DL for Protein Structure Prediction

Iaroslav Geraskin

Who I am

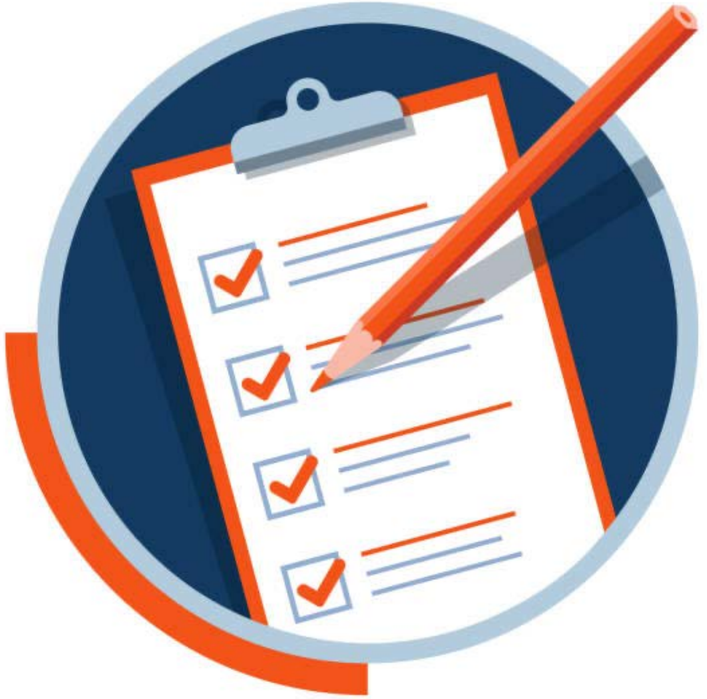
Iaroslav Geraskin

BIOCAD
Biotechnology Company

- 4 years in ML research and infrastructure
- MS in Computational Biology
- Worked on antibody structure prediction ML for drug discovery

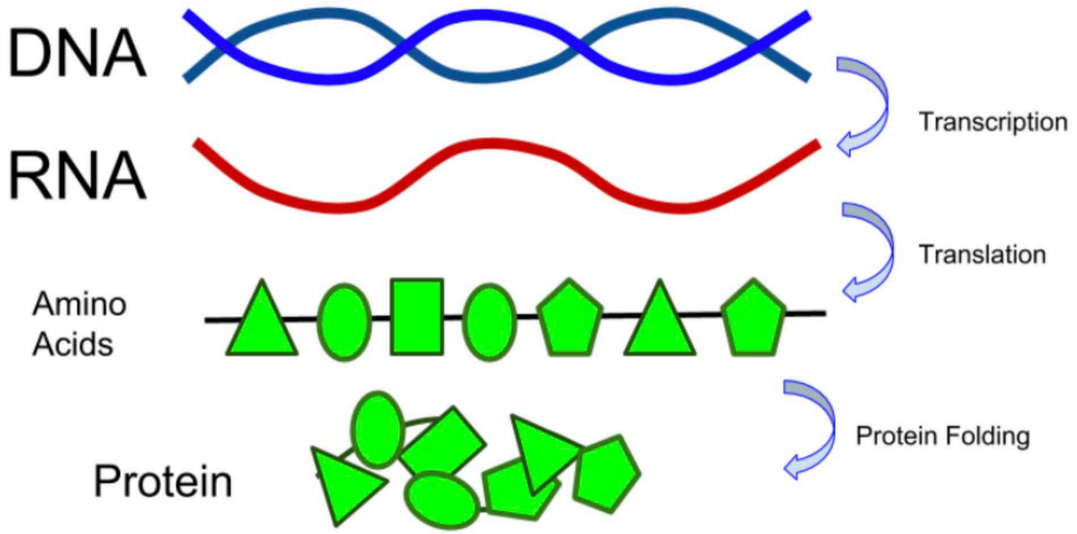
Agenda

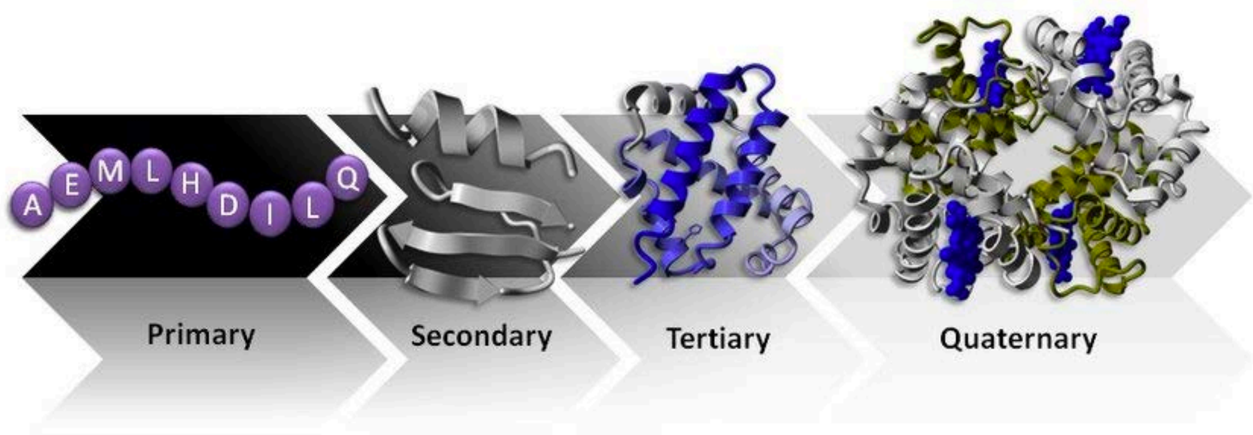
- Proteins and why we need them
- Evolutionary information and MSA
- Structure prediction methods
- Physics
- Statistics
- Deep Learning



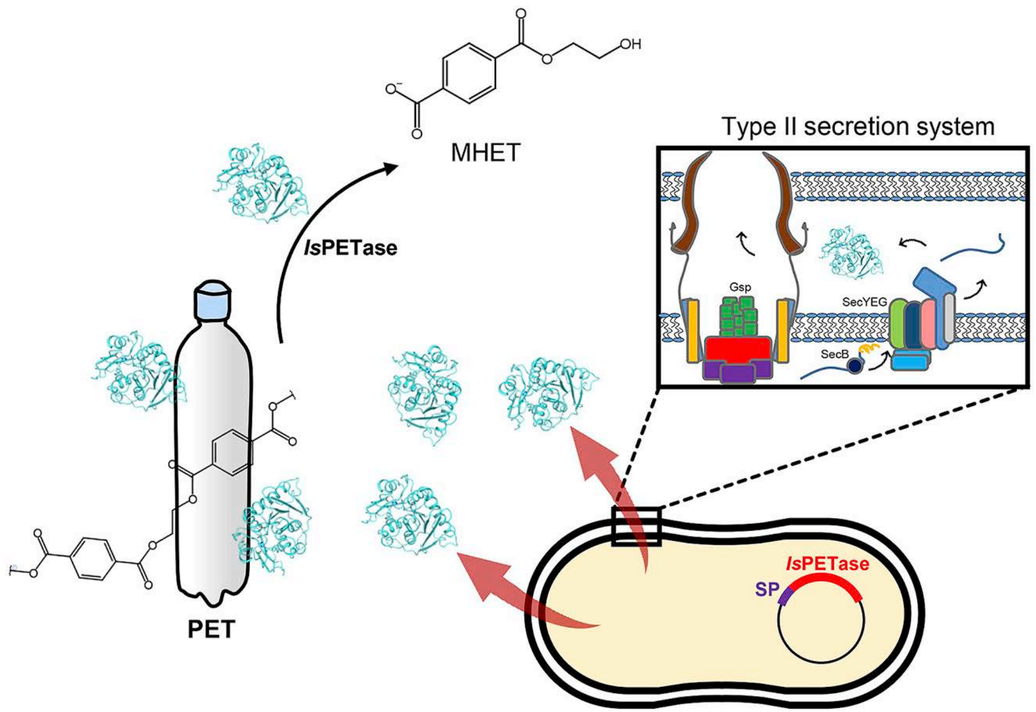
Proteins

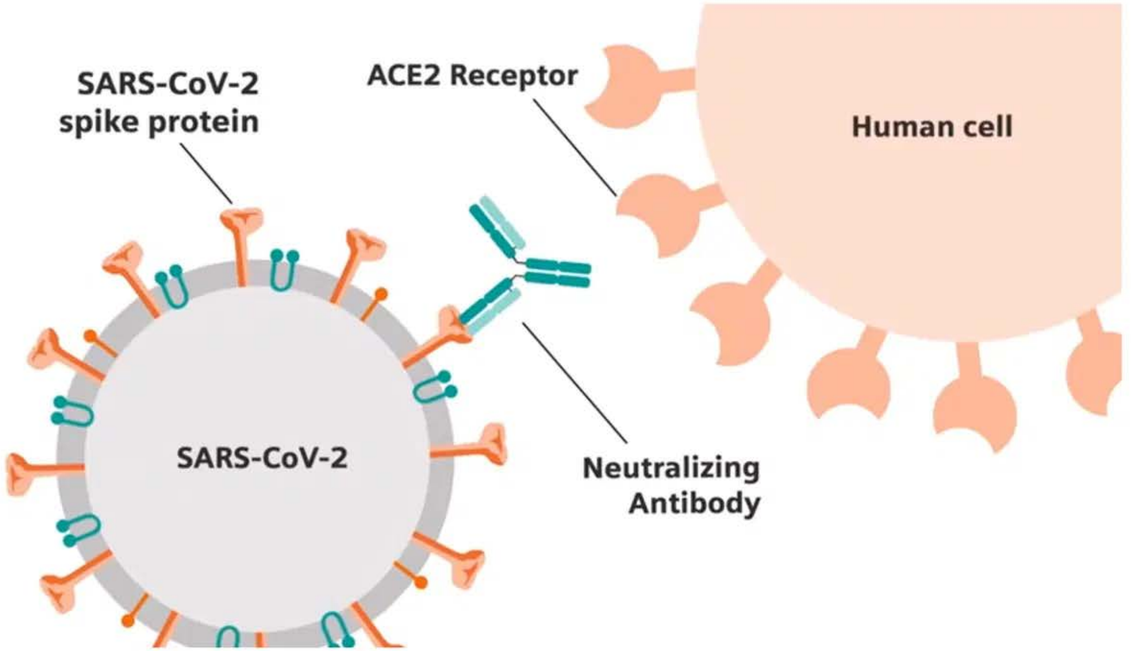
- The task of structure prediction is one of the most important tasks in bioinformatics
- Data obtained through prediction are used in medicine and biotechnology (e.g. creating new enzymes and drugs)





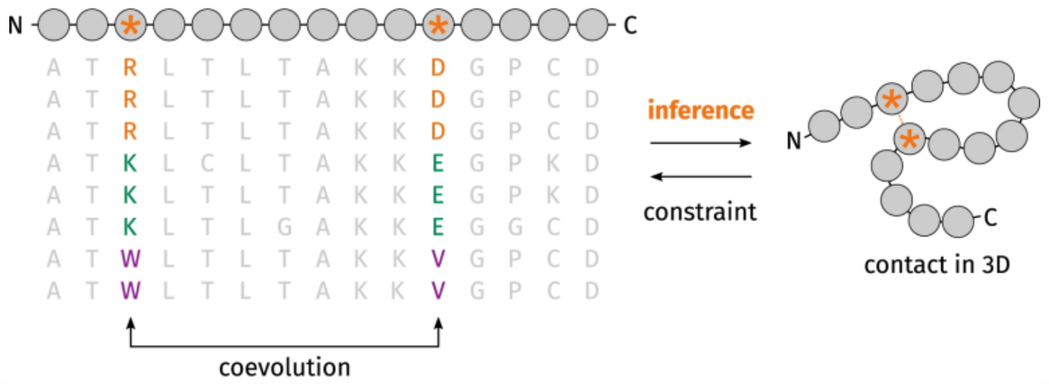
Why do we need proteins?

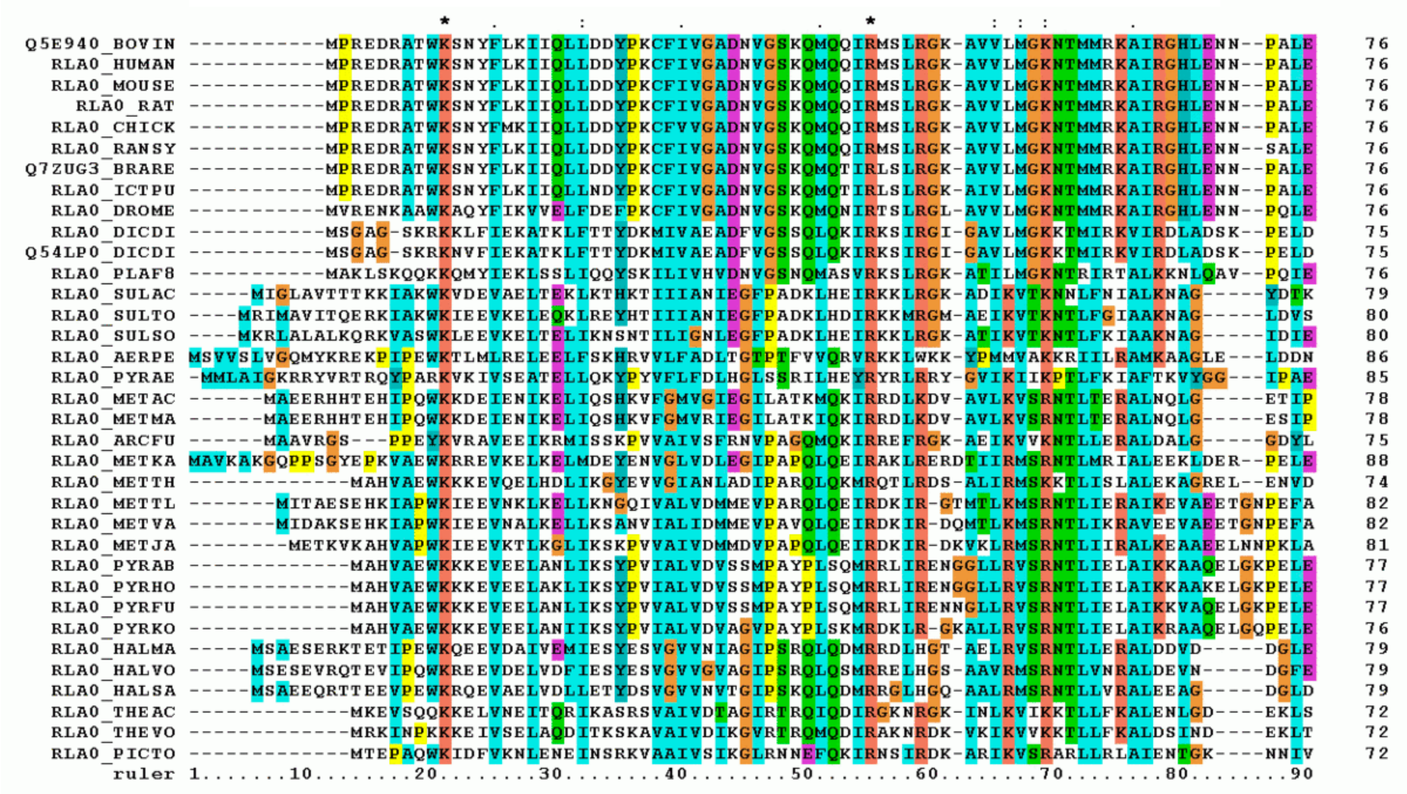




MSA

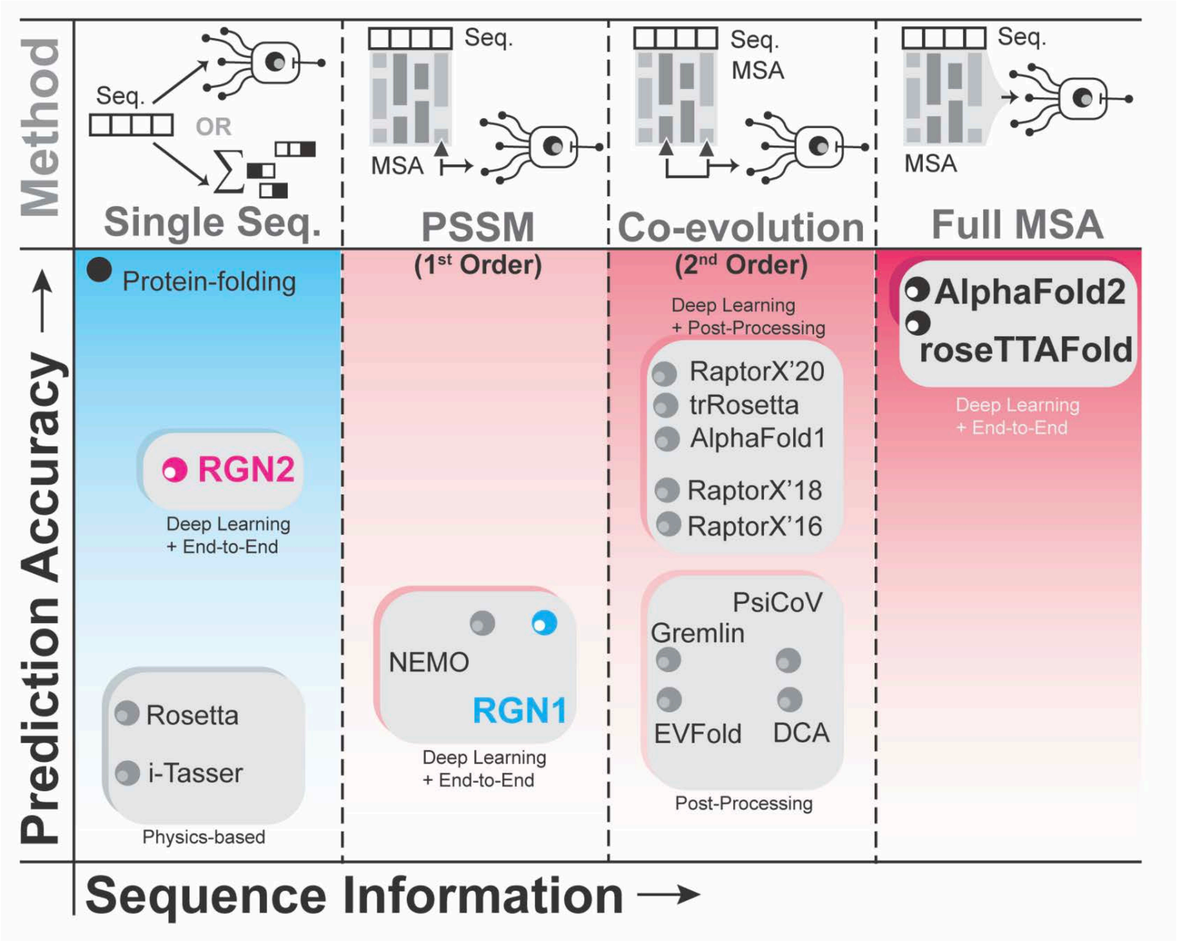
- Multiple alignment allows you to extract evolutionary information from multiple sequences, determine which positions depend on which positions, which positions are conserved and which are not





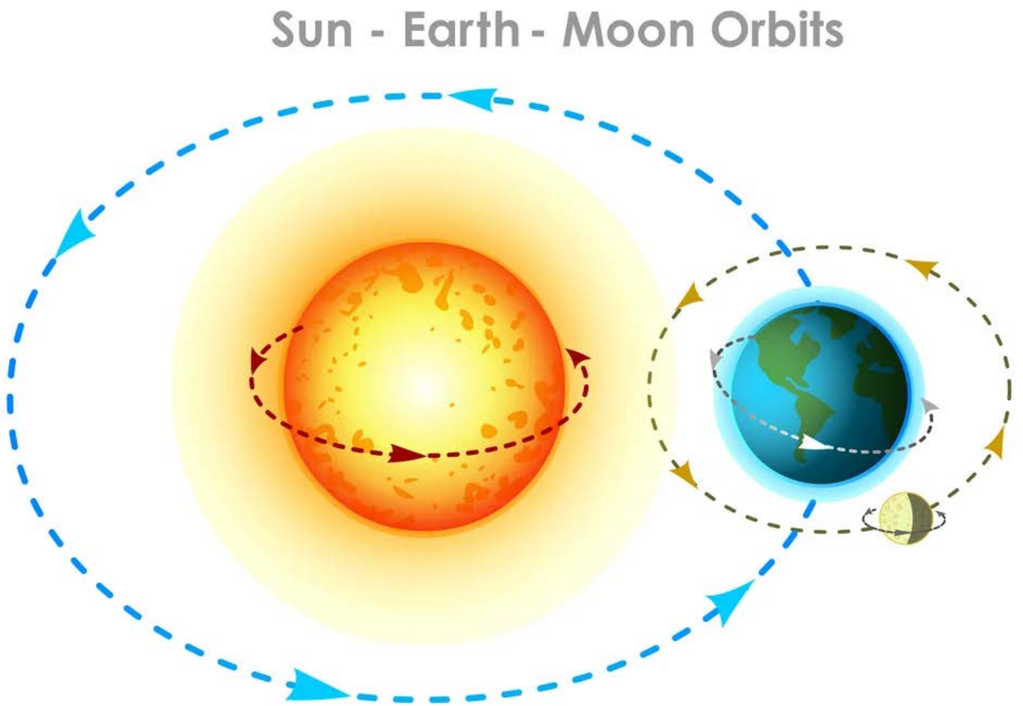
Structure prediction methods

- Physical methods such as Rosetta are computationally complex
- Methods that use machine learning and more evolutionary data perform better on average
- End-to-End methods are faster because do not use long iterative simulations (unlike physical methods)
- For some classes of proteins for which evolutionary information is not available, methods that rely on it perform less well.



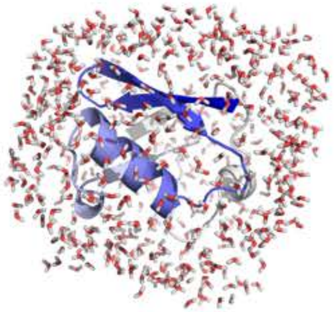
Three-body problem

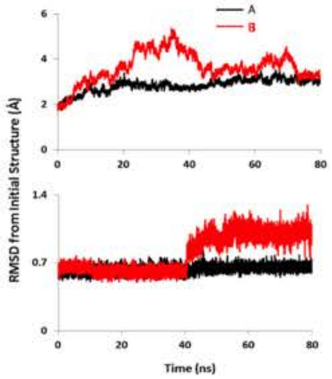
- Three-body problem does not have a general solution that can be expressed in terms of a finite number of standard mathematical operations. Moreover, the motion of three bodies is generally non-repeating, except in special cases.



Molecular dynamics

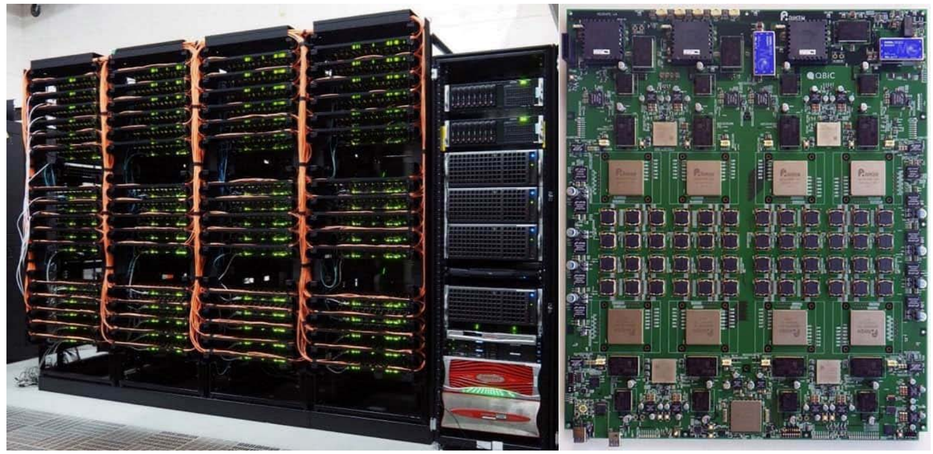
Simulation, HPC, trajectories

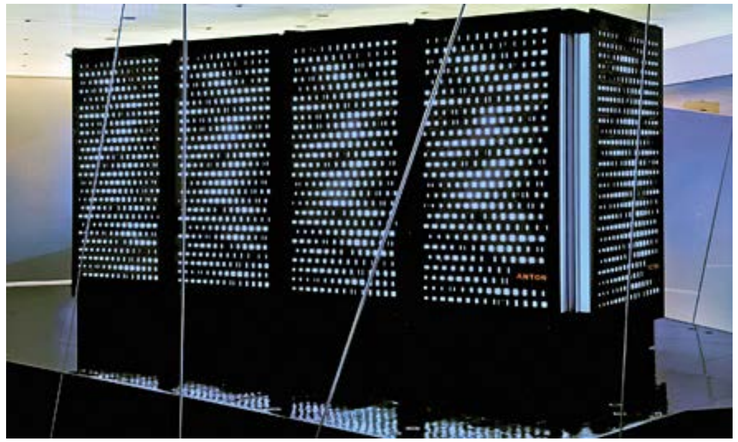




Structure Check
Protonation State
Solvation
Adding Salts

1. Starting Structure → 2. Preparation of the Simulation System → Select Simulation Protocol → 3. Simulation Run → Post-processing → 4. Trajectory Analysis

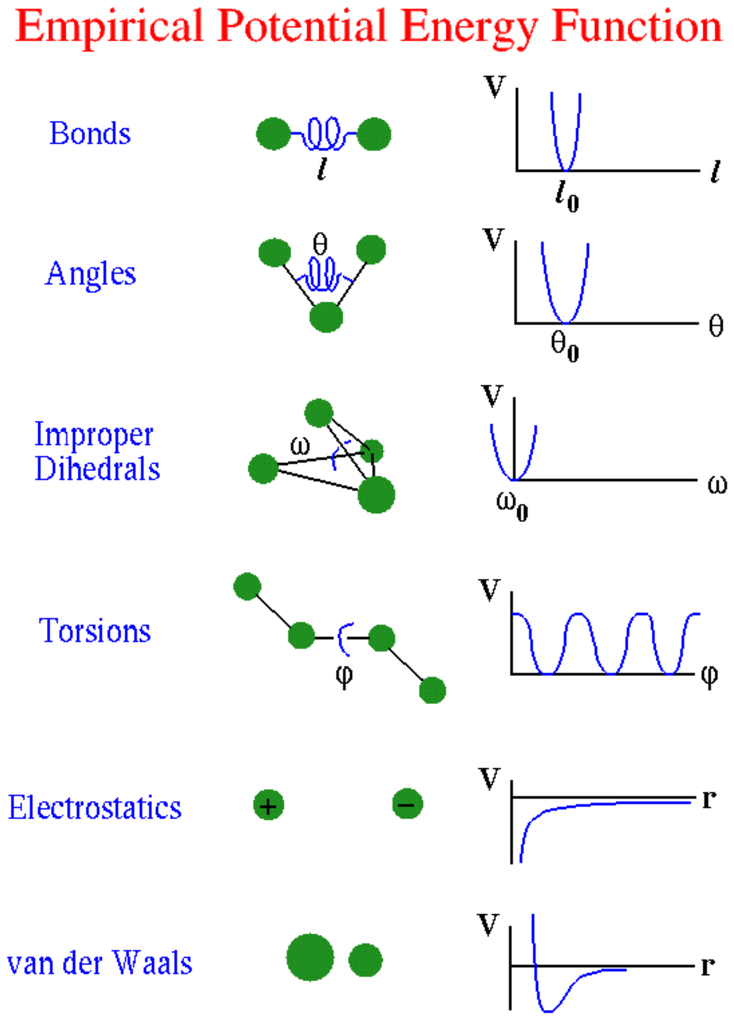




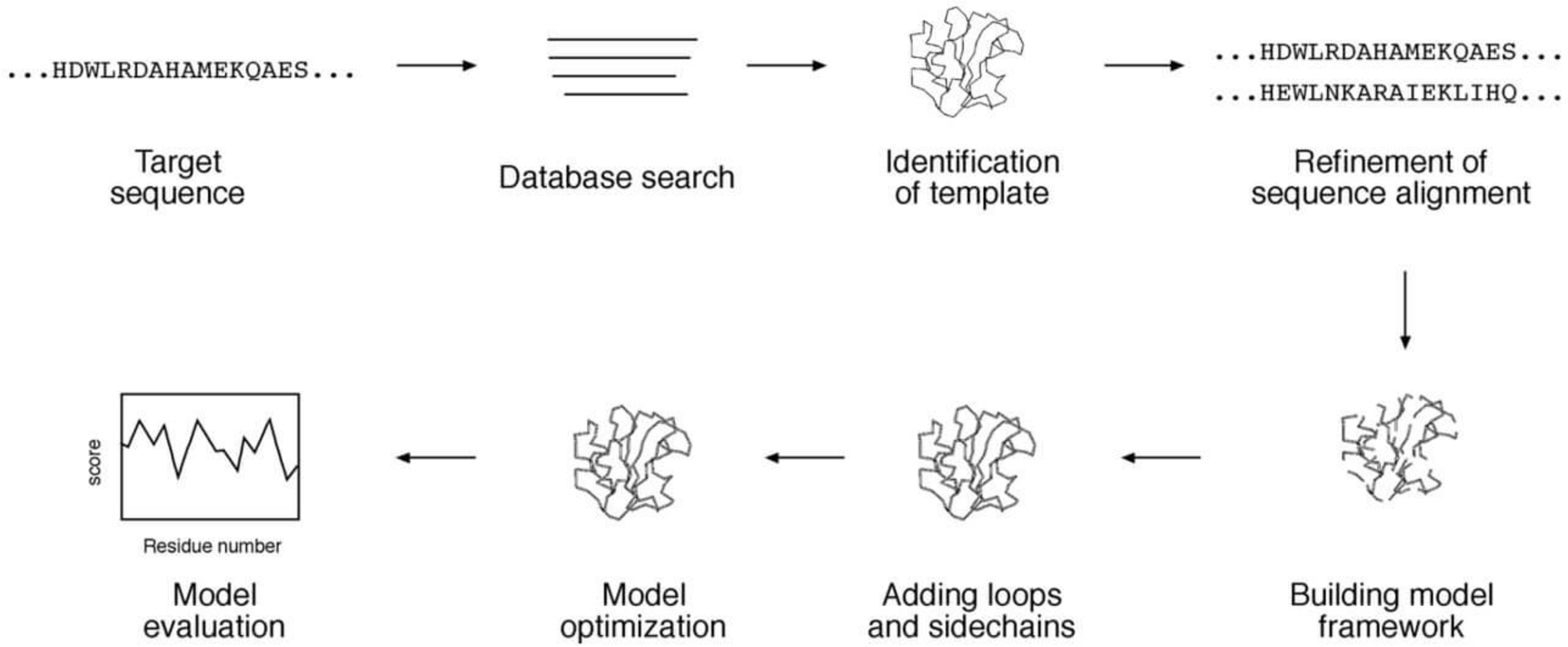
Molecular dynamics

Physical potential

- Compute new positions for particles based on forces acting on them
- Take a lot of small steps
- Can we get a good approximation first?
- Why follow physics when we have statistics?



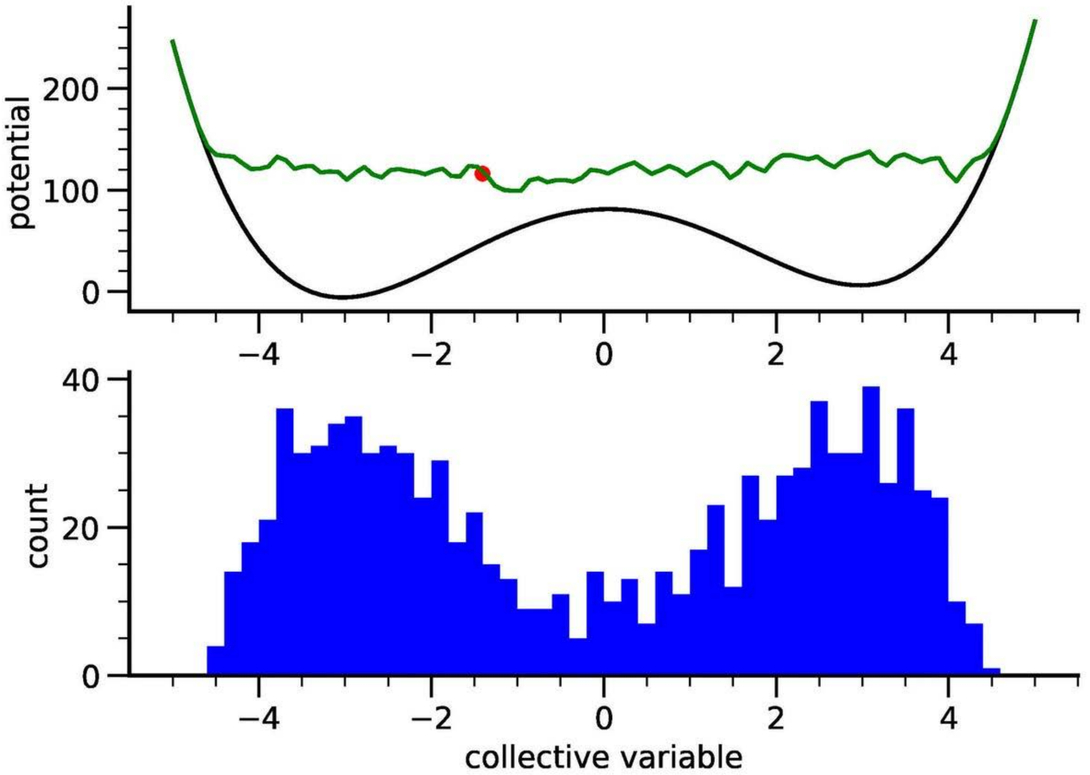
Homology modeling



Molecular dynamics

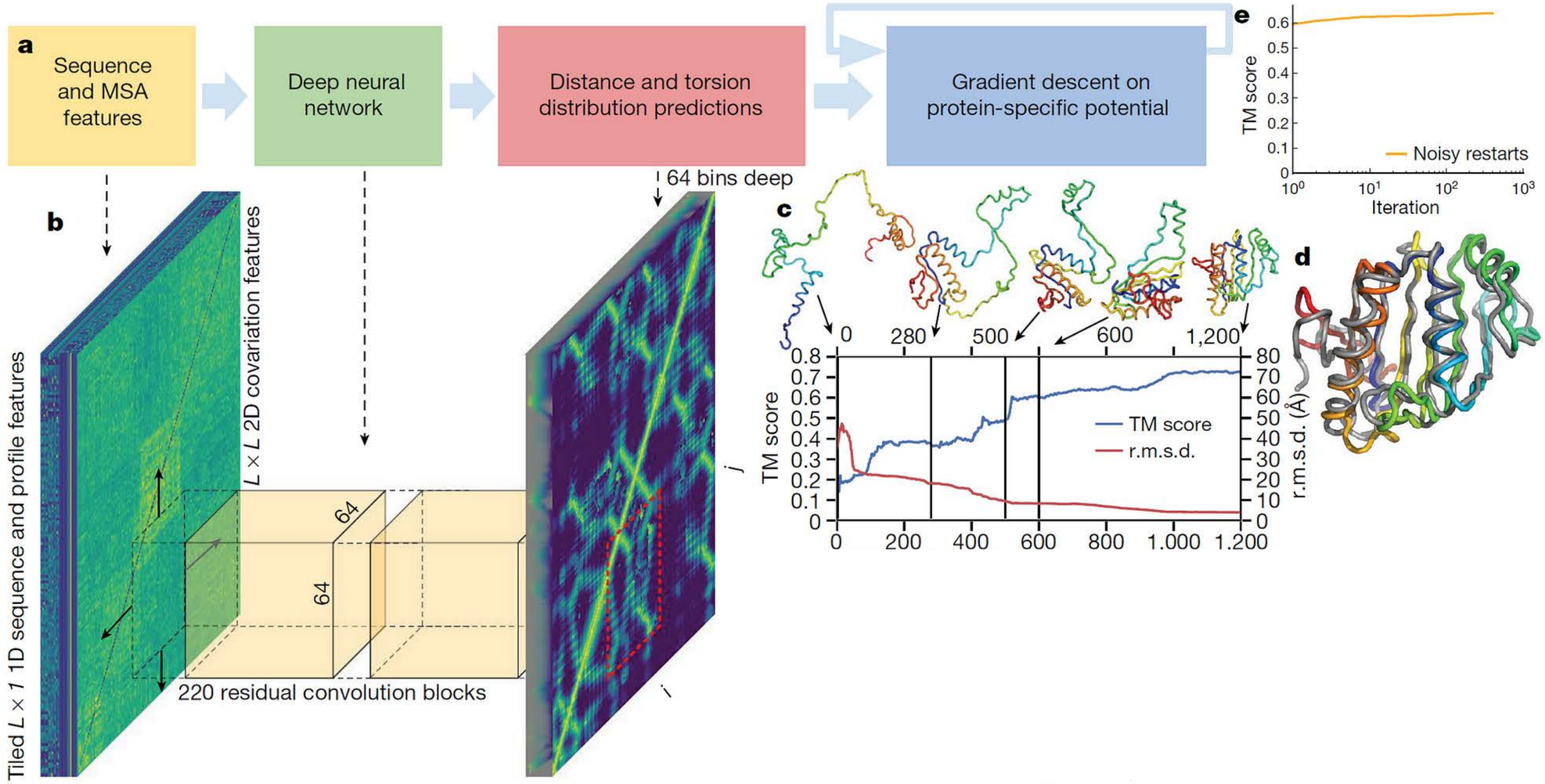
Statistical potential

- If we are moving into an unlikely position apply opposite forces
- Estimate likeliness of different positions based on known structures
- Lower potential of likely positions, make unlikely positions have higher potential
- Need to know likelihood of positions



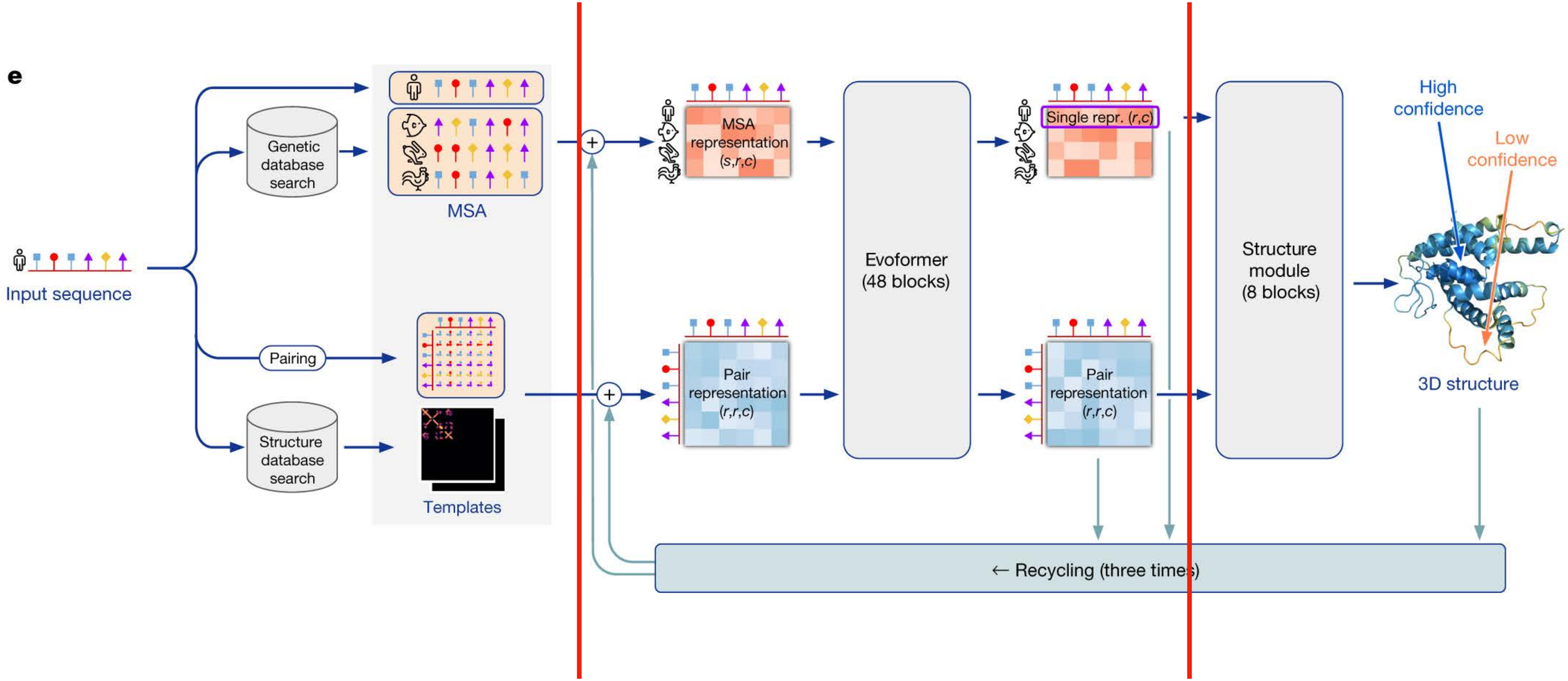
Alphafold 1

Predict the potential, convolution, MSA features



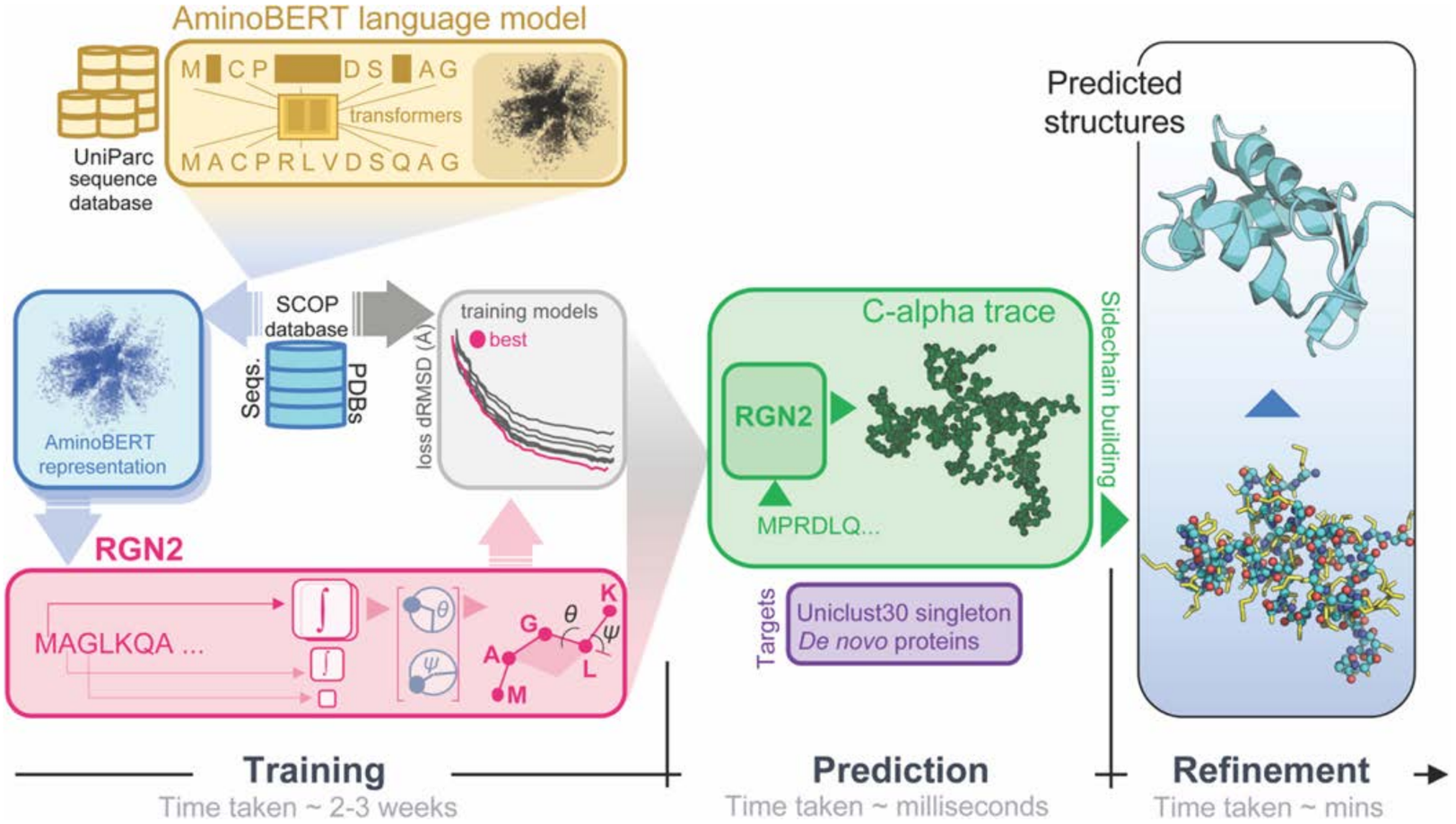
Alphafold 2

Use structure templates + transformers + end-to-end structure



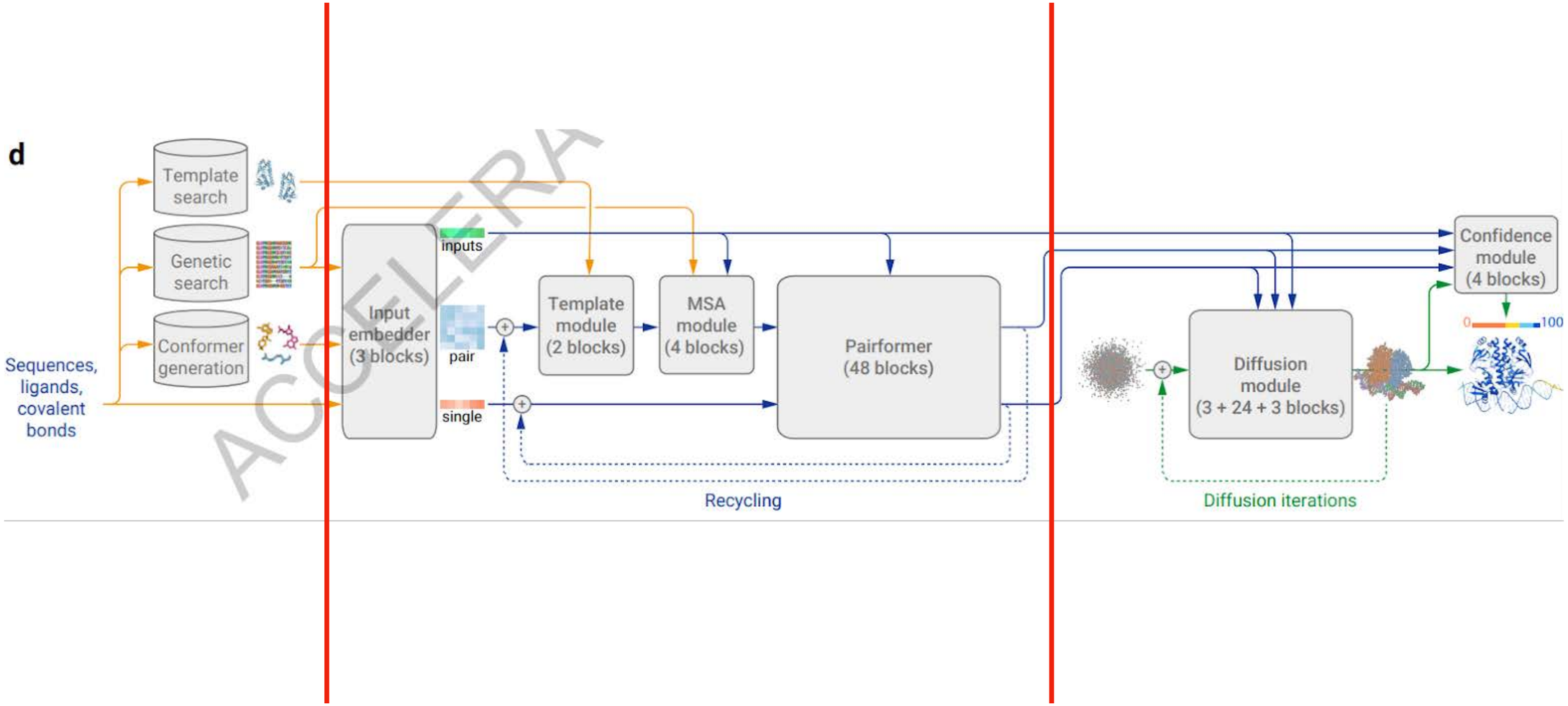
Single Seq

- Compress evolutionary information from a large database into a LLM
- Train a small network to predict geometric features
- Use physics based methods to refine



Alphafold 3

Diffusion, ligands and more



Conclusion

- Physics based methods require a lot of compute
- Statistics can help find good approximation
- DL can help find good approximation faster/better
- Getting more data sources into a model is challenging
- Pre-trained language models can help encode protein knowledge
- End-to-end methods allow to use DL for every step of structure prediction
- Physics based refiner can be used to obtain more "natural" result
- New DL methods such as transformers and diffusion models trickle down into biology

Questions? Reach out!

linkedin