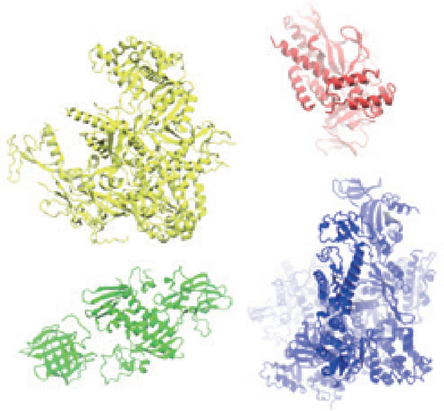


Kaufmann B et al. PNAS 2006;103:12400-12404

Integrative Modeling

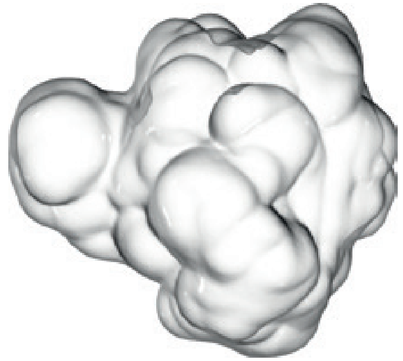
Individual subunits

- X-ray crystallography
- NMR
- Cryo-EM
- Homology models



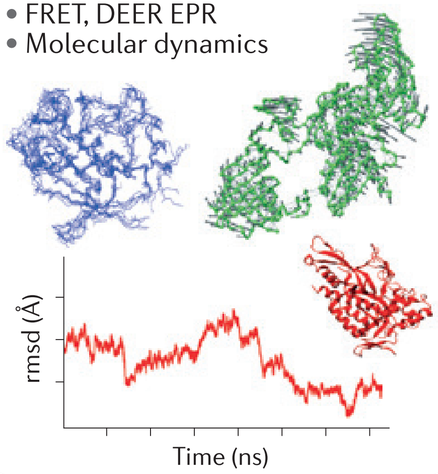
Volumetric maps

- Cryo-EM
- Electron tomography
- SAXS, SANS
- AFM



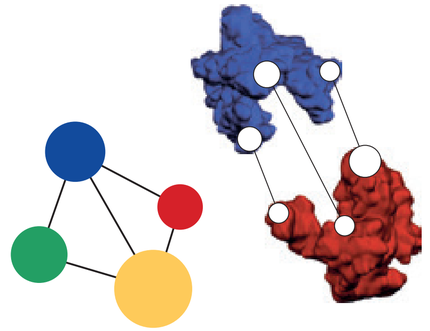
Structural flexibility

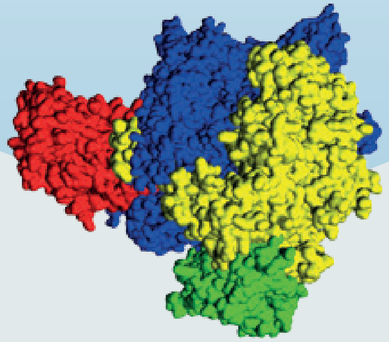
- Side-chain and backbone sampling
- Elastic network models
- NMR ensembles
- FRET, DEER EPR
- Molecular dynamics



Spatial connectivity

- Mutagenesis
- Evolutionary couplings
- Chemical crosslinking
- Proteomics
- H/D exchange
- ChIP–seq and ChIP–exo
- 3C, 4C, 5C and Hi-C



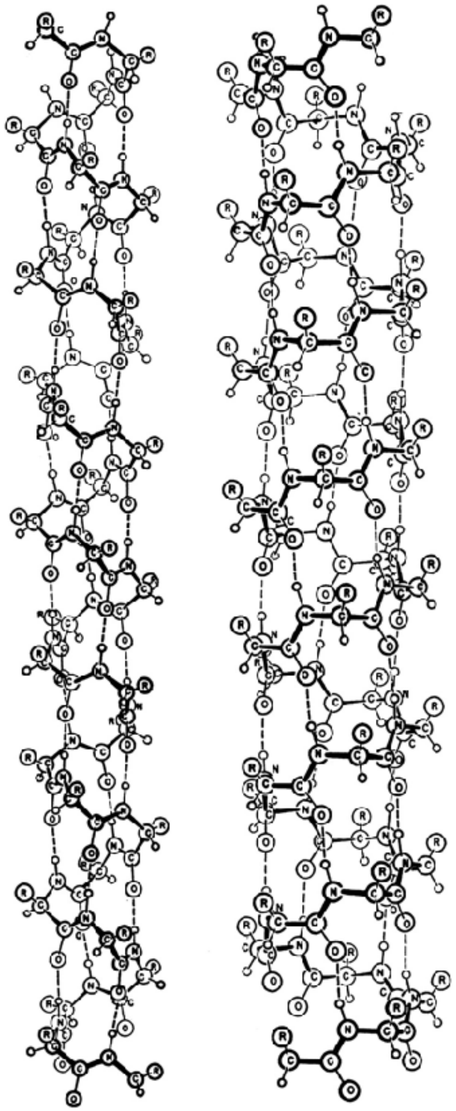


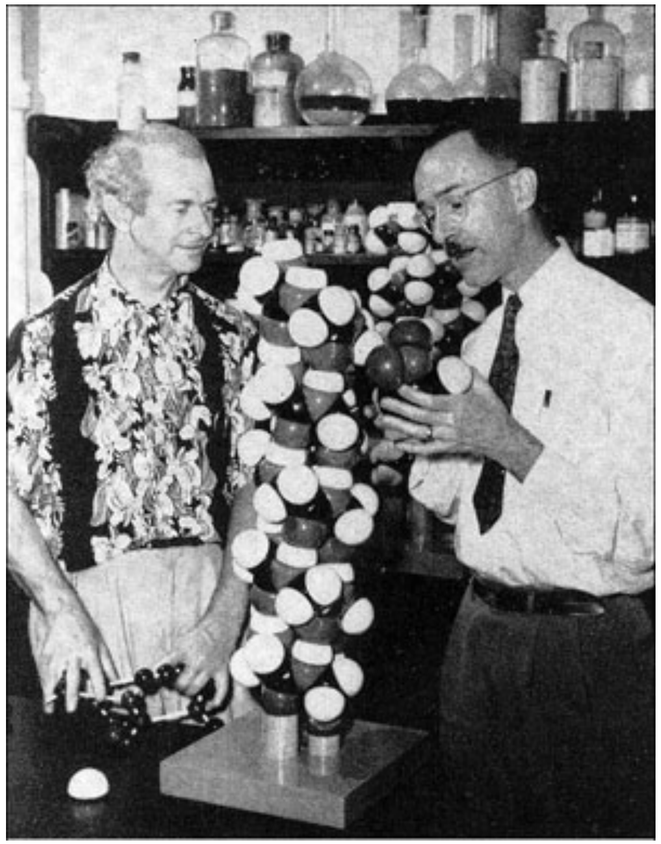
Near-atomic-resolution structure of supramolecular assemblies

3C, chromatin conformation capture; 4C, circularized 3C; 5C, carbon-copy 3C; AFM, atomic force microscopy; ChIP–exo, ChIP–seq with an exonuclease sample preparation step; ChIP–seq, chromatin immunoprecipitation followed by sequencing; DEER EPR, double electron–electron resonance electron paramagnetic resonance; FRET, fluorescence resonance energy transfer; H/D exchange, hydrogen–deuterium exchange; NMR, nuclear magnetic resonance; Hi-C, genome-wide 3C; rmsd, root-mean-square deviation; SANS, small-angle neutron scattering; SAXS, small-angle X-ray scattering.

Dal Peraro et al. **Nat. Rev. Microbiol.** 2016

(the dawn of) Integrative modeling

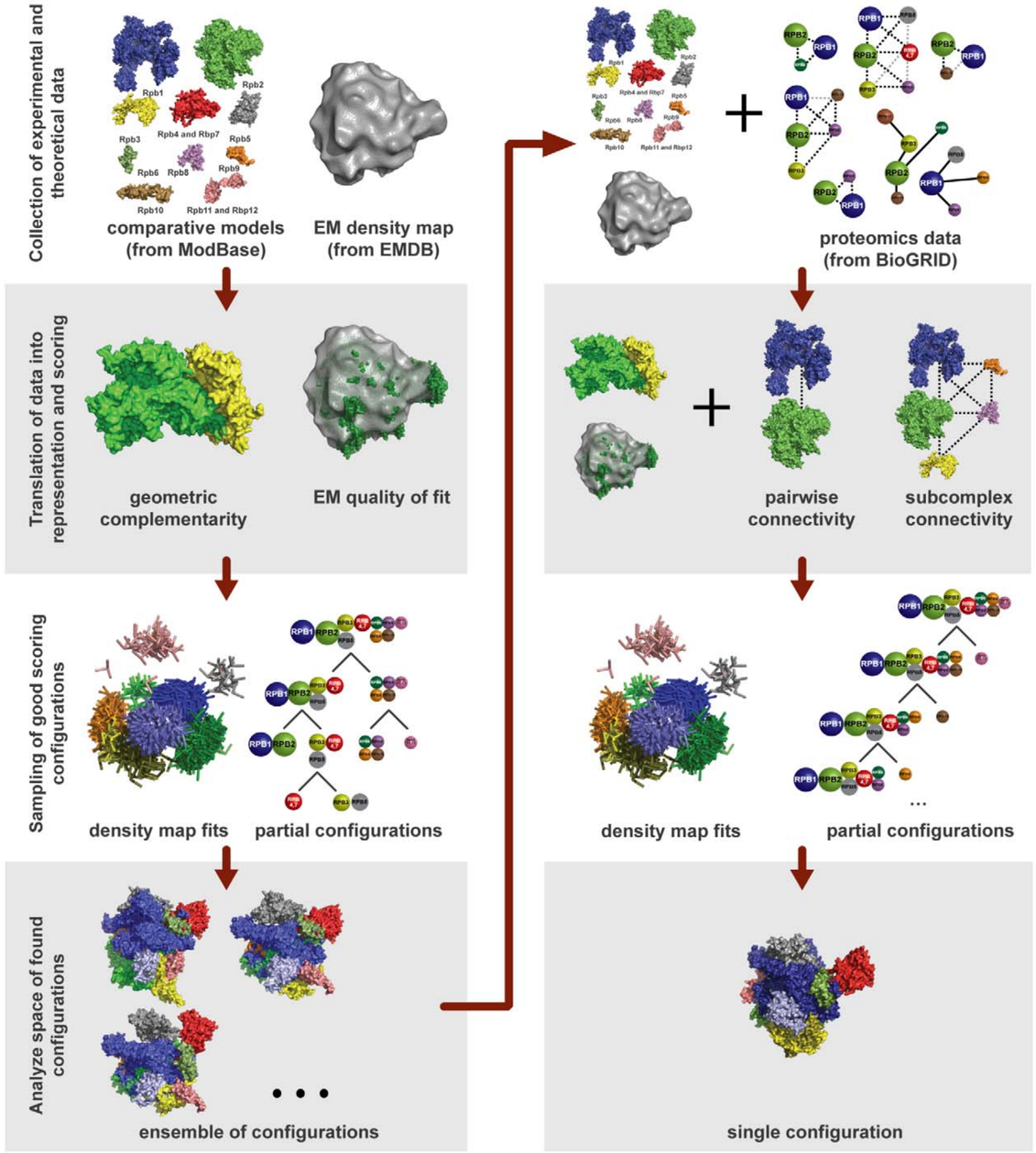




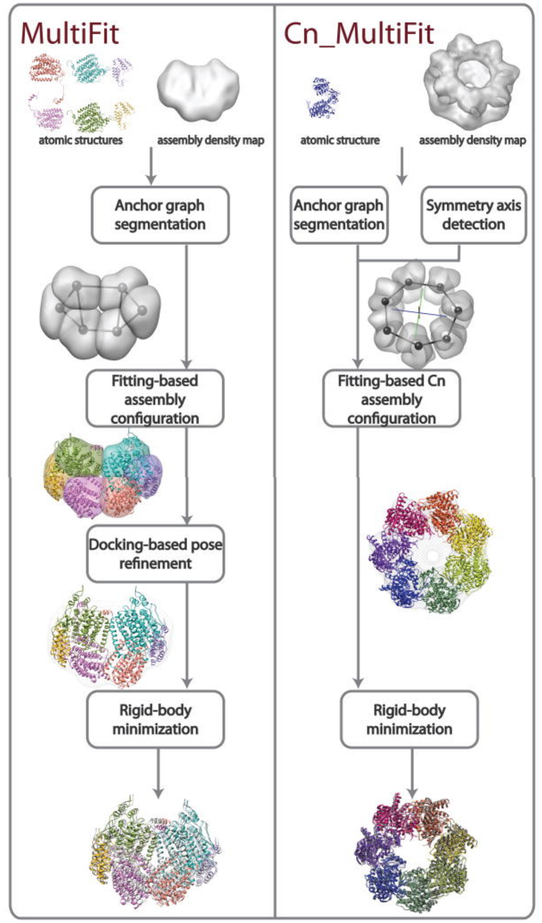
Pauling and Corey protein structure papers (1951):

X-ray myoglobin 1959 (Kendrew and Perutz)

Integrative modeling



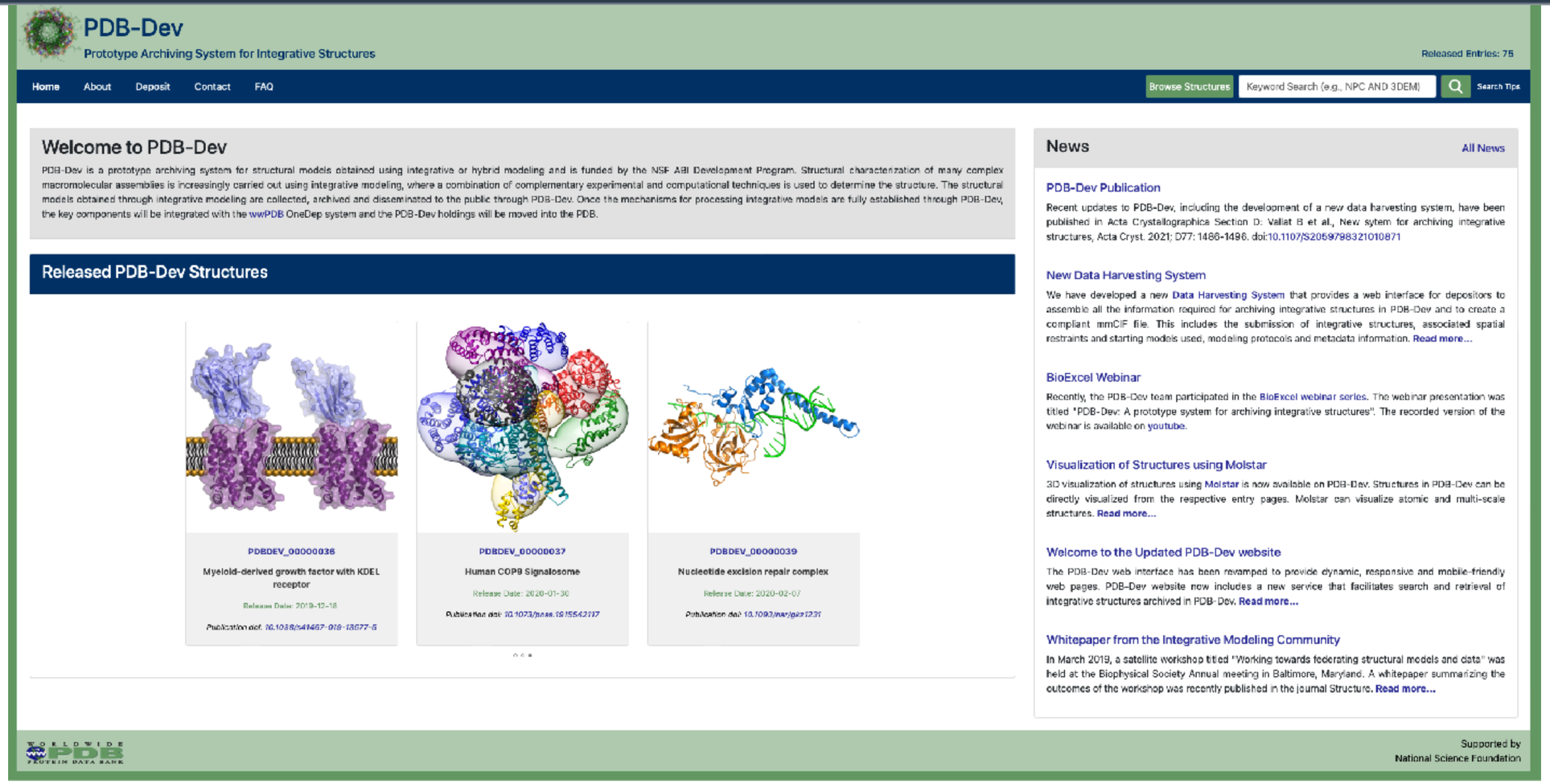
IMP: integrative modeling platform
http://salilab.org/imp/



https://modbase.compbio.ucsf.edu/multifit/

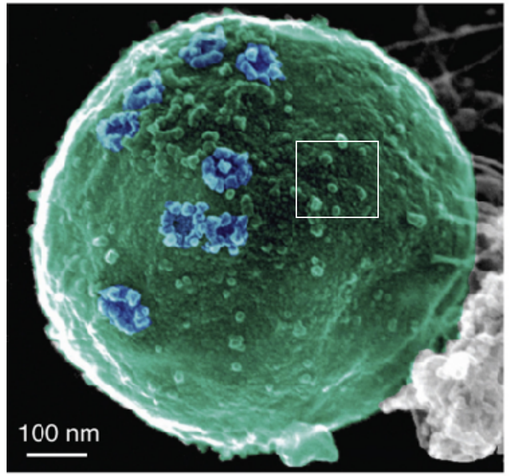
Russel et al. (2012) Putting the Pieces Together: Integrative Modeling Platform Software for Structure Determination of Macromolecular Assemblies.
PLoS Biol 10(1): e1001244. doi:10.1371/journal.pbio.1001244

https://pdb-dev.wwpdb.org

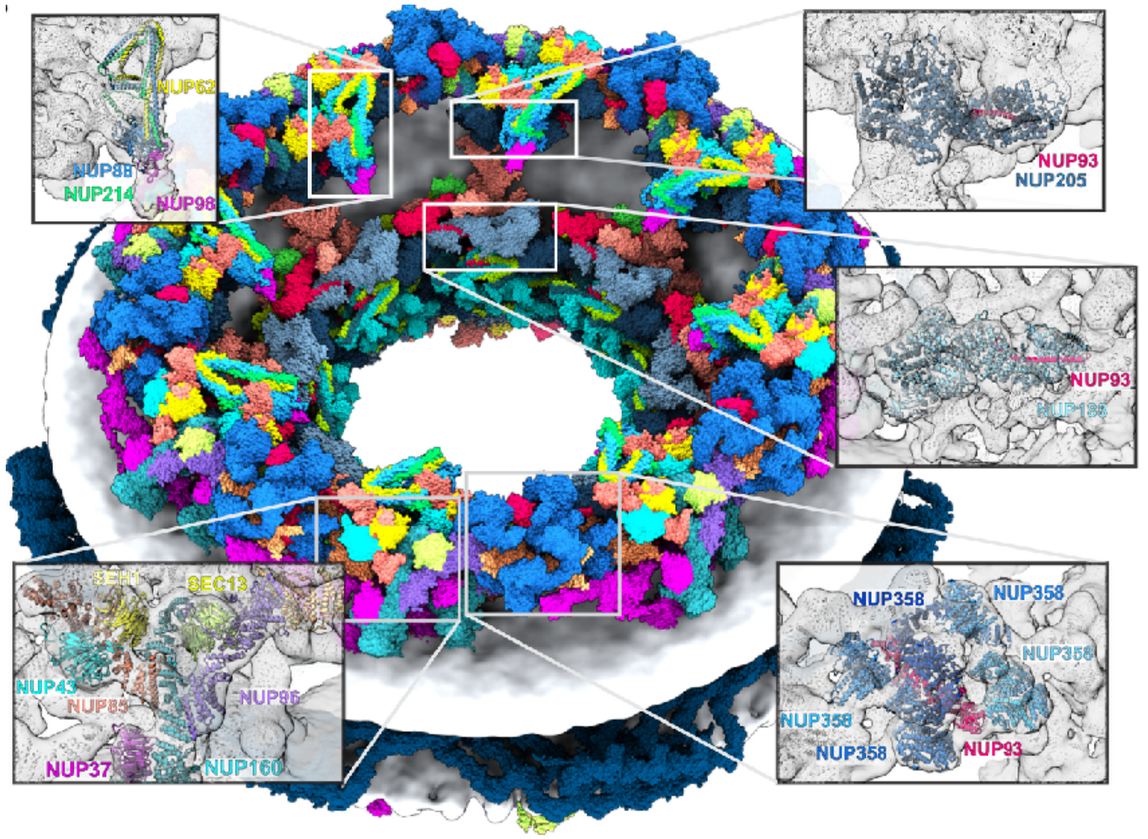


Sali *et al.*, **Structure** 2015

e.g., the Nuclear Pore Complex … combining AF and cryoEM/ET



yeast NPC :
~52 MDa complex
~550 protein subunits of ~30 different types



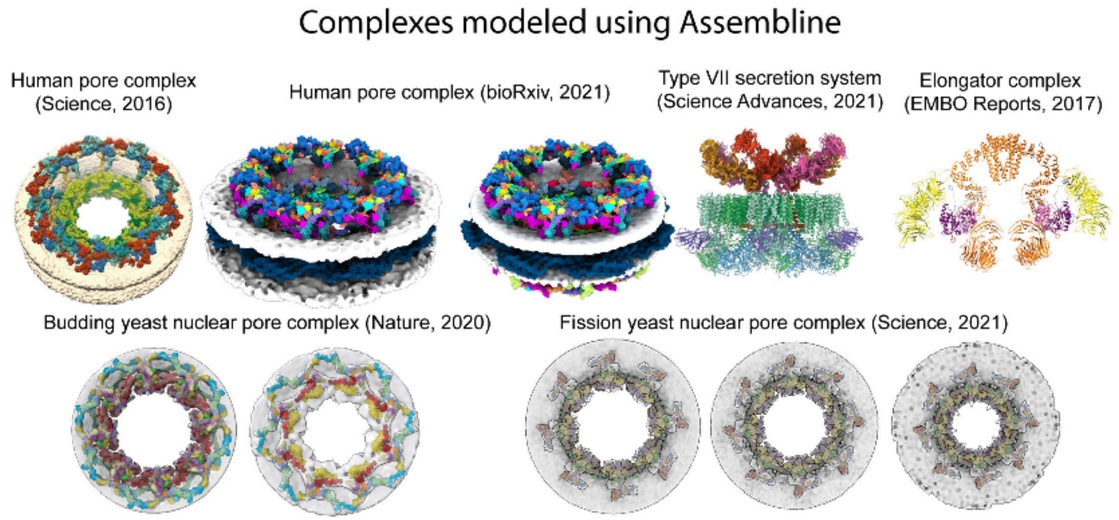
DOI: 10.1126/science.abm9506

EMBL

Assembline

Assembline is an assembly line of macromolecular assemblies!

Assembline is a multi-step protocol for integrative structural modeling of macromolecular complexes based on electron microscopy, cross-linking mass spectrometry and other data. The protocols is based on our Xlink Analyzer and external software: Integrative Modeling Platform (IMP), Python Modeling Interface (PMI), UCSF Chimera, and imp-sampcon. Comparing to other methods, Assembline enables efficient sampling of conformational space through a multi-step procedure, provides new modeling restraints, and includes a unique configuration system for setting up the modelling project.



Code

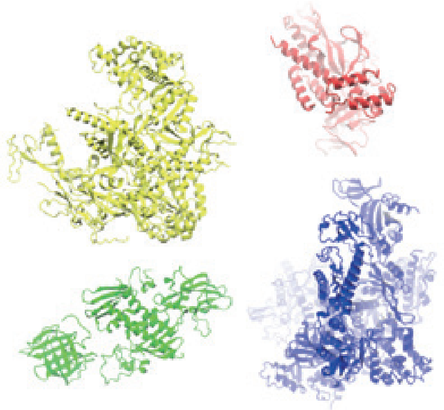
- Assembline installation package for Anaconda
- Assembline code
- Installation instructions and manual

Rantos V., Karius K., and Kosinski J. Integrative structural modeling of macromolecular complexes using Assembline, Nature Protocols, 2021

Integrative Modeling

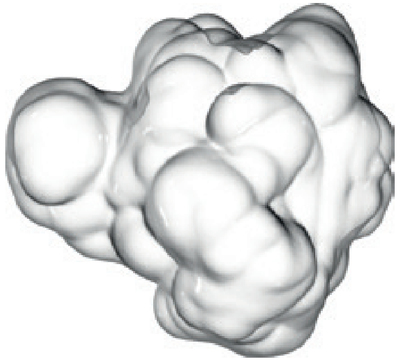
Individual subunits

- X-ray crystallography
- NMR
- Cryo-EM
- Homology models



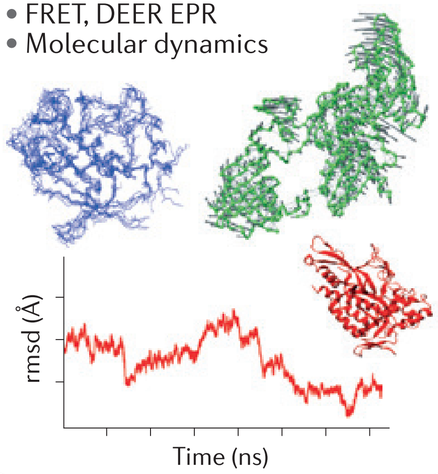
Volumetric maps

- Cryo-EM
- Electron tomography
- SAXS, SANS
- AFM



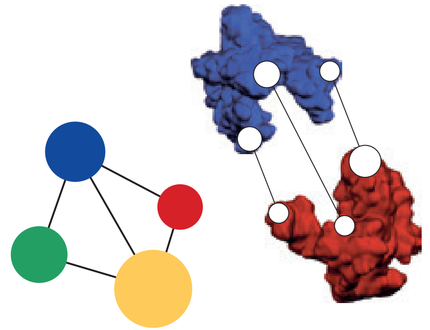
Structural flexibility

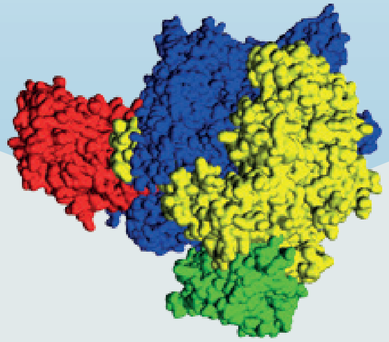
- Side-chain and backbone sampling
- Elastic network models
- NMR ensembles
- FRET, DEER EPR
- Molecular dynamics



Spatial connectivity

- Mutagenesis
- Evolutionary couplings
- Chemical crosslinking
- Proteomics
- H/D exchange
- ChIP–seq and ChIP–exo
- 3C, 4C, 5C and Hi-C



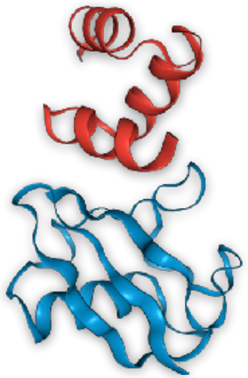


Near-atomic-resolution structure of supramolecular assemblies

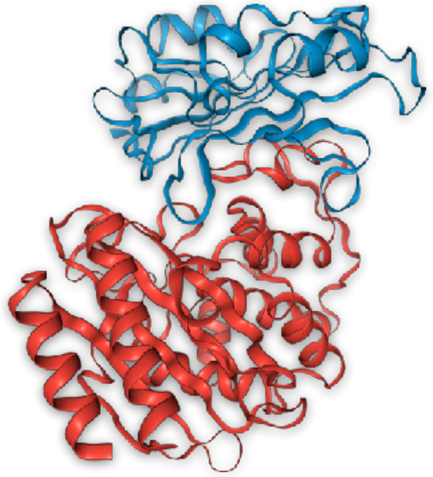
3C, chromatin conformation capture; 4C, circularized 3C; 5C, carbon-copy 3C; AFM, atomic force microscopy; ChIP–exo, ChIP–seq with an exonuclease sample preparation step; ChIP–seq, chromatin immunoprecipitation followed by sequencing; DEER EPR, double electron–electron resonance electron paramagnetic resonance; FRET, fluorescence resonance energy transfer; H/D exchange, hydrogen–deuterium exchange; NMR, nuclear magnetic resonance; Hi-C, genome-wide 3C; rmsd, root-mean-square deviation; SANS, small-angle neutron scattering; SAXS, small-angle X-ray scattering.

Dal Peraro et al. **Nat. Rev. Microbiol.** 2016

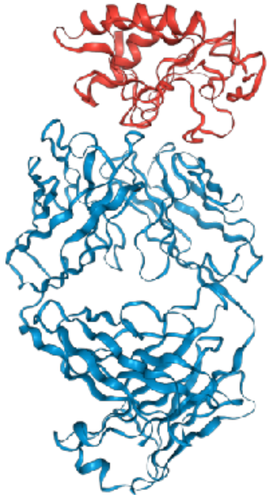
Protein-protein interaction interfaces



Enzyme-Substrate (2OOB)
Ubiquitin bound to ubiquitin ligase



Enzyme-Inhibitor (1JTG)
Beta-lactamase bound to beta-lactamase inhibitor

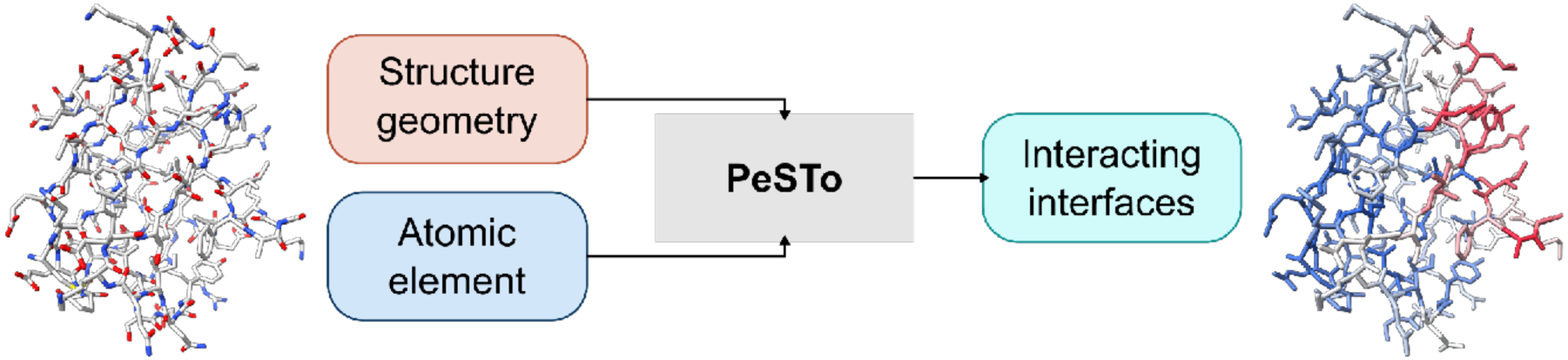


Antibody-Antigen (3MXW)
Sonic hedgehog bound to the 5E1 fab fragment

Berman H. M. et al., *The Protein Data Bank,* Nucleic Acids Research, 2000

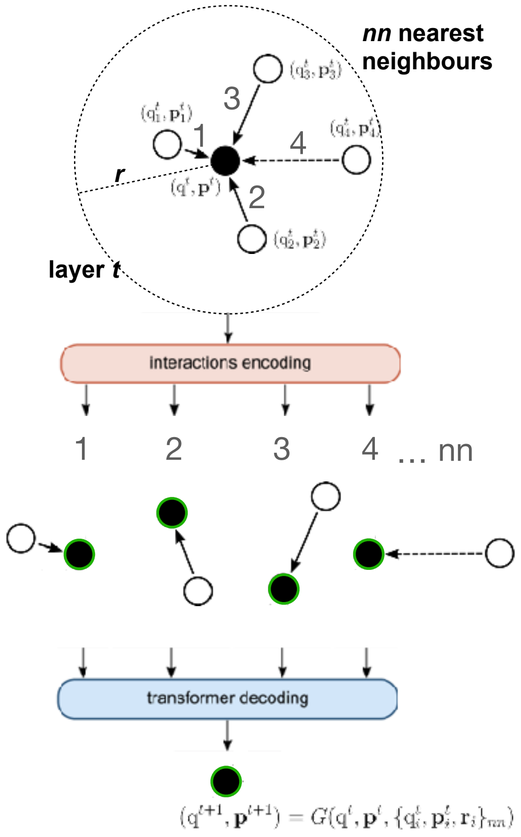
PeSTo: **P**rot**e**in **S**tructure **T**rans**f**o**r**mer

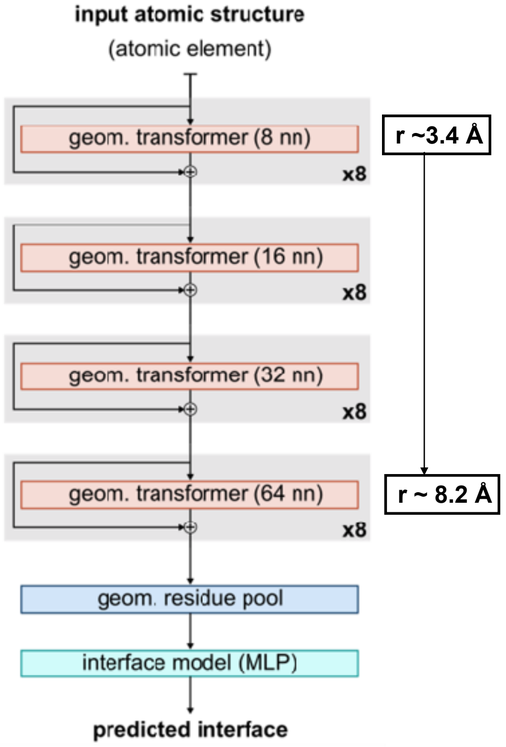
Lucien Krapp

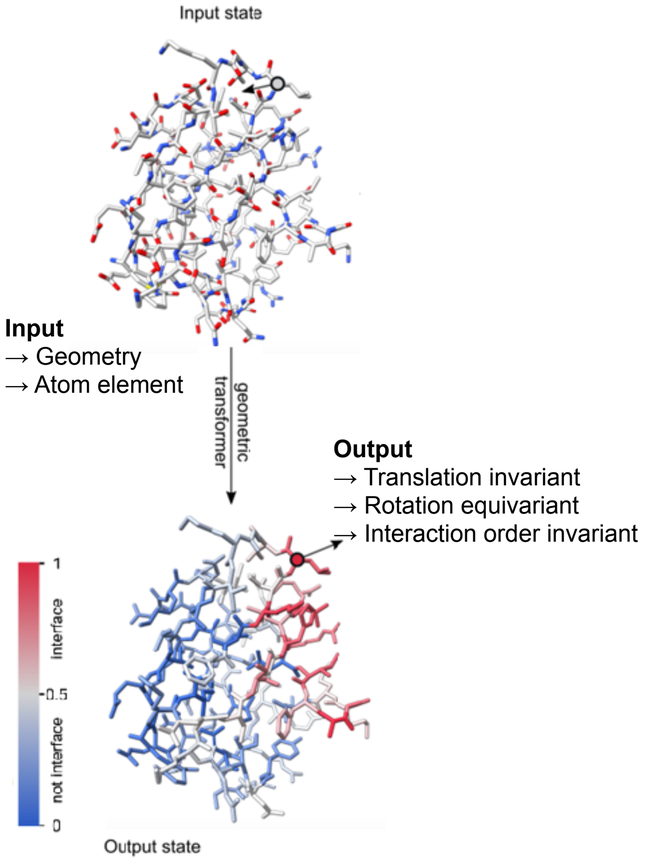


PeSTo: parameter-free geometric deep learning for accurate prediction of protein binding interfaces
Krapp L. F. *et al.* Nature Communications, 2023

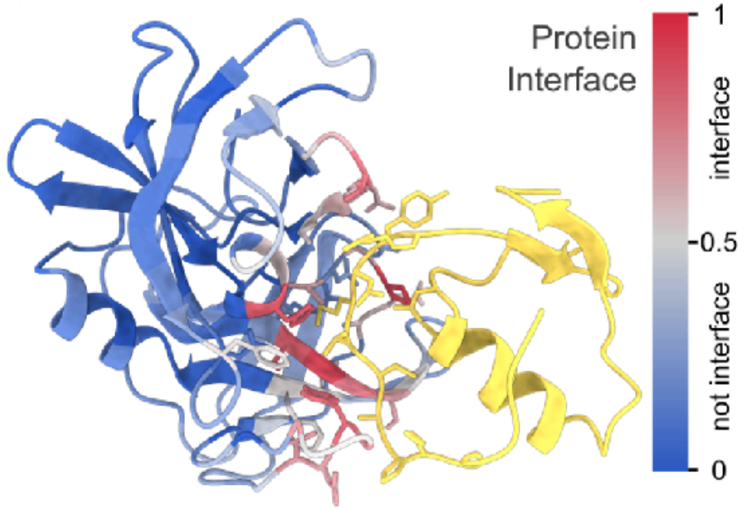
PeSTo: Protein Structure Transformer



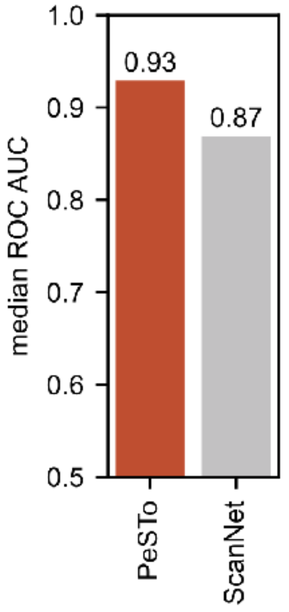


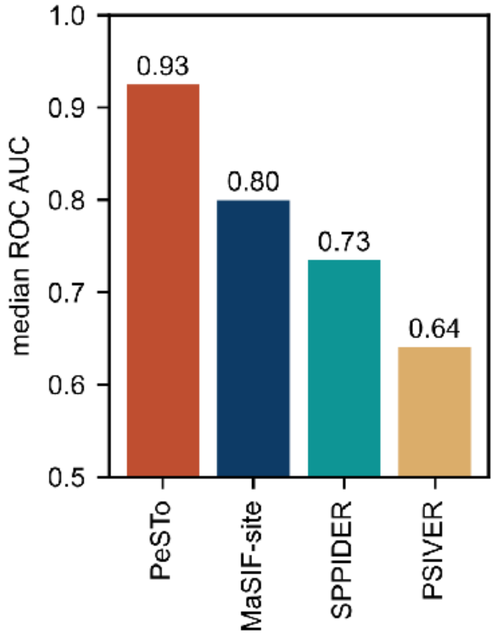


PeSTo for protein-protein interfaces prediction … comparison with state of the art



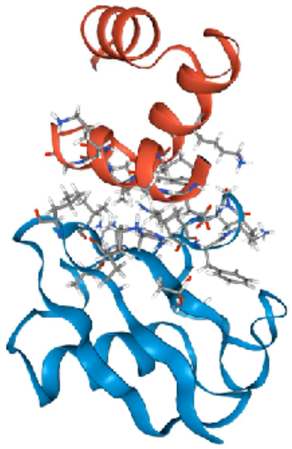
Streptogrisin B with ovomucoid - unbound conformation (0.93 Å RMSD) with a ROC AUC of 96%



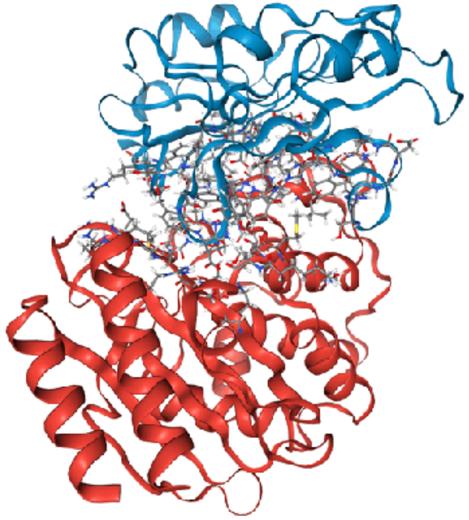


Gainza P. et al., *Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning,* Nature Methods, 2020
Tubiana J. et al., *ScanNet: an interpretable geometric deep learning model for structure-based protein binding site prediction,* Nature Methods, 2022

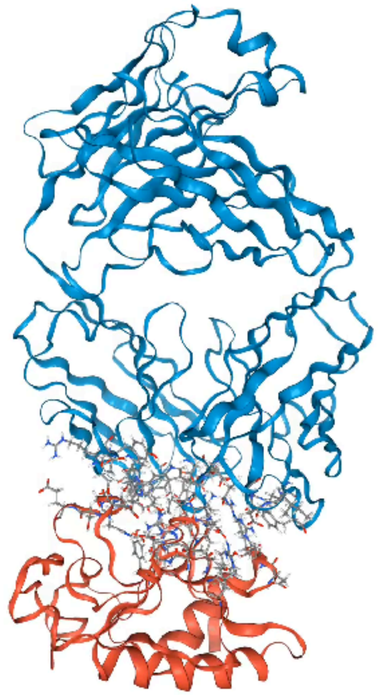
Interfaces are dynamic



Enzyme-Substrate (2OOB)
Ubiquitin bound to ubiquitin ligase



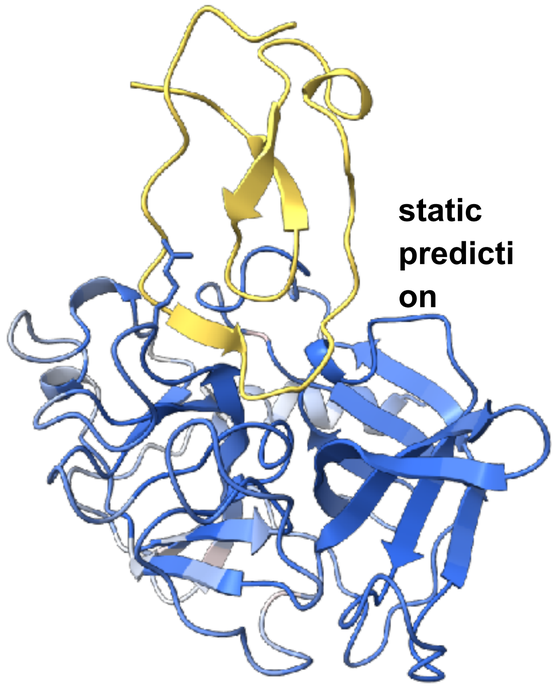
Enzyme-Inhibitor (1JTG)
Beta-lactamase bound to beta-lactamase inhibitor



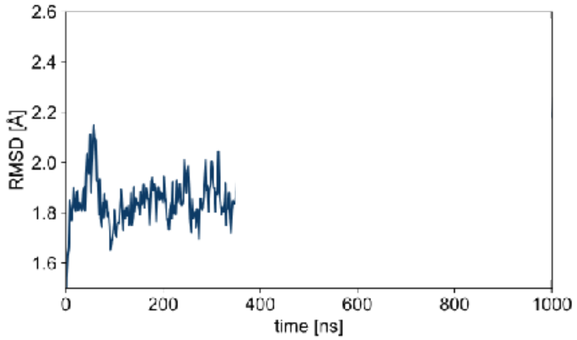
Antibody-Antigen (3MXW)
Sonic hedgehog bound to the 5E1 fab fragment

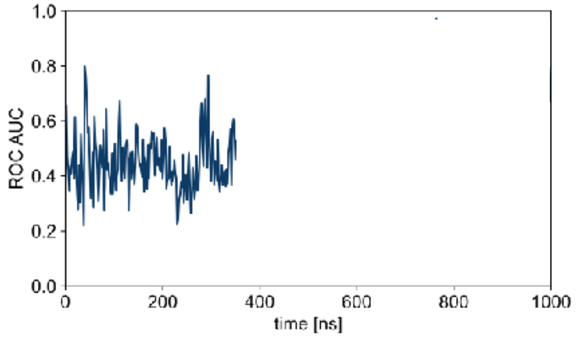
Interface recovery through molecular simulation

Elafin complexed with porcine pancreatic elastase (1FLE)



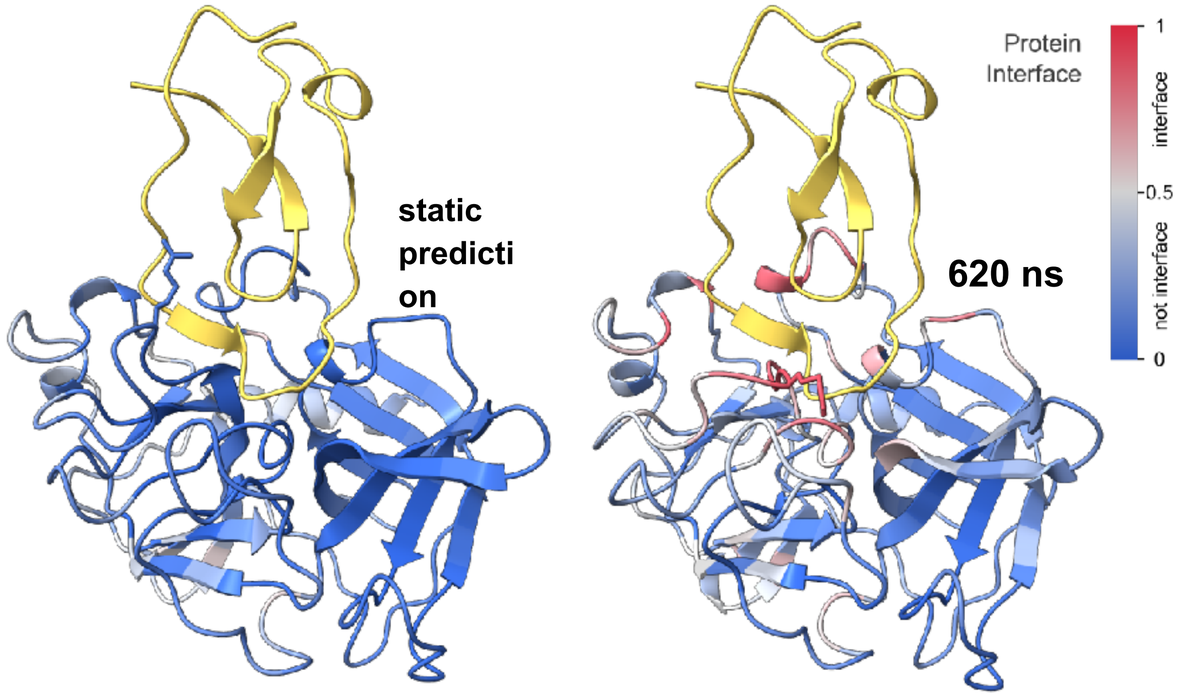
unbound porcine pancreatic elastase

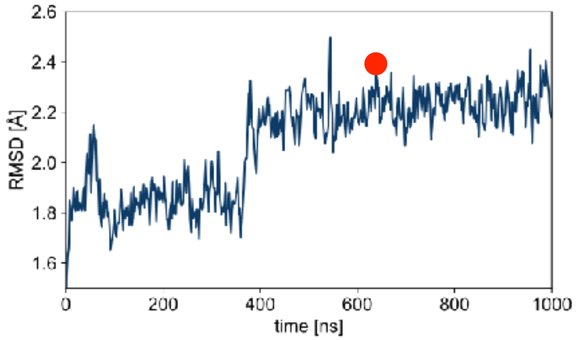


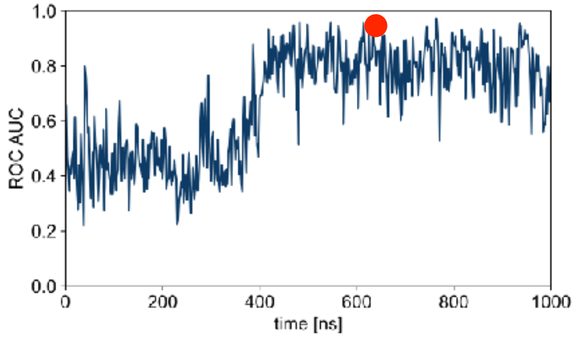


Interface recovery through molecular simulation

Elafin complexed with porcine pancreatic elastase (1FLE)







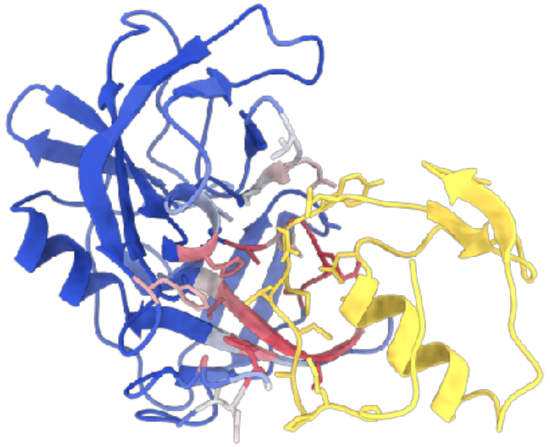
unbound porcine pancreatic elastase **dynamic prediction**

(after 1 μs of MD simulation)

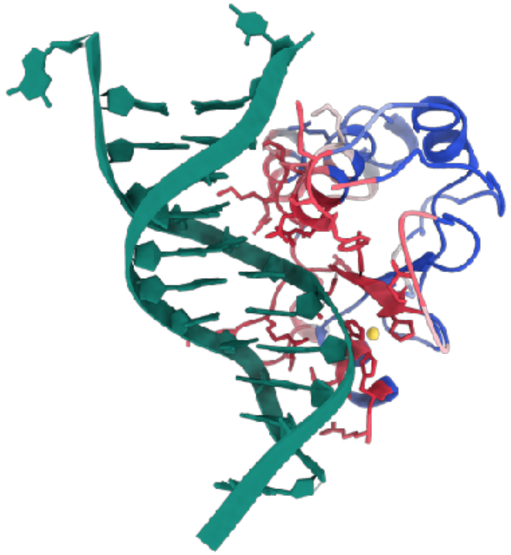
revealing cryptic binding sites
and potential allosteric mechanisms

Other binding interfaces of proteins

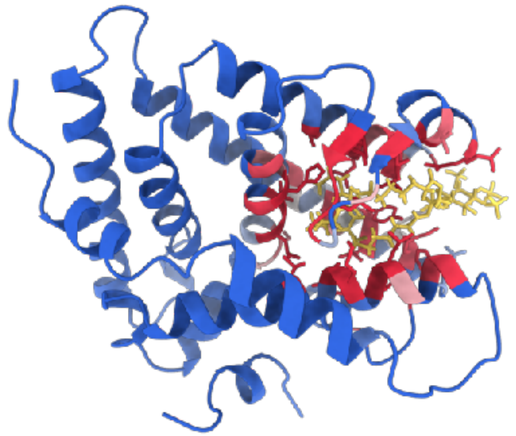
Protein-protein



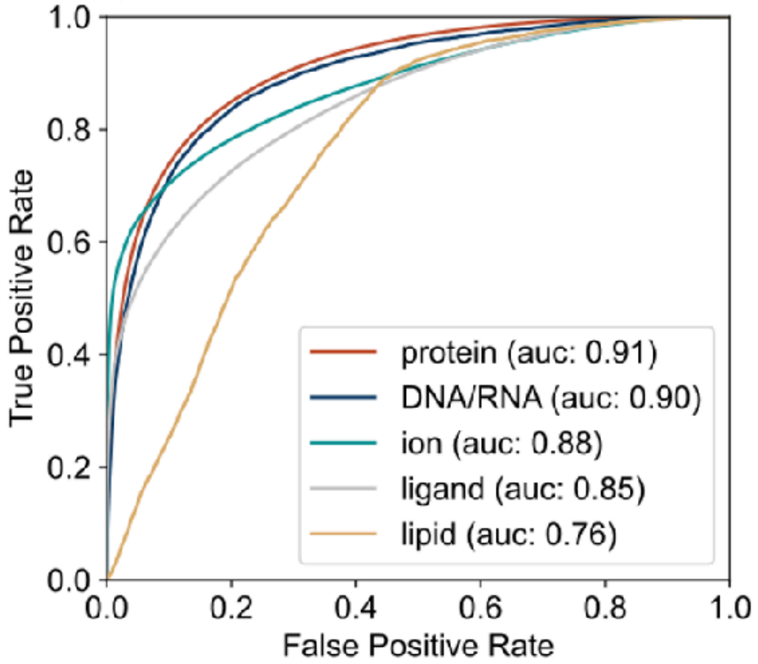
Protein-DNA-ions



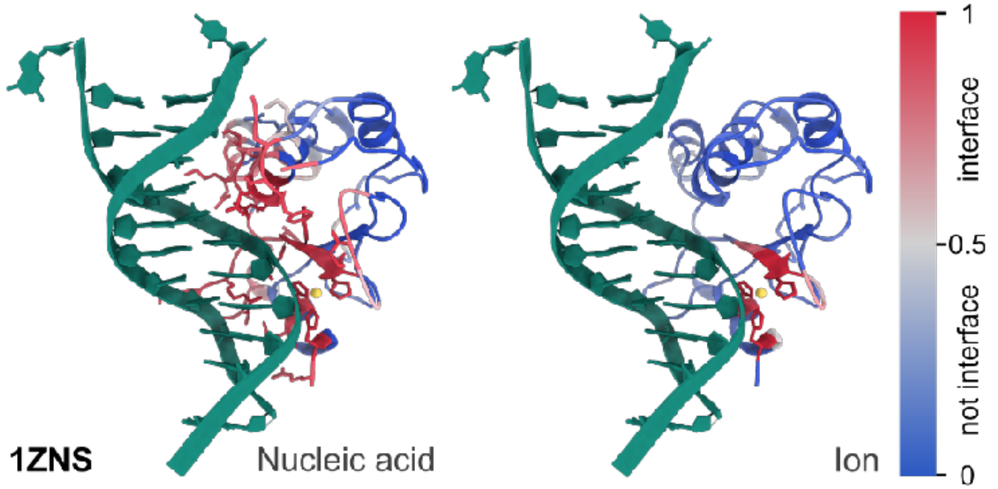
Protein-lipid



Extending prediction to other interacting interfaces

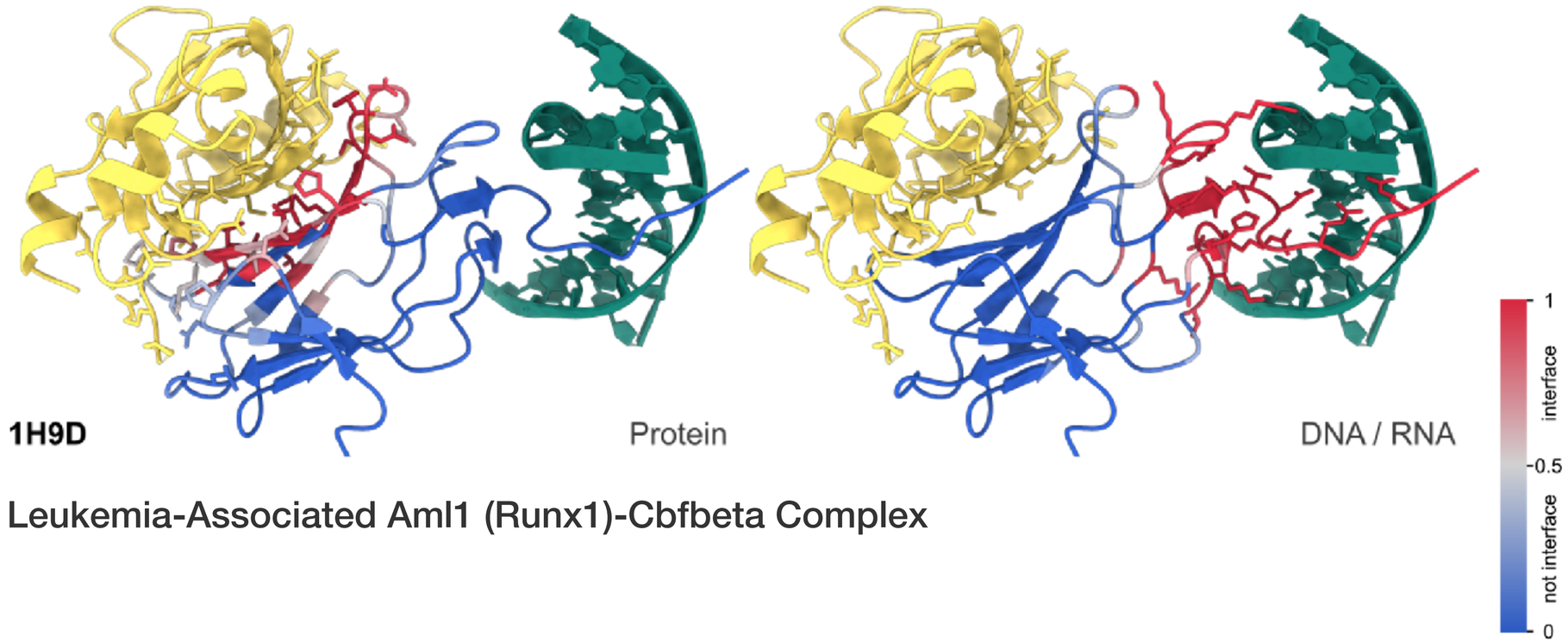


nucleic acid & ion interface



ColE7 endonuclease domain

Protein-nucleic acid interface predictions



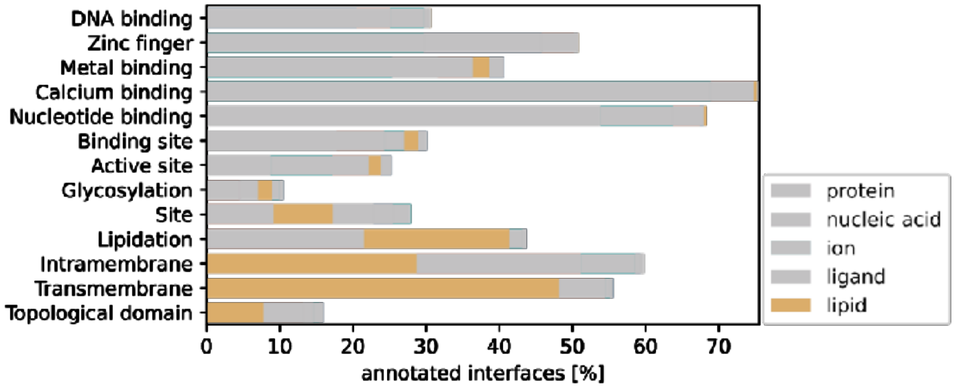
Leukemia-Associated Aml1 (Runx1)-Cbfbeta Complex

Proteome-wide interface prediction - aka *interfaceome*

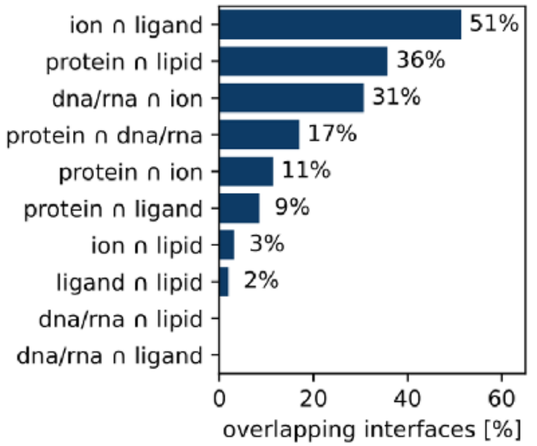
Human proteome prediction using the AF-EBI database

(only high-quality models, 7464 of 20504)

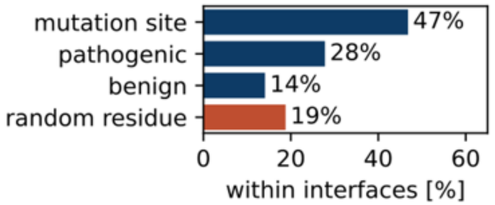
correlation with protein function and features



interface cross-talk



correlation with protein mutations



@ http://pesto.epfl.ch

Fabio Cortés

PeSTo

PeSTo (Protein Structure Transformer) is a parameter-free geometric deep learning method to predict protein interaction interfaces from a protein structure. It is available for free without registration as an online tool. A manuscript of the method is in preparation and will be available soon.

Learn more about this project in this preprint at Biorxiv.

How to use

Copy-paste your atomic coordinates in PDB format, or upload a PDB file from your drive, or fetch a protein structure/model from:

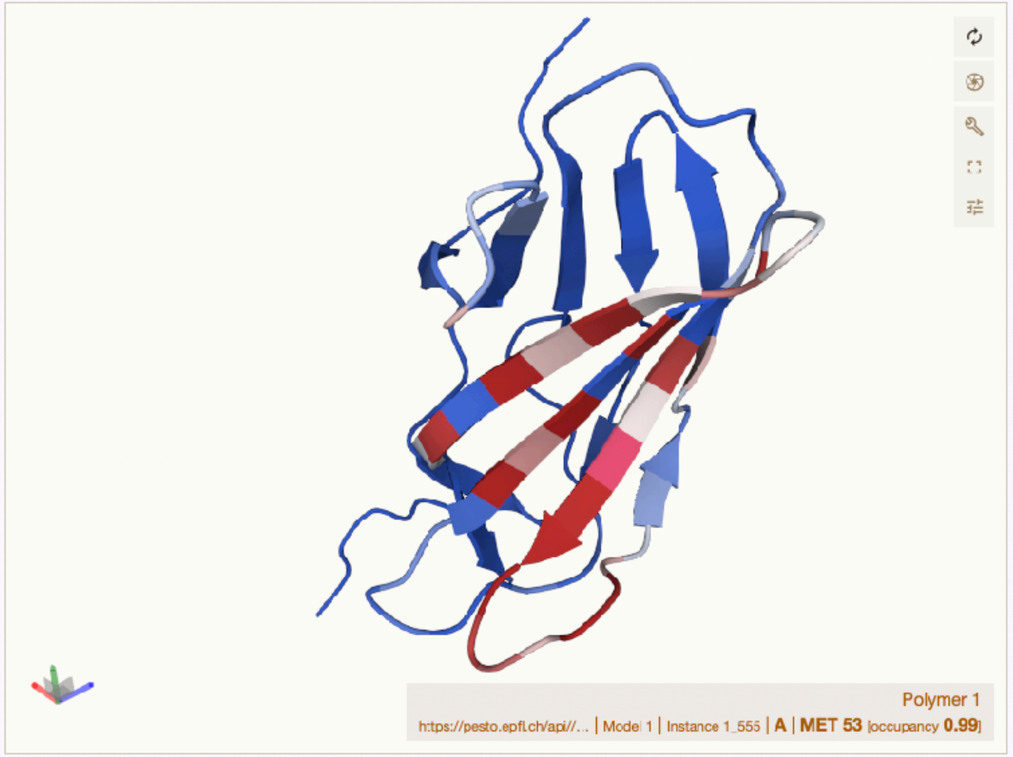
- The protein data bank by typing a PDB ID. Example: 2CUA
- The AlphaFold-EBI database by typing a Uniprot ID. Example: P27695
- Upload your own PDB formated structure

Then click "Detect chains", select one or more, and submit your job to run the prediction. Your results should be available in less than a minute. If an error occurs, the PDB file might be not correctly formated or the input structure is too big

2CUA | Fetch PDB/AF-EBI | Upload PDB

Copy-Paste molecule here

Protein | DNA-RNA | Lipid | Ligand | Ion



Chain	Res Name	Res ID	Prediction
A:0	ASP	13	0.64
A:0	THR	51	0.91
A:0	VAL	52	0.54
A:0	MET	53	0.99
A:0	ALA	54	0.96
A:0	GLY	55	0.97
A:0	ASN	56	0.98
A:0	ASP	57	0.91

Paths to drug discovery and the role of computational chemistry

1. The process of Drug Discovery (& Development)

2. Structure-based Drug Design

courtesy of Marco De Vivo, PhD
Laboratory of Molecular Modeling and Drug Discovery
Istituto Italiano di Tecnologia- Genoa, Italy

Drug Discovery & Development

- Some facts -

FDA approvals in 2020

53 New Drugs Approved

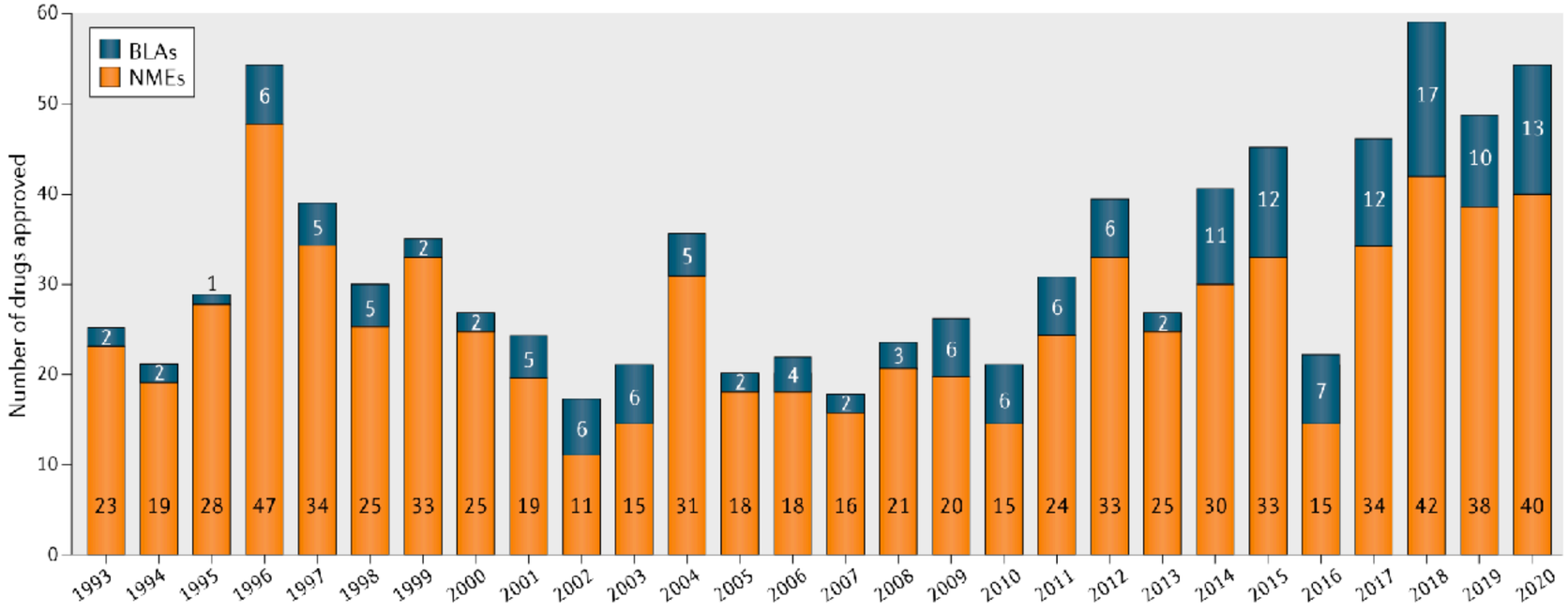


Fig. 1 | **Novel FDA approvals since 1993.** Annual numbers of new molecular entities (NMEs) and biologics license applications (BLAs) approved by the FDA's Center for Drug Evaluation and Research (CDER). See TABLE 1 for new approvals in 2020. Approvals by the Center for Biologics Evaluation and Research (CBER), for products such as vaccines and gene therapies, are not included in this drug count (see TABLE 2). Source: FDA.

Source:

Asher Mullard,

Nature Reviews Drug Discovery, February (**2021**)

Drug Discovery & Development

- Some facts -

Approval by therapeutic area 2020

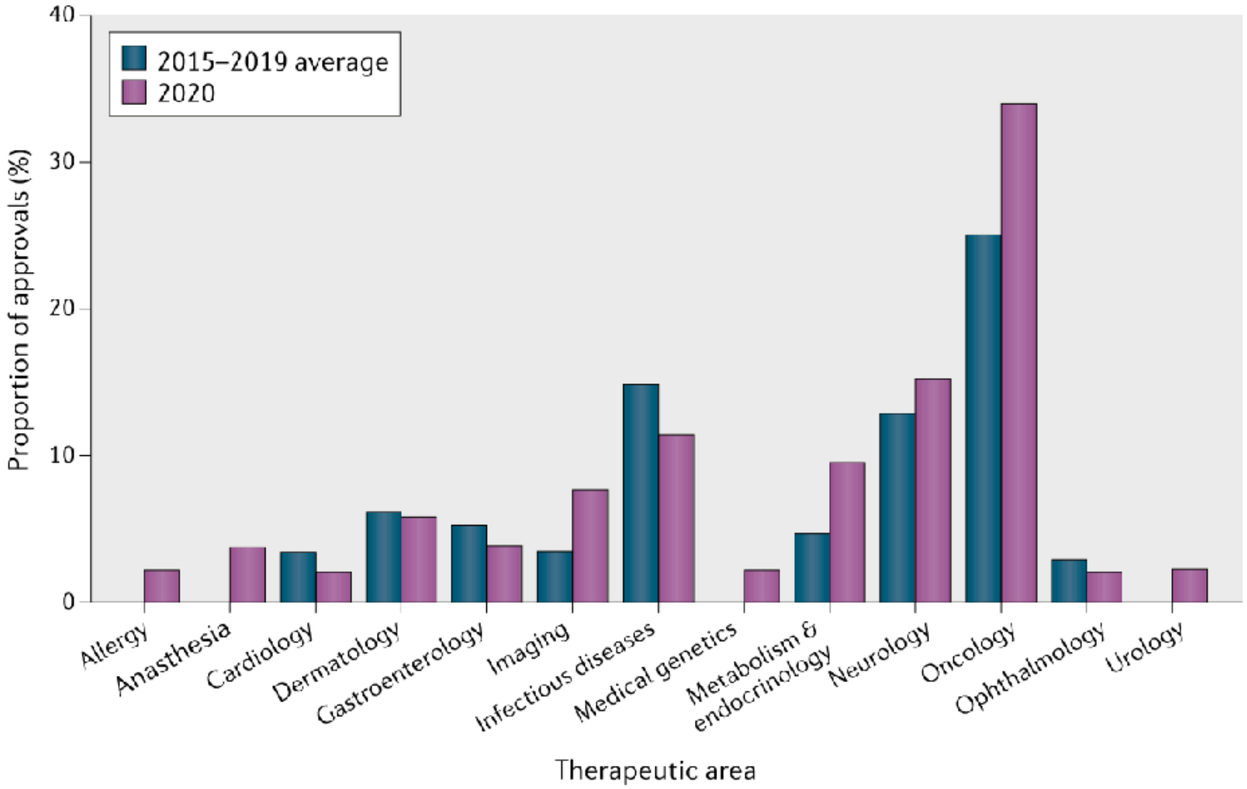
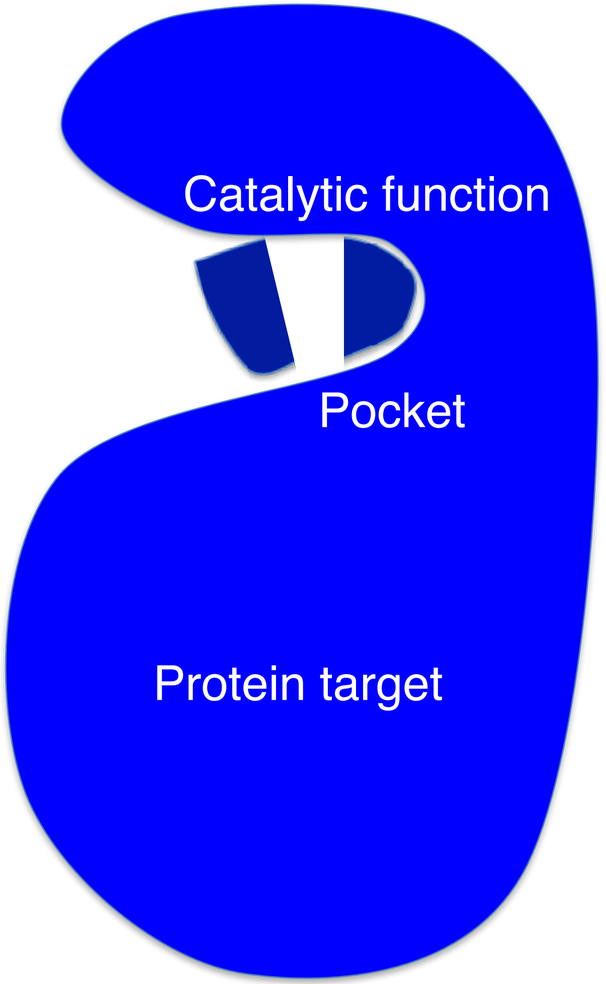


Fig. 2 | **CDER approvals by selected therapeutic areas.** Source: *Nature Reviews Drug Discovery*, FDA.

Source:
Asher Mullard,
Nature Reviews Drug Discovery, February (**2021**)

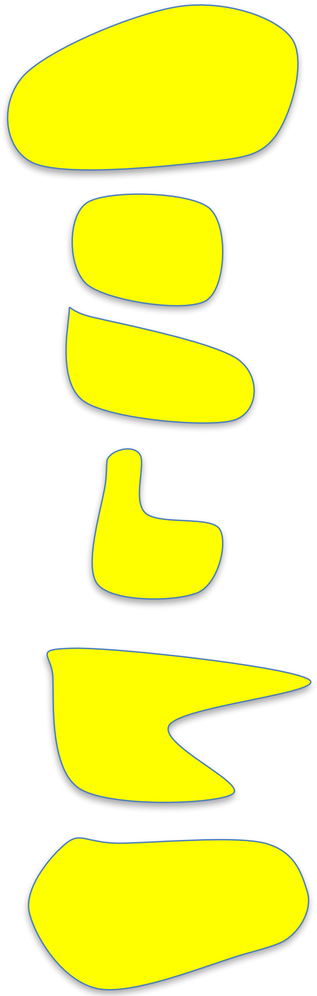
Enzymes and catalytic activity

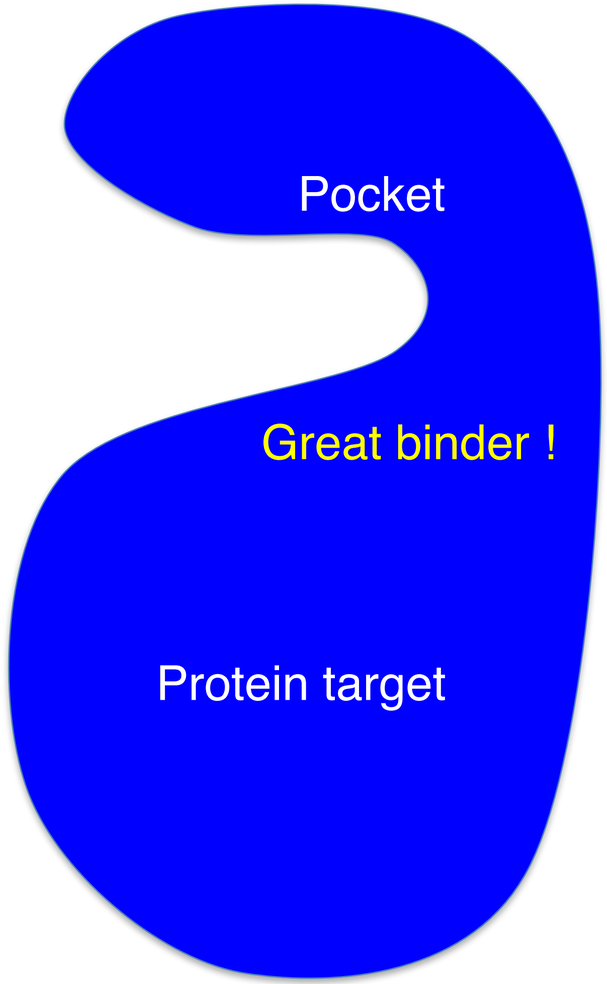
Protein substrate



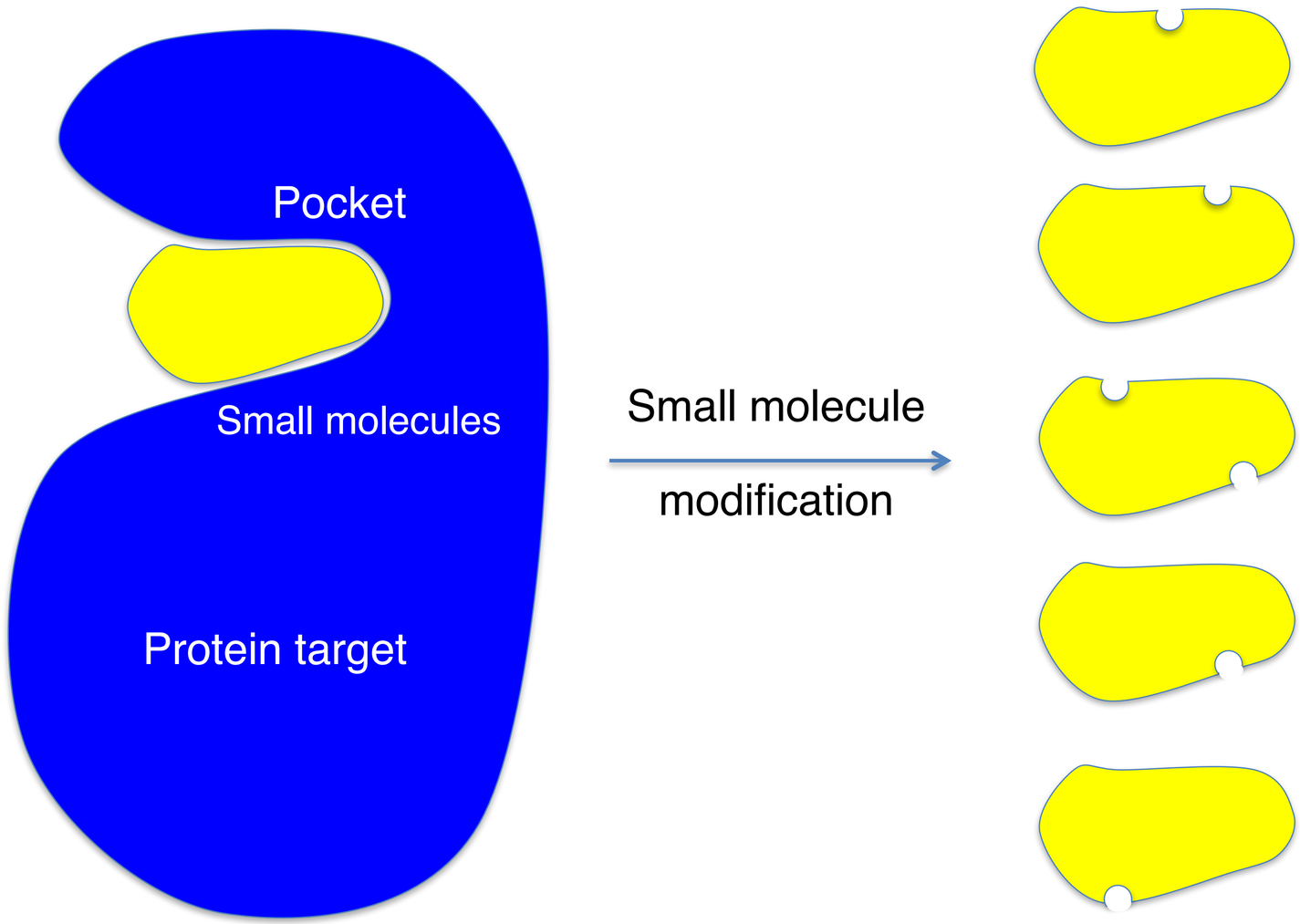
Protein binding and inhibition

Small molecules



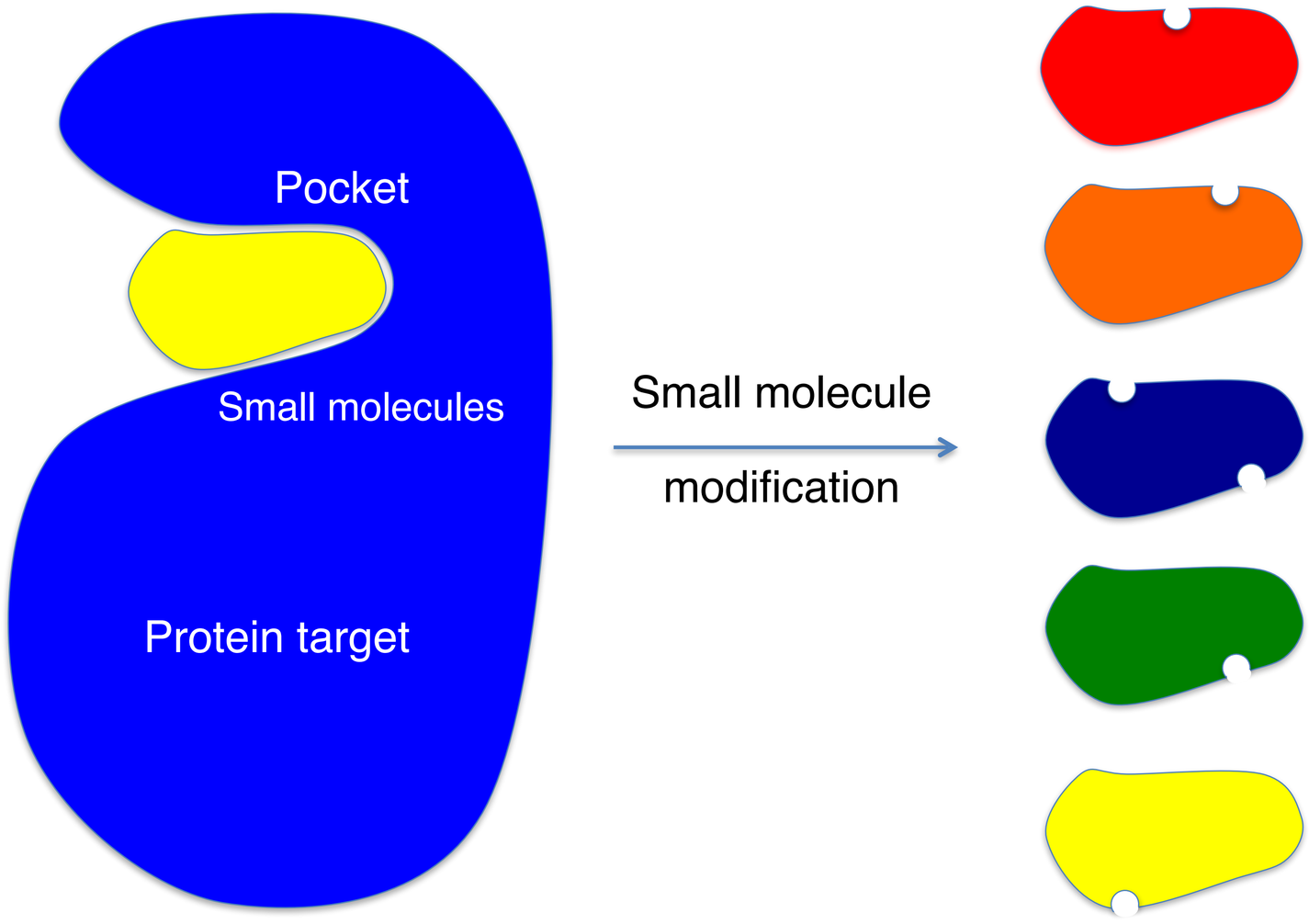


Drug design



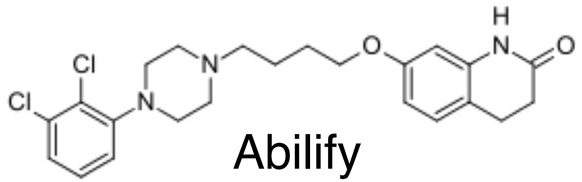
From a good inhibitor to a potential drug

Drug design

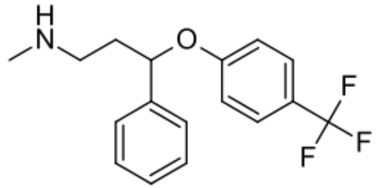


Similar binding but DIFFERENT properties!

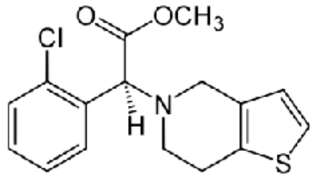
Small molecules drugs



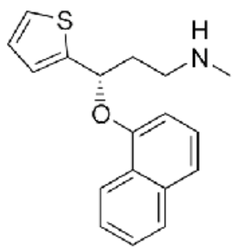
Abilify
(antipsychotic)



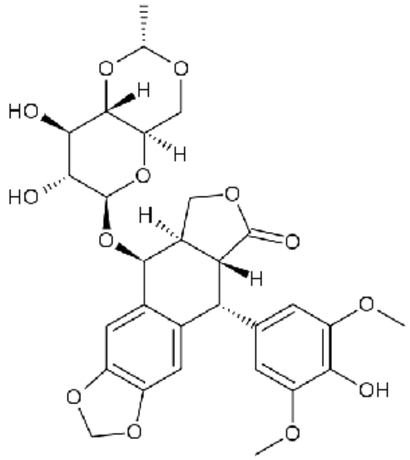
Prozac
(antidepressant)



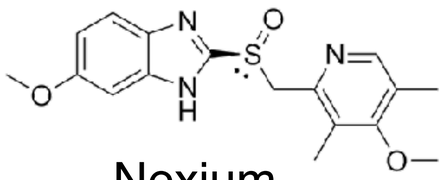
Plavix
(antiplatelet agent)



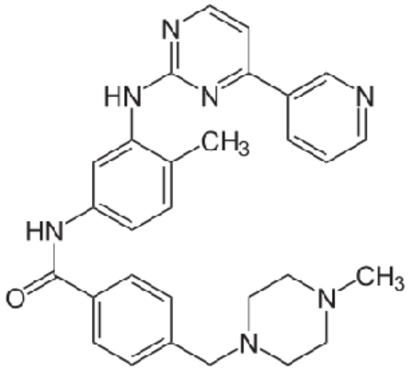
Cymbalta
(pain and anxiety)



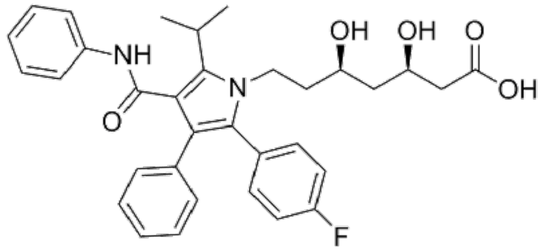
Etoposide
(cancer)



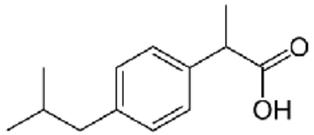
Nexium
(gastric acid)



Gleevec
(cancer)



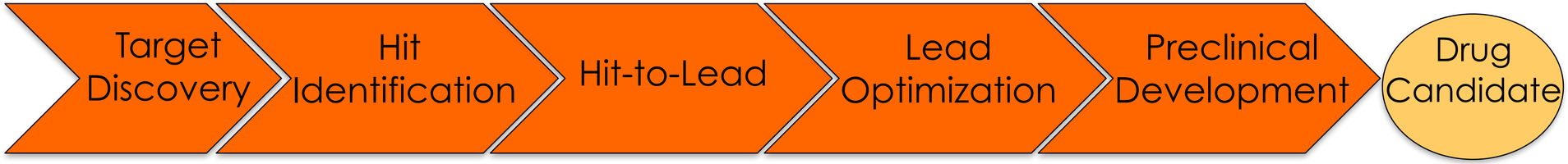
Lipitor
(Cholesterol)

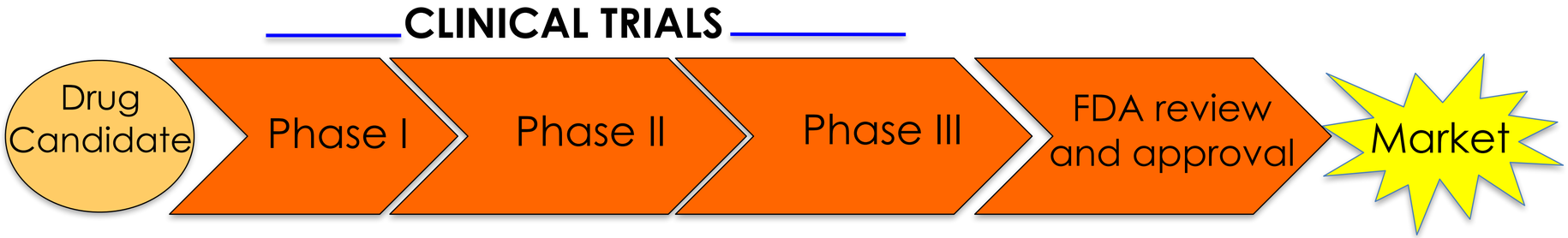


Ibuprofen
(inflammation)

Drug Discovery & Development

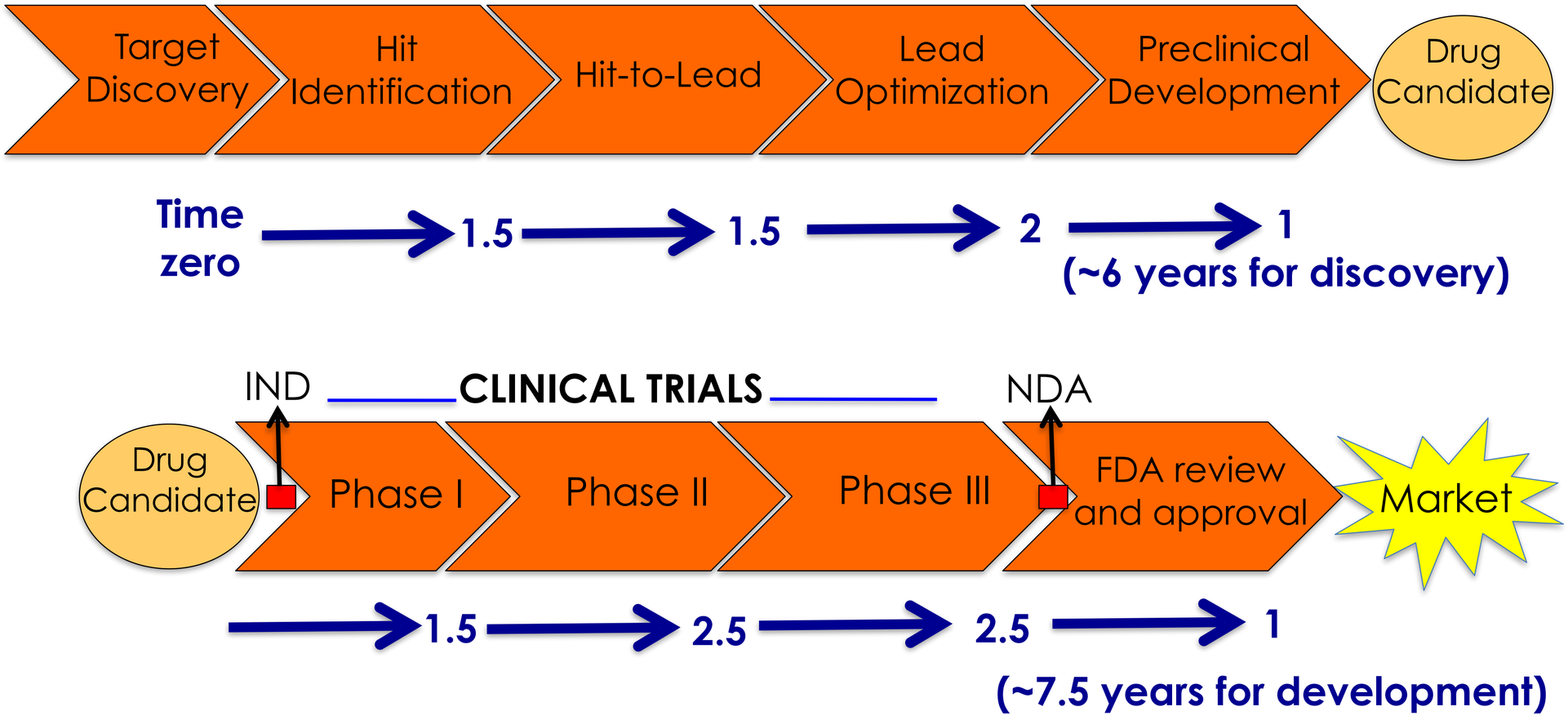
Sequence of steps





Drug Discovery & Development

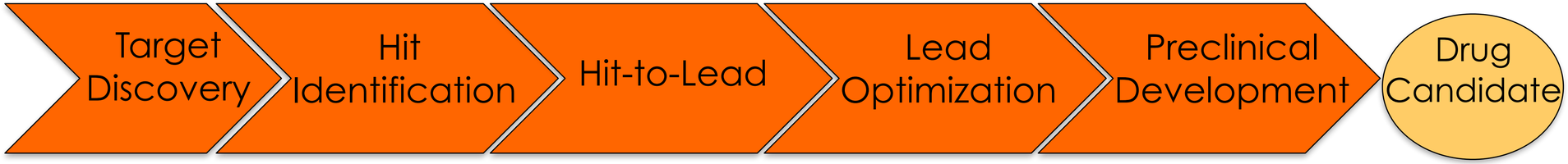
Average time requested for it



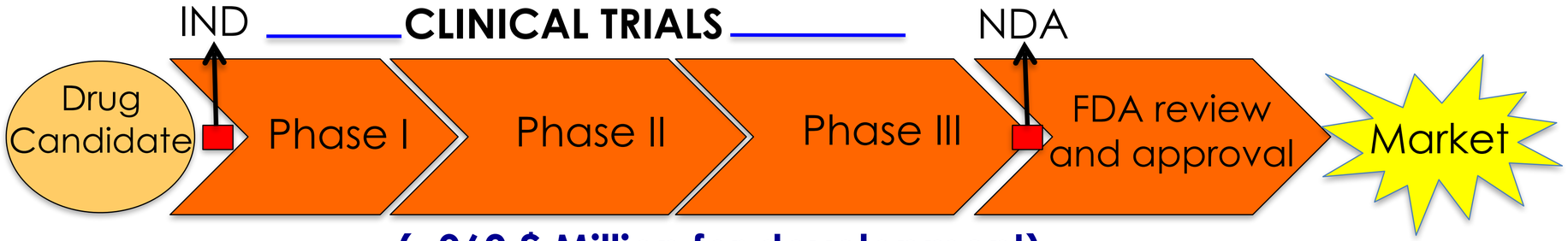
Total time (on average) = ~13.5 years

Drug Discovery & Development

Average <u>cost</u> requested for it



(~820 $ Million for discovery)



(~960 $ Million for development)

Total cost on average = ~1.78 $ Billion for one NME

Source: How to improve R&D productivity: the pharmaceutical industry's grand challenge
Steven M. Paul, *Nat. Review Drug Discovery* March Vol. 9 **2010**

Computational Drug Design

Two approaches:

- **Structure-Based Drug Design - SBDD**

  Drug design based on the interaction of the **ligand** with the 3D dimensional structure of the **receptor**

- **Ligand-Based Drug Design - LBDD**

  Unknown structure of the receptor.
  Drug design based on the key features of active **compounds.**

  Hypothesis:
  Ligands similar to an active ligand are more likely to be active than random ligands. **(pharmacophore models)**

Computational Drug Design

Approaches and methods:

◇ Structure-Based Drug Design - SBDD

Docking (Glide, Dock, Autodock, ICM… etc)
Kitchen D., *Nat. Review Drug Discovery* Vol. 3 Nov. **2004**

De novo design (BOMB, SMoG, BREED.. etc)
Gisbert Schneider and Uli Fechner, , *Nat. Review Drug Discovery* Vol. 4 Aug. **2005**

◇ Ligand-Based Drug Design - LBDD

Quantitative Structure-Activity Relationship (QSAR)
Markus A. Lill, Drug Discovery Today, Vol. 12 Dec. **2007**

Ligand similarity approaches (2D or 3D)
Johann Gasteiger, J. Med. Chem. 49, 22, **2006**.

Structure-Based Drug Design

MUST: 3-dimensional structure of the target.

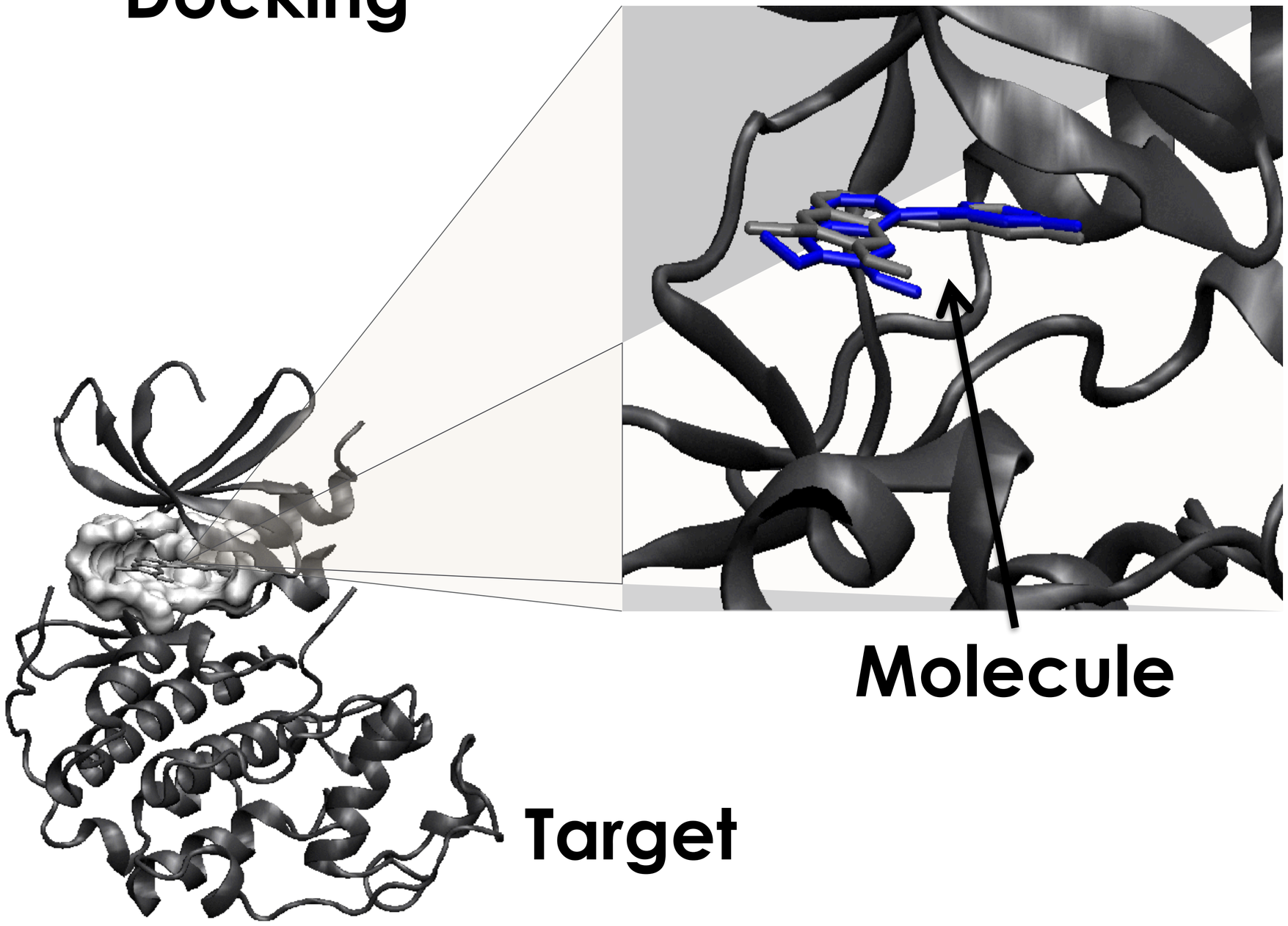
Sources of structures:

1. Crystallography
2. NMR structures
3. Cryo-EM
4. Homology or AF structures

```
AAB24882        TYHMCQFHCRYVNNHSGEKLYECNERSKAFSCPSHLQCHKRRQIGEKTHEHNQCGKAFPT 60
AAB24881        --------------------YECNQCGKAFAQHSSLKCHYRTHIGEKPYECNQCGKAFSK 40
                                    ****: .***:  * *:** * :****.:* *******..

AAB24882        PSHLQYHERTHTGEKPYECHQCGQAFKKCSLLQRHKRTHTGEKPYE-CNQCGKAFAQ- 116
AAB24881        HSHLQCHKRTHTGEKPYECNQCGKAFSQHGLLQRHKRTHTGEKPYMNVINMVKPLHNS 98
                 **** *:***********:***:**.: .***************    :  *.: :
```

Docking

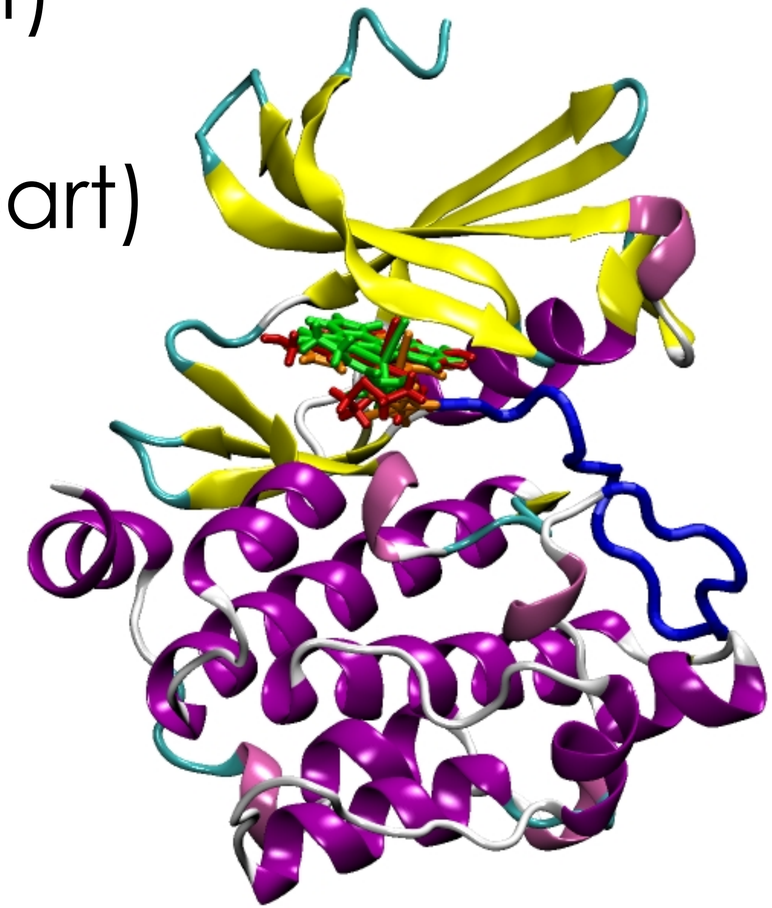


Ligand Docking

Two problem to solve:

1. Posing (the easy part)

1. Scoring (the tough part)



Docking and scoring in virtual screening for drug discovery: methods and applications
Douglas B. Kitchen et al *Drug Discovery, Vol. 3, Nov* ***2004***

Ligand Docking - Scoring

The quantitative modeling of receptor – ligand interactions can be achieved by determining the equilibrium binding constant K_{eq}. The binding constant K_{eq} is directly related to the Gibbs free energy:

$$\Delta G_{bind} = -RT \ln K_{eq}$$

$$\Delta G_{bind} = \Delta H - T\Delta S = G_{complex} - (G_{receptor} + G_{ligand}).$$

Why it is so difficult to score compounds:

Experimental range of binding affinities: from 10^{-2} M (mM) to 10^{-12}M (pM)

At T=298K the enthalpic contribution to the $\Delta G_{binding}$ is between -2.4 kcal/mol and -16.7 kcal/mol

In other words, a change in binding (free) energy of ~1.5 kcal/mol alters the binding affinity of one order of magnitude (T=298) !!!!

Ligand Docking - Scoring

Scoring Functions FOR DOCKING CALCULATIONS:

- Force-Field Based (Physics-based)
- Empirical
- Knowledge-based
- Descriptor-based

Classification of Current Scoring Functions
Liu and Wang, JCIM **2015**, *55* (3), pp 475–482

Consensus Scoring

Consensus scoring are more often applied in **Virtual Screening**

Ligand Docking - Scoring

$$\Delta G_{bind} = E_{MM} - T\Delta S_{solute} + \Delta G_{solvent}$$

E_{MM} from FF:

$$E_{MM} = \sum_{bonds} K_r(r - r_{eq})^2 + \sum_{angles} K_\vartheta(\vartheta - \vartheta_{eq})^2 + \sum_{dihedrals} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)]$$
$$+ \sum_{i<j} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\varepsilon R_{ij}} \right]. \qquad 3.$$

ΔS_{solute}: The solute entropy consists of four terms, namely translational, rotational, vibrational, and conformational entropy.

$\Delta g_{solvent}$: The solvent free energy consists of the two terms: 1) a nonpolar and a polar term.

Ligand Docking - Scoring

DOCK (v4.0)	$E_{vdW} + E_{electrostatic} = \sum_{prot}\sum_{lig}\left[\left(\frac{A_{ij}}{d_{ij}^{a}} + \frac{B_{ij}}{d_{ij}^{b}}\right) + 332.0\frac{q_i q_j}{\in (d_{ij}) d_{ij}}\right]$

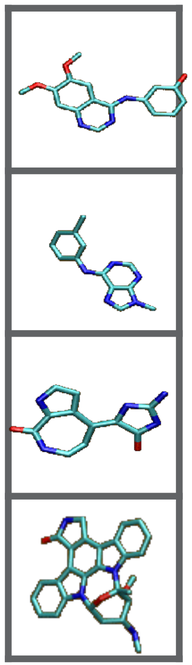
Total energy is given by the sum of energy terms.

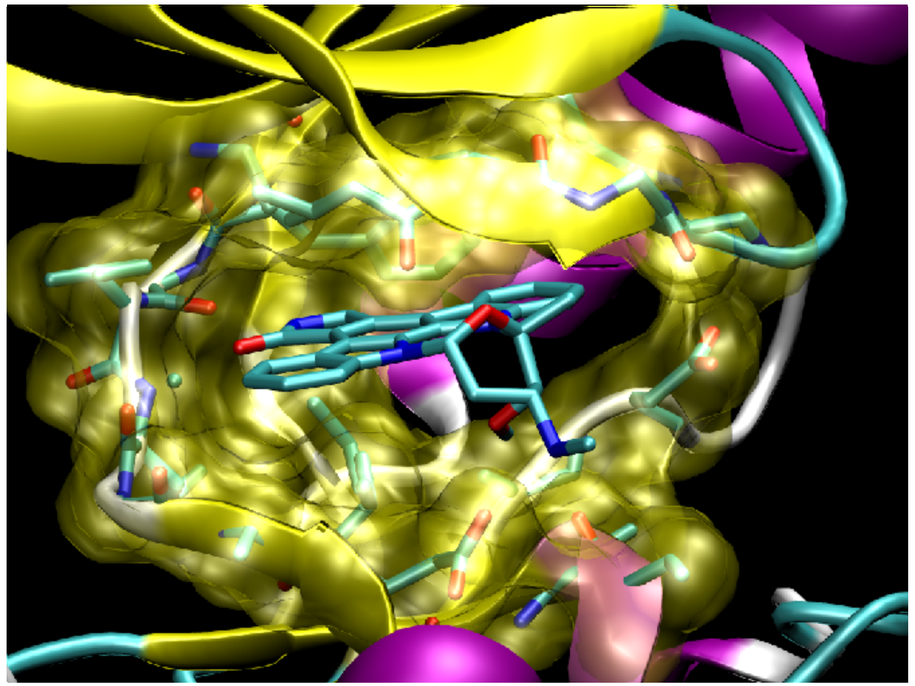
For two atoms *i* and *j*, A_{ij} and *Bij* are van der Waals parameters for given atom types, d_{ij} is the interatomic distance, q_i and q_j are atomic partial charges, and $\varepsilon(d_{ij})$ is a distance-dependent dielectric function.

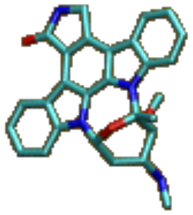
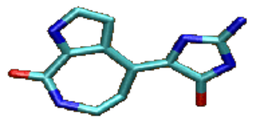
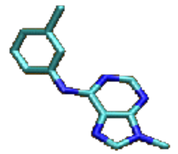
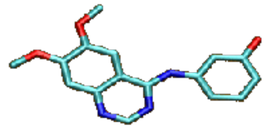
Virtual screening

An exercise carried out by computational means aimed at predicting which molecules from an ensemble will likely display some activity against a target.

VLS is usually implemented as an iterative docking simulation at the target binding site.





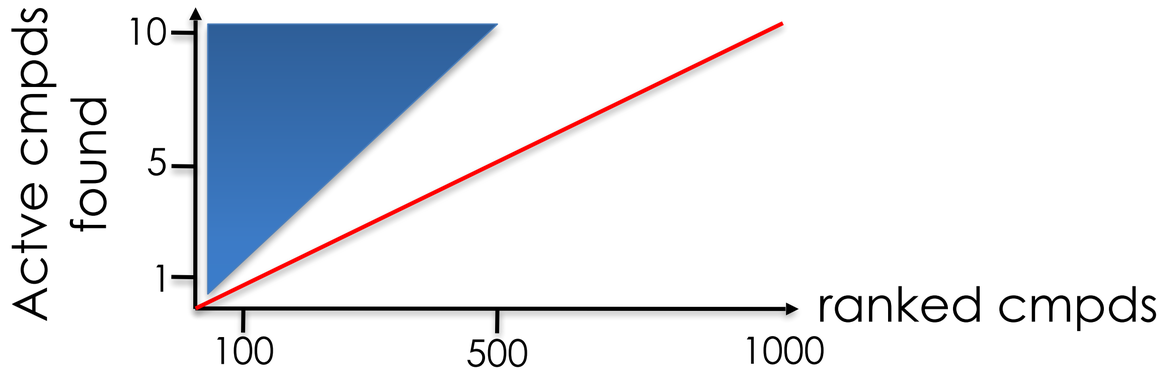
1st -37.9 kcal/mol	2nd -29.7 kcal/mol	3rd -27.3 kcal/mol	4th -9.8 kcal/mol
			

How can we evaluate the virtual screening performance?

Enrichment factor **:**

Example:
Suppose to have 10 active cmpds in a database of 1000 cmpds (1% of active compounds)

Random pick: 1 out of 100 should be active. (1% chance)



Ligand Docking - Scoring

Speed

Docking
Virtual Screning
Etc ..

Accuracy

Molecular Dynamics
Quantum Mechanics
Etc..

New directions

Published as a conference paper at ICLR 2023

DIFFDOCK: DIFFUSION STEPS, TWISTS, AND TURNS FOR MOLECULAR DOCKING

Gabriele Corso*, **Hannes Stärk***, **Bowen Jing***, **Regina Barzilay** & **Tommi Jaakkola**
CSAIL, Massachusetts Institute of Technology

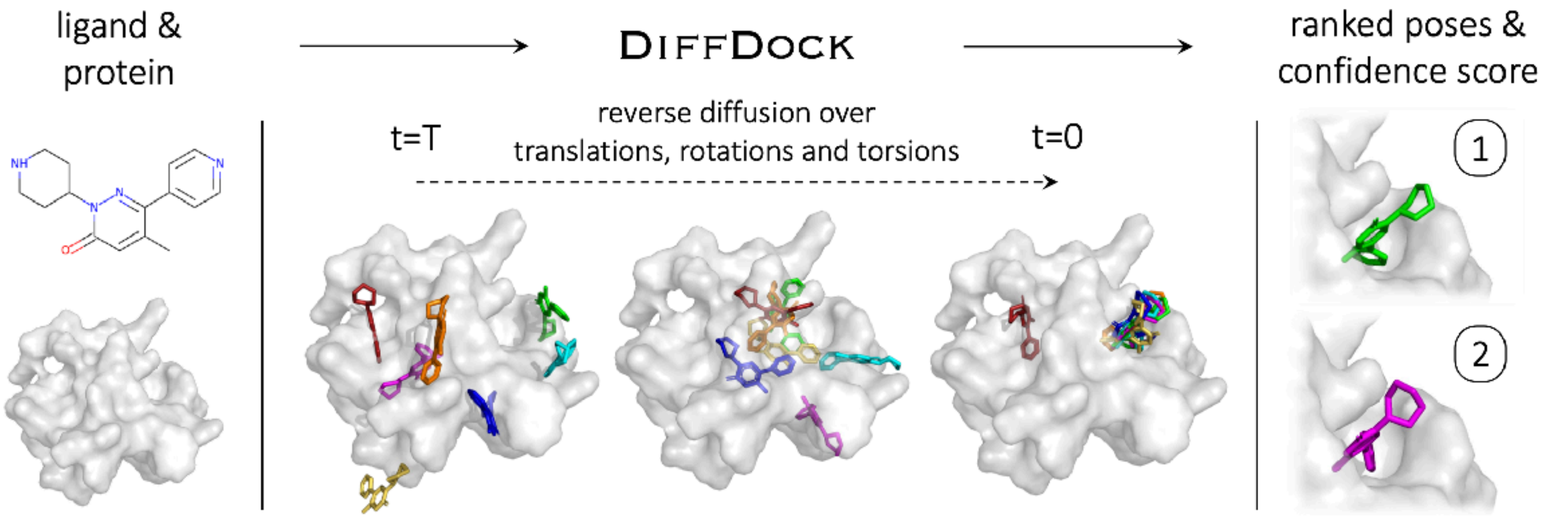
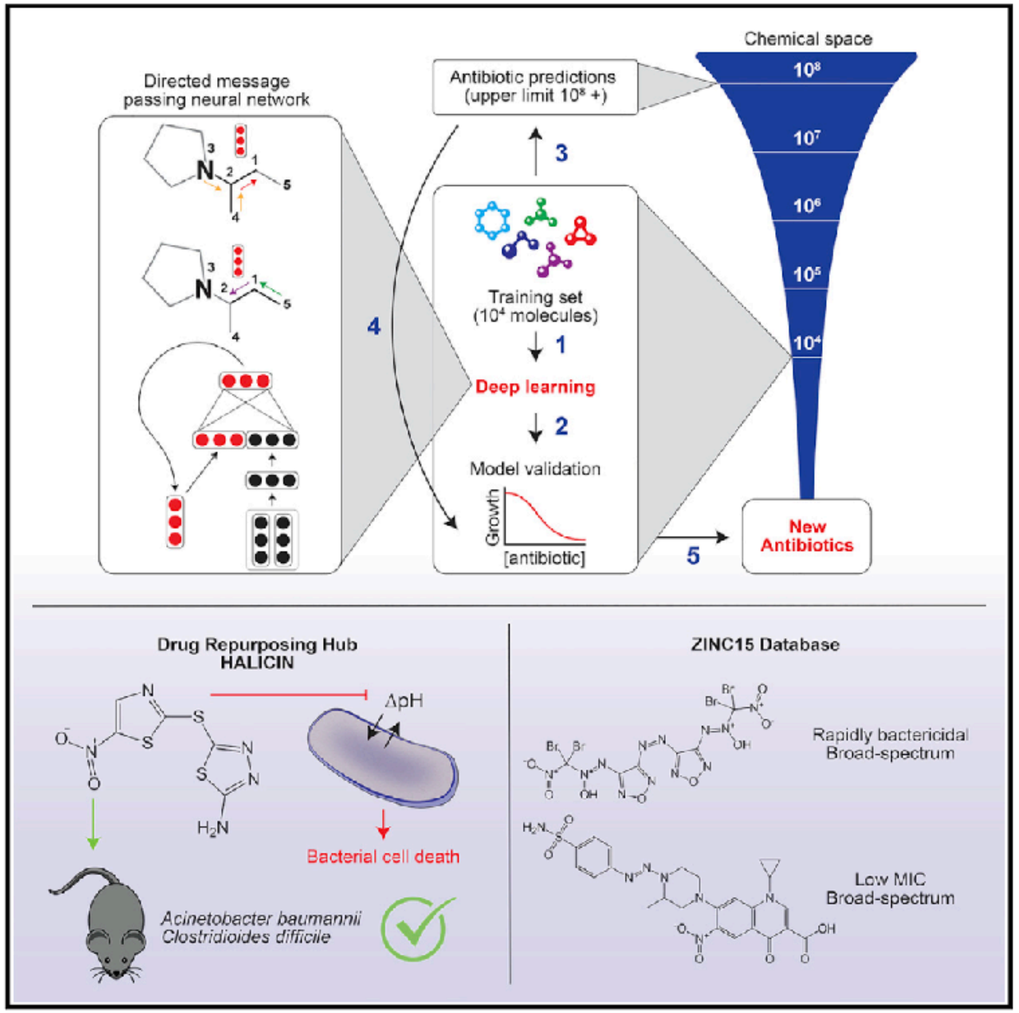


Figure 1: Overview of DIFFDOCK. *Left*: The model takes as input the separate ligand and protein structures. *Center*: Randomly sampled initial poses are denoised via a reverse diffusion over translational, rotational, and torsional degrees of freedom. *Right:*. The sampled poses are ranked by the confidence model to produce a final prediction and confidence score.

Cell

A Deep Learning Approach to Antibiotic Discovery

Graphical Abstract



Authors

Jonathan M. Stokes, Kevin Yang, Kyle Swanson, ..., Tommi S. Jaakkola, Regina Barzilay, James J. Collins

Correspondence

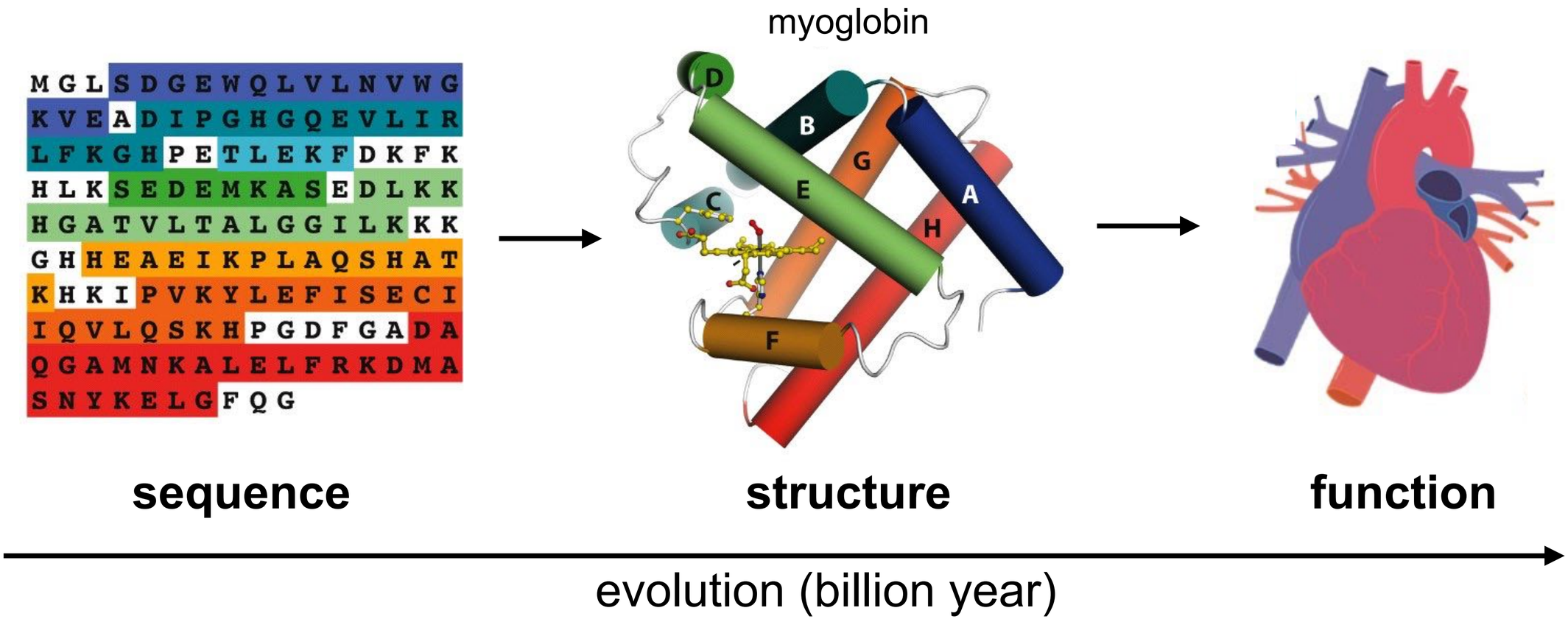
regina@csail.mit.edu (R.B.), jimjc@mit.edu (J.J.C.)

In Brief

A trained deep neural network predicts antibiotic activity in molecules that are structurally different from known antibiotics, among which Halicin exhibits efficacy against broad-spectrum bacterial infections in mice.

Matteo Dal Peraro

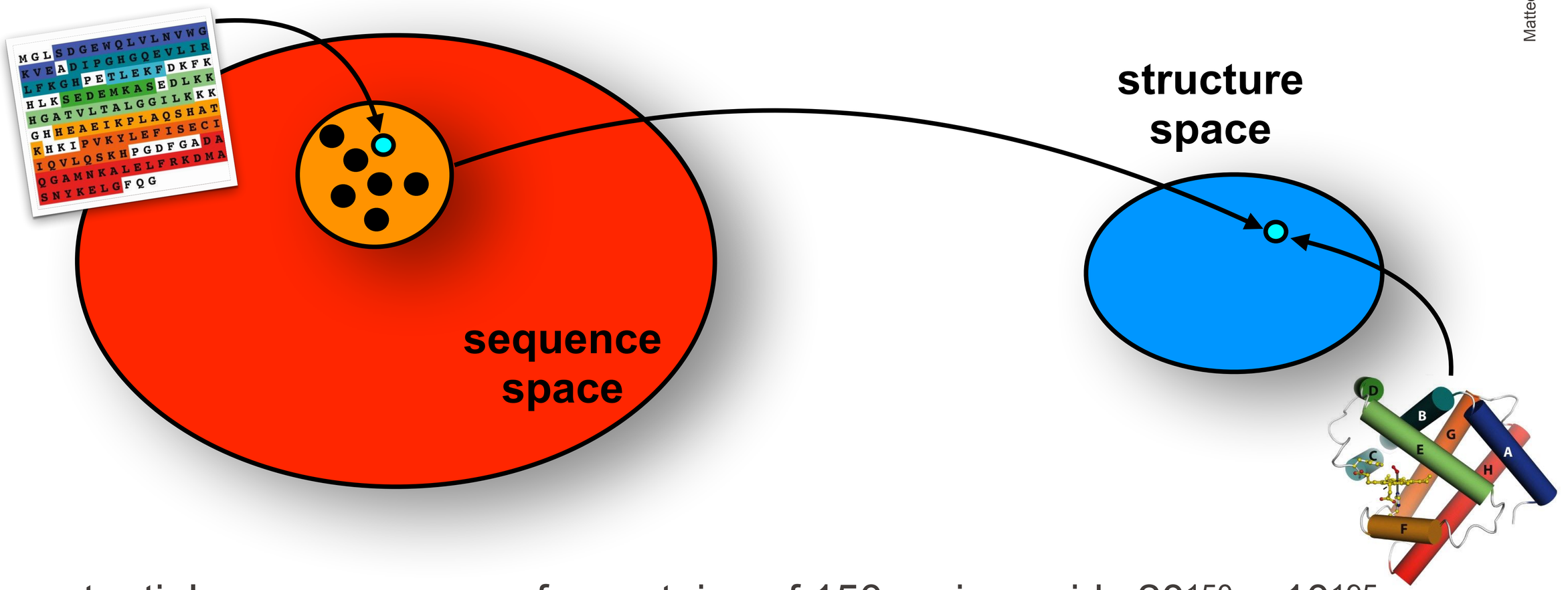
The folding paradigm



- Prediction of final structure and binding helps discovering new biology
- Not all the questions are answered though by AF2 !!

Matteo Dal Peraro

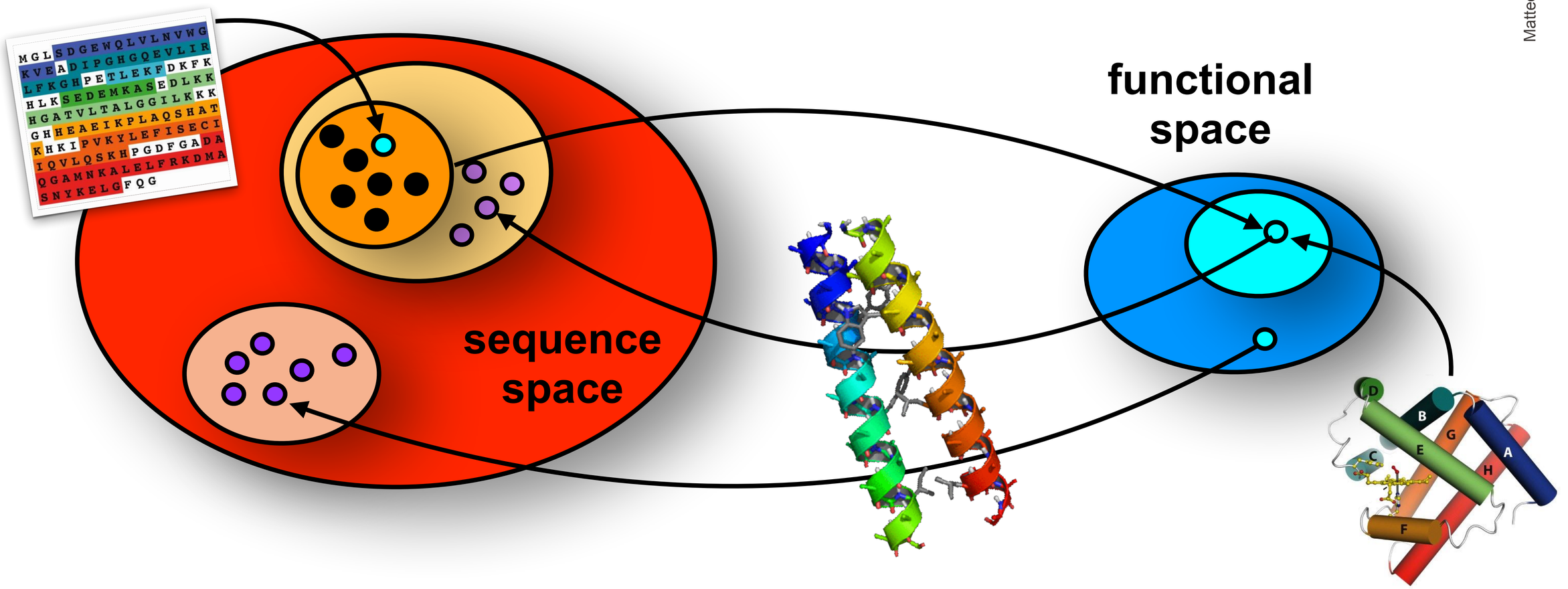
The sequence space is enormous



- potential sequence space for proteins of 150 amino acids $20^{150} \sim 10^{195}$
- atoms in the observed universe $\sim 10^{80}$
- the sequences explored by evolution are much less ($\sim 10^{10\text{-}20}$), structures lesser

Matteo Dal Peraro

The inverse folding problem — design

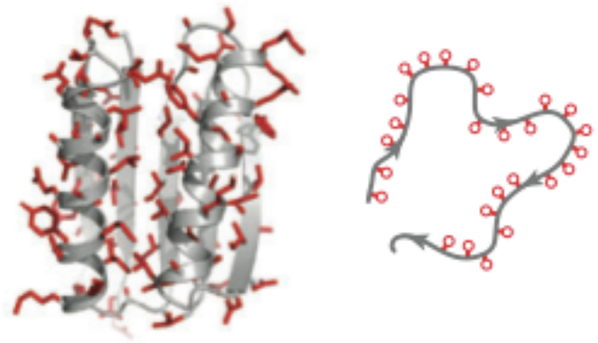


- Application to study protein evolution and function
- Protein engineering for therapeutics, synthetic biology and (bio)technology

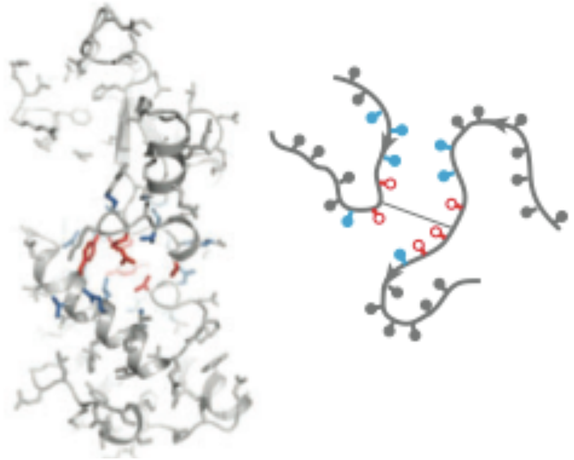
Multiple tasks for protein design

- create de novo proteins
- explore new folds
- embed new functions

Protein design



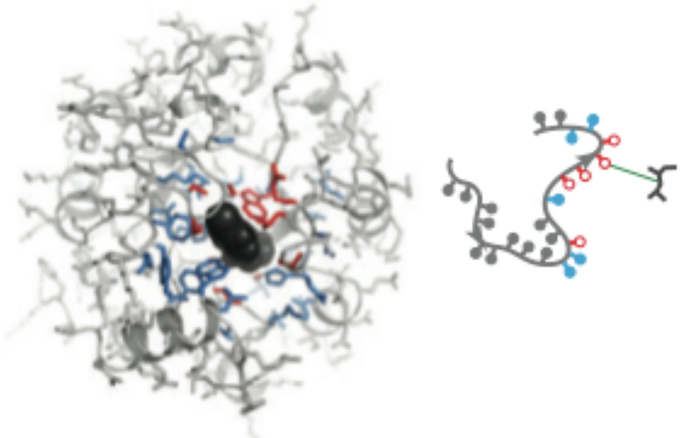
Protein-protein interface design



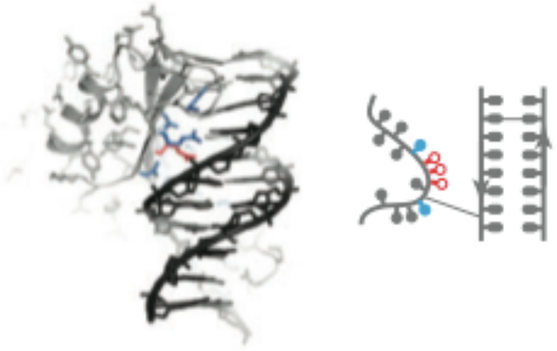
- create high affinity binders
- therapeutic biologics
- artificial sensors/probes

- tailor enzymatic function
- improve thermostability
- catalyse new reactions

Enzyme design



Protein-DNA interface design



- explore DNA interactions
- new therapeutic solutions

• Filled colored circles - flexible side chains
○ empty colored circles – flexible amino acid: design

The origins: the Paracelsus challenge (‘94)

- Rose and Creamer: convert a protein to another fold changing no more than 50% of its sequence

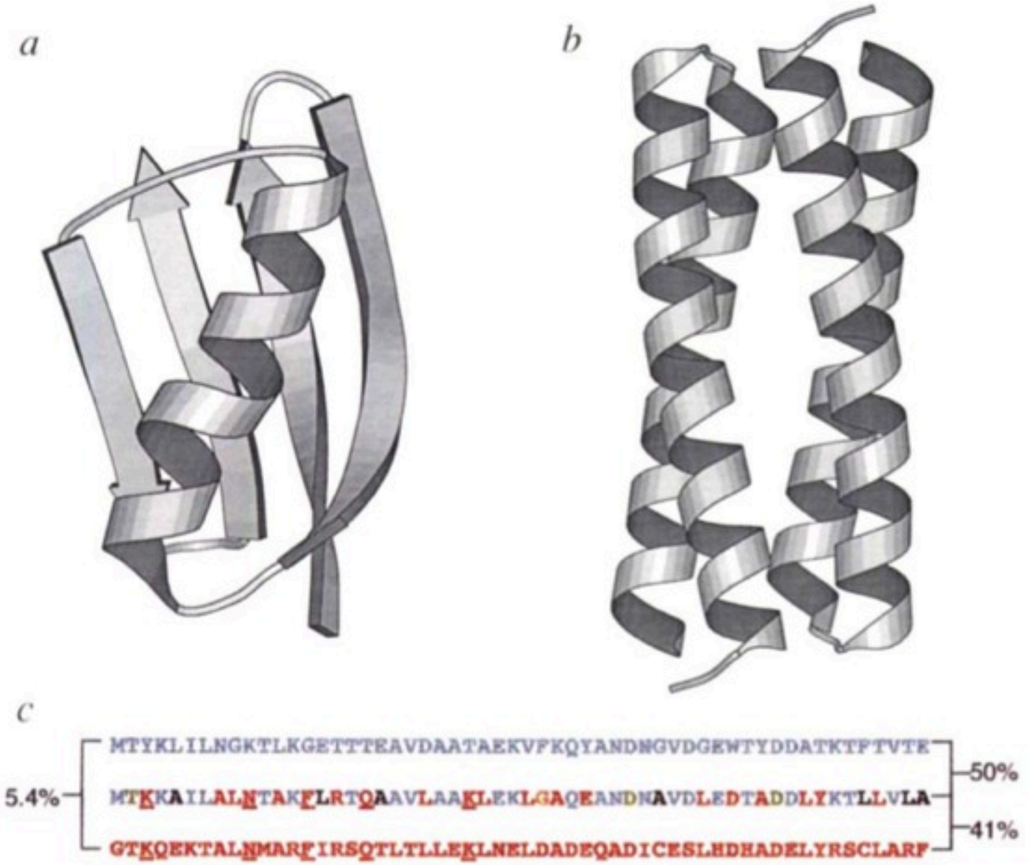


Fig. 1 Ribbon representation[29] of the folds of ***a***, the B1 domain of IgG-binding protein G[5] and ***b***, Rop[6]. ***c***, An alignment of the sequences of the B1 domain (blue), Rop (red) and Janus. Residues in Janus are coded as follows: blue, residues from B1; red, residues from Rop; underlined red, RNA-binding residues in Rop[13]; green, residues that are conserved in both Rop and B1; black, ‘a’ and ‘d’ position residues that are different from those in wild-type Rop; orange, the first residue of the turn between Helix 1 and Helix 2. The D30G mutation was introduced in the turn of Janus because a previous study demonstrated that this point mutation increases the stability of Rop[30]. The percent identity between the different sequences are indicated. The seven amino acid, unstructured C-terminal tail of Rop (Gly–Asp–Asp–Gly–Glu–Asn–Leu) extends beyond the sequence depicted for both Rop and Janus and is also not shown in (*b*). It was retained in Janus because it increases the solubility of wild type Rop[31].

Dalal et al., Protein alchemy: Changing β-sheet into α-helix, Nature Structural & Molecular Biology 1997

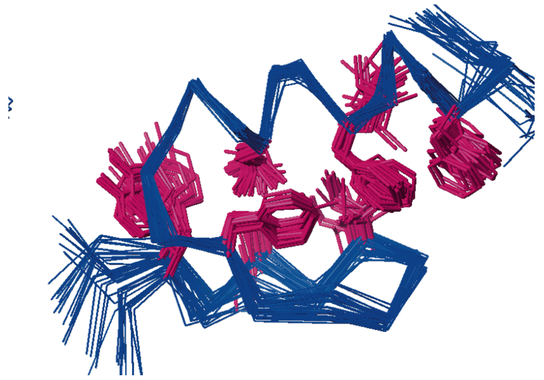
Design of a zinc-less zinc finger

De Novo Protein Design: Fully Automated Sequence Selection

Bassil I. Dahiyat† and Stephen L. Mayo*

The first fully automated design and experimental validation of a novel sequence for an entire protein is described. A computational design algorithm based on physical chemical potential functions and stereochemical constraints was used to screen a combinatorial library of 1.9×10^{27} possible amino acid sequences for compatibility with the design target, a $\beta\beta\alpha$ protein motif based on the polypeptide backbone structure of a zinc finger domain. A BLAST search shows that the designed sequence, full sequence design 1 (FSD-1), has very low identity to any known protein sequence. The solution structure of FSD-1 was solved by nuclear magnetic resonance spectroscopy and indicates that FSD-1 forms a compact well-ordered structure, which is in excellent agreement with the design target structure. This result demonstrates that computational methods can perform the immense combinatorial search required for protein design, and it suggests that an unbiased and quantitative algorithm can be used in various structural contexts.

Dahiyat, BI, and SL Mayo.
Science 278, 5335 (3 October 1997): 82-7



FSD-1 NMR determination

Zif268

FSD-1 design

comparison of FSD-1 designed and NMR

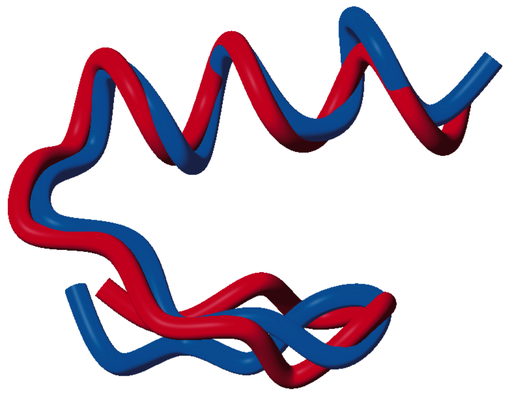
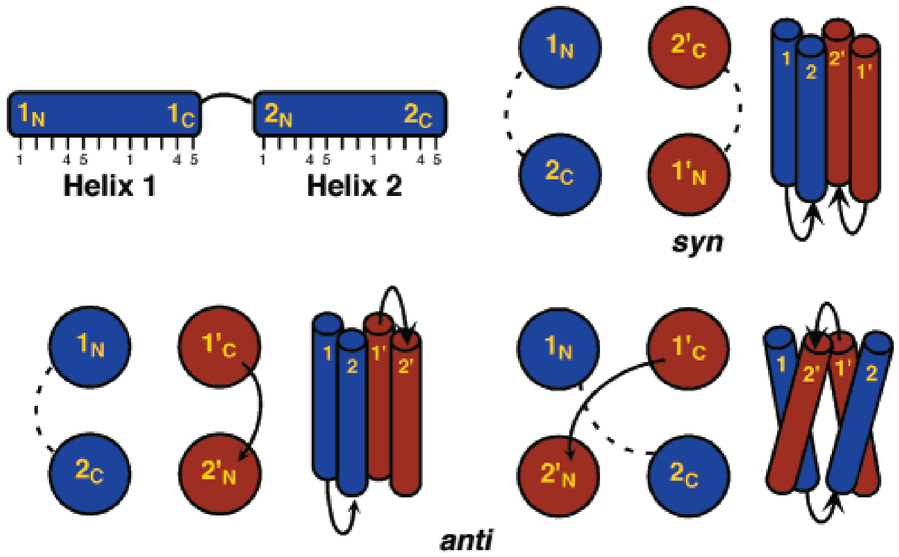


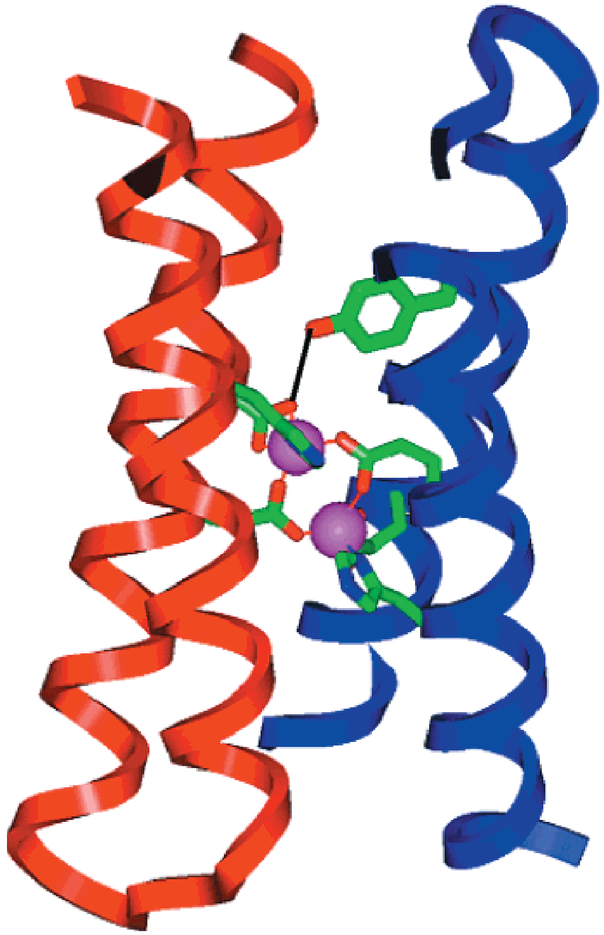
Table 2. Comparison of the FSD-1 experimentally determined structure and the design target structure. The FSD-1 structure is the restrained energy minimized average from the NMR structure determination. The design target structure is the second DNA binding module of the zinc finger Zif268 (*9*)

Atomic rms deviations (Å)		
Backbone, residues 3 to 26	1.98	
Backbone, residues 8 to 26	0.98	
*Super-secondary structure parameters**	FSD-1	Design target
h (Å)	9.9	8.9
θ (degrees)	14.2	16.5
Ω (degrees)	13.1	13.5

De novo design of helical bundles

- use knowledge on natural occurring helical bundles to assemble and functionalize new folds and functions



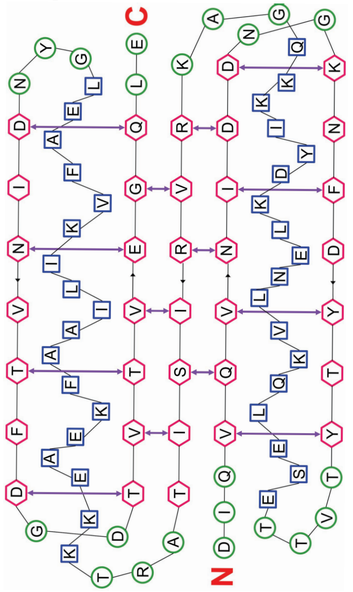


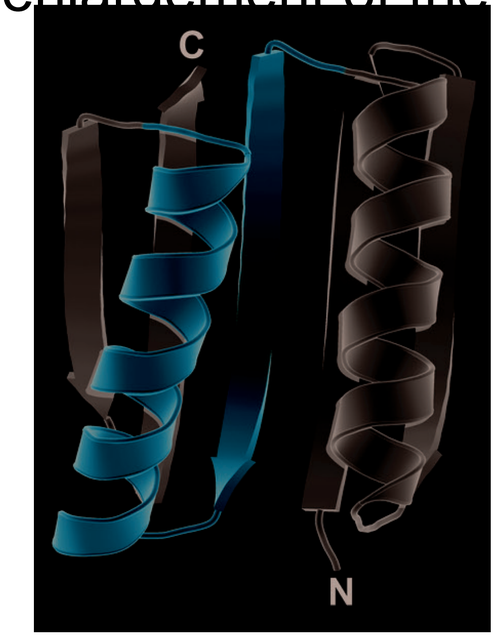
The 'Due Ferri' (two-iron; DF) family

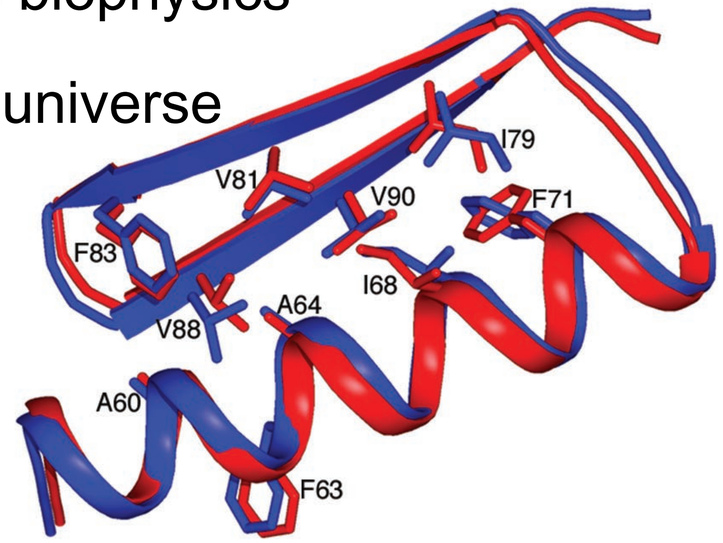
DeGrado and coworkers, ACCOUNTS OF CHEMICAL RESEARCH / VOL. 33, NO. 11, 2000

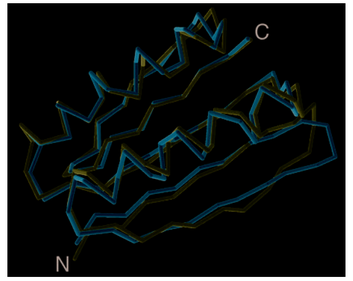
Design of a new protein fold Top7

- define a new topology > design sequence to fit backbone > relax the backbone to fit the sequence > iterate (using ROSETTA)
- top7 is soluble, stable, monomeric, unfolds cooperatively sequence is unique (no good homolog by BLAST search)
- energy function is good for representing its biophysics
- implications for enlargement of the protein universe





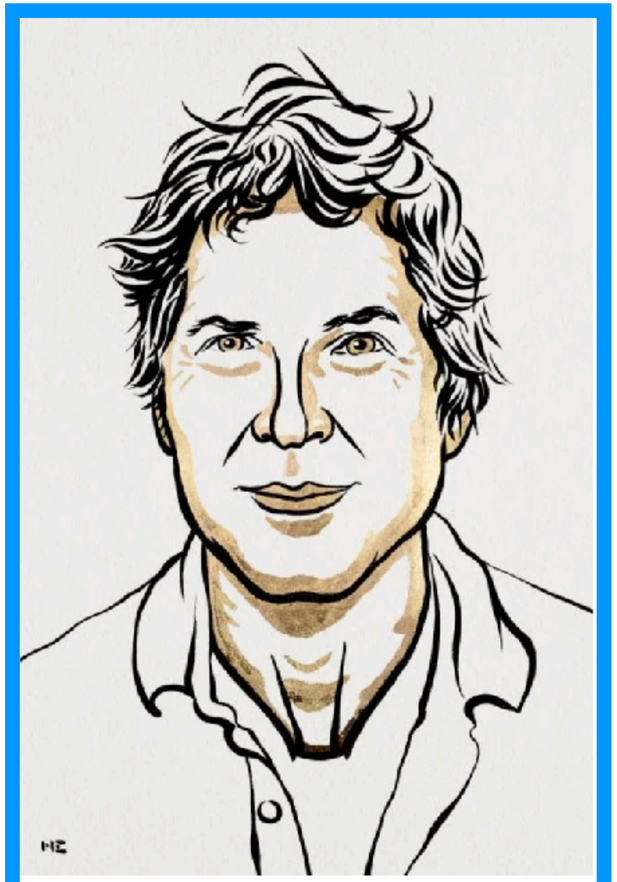




Kuhlman, B, G Dantas, GC Ireton, G Varani, BL Stoddard, and D Baker. "Design of A Novel Globular Protein Fold with Atomic-level Accuracy." Science 302, no. 5649 (21 November 2003): 1364-8.

RMSD: 1.2 Å (much better than prediction)

The Nobel Prize in Chemistry 2024 was divided, one half awarded to David Baker "for computational protein design", the other half jointly to Demis Hassabis and John M. Jumper "for protein structure prediction"



Ill. Niklas Elmehed © Nobel Prize Outreach

David Baker

Prize share: 1/2

Ill. Niklas Elmehed © Nobel Prize Outreach

Demis Hassabis

Prize share: 1/4

Ill. Niklas Elmehed © Nobel Prize Outreach

John M. Jumper

Prize share: 1/4

Machine learning for protein design

RESEARCH ARTICLES

Cite as: J. Dauparas *et al.*, *Science* 10.1126/science.add2187 (2022).

Science

Robust deep learning–based protein sequence design using ProteinMPNN

Chroma - **Generate**:Biomedicines

Article

Illuminating protein space with a programmable generative model

https://doi.org/10.1038/s41586-023-06728-8

Received: 20 December 2022

Accepted: 6 October 2023

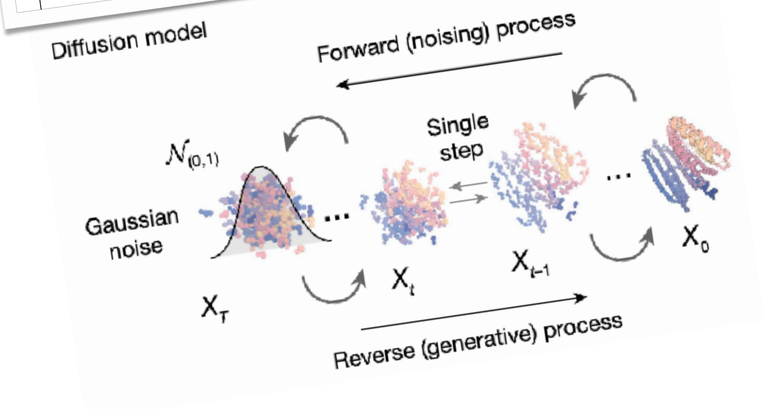
Published online: 15 November 2023

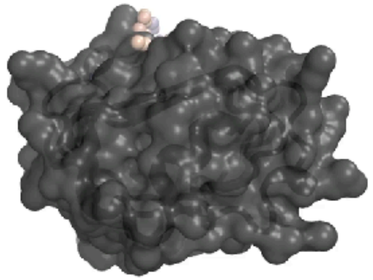
John B. Ingraham[1], Max Baranov[1], Zak Costello[1], Karl W. Barber[1], Wujie Wang[1], Ahmed Ismail[1], Vincent Frappier[1], Dana M. Lord[1], Christopher Ng-Thow-Hing[1], Erik R. [illegible], Vincent Xue[1], Sarah C. Cowles[1], Alan Leung[1], João V. Rodr[illegible], Alex M. Ayoub[1], Robin Green[1], Katherine P[illegible], Fritz Obermeyer[1], Adam R. P[illegible]

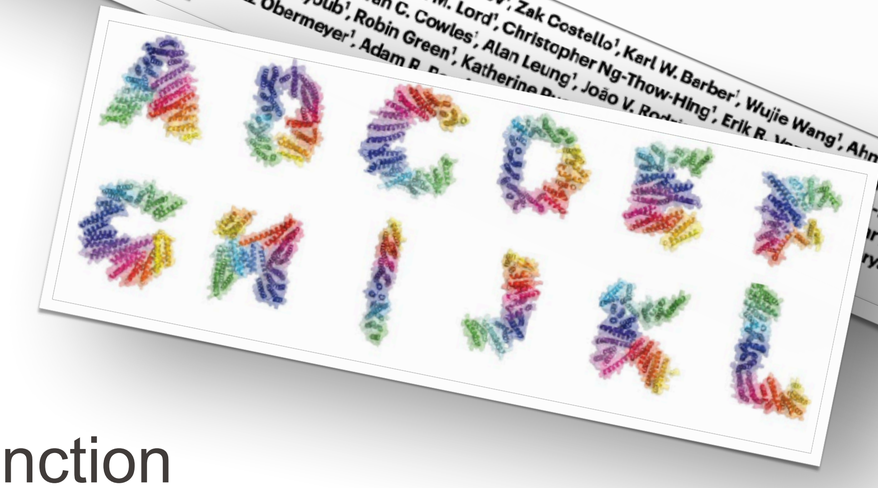
Article

De novo design of protein structure and function with RFdiffusion

Baker lab - UW





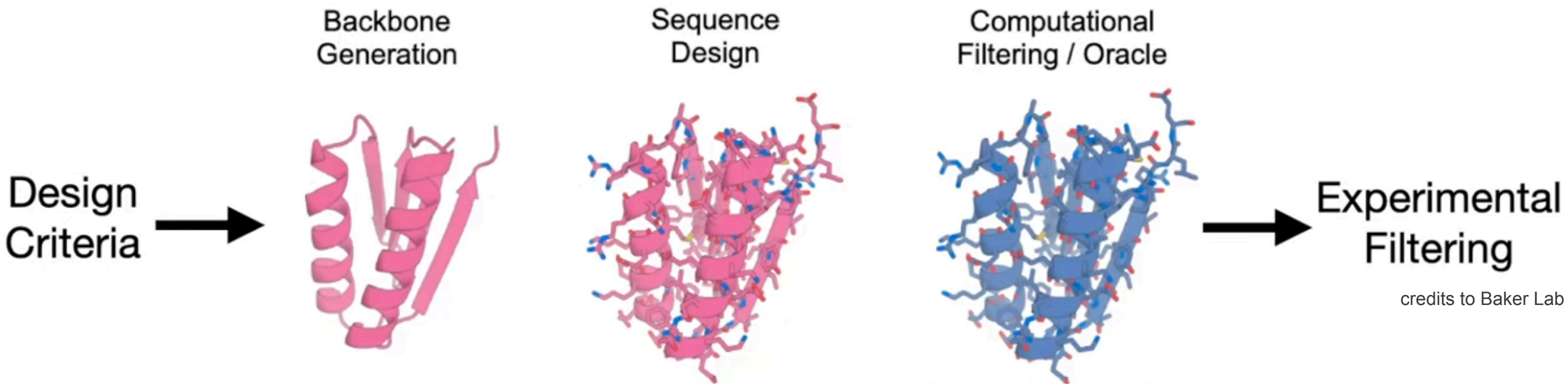


- Application to study protein evolution and function
- Protein engineering for therapeutics, synthetic biology and biotechnology

… leading to molecular engineering

- Application to study protein evolution and function
- Protein engineering for therapeutics, synthetic biology and biotechnology

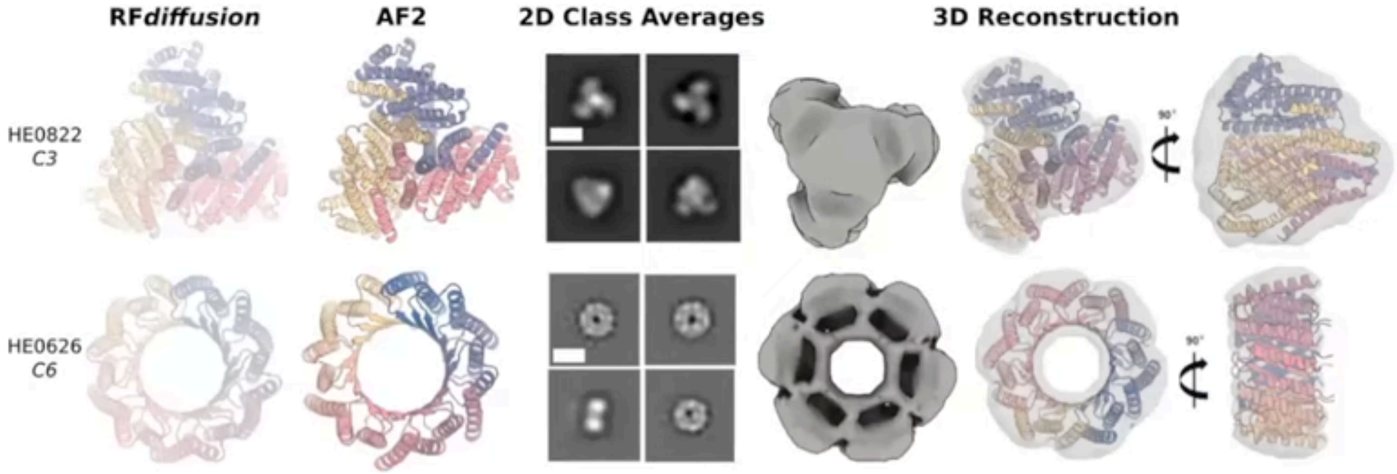
Pipeline of today's protein design



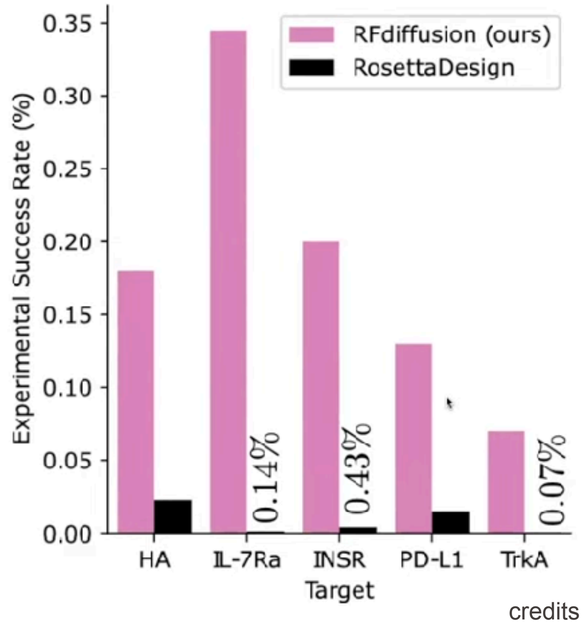
- RFdiffusion
- ProteinMPNN
- AF2

- AF2 has been key to filter potentially good protein designs
- Experimental testing is the ultimate validation of designs
- AI methods enhanced the experimental rate of success
- Protein engineering is now feasible for therapeutics, synthetic biology and biotechnology

Pipeline of today's protein design



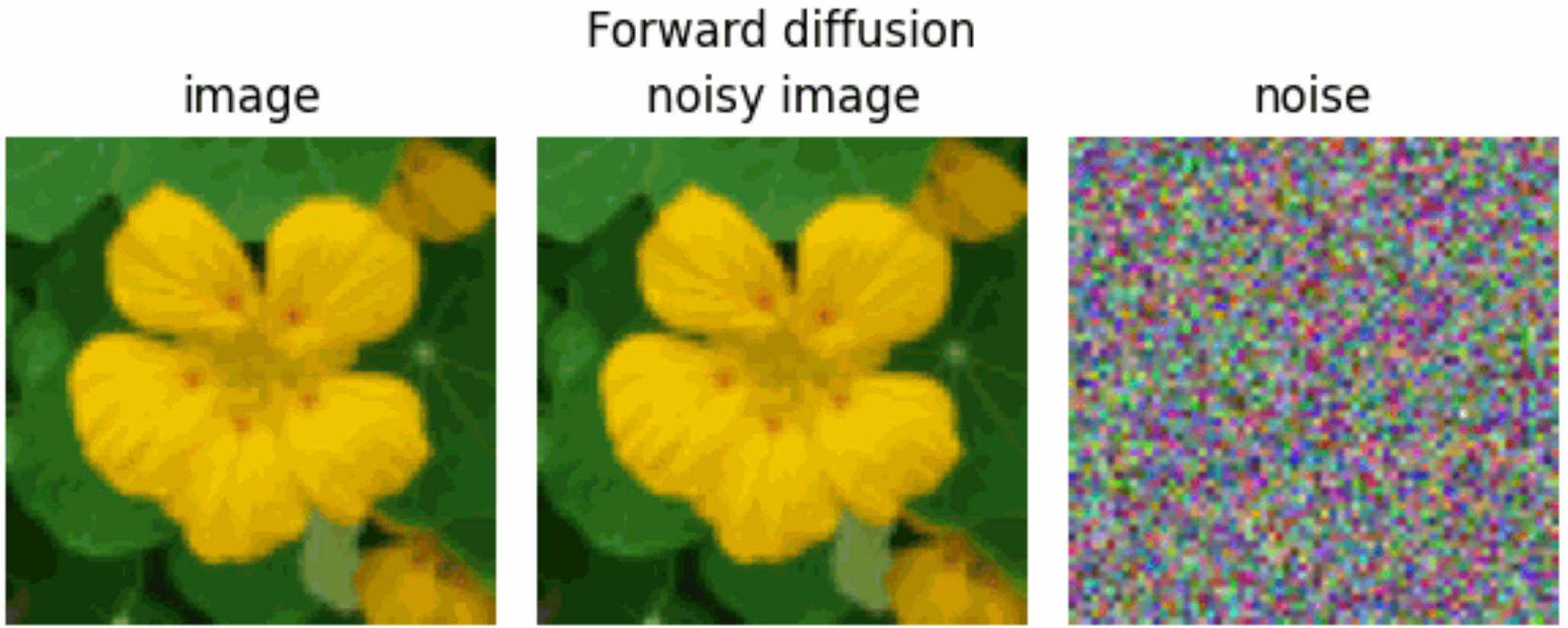
validation using cryoEM



credits to Baker Lab

- AF2 has been key to filter potentially good protein designs
- Experimental testing is the ultimate validation of designs
- AI methods enhanced the experimental rate of success
- Protein engineering is now feasible for therapeutics, synthetic biology and biotechnology

Machine learning for protein design



András Béres

- Denoising Diffusion Models - as those used in DALL-E
- Trained to denoise noisy images, they can generate images by iteratively denoising pure noise

Article

De novo design of protein structure and function with RFdiffusion

https://doi.org/10.1038/s41586-023-06415-8

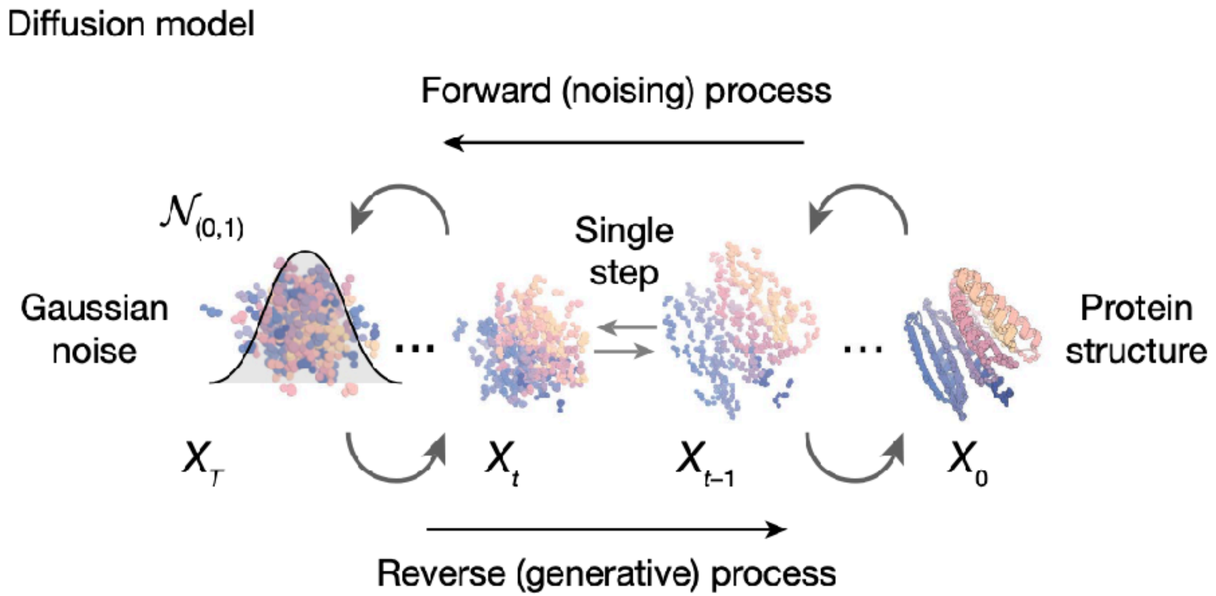
Received: 14 December 2022

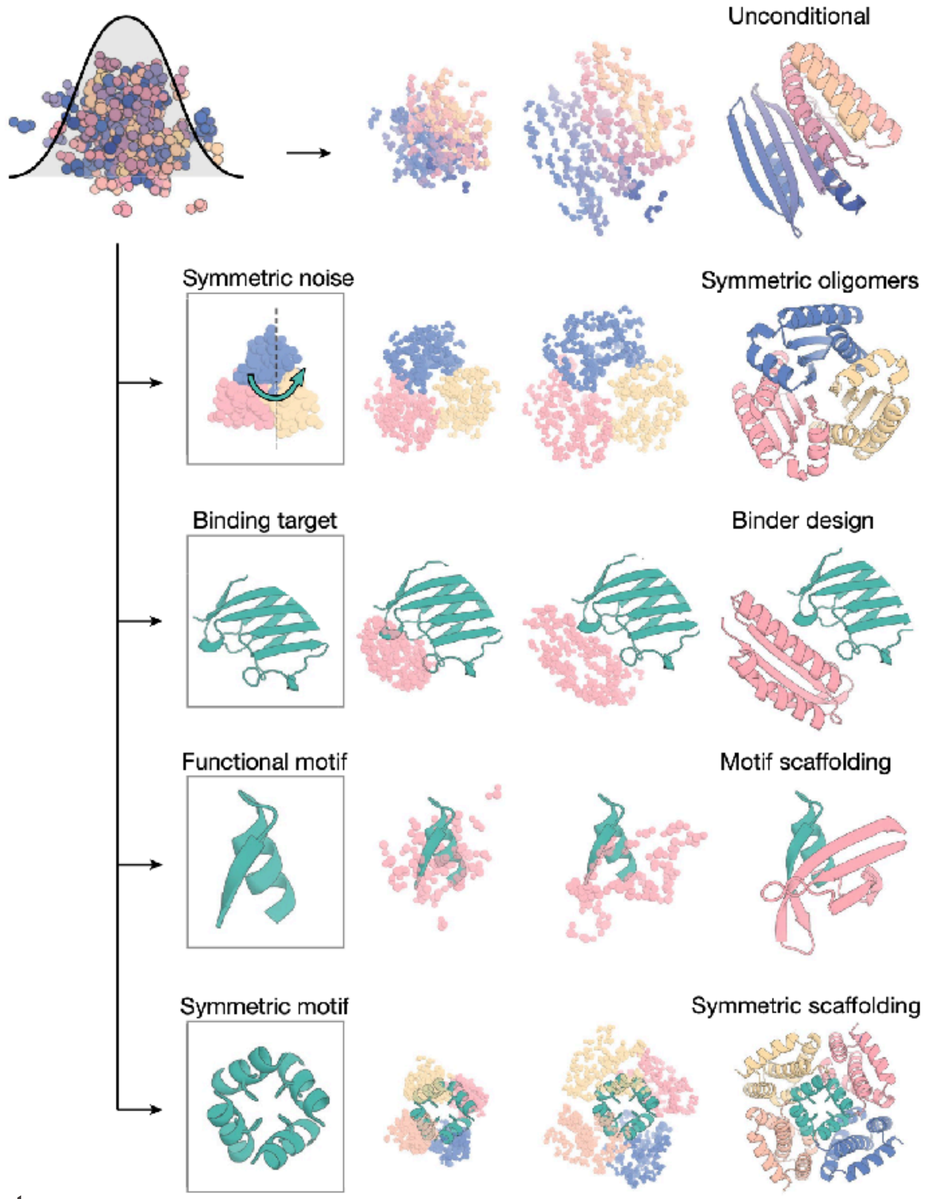
Accepted: 7 July 2023

Published online: 11 July 2023

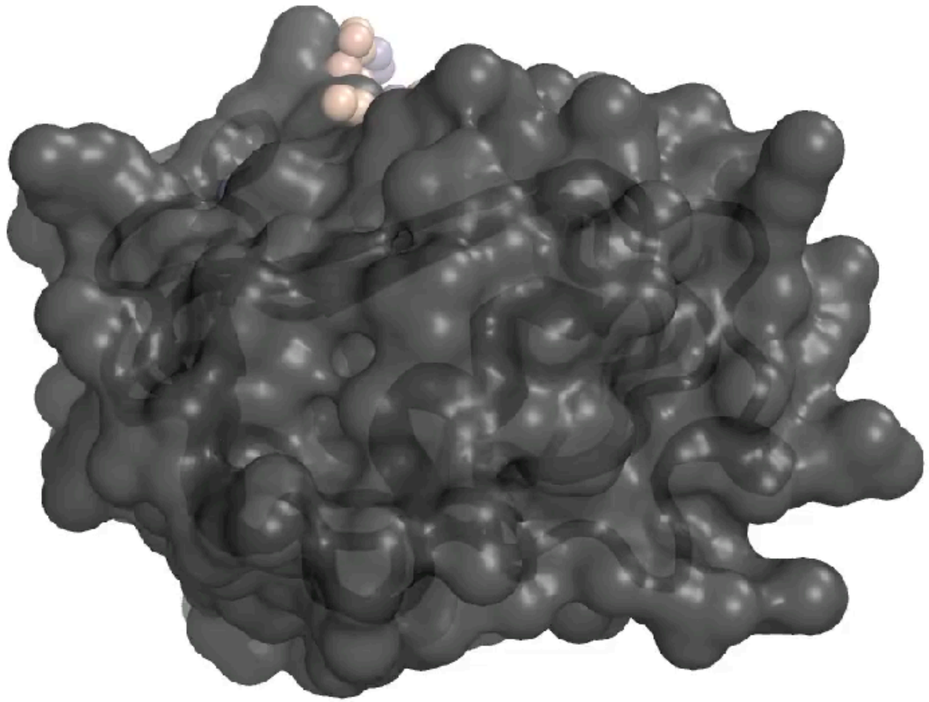
Open access

Joseph L. Watson[1,2,15], David Juergens[1,2,3,15], Nathaniel R. Bennett[1,2,3,15], Brian L. Trippe[2,4,5,15], Jason Yim[2,6,15], Helen E. Eisenach[1,2,15], Woody Ahern[1,2,7,15], Andrew J. Borst[1,2], Robert J. Ragotte[1,2], Lukas F. Milles[1,2], Basile I. M. Wicky[1,2], Nikita Hanikel[1,2], Samuel J. Pellock[1,2], Alexis Courbet[1,2,8], William Sheffler[1,2], Jue Wang[1,2], Preetham Venkatesh[1,2,9], Isaac Sappington[1,2,9], Susana Vázquez Torres[1,2,9], Anna Lauko[1,2,9], Valentin De Bortoli[8], Emile Mathieu[10], Sergey Ovchinnikov[11,12], Regina Barzilay[6], Tommi S. Jaakkola[6], Frank DiMaio[1,2], Minkyung Baek[13] & David Baker[1,2,14]✉





- the reverse process is learned using a neural network
- its loss function encourages the reverse process to accurately estimate how the data transitions from one noisy step to the previous step.



https://www.bakerlab.org/2022/11/30/diffusion-model-for-protein-design/

Cite as: J. Dauparas *et al.*, *Science* 10.1126/science.add2187 (2022).

Robust deep learning–based protein sequence design using ProteinMPNN

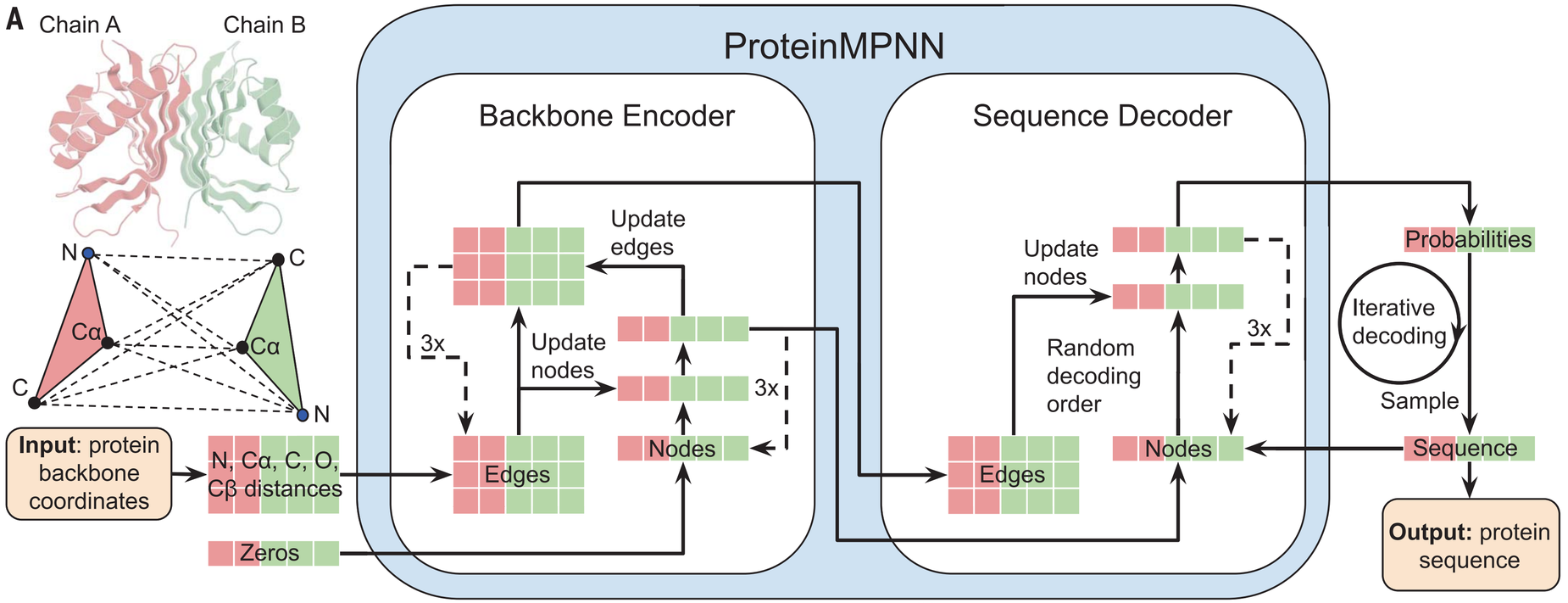
J. Dauparas[1,2], I. Anishchenko[1,2], N. Bennett[1,2,3], H. Bai[1,2,4], R. J. Ragotte[1,2], L. F. Milles[1,2], B. I. M. Wicky[1,2], A. Courbet[1,2,4], R. J. de Haas[5], N. Bethel[1,2,4], P. J. Y. Leung[1,2,3], T. F. Huddy[1,2], S. Pellock[1,2], D. Tischer[1,2], F. Chan[1,2], B. Koepnick[1,2], H. Nguyen[1,2], A. Kang[1,2], B. Sankaran[6], A. K. Bera[1,2], N. P. King[1,2], D. Baker[1,2,4]*

[1]Department of Biochemistry, University of Washington, Seattle, WA, USA. [2]Institute for Protein Design, University of Washington, Seattle, WA, USA. [3]Molecular Engineering Graduate Program, University of Washington, Seattle, WA, USA. [4]Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA. [5]Department of Physical Chemistry and Soft Matter, Wageningen University and Research, Wageningen, Netherlands. [6]Berkeley Center for Structural Biology, Molecular Biophysics and Integrated Bioimaging, Lawrence Berkeley Laboratory, Berkeley, CA, USA.

*Corresponding author. Email: dabaker@uw.edu

While deep learning has revolutionized protein structure prediction, almost all experimentally characterized de novo protein designs have been generated using physically based approaches such as Rosetta. Here we describe a deep learning–based protein sequence design method, ProteinMPNN, with outstanding performance in both in silico and experimental tests. On native protein backbones, ProteinMPNN has a sequence recovery of 52.4%, compared to 32.9% for Rosetta. The amino acid sequence at different positions can be coupled between single or multiple chains, enabling application to a wide range of current protein design challenges. We demonstrate the broad utility and high accuracy of ProteinMPNN using X-ray crystallography, cryoEM and functional studies by rescuing previously failed designs, made using Rosetta or AlphaFold, of protein monomers, cyclic homo-oligomers, tetrahedral nanoparticles, and target binding proteins.

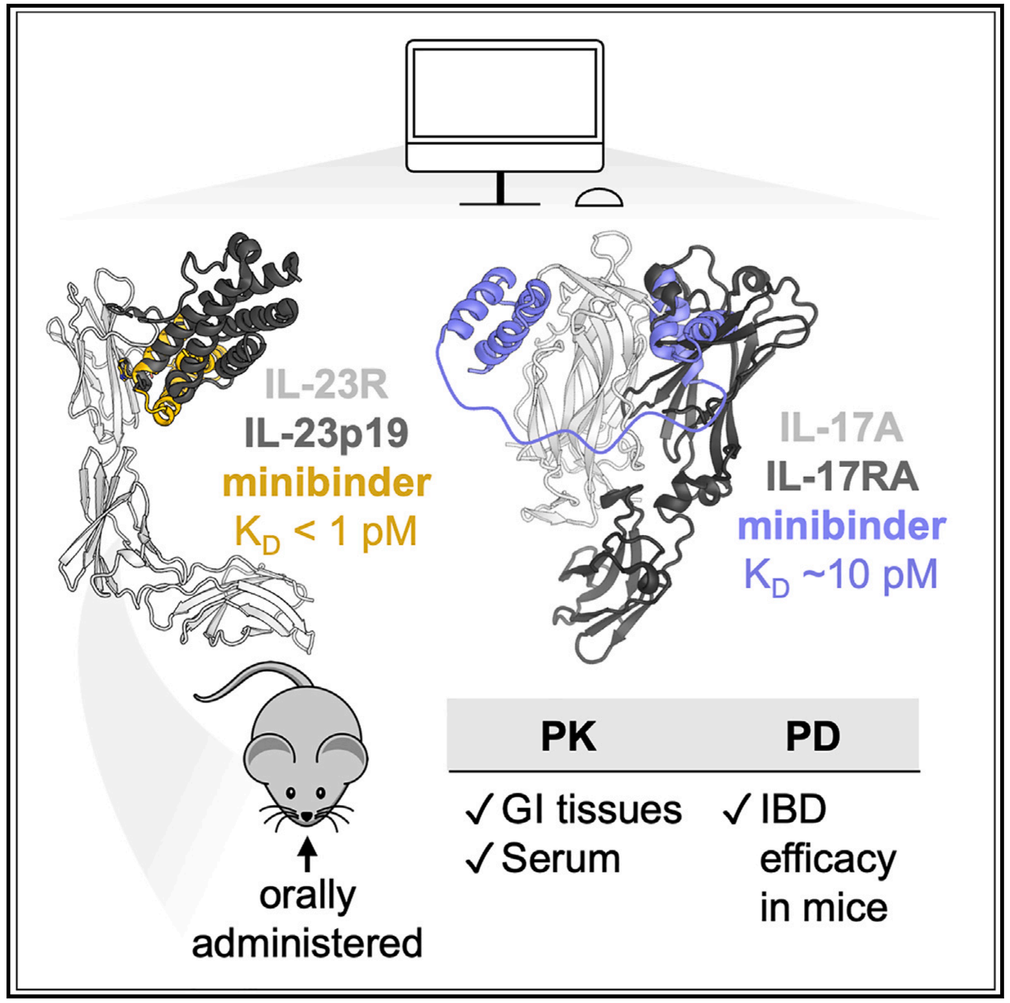
- Backbone distances are encoded and processed using a message-passing neural network (Encoder) to obtain graph node and edge features.
- The encoded features, together with a partial sequence, are used to generate amino acids iteratively in a random decoding order.



Cell

Preclinical proof of principle for orally delivered Th17 antagonist miniproteins

Graphical abstract



Authors

Stephanie Berger, Franziska Seeger, Ta-Yi Yu, ..., Matthias Siebeck, Roswitha Gropp, David Baker

Correspondence

berger389@gmail.com (S.B.), dabaker@uw.edu (D.B.)

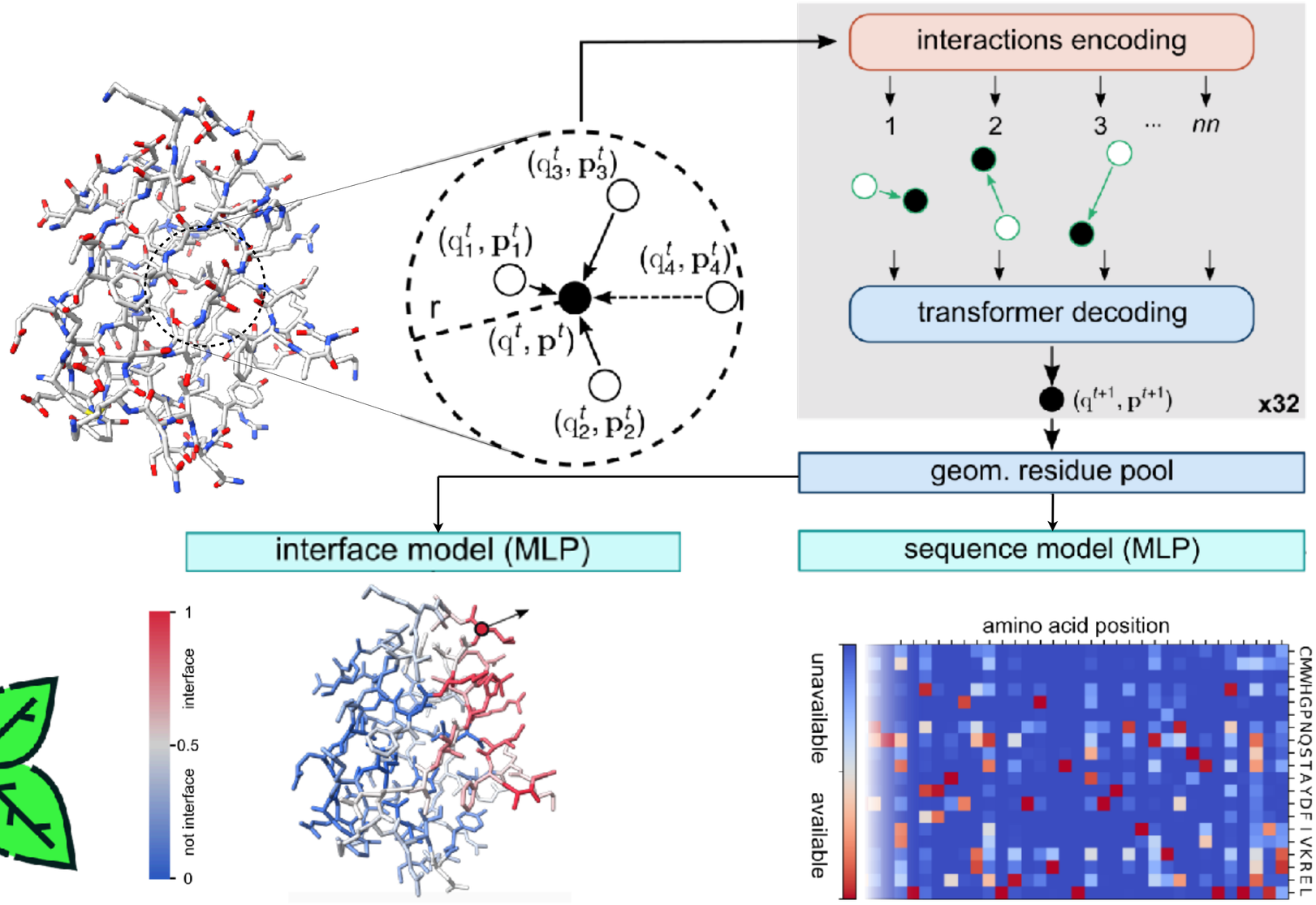
Highlights

- Computational design yielded low- and sub-pM minibinders of IL-17A and IL-23R
- IL-23R minibinders are extremely resistant to heat, acid, and proteolysis
- Oral IL-23R minibinder is as effective as a clinical mAb in mouse colitis

Berger et al., 2024, Cell *187*, 4305–4317
August 8, 2024
https://doi.org/10.1016/j.cell.2024.05.052

Matteo Dal Peraro

Protein Structure Transformer @LBM



PeSTo: binding interfaces

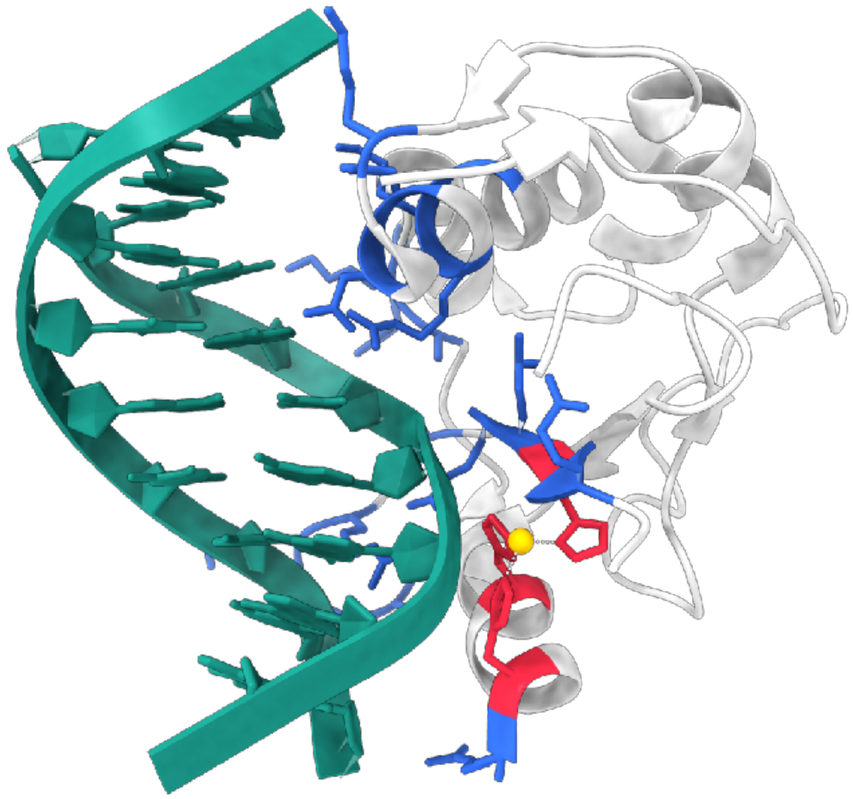
Krapp et al. *Nat Comms* 2023

CARBonAra: molecular design

Krapp et al. *Nat Comms* 2024

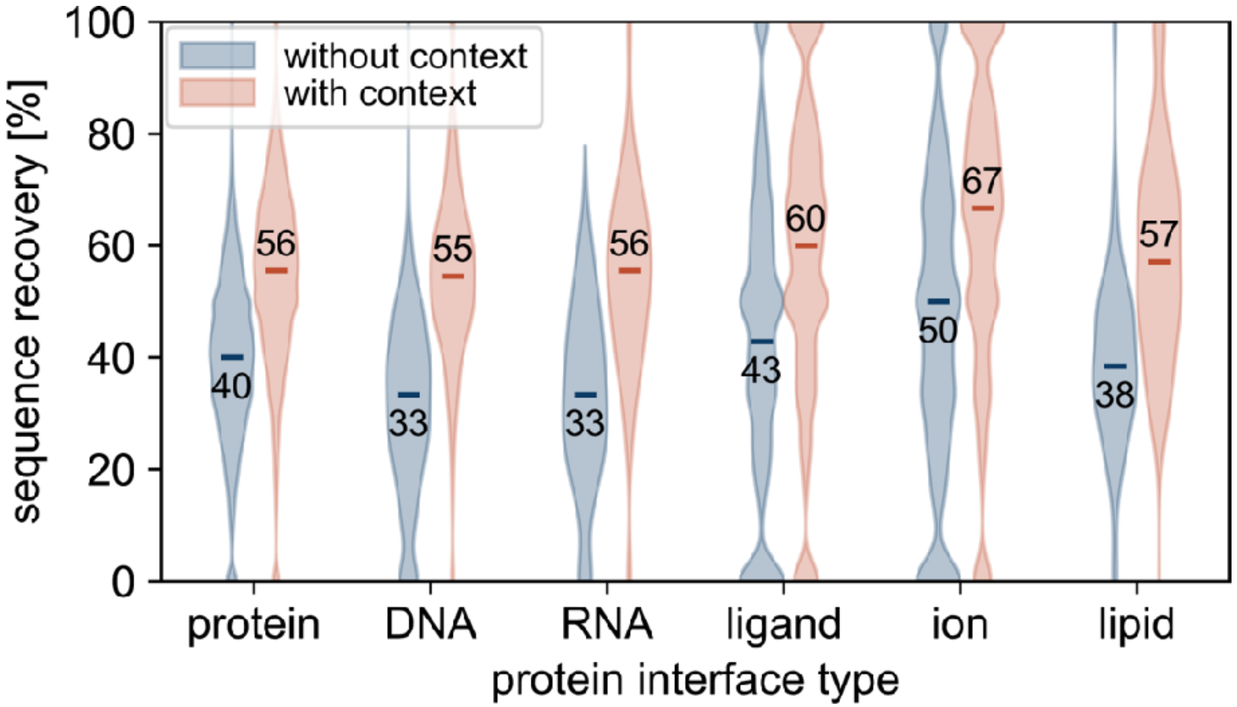
Unique ability — context awareness

- example with context



colicin E7

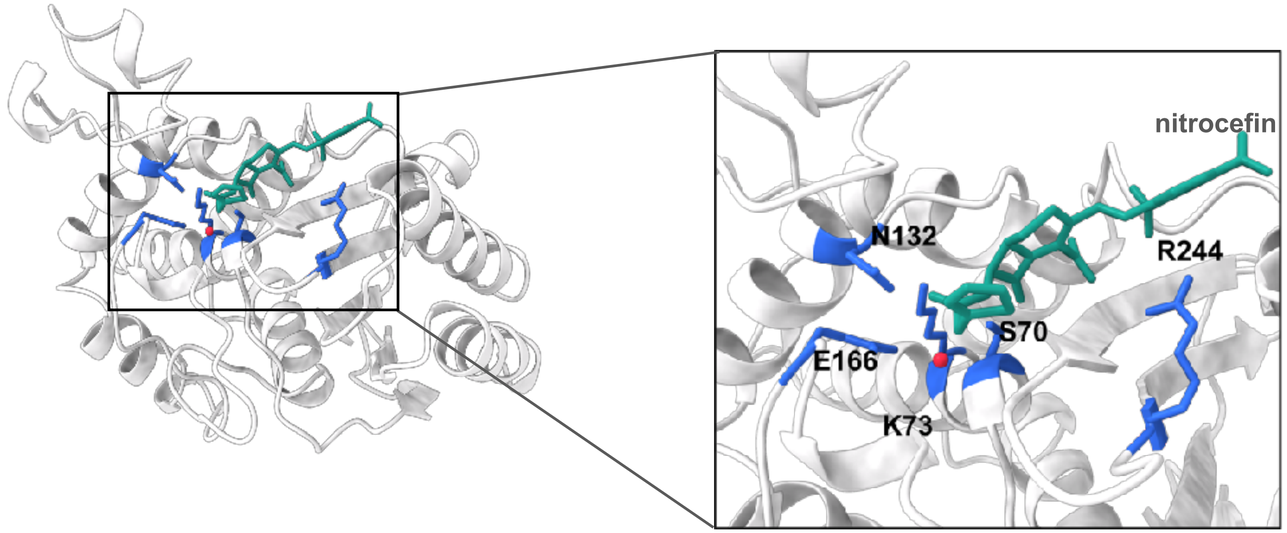
- large-scale benchmark



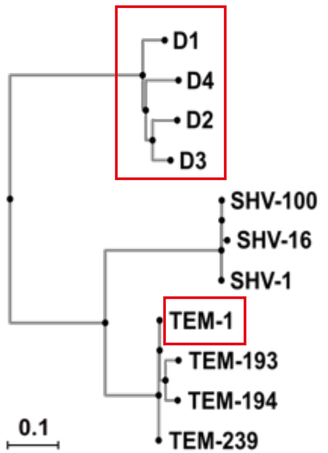
1000 structures sampled with maximum 30% sequence identity and separate C.A.T.H. classification from training set

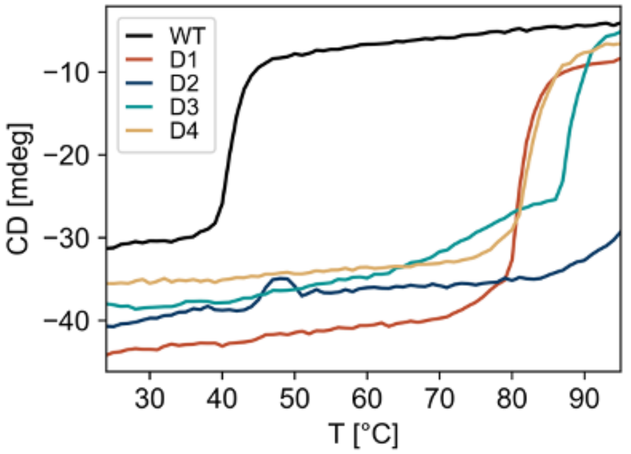
Can we re-engineer an enzyme?

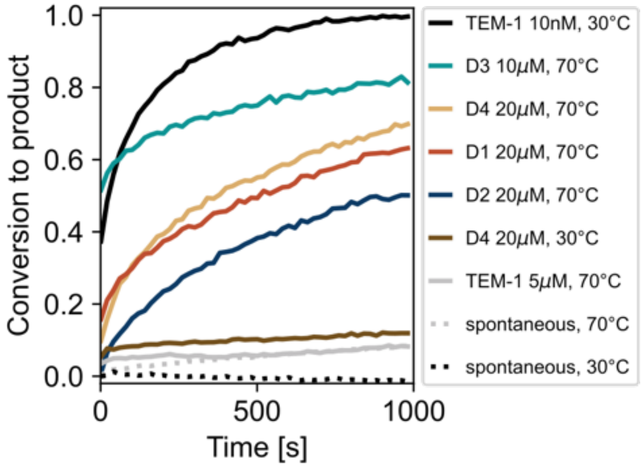
- TEM-1 serine β-lactamase



only 50% sequence identity







- sequences generation with substrate as constraint
- selected 10 top-ranked predictions based on plDDT
- 4/10 designs are soluble and monomeric
- they are folded and more thermostable than wild-type TEM-1
- catalytically active at high T - not as the wild-type yet
- represent a separate subclass of β-lactamases

Krapp et al. *Nat Comms* 2024

The future is bright and exciting …

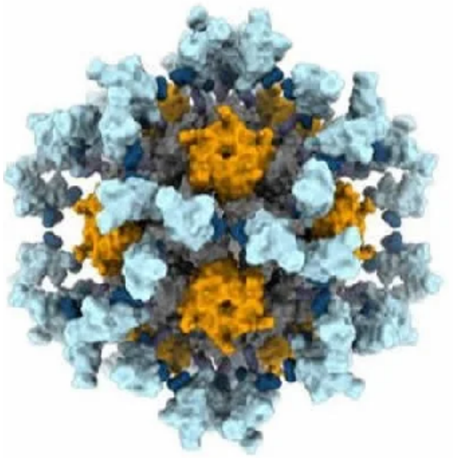
… biomolecular design will address many societal needs

- **Medicine**

  vaccines & antivirals

  smart medicines

  drug delivery



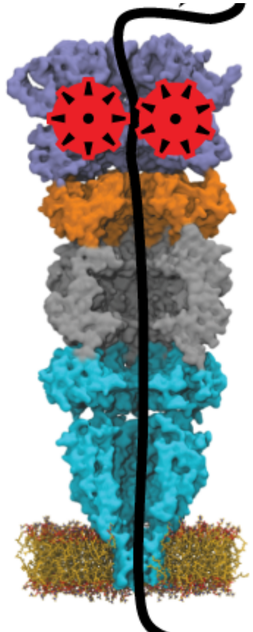
SARS-CoV-2 RBD
nanoparticle immunogen (Cell 2020)

- **Biotechnology**

  protein-silicon devices

  bio-based computers

  nanoscale manufacturing



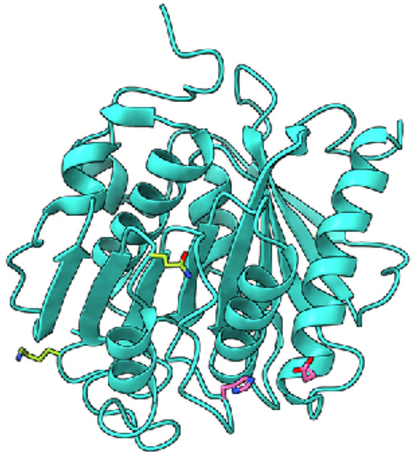
SM proteomics with
biological nanopores
(Nat Chem 2021)

- **Sustainability**

  artificial photosynthesis

  CO_2 sequestration

  plastic degradation



FAST-PETase
(Nature 2022)

Matteo Dal Peraro

We weren’t alone !!!!

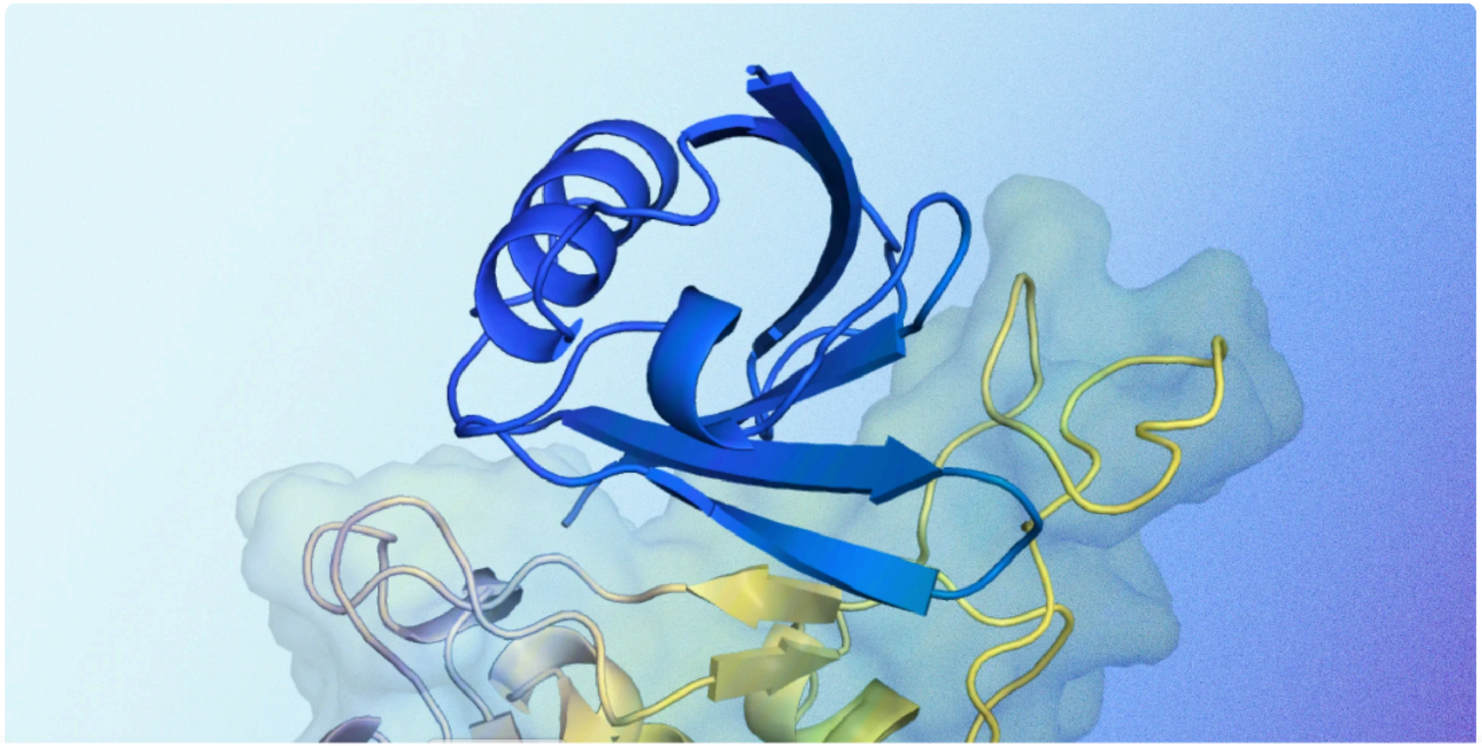
Google DeepMind About Research Technologies Impact Discover

AlphaProteo generates novel proteins for biology and health research

5 SEPTEMBER 2024

Protein Design and Wet Lab teams

Share



Designing Life with AI

https://www.designinglifewithai.com/

contact the MAKE team for ongoing projects offered by labs

We're thrilled to introduce "Designing Life with AI" at EPFL, where AI and protein design intersect, involving faculty, professors, and 40 students collaborating on topics like binder design and phosphosite engineering to kinase remodeling. After a year of innovative research, our projects are now being tested in the wet-lab, and we're working on creating a pipeline and resources for new students, aiming to expand our project and make EPFL a hub for protein design.