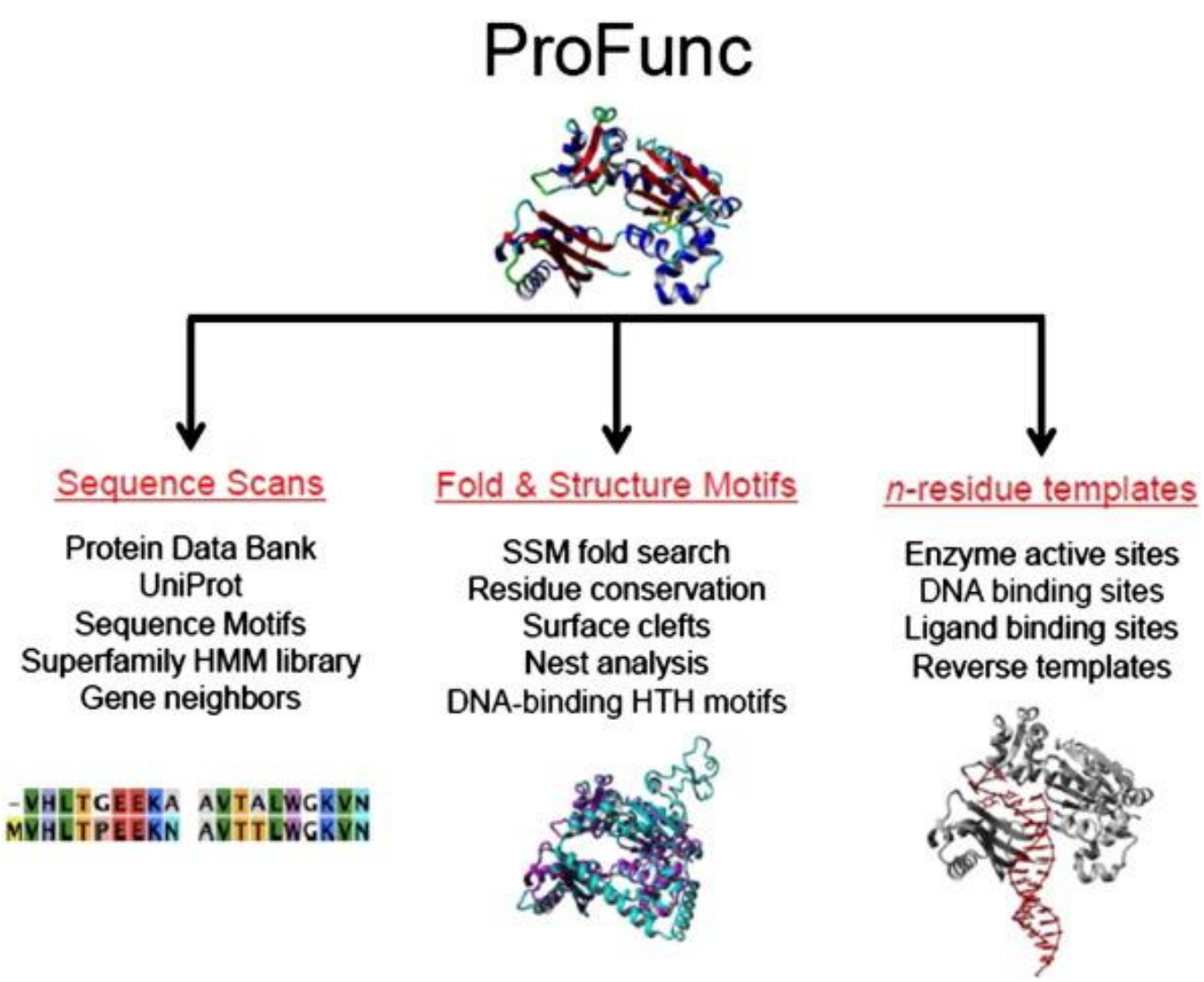
Lezione 7

Andrea Ilari

Relazione – struttura attività nelle proteine

Analisi strutturale

OK
ho una struttura…
…e adesso?

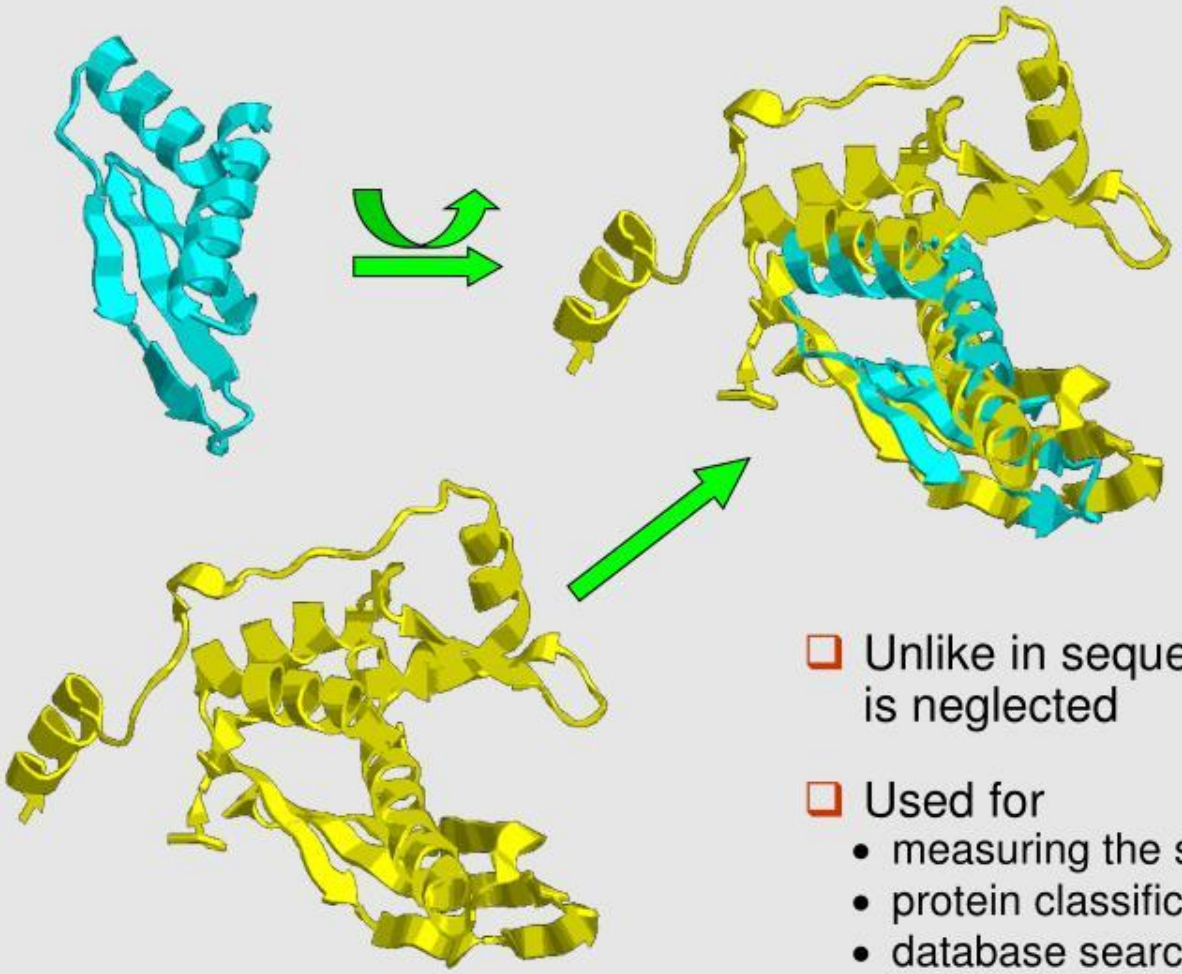


Biochemical functional predictions for protein structures of unknown or uncertain function

https://www.sciencedirect.com/science/article/pii/S2001037015000070#f0015

https://www.ebi.ac.uk/msd-srv/ssm/

Structure alignment



Structure alignment may be defined as identification of residues occupying "equivalent" geometrical positions

- Unlike in sequence alignment, residue type is neglected
- Used for
  - measuring the structural similarity
  - protein classification and functional analysis
  - database searches

https://www.ebi.ac.uk/msd-srv/ssm/

Sequence and Structure Alignments

Sequence alignment

Based on residue identity, sometimes with a modified alphabet

```
--AARNEDDDGKMPSTF-L
E-AARNFG-DGK--STFIL
```

Used for:

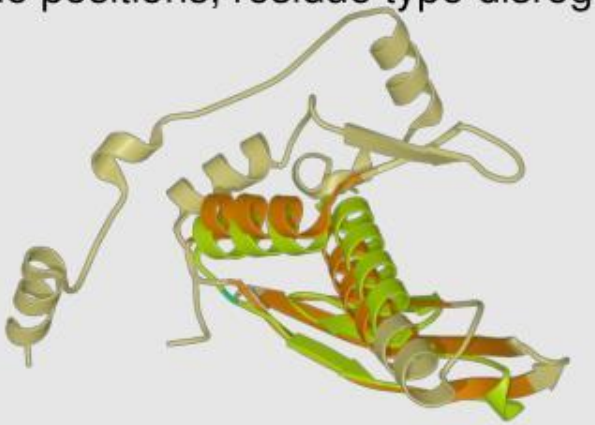
- evolution studies
- protein function analysis
- guessing on structure similarity

Algorithms: Dynamic programming + heuristics

Applications: BLAST, FASTA, FLASH and others

Structure alignment

Based on geometrical equivalence of residue positions, residue type disregarded



Used for:

- protein function analysis
- some aspects of evolution studies

Algorithms: Dynamic programming, graph theory, MC, geometric hashing and others

Applications: DALI, VAST, CE, MASS, SSM and others

https://www.ebi.ac.uk/msd-srv/ssm/

Methods

Many methods are known:

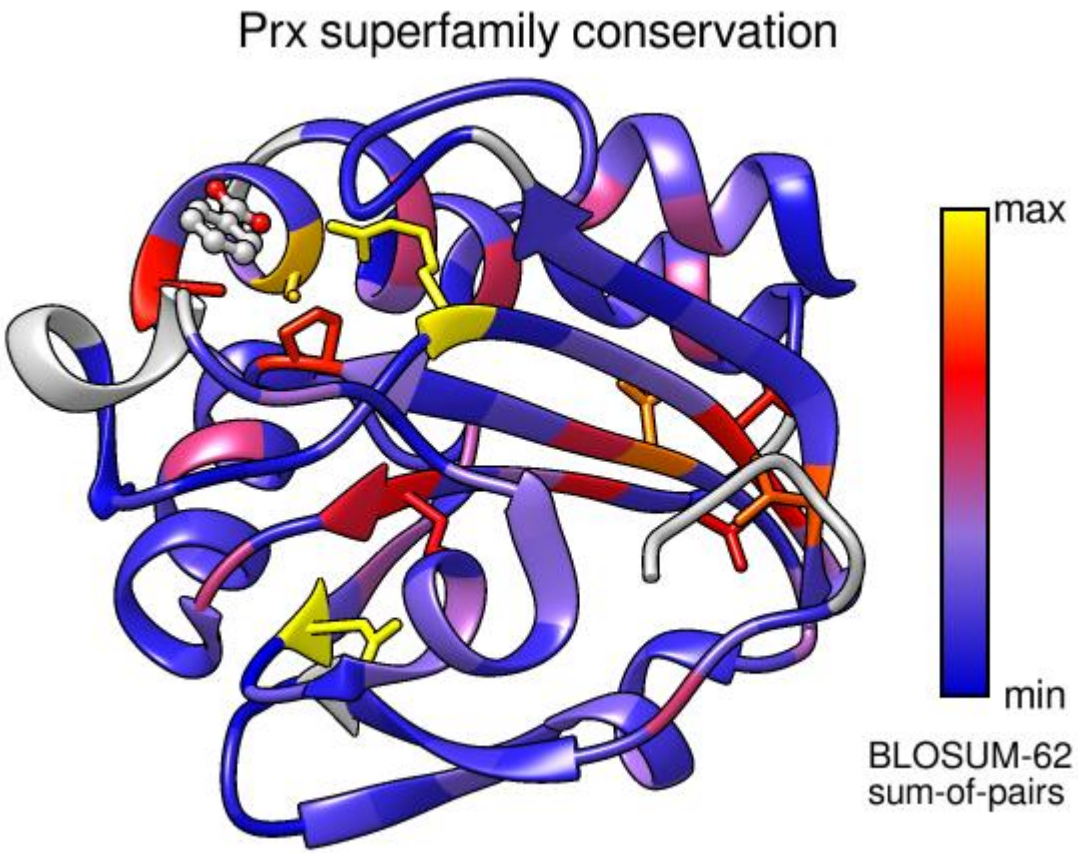
- Distance matrix alignment (**DALI**, Holm & Sander, EBI)
- Vector alignment (**VAST**, Bryant et. al. NCBI)
- Depth-first recursive search on SSEs (**DEJAVU**, Madsen & Kleywegt, Uppsala)
- Combinatorial extension (**CE**, Shindyalov & Bourne, SDSC)
- Dynamical programming on C_α (Gerstein & Levitt)
- Dynamical programming on SSEs (**SSA**, Singh & Brutlag, Stanford University)
- many more …

SSM employs a 2-step procedure:

A Initial structure alignment and superposition using SSE graph matching

B C_α - alignment

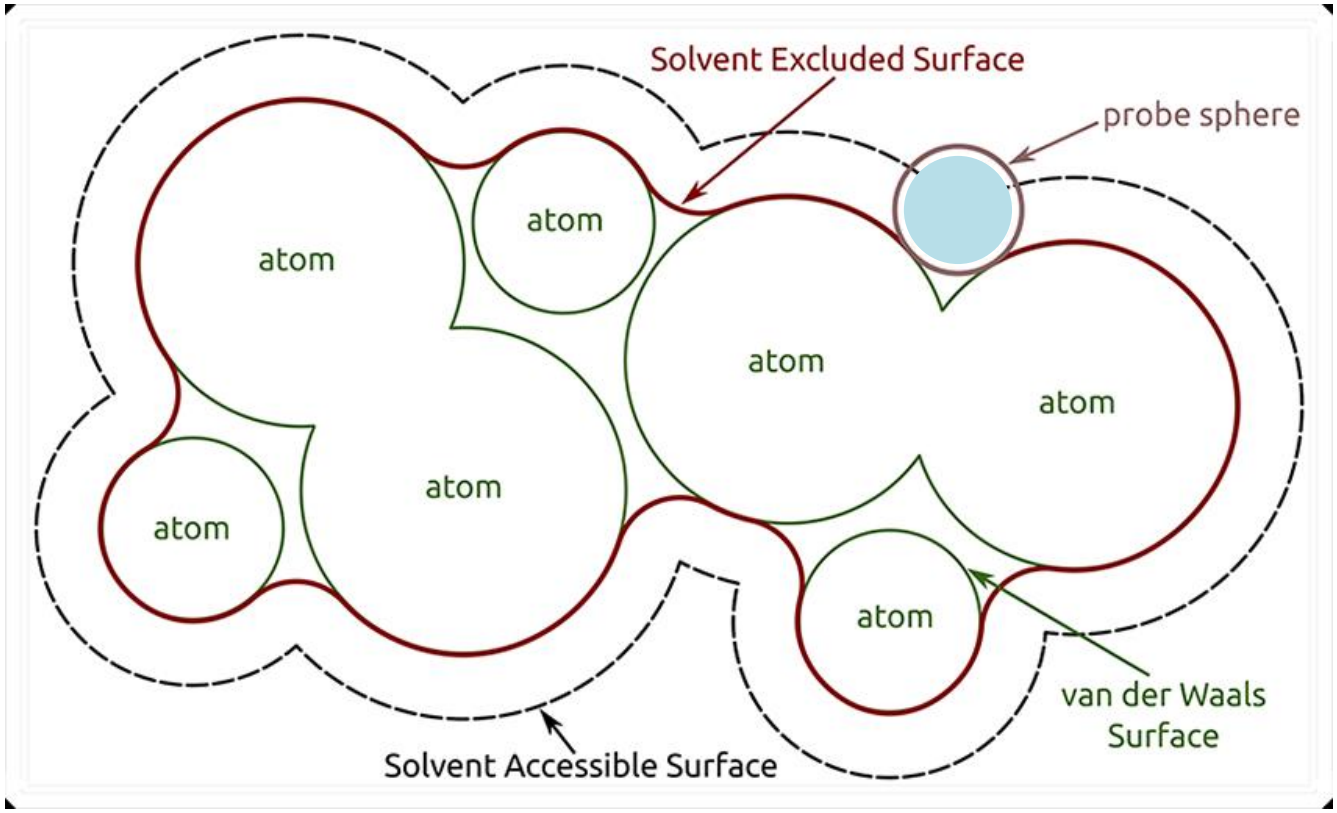
1. Colorare per CONSERVAZIONE



Tutorial in Chimera

https://www.cgl.ucsf.edu/chimera/data/tutorials/systems/outline.html

2. Calcolo della SUPERFICIE

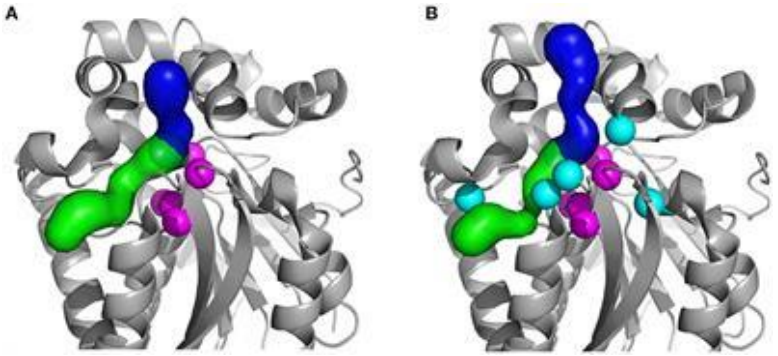


Solitamente si utilizza l'algoritmo di Shrake e Rupley: Si attribuiscono i raggi di VDW
agli atomi della proteina e poi si fa «*rotolare*» una sfera di raggio fisso compatibile con il raggio di una molecola d'acqua (1.4 Å)

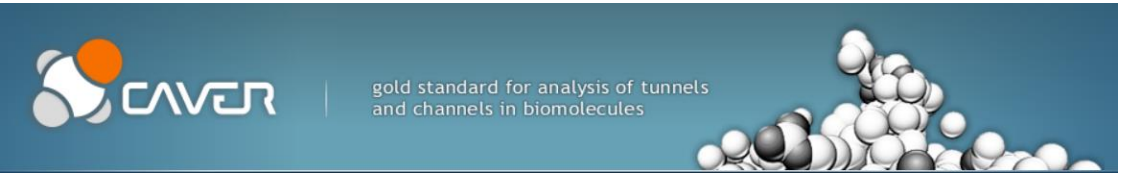
Shrake A, Rupley JA. (1973). Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *J Mol Biol* **79**(2):351-71. DOI: 10.1016/0022-2836(73)90011-9

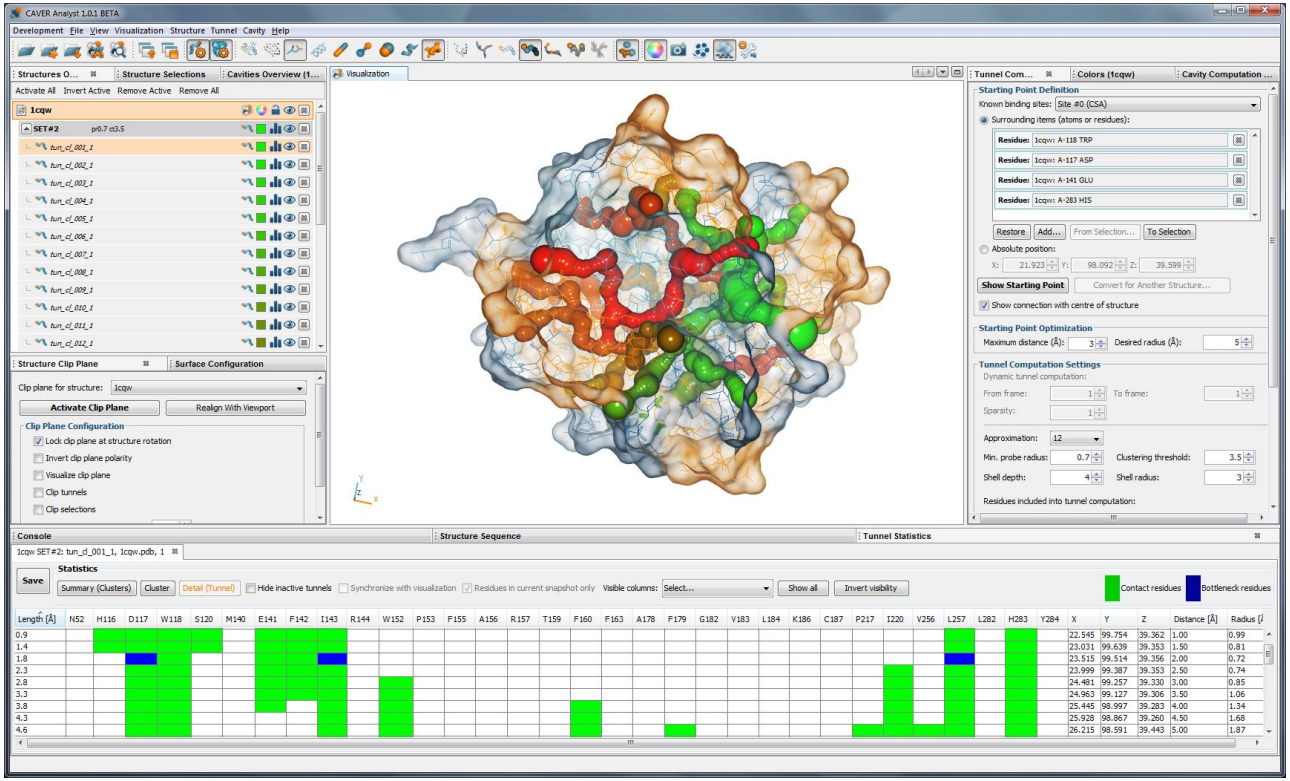
3. Calcolo delle CAVITA'

Sono interessato soprattutto alle **cavità e tunnel**, che sono possibili siti di binding del substrato o per potenziali farmaci



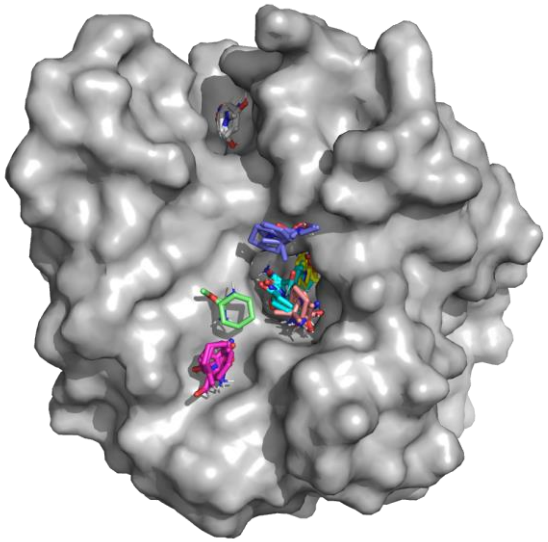
Caver: http://www.caver.cz/index.php?sid=100



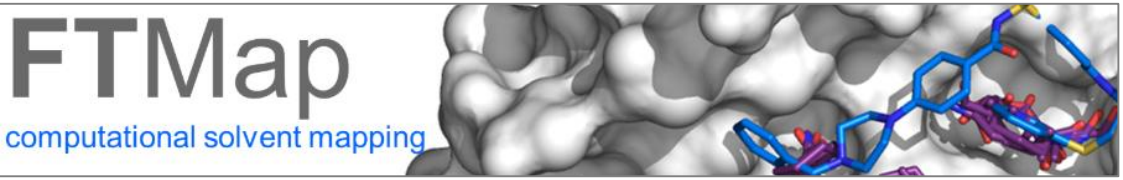


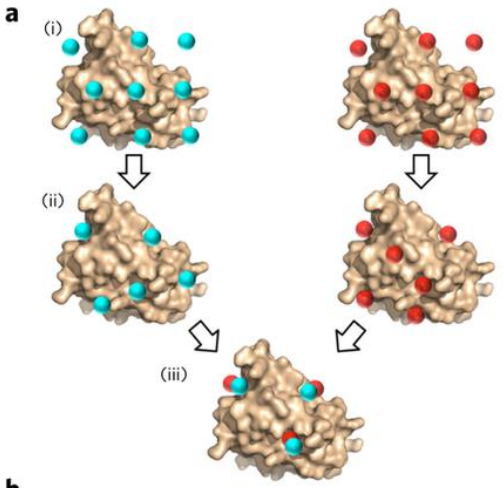
3. Calcolo delle CAVITA'

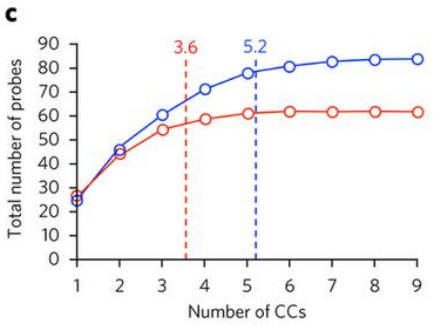
Sono interessato soprattutto alle **cavità e tunnel**, che sono possibili siti di binding del substrato o per potenziali farmaci

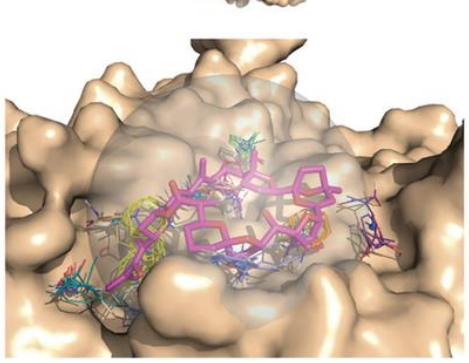


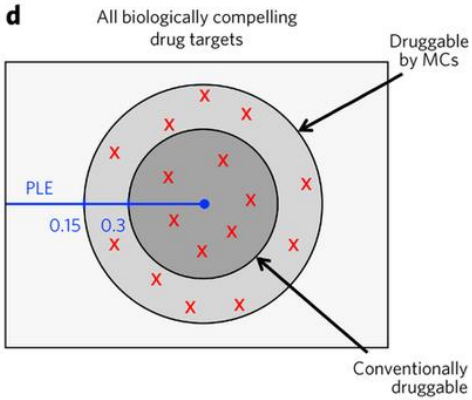
Hotspot analysis (FTMap) – http://ftmap.bu.edu



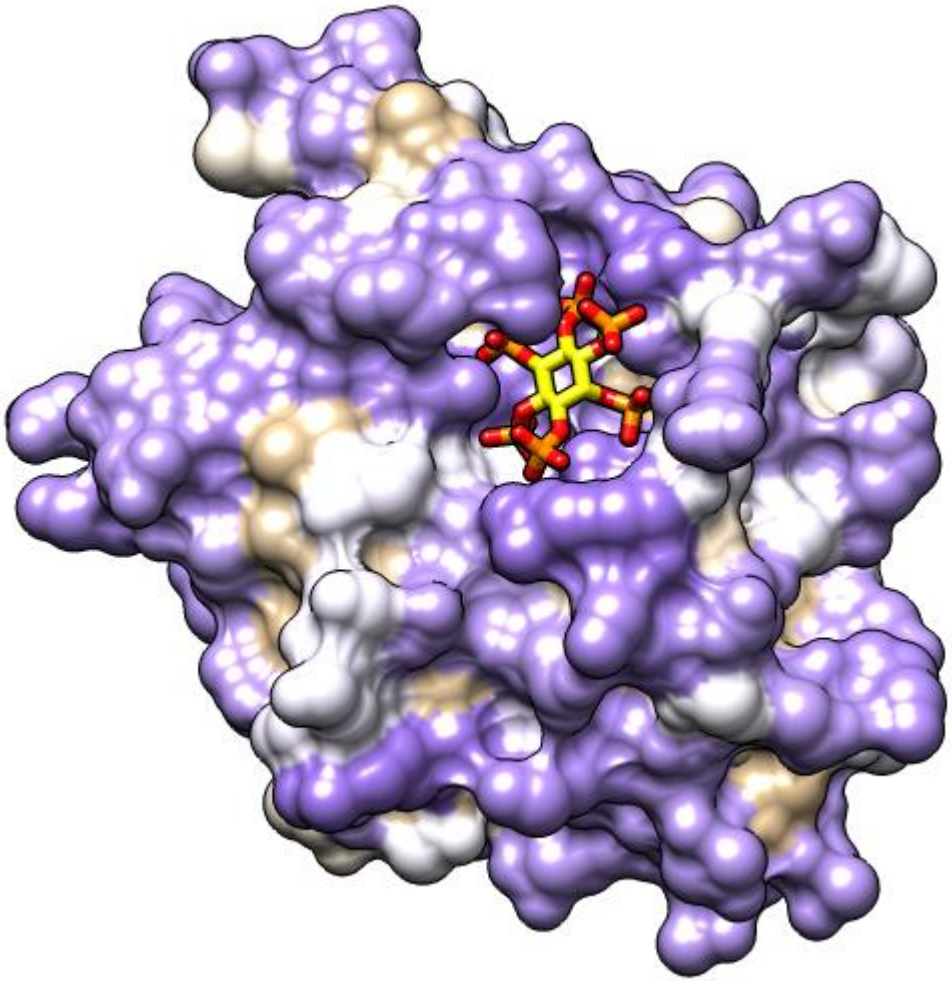








4. Colorare la superficie per IDROFOBICITA’



Tutorial in Chimera (superfici di potenziale e idrofobicità)

https://www.cgl.ucsf.edu/chimera/current/docs/UsersGuide/tutorials/surfprop.html

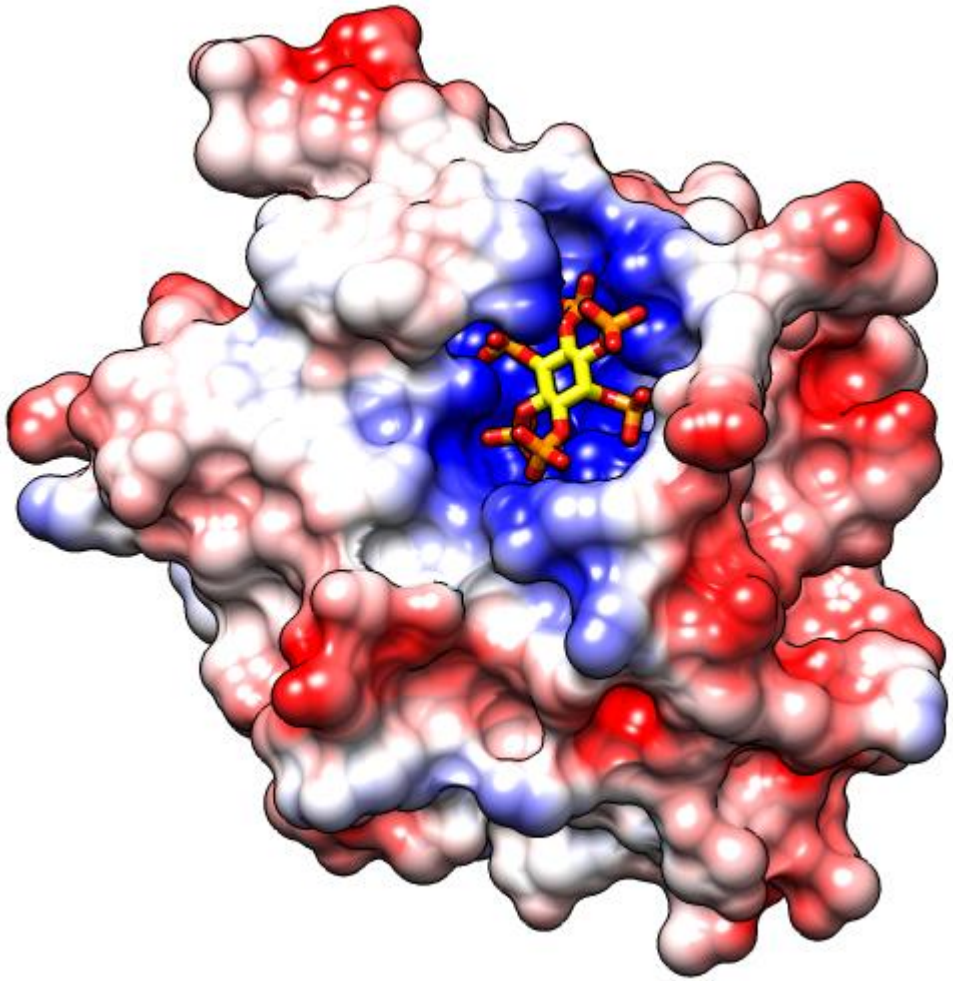
5. Colorare la superficie per CARICA

1. Coulombiano:
potenziale calcolato
Secondo la legge di Coulomb

$$\varphi = \Sigma \left[q_i / (\varepsilon d_i) \right]$$

Dove ***q*** è la carica e ***d*** la distanza dall'atomo ***i*** mentre ***ε*** la costant dielettrica del mezzo

2. Poisson-Boltzmann:
Più accurata perché tiene conto
Della variazione di costante dielettrica
Dovuta alla forma della superficie

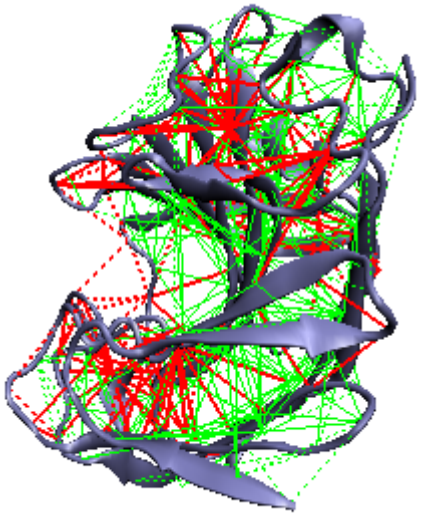


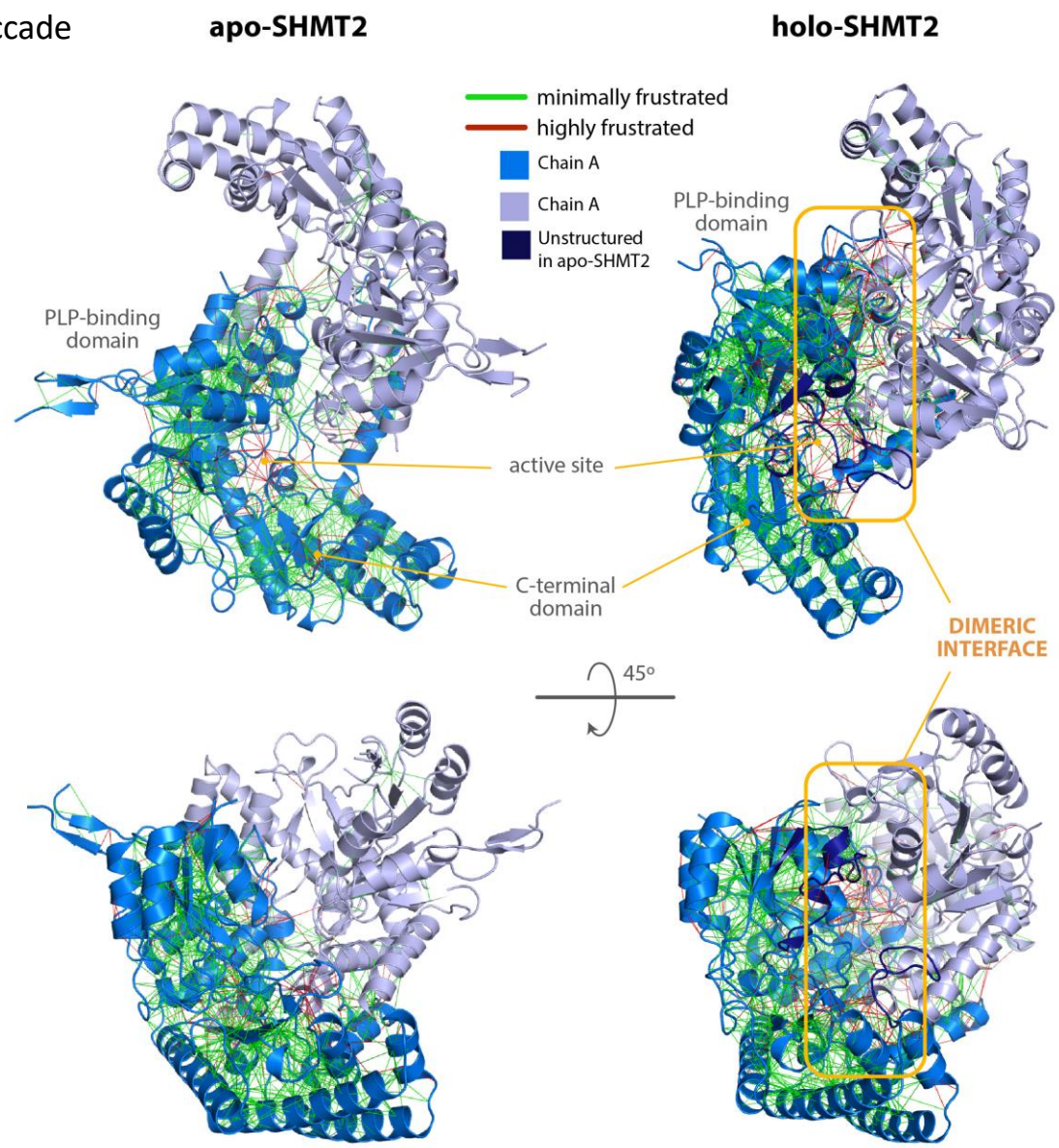
Tutorial in Chimera (superfici di potenziale e idrofobicità)

https://www.cgl.ucsf.edu/chimera/current/docs/UsersGuide/tutorials/surfprop.html

5. Grado di FRUSTRAZIONE

Frustrazione. Normalmente le proteine acquistano una conformazione che rappresenta un minimo di energia. Ma non sempre è così, quando questo accade si dice che quella zona della proteina ha un certo grado di frustrazione.
Spesso la frustrazione è conservata per una certa «funzione»

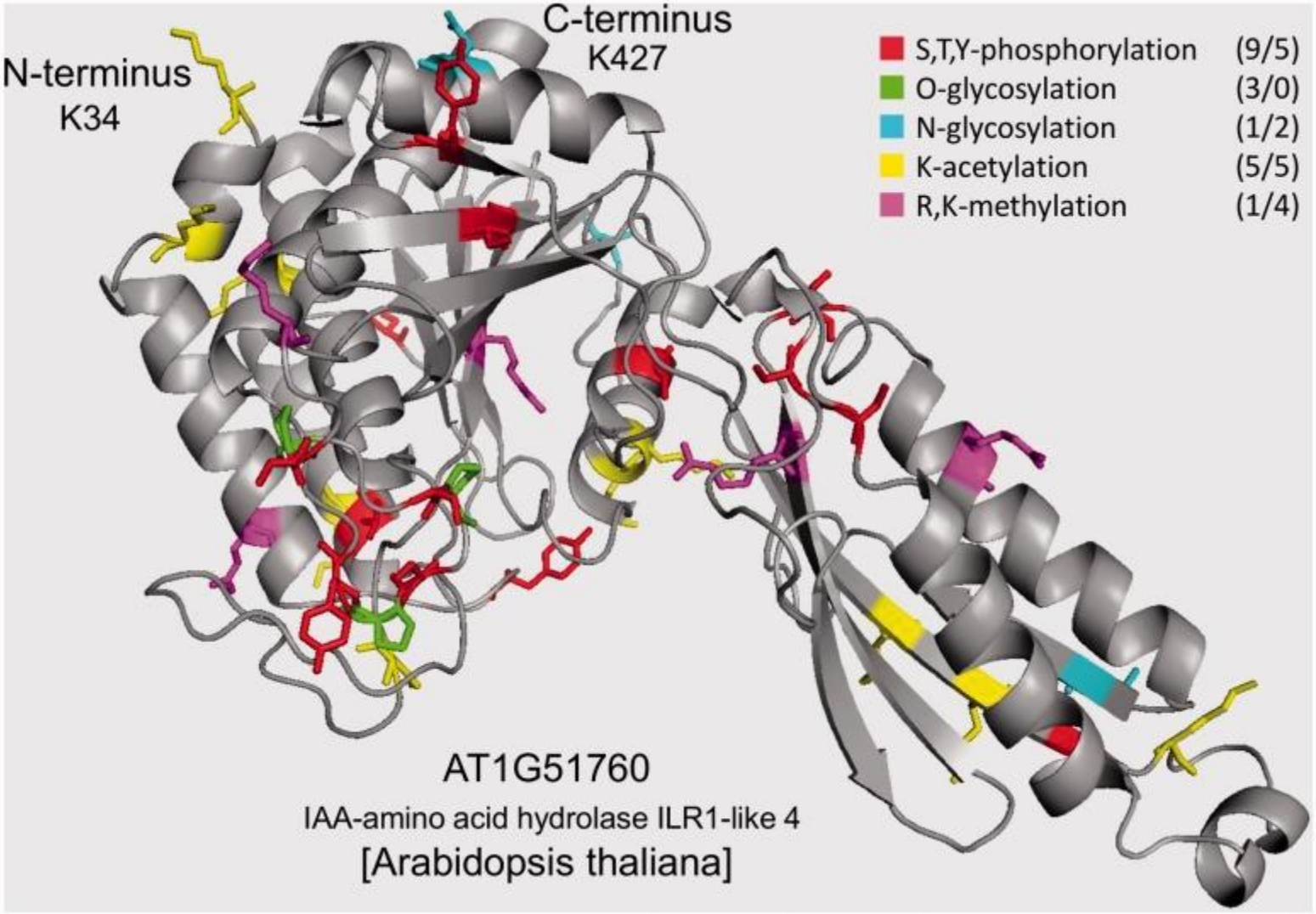




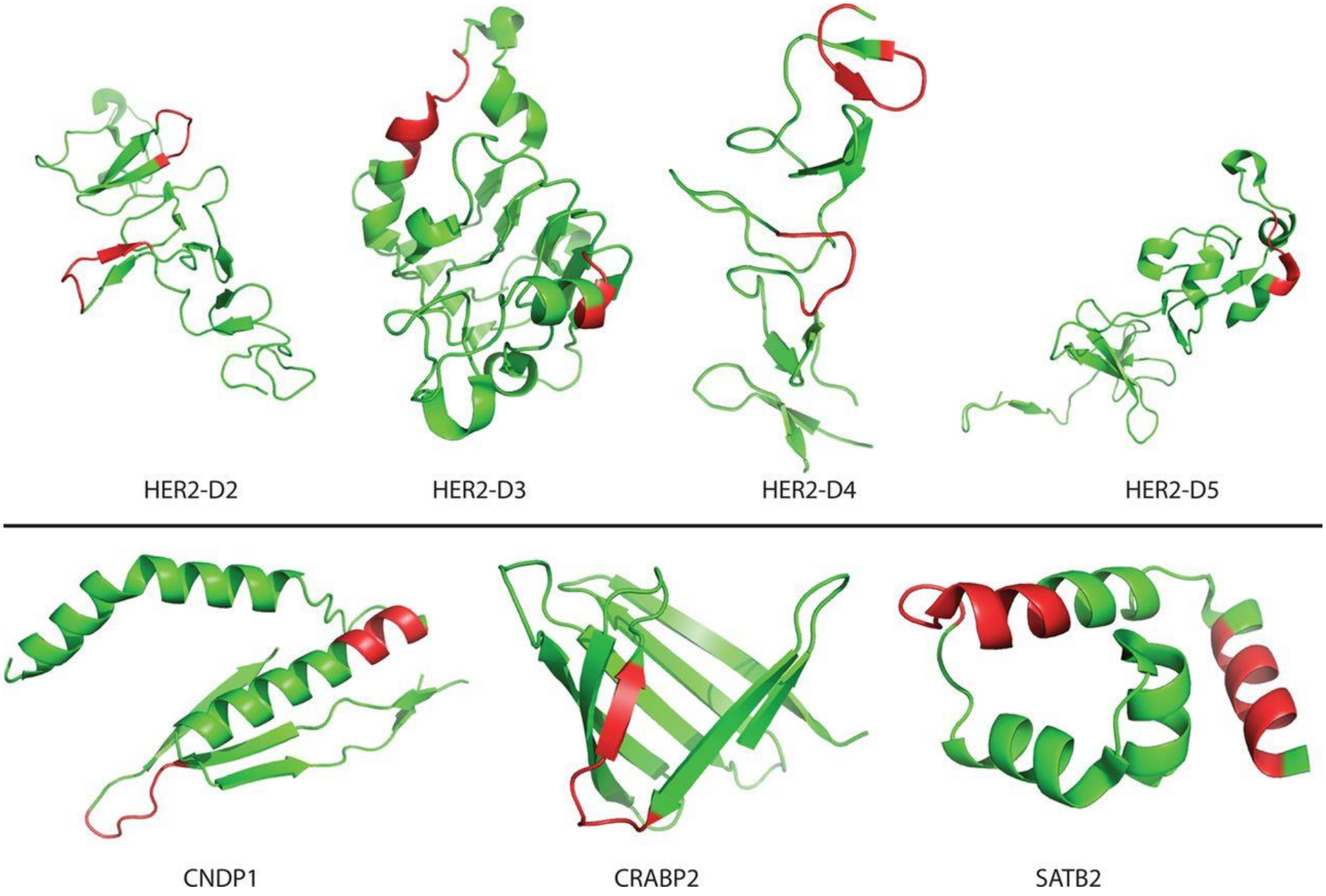
Frustratometer server:

http://frustratometer.qb.fcen.uba.ar/

6. Mappare le modificazioni post-traduzionali (PTM)

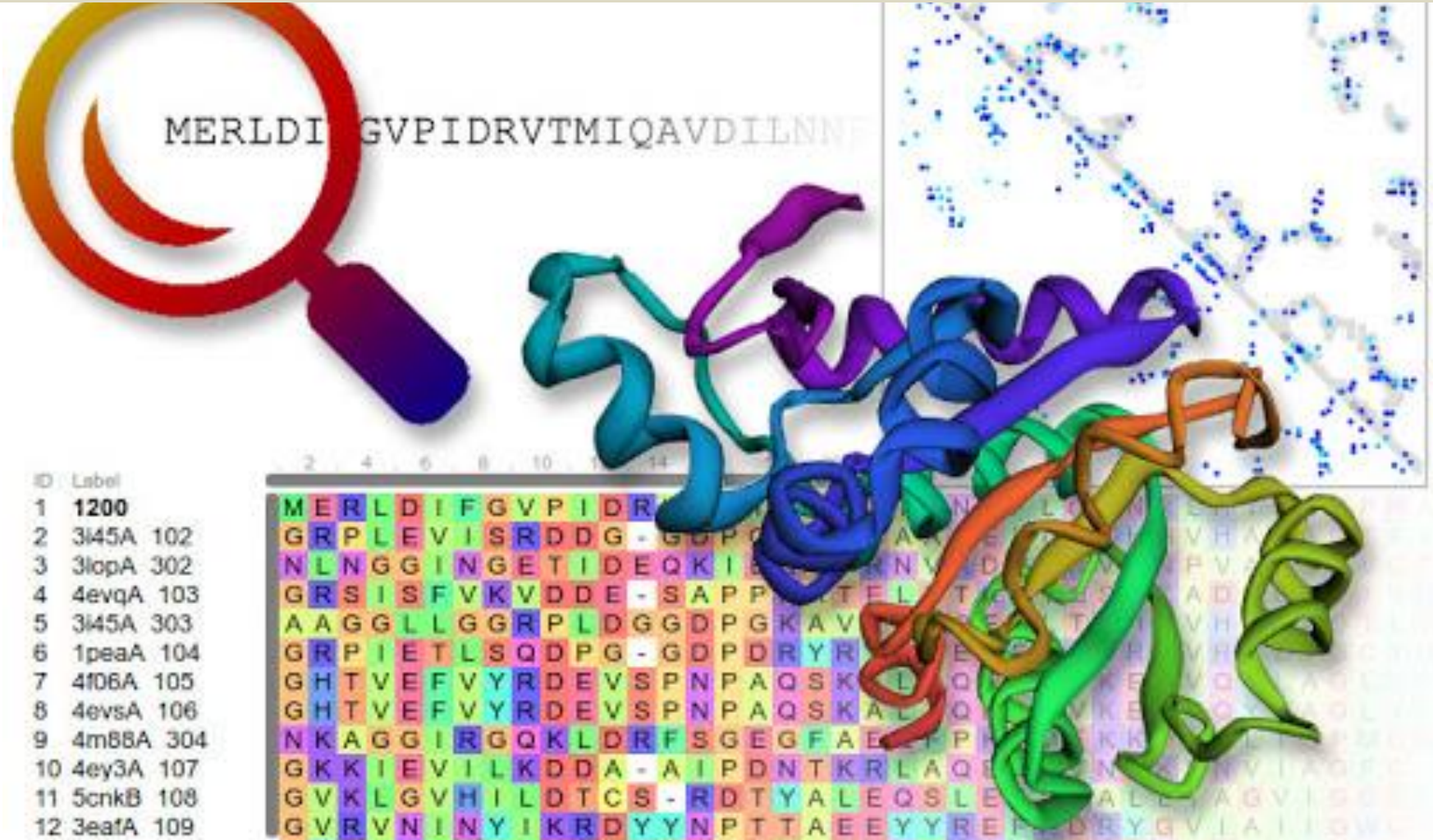


7. Mappare gli epitopi

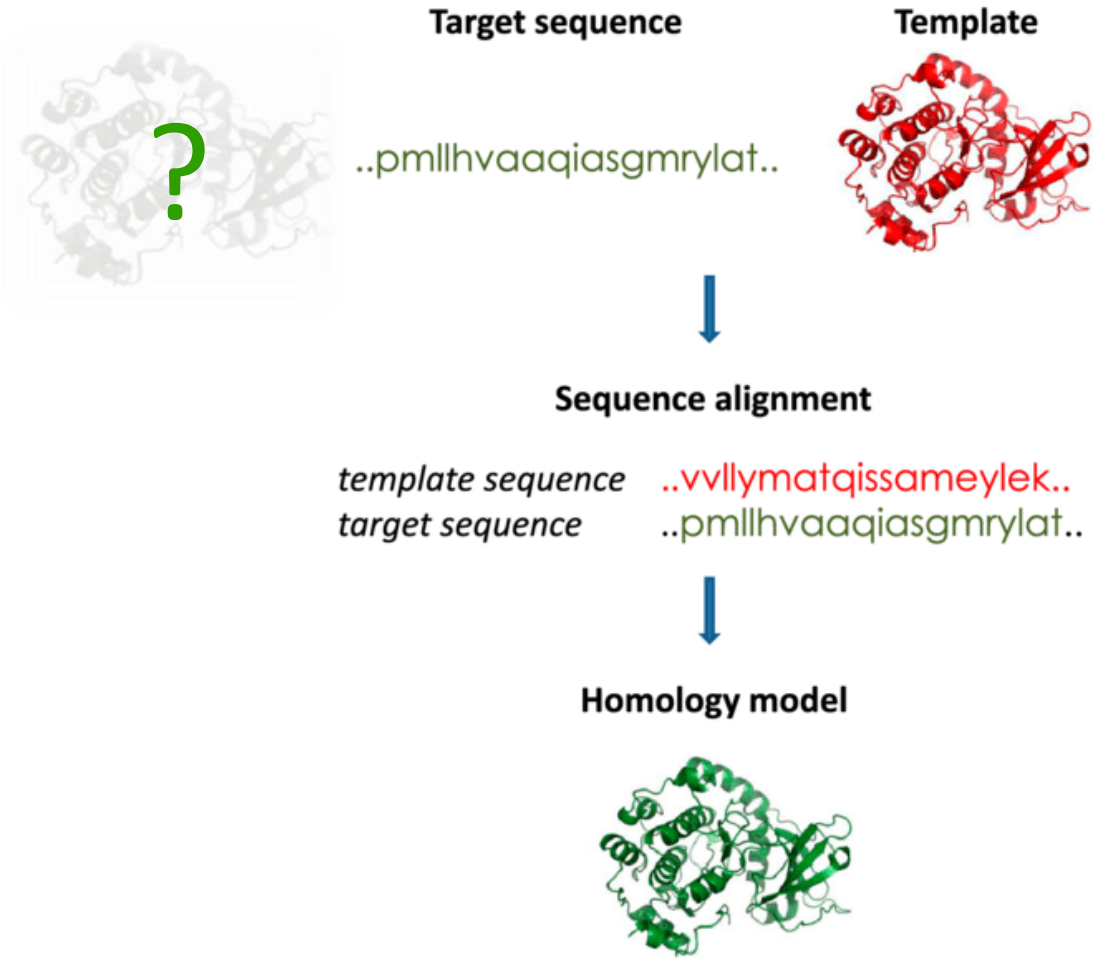


Se NON ho
La struttura?

Homology Modeling

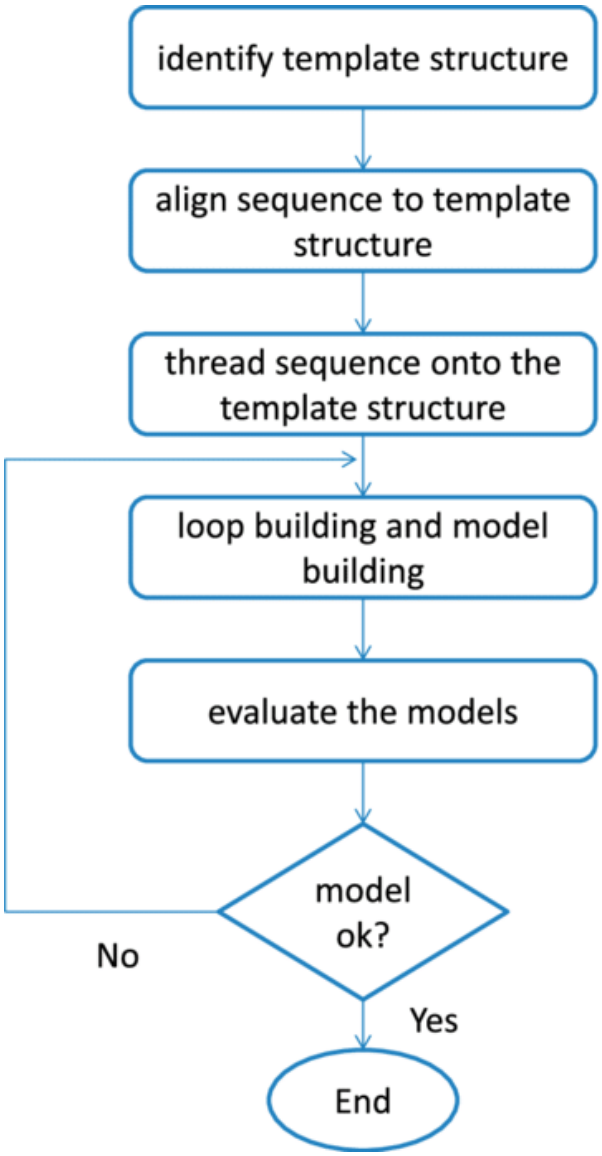


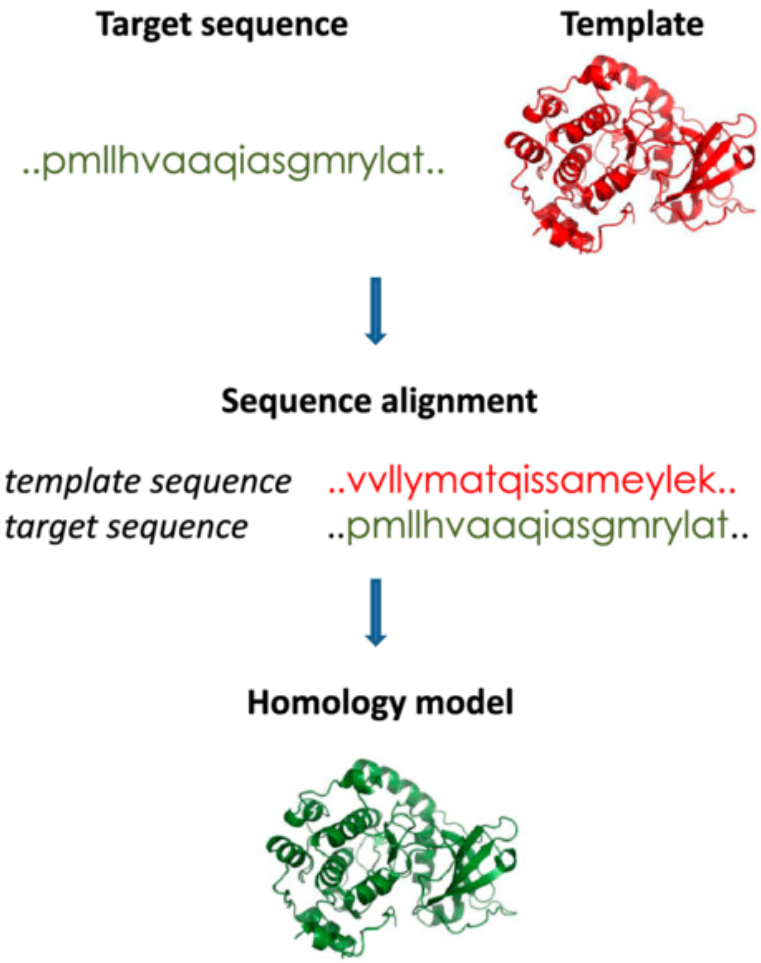
Homology modelling



https://pubmed.ncbi.nlm.nih.gov/24381236/

Homology modelling





https://pubmed.ncbi.nlm.nih.gov/24381236/

Homology Modeling Software

Databases
- ModBase
- Genomic Threading Database

Proprietary
- Orchestrar (*Tripos*)
- Composer (*Tripos*)
- Modeller (*Accelrys*)
- Prime (*Schrödinger*)
- MOE (*Chemical Computing Group*)
- ICM (*MolSoft*)

Free or Open-Source
- MolIDE
- DeepView
- Mint
- Permol
- WhatIf
- Rosetta

Web Servers (sequence-based)
- PredictProtein
- **SWISS-MODEL**
- **(PS)2**
- CPHmodels 2.0
- 3D-JIGSAW
- ESyPred3D
- PROTEUS2
- HOMODELLER

Web Servers (sequence, secondary structure, other criteria)
- **PHYRE**
- **LOOPP**
- **LOMETS**
- MUSTER

Homology modelling

NEWS · 30 NOVEMBER 2020

'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures

Google's deep-learning program for determining the 3D shapes of proteins stands to transform biology, say scientists.

Nature **588**, 203-204 (2020)

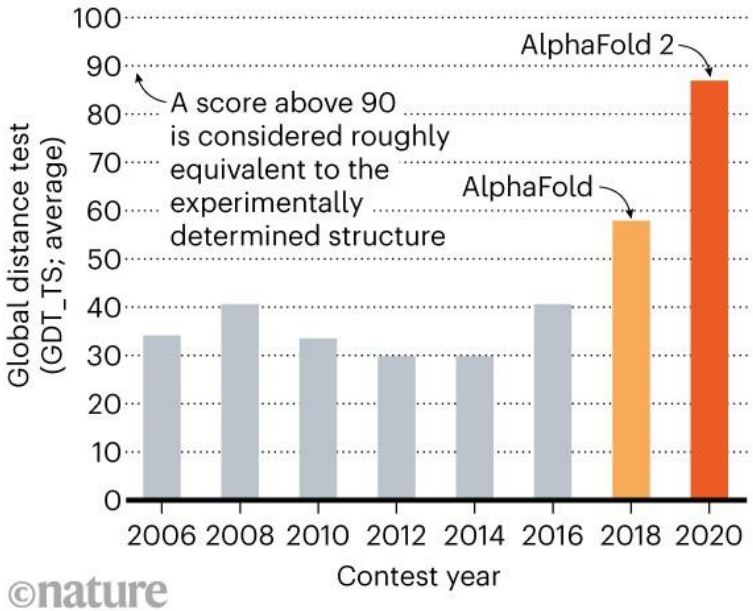
DeepMind's 2018 performance at CASP13 startled many scientists in the field, which has long been the bastion of small academic groups. But its approach was broadly similar to those of other teams that were applying AI, says Jinbo Xu, a computational biologist at the University of Chicago, Illinois.

The first iteration of AlphaFold applied the AI method known as deep learning to structural and genetic data to predict the distance between pairs of amino acids in a protein. In a second step that does not invoke AI, AlphaFold uses this information to come up with a 'consensus' model of what the protein should look like, says John Jumper at DeepMind, who is leading the project.

The team tried to build on that approach but eventually hit the wall. So it changed tack, says Jumper, and developed an AI network that incorporated additional information about the physical and geometric constraints that determine how a protein folds. They also set it a more difficult, task: instead of predicting relationships between amino acids, the network predicts the final structure of a target protein sequence. "It's a more complex system by quite a bit," Jumper says.

STRUCTURE SOLVER

DeepMind's AlphaFold 2 algorithm significantly outperformed other teams at the CASP14 protein-folding contest — and its previous version's performance at the last CASP.



Banche dati homology models

ModBase - database of comparative protein structure models

https://modbase.compbio.ucsf.edu

Uses ModPipe, automated modeling pipeline relying on the programs PSI-BLAST and MODELLER

>30% sequence ID, >4 million models, >1 million sequences

Genomic Threading Database - for detecting remote homology between protein sequences and known folds

http://bioinf.cs.ucl.ac.uk/GTD

seq ID 10-30%, > 1 million sequences

Factors determining model quality

- % sequence identity to templates

% ID	Confidence?
> 30	good to great
25 - 30	low to maybe?
< 25	low

- coverage
- steric or electrostatic clashes
- agreement with bench data
- agreement with general protein structure knowledge
- scoring (RMSD, C-score, Tm-score, others….)

root-mean-square deviation (RMSD)

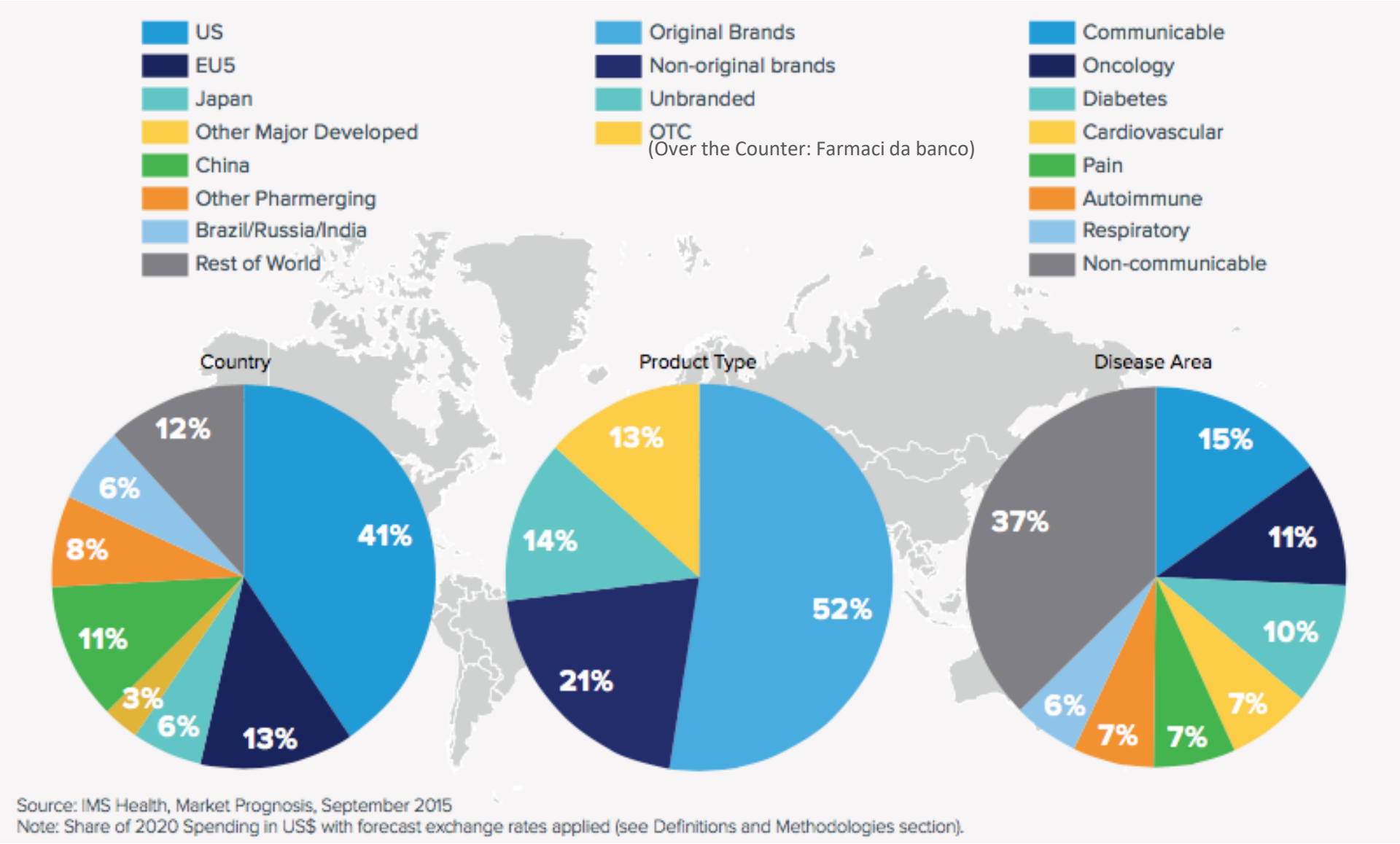
the root-mean-square deviation of atomic positions is the measure of the average distance between the atoms of superimposed proteins

$$\mathrm{RMSD}(\mathbf{v}, \mathbf{w}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \|v_i - w_i\|^2}$$

$$= \sqrt{\frac{1}{n}\sum_{i=1}^{n} ((v_{ix} - w_{ix})^2 + (v_{iy} - w_{iy})^2 + (v_{iz} - w_{iz})^2)}$$

Drug Development

Il Mercato



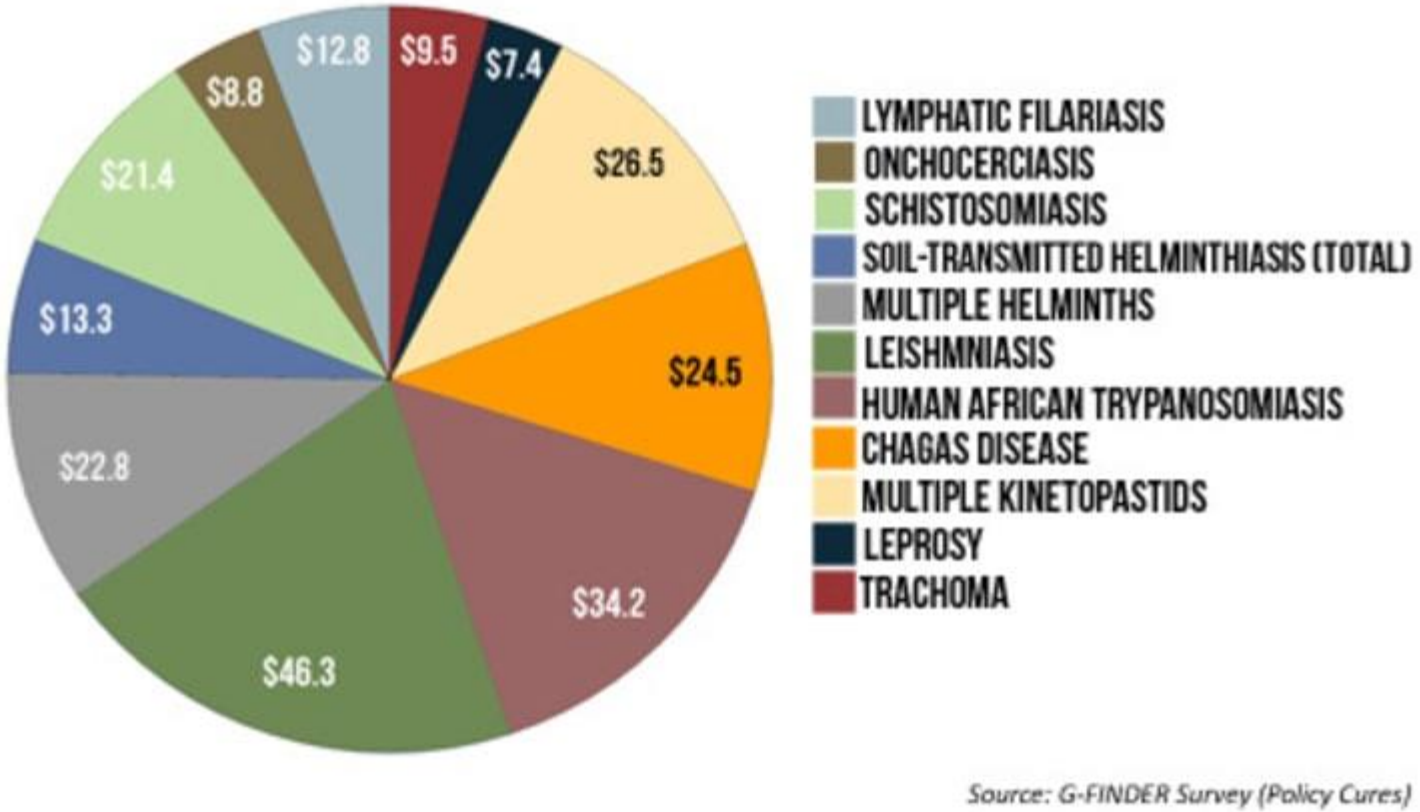
Source: IMS Health, Market Prognosis, September 2015
Note: Share of 2020 Spending in US$ with forecast exchange rates applied (see Definitions and Methodologies section).

Il Mercato

Sr. No.	Company	R & D spend($bn),2010
1	Novartis	7.9
2	Merck & Co	8.1
3	Roche	7.8
4	GlaxoSmithKline	5.7
5	Sanofi	5.8
6	Pfizer	9.1
7	Johnson & Johnson	4.5
8	Eli Lilly	4.7
9	AstraZeneca	4.2
10	Takeda	3.4
11	Bayer	2.3
12	Bristol-Myers Squibb	3.3
13	Boehringer Ingelheim	3.1
14	Amgen	2.8
15	Novo Nordisk	1.7

NEGLECTED TROPICAL DISEASES 2011

RESEARCH AND DEVELOPMENT FUNDING BY DISEASE IN U.S. MILLION DOLLARS

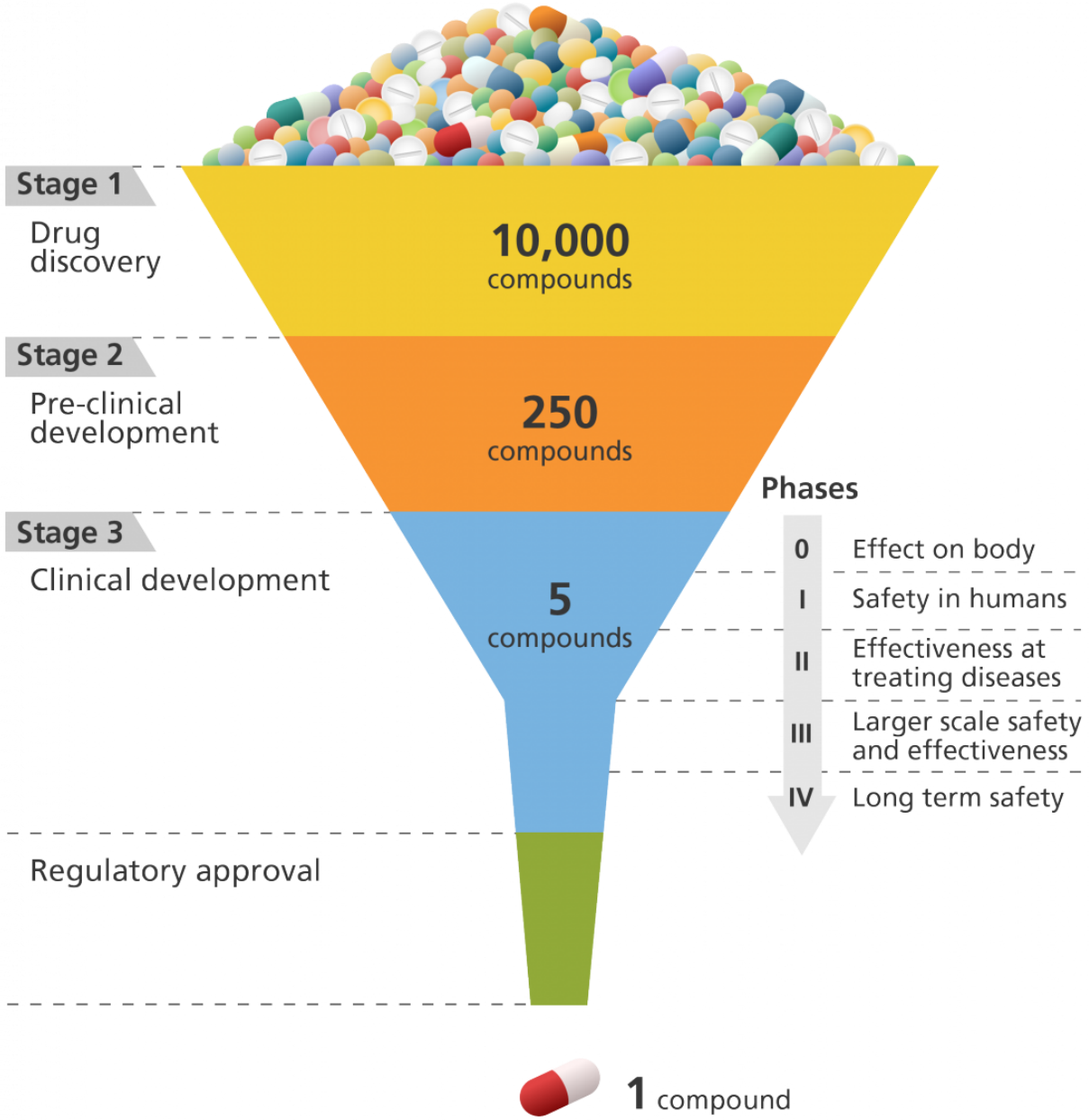


Source: G-FINDER Survey (Policy Cures)

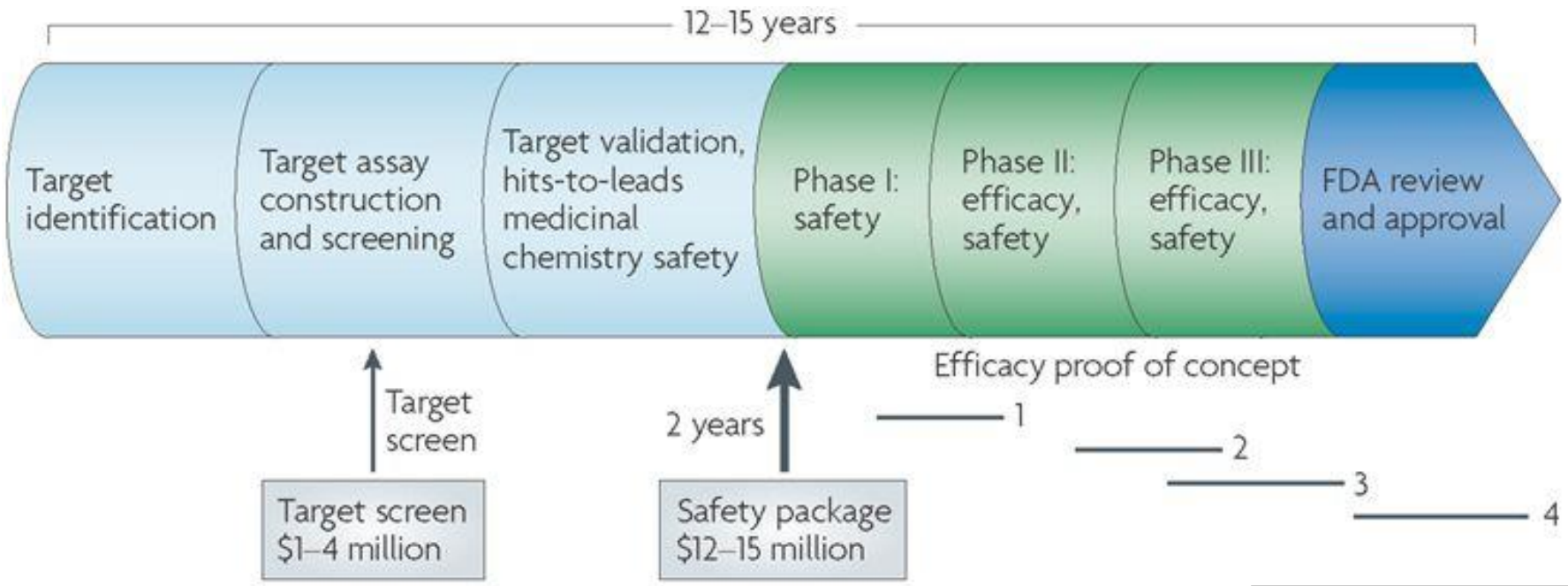
Il Mercato

R&D Function	%
Discovery/Basic Research	
Synthesis & Extraction	10.0
Biological Screening & testing	14.2
Preclinical Testing	
Toxicology & Safety testing	4.5
Pharmaceutical Dosage Formulation	7.3
Clinical Trials	
Phase I, II, III	29.1
Phase IV	11.7
Manufacturing & QC	8.3
IND & NDA	4.1
Bioavailability	1.8
Others	9.0
Total	100.0

The development funnel



COSTI e TEMPI nello sviluppo di farmaci



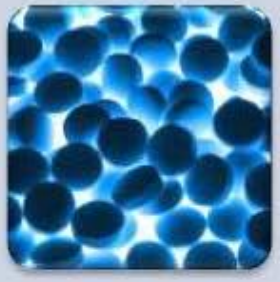
$500 – 800 milioni

https://www.nature.com/articles/nrd2593

LE FASI nello sviluppo di farmaci





Target Selection

Structural biology

- Cellular and Genetic Targets
- Genomics
- Proteomics
- Bioinformatics

Lead Discovery

Structure based drug design

- Synthesis and Isolation
- Combinatorial Chemistry
- Assay development
- High-Throughput Screening

Medicinal Chemistry

- Library Development
- SAR Studies
- In Silico Screening
- Chemical Synthesis

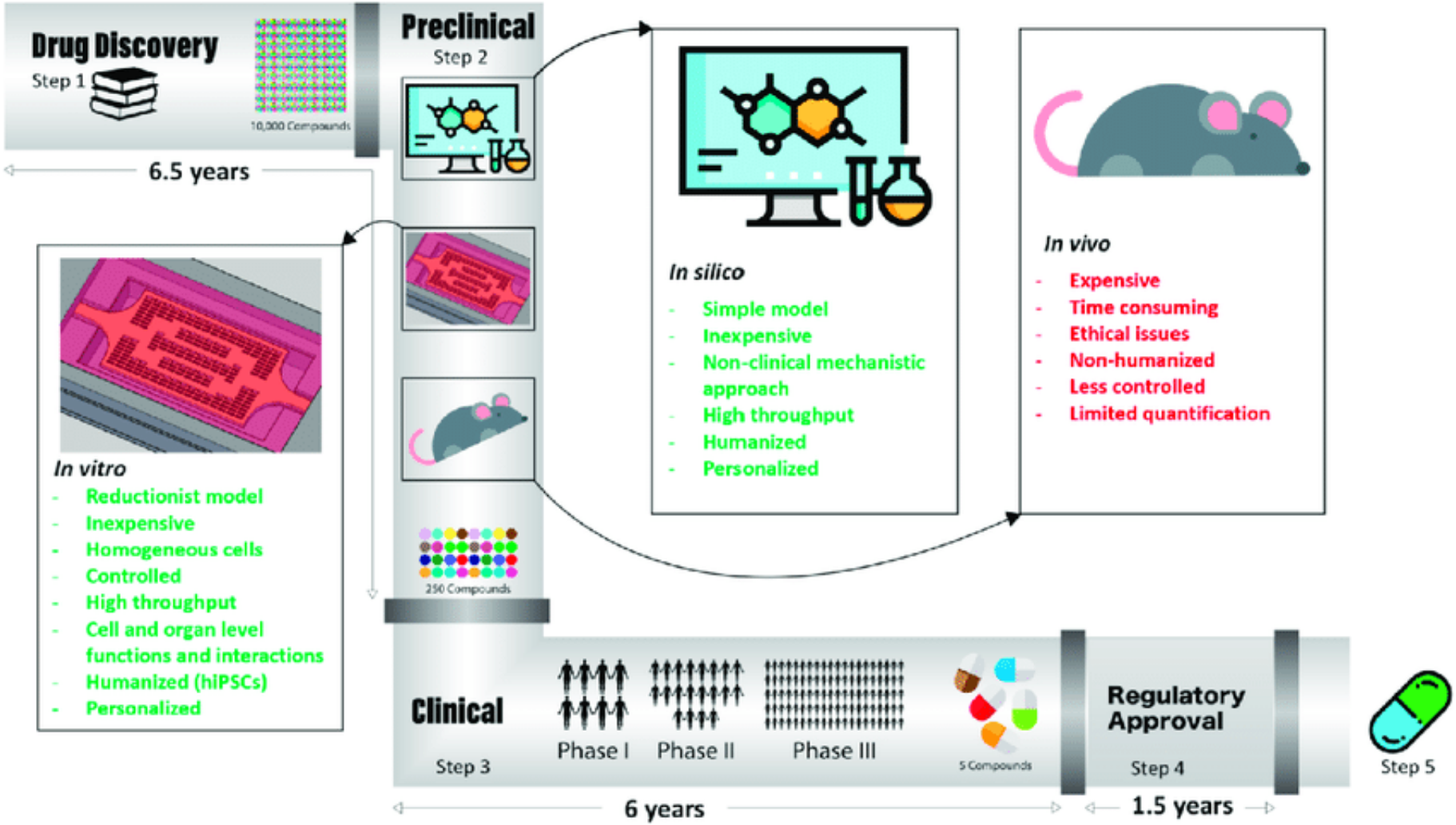
In Vitro Studies

- Drug Affinity and Selectivity
- Cell Disease Models
- MOA
- Lead Candidate Refinement

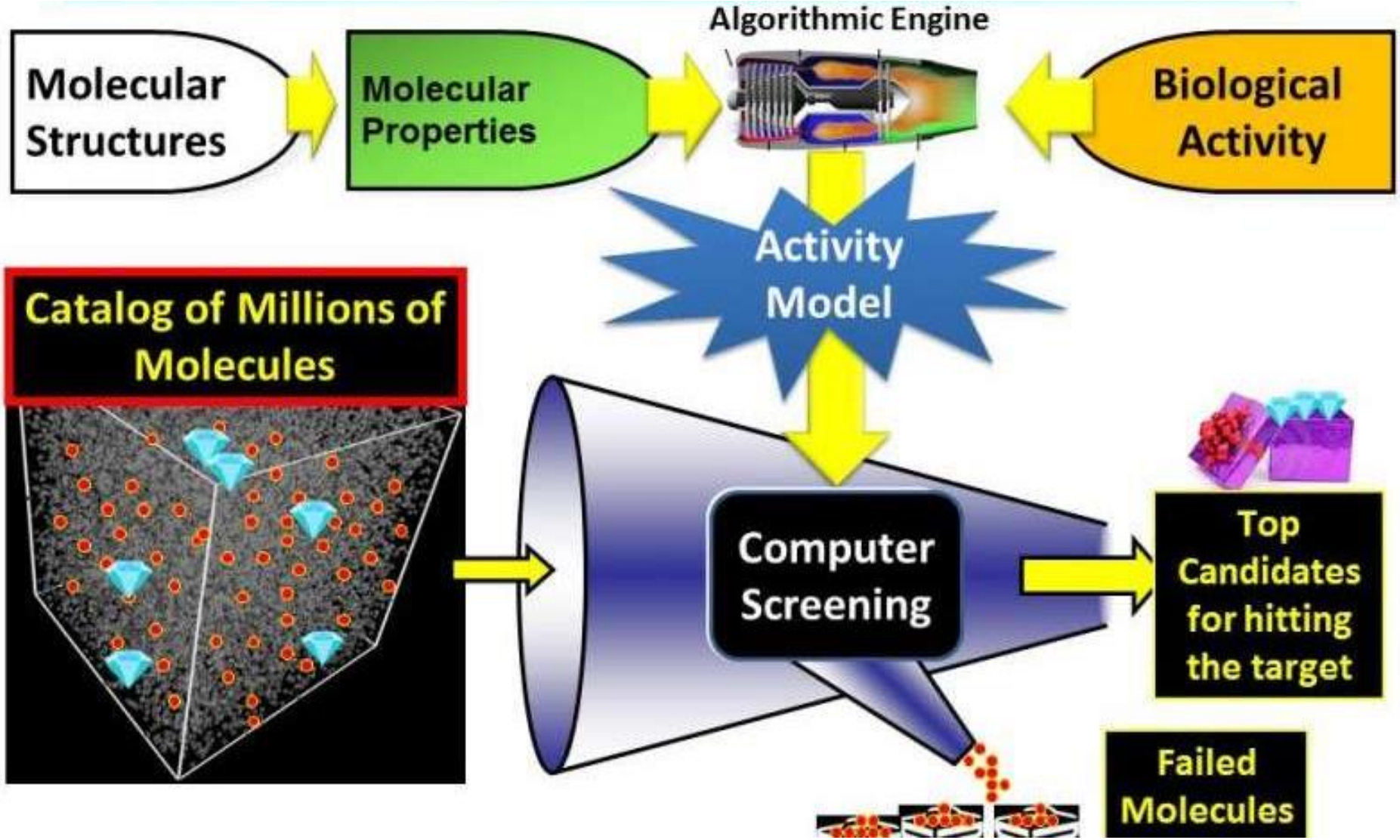
In Vivo Studies

- Animal models of Disease States
- Behavioural Studies
- Functional Imaging
- Ex-Vivo Studies

Clinical Trials and Therapeutics



https://www.sciencedirect.com/science/article/pii/S0142961218306811?via%3Dihub



https://phys.org/news/2017-06-algorithm-drug-discovery-methods.html