

DL for Protein Structure Prediction

Iaroslav Geraskin

Who I am

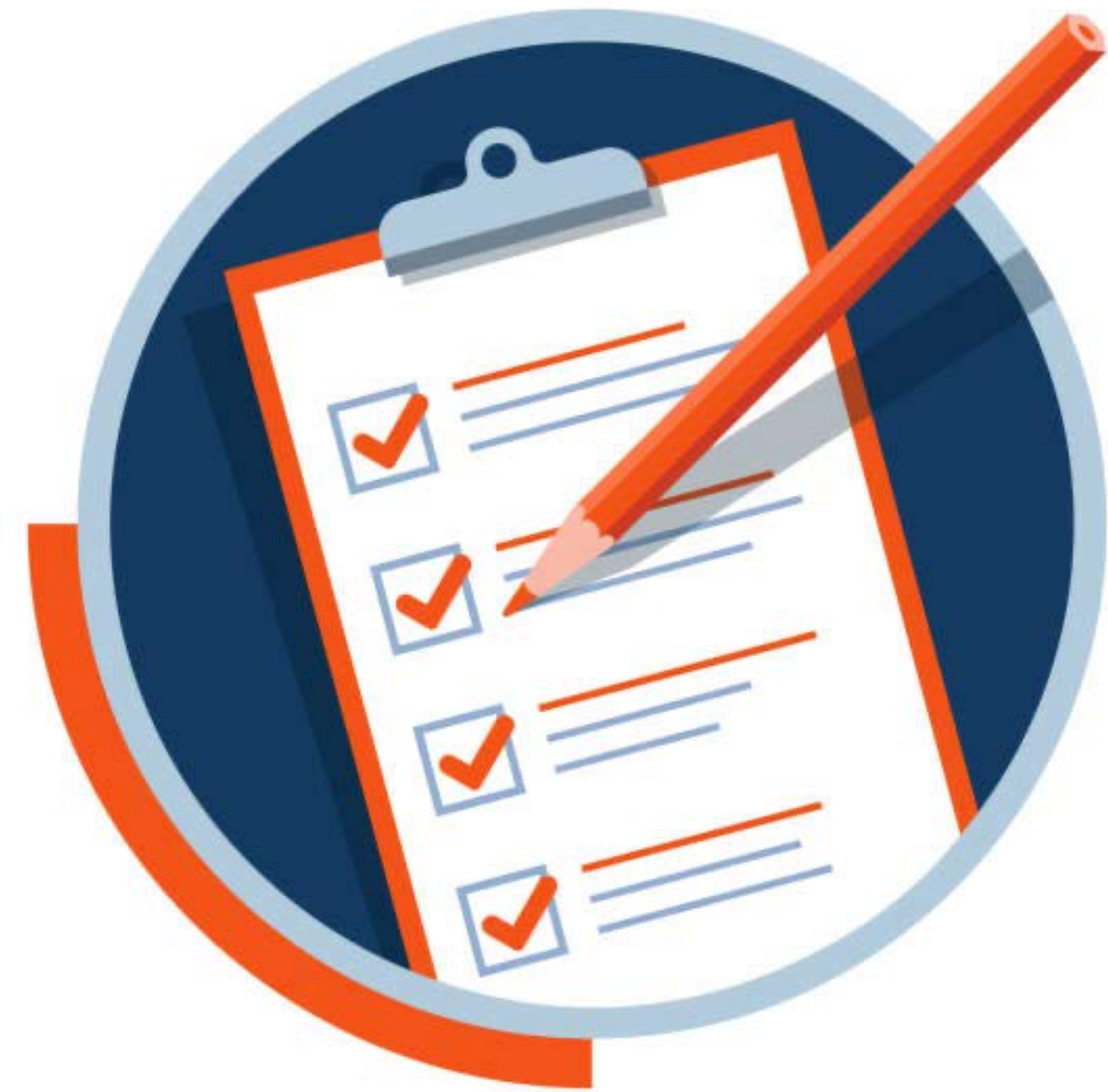
Iaroslav Geraskin

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



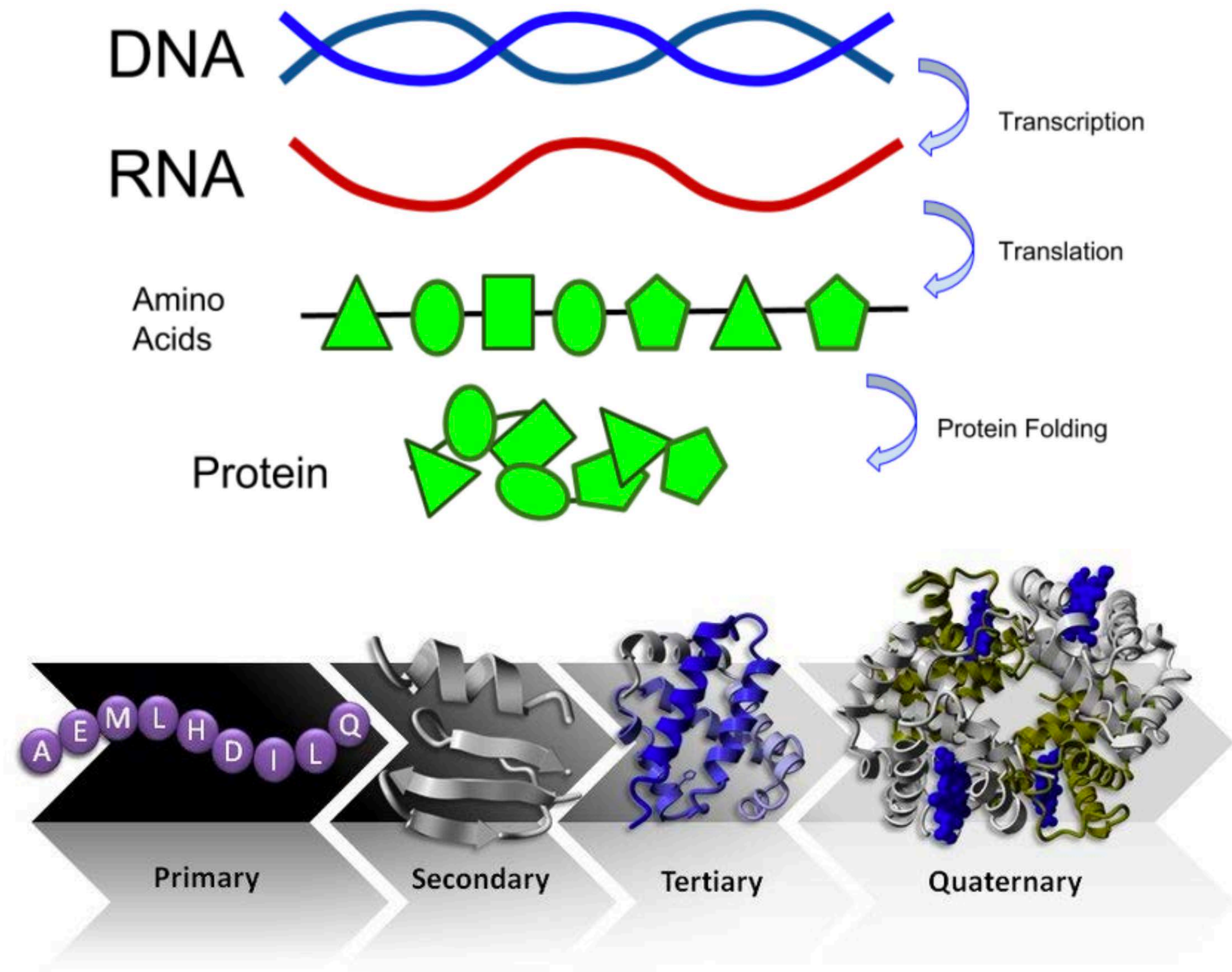
Agenda

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

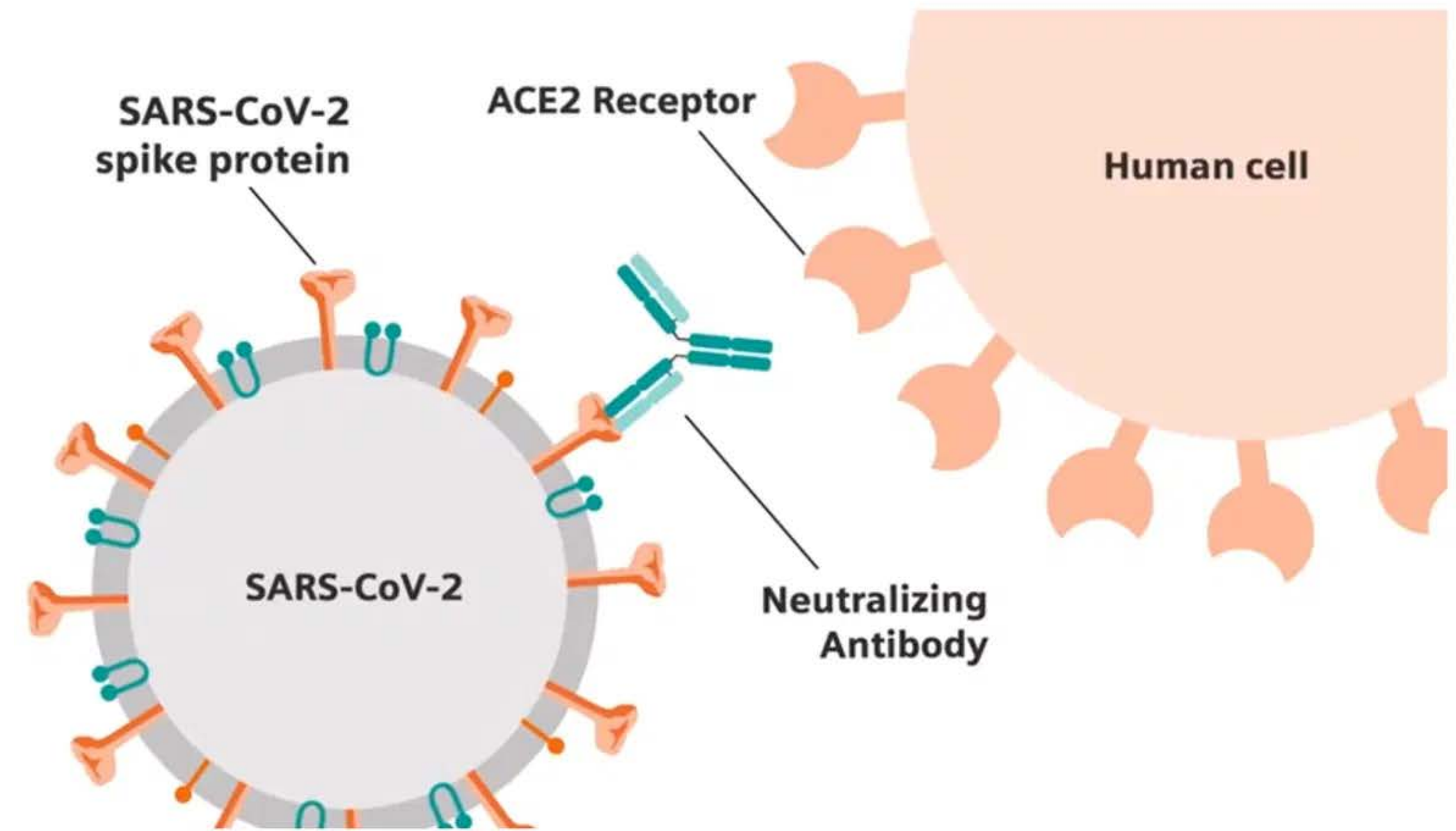
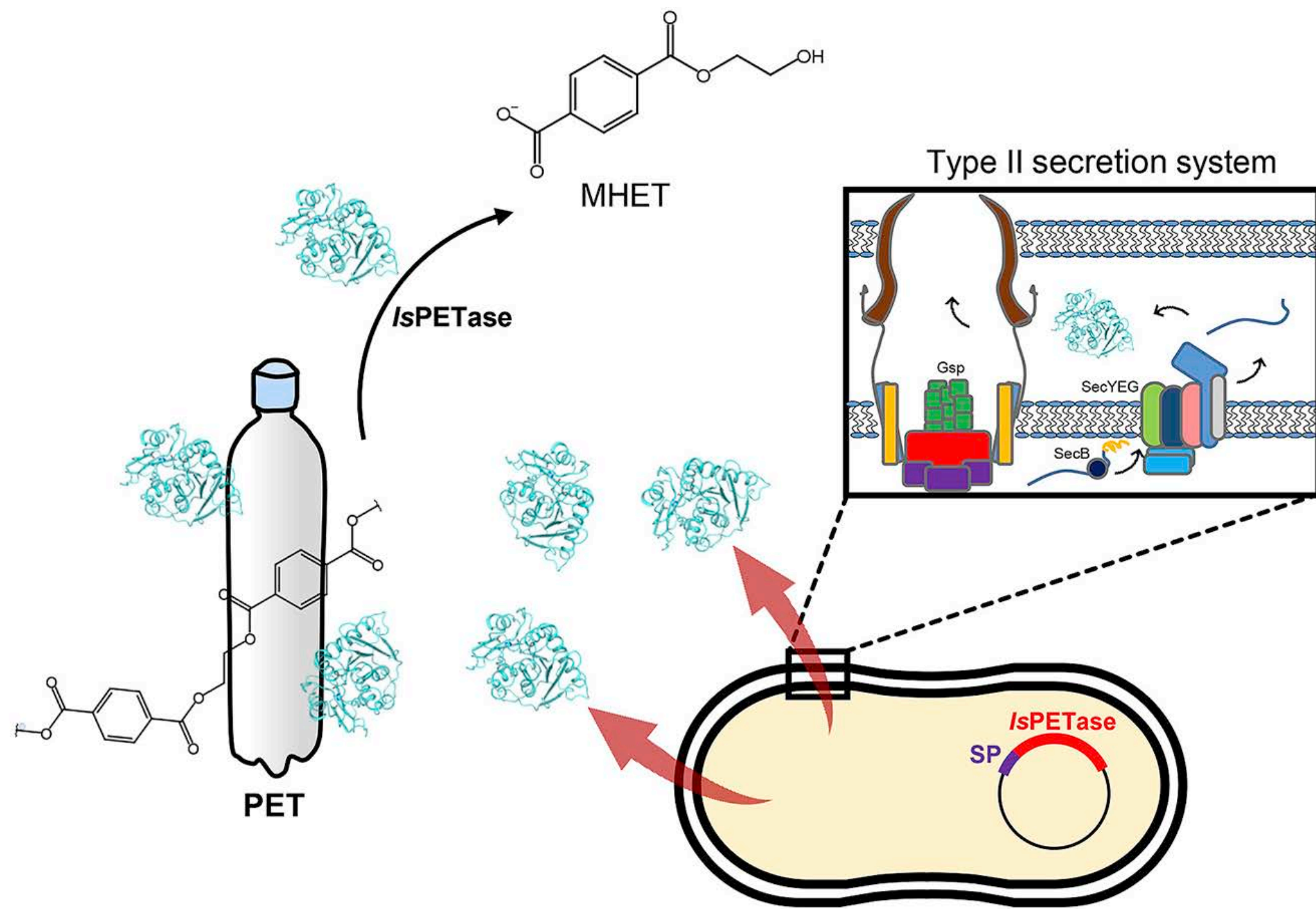


Proteins

- 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)

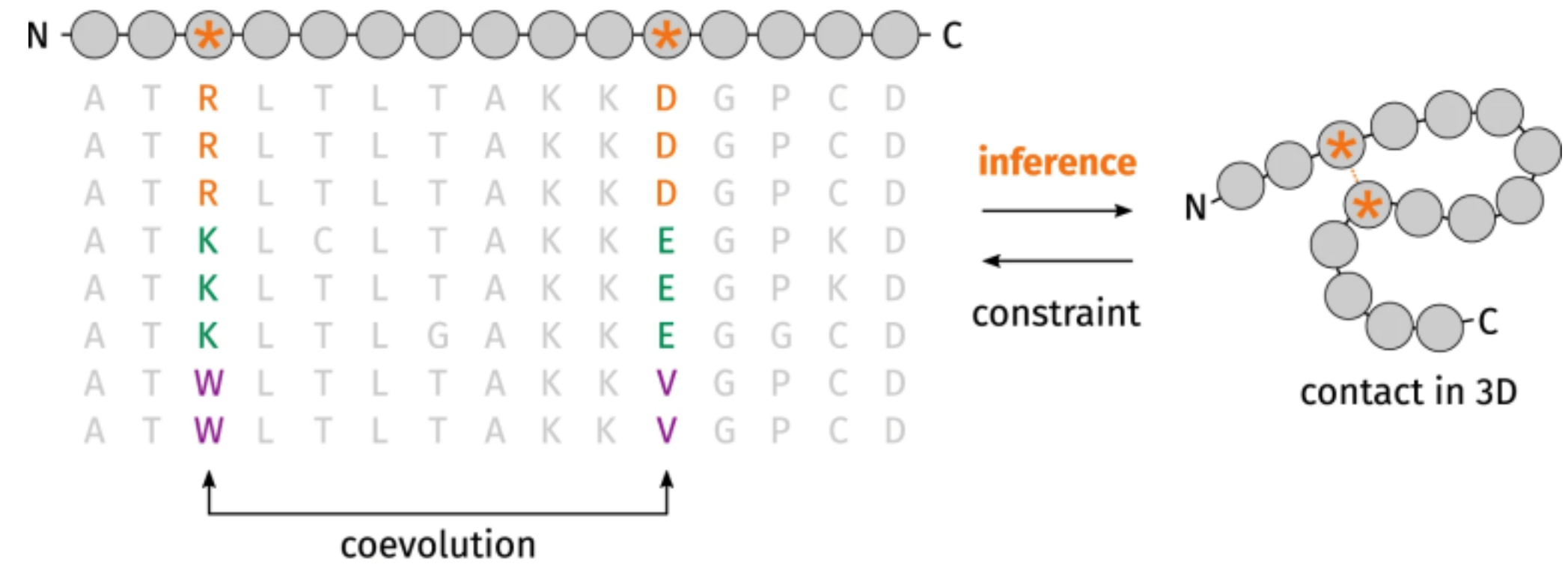


Why do we need proteins?



MSA

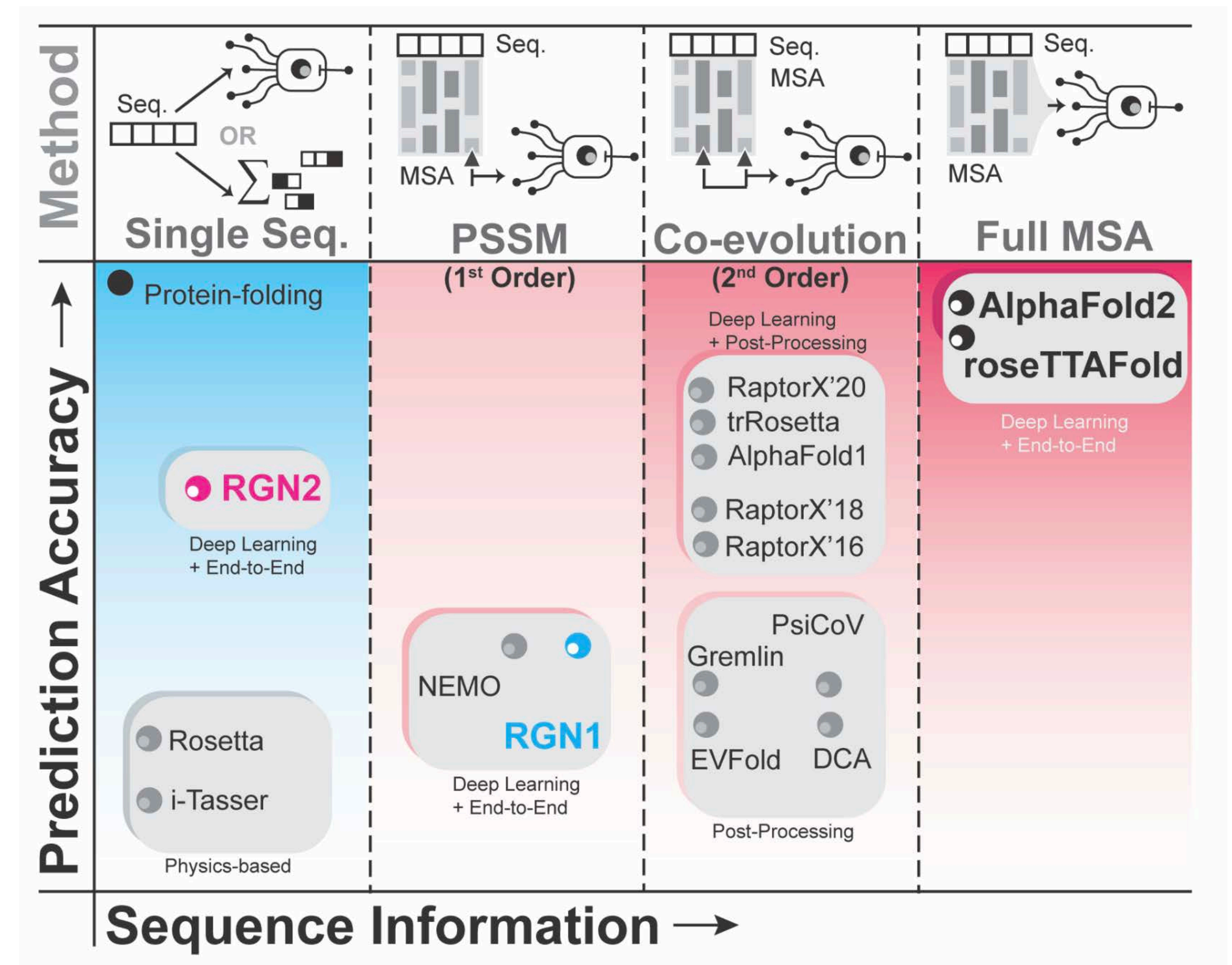
- Multiple alignment allows you to extract evolutionary information from multiple sequences, determine which positions depend on which positions, which positions are conserved and which are not



Q5E940_BOVIN	-----MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKOMQIRMSLRGK-AVVLGKNTMMRKAIRGHLENN--PALE	76
RLA0_HUMAN	-----MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKOMQIRMSLRGK-AVVLGKNTMMRKAIRGHLENN--PALE	76
RLA0_MOUSE	-----MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKOMQIRMSLRGK-AVVLGKNTMMRKAIRGHLENN--PALE	76
RLA0_RAT	-----MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKOMQIRMSLRGK-AVVLGKNTMMRKAIRGHLENN--PALE	76
RLA0_CHICK	-----MPREDRATWKSNYFMKIIQLLDDYPKCFVVGADNVGSKOMQIRMSLRGK-AVVLGKNTMMRKAIRGHLENN--PALE	76
RLA0_RANSY	-----MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKOMQIRMSLRGK-AVVLGKNTMMRKAIRGHLENN--SALE	76
Q7ZUG3_BRARE	-----MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMOTIRLSLRGK-AVVLGKNTMMRKAIRGHLENN--PALE	76
RLA0_ICTPU	-----MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMOTIRLSLRGK-AIVLMGKNTMMRKAIRGHLENN--PALE	76
RLA0_DROME	-----MVRENKAAWKAQYFIKVVLEFDFPKCFIVGADNVGSKOMQIRMSLRGL-AVVLGKNTMMRKAIRGHLENN--PQLE	76
RLA0_DICDI	-----MSGAG-SKRKKLFIEKATKLFYTDKMIVAEADFGVSSQLQKIRKSIRGI-GAVLMGKNTMIRKVIKIRDLADSK--PELD	75
Q54LP0_DICDI	-----MSGAG-SKRKNVFIEKATKLFYTDKMIVAEADFGVSSQLQKIRKSIRGI-GAVLMGKNTMIRKVIKIRDLADSK--PELD	75
RLA0_PLAF8	-----MAKLSKQKQKQMYIEKLSLIQQYSKILIVHVDNVGSKNOMASVRKSIRGK-ATILMGKNTIRIRALKKNLQAV--PQIE	76
RLA0_SULAC	----MIGLAVTTTKIAKWKVDEVAELTEKLTHTKTIIANIEGFPADKLHEIRKKLRGK-ADIKVTKNNLNFNIALKNAG----YDTK	79
RLA0_SULTO	----MRIMAVITQERKIAKWKIEEVKELEKLRHYHTIIIANIEGFPADKLHDIRKKMRGM-AEIKVTKNTLFGIAAKNAG----LDVS	80
RLA0_SULSO	----MKRLALALKQRKVASWKLEEVKELTELKNSNTILIGNLEGFPADKLHEIRKKLRGK-ATIKVTKNTLFGIAAKNAG----IDIE	80
RLA0_AERPE	MSVVSIVGQMYKREKPIPEWKTLMLELELFSKHRVLFADLTGTPFVVRVRKKLWKK-YPMMVAKKRIILRAMKAAGLE--LDDN	86
RLA0_PYRAE	-MMLAIGKRRYVRTQYPARKVKIVSEATELLQKYPYVFLFDLHGLSSRILHEYRYRLRY-GVIKIKPTLFLKIAFTKVYGG--IPAE	85
RLA0_METAC	-----MAEERHTEHIPQWKDEIENIKELIQSHKVFVGMVIEGILATKMKIRRDLDKV-AVLKVSNTLTERALNQLG--ETIP	78
RLA0_METMA	-----MAEERHTEHIPQWKDEIENIKELIQSHKVFVGMVIEGILATKMKIRRDLDKV-AVLKVSNTLTERALNQLG--ESIP	78
RLA0_ARCFU	-----MAAVRGS--PPEYKVRAVEEIKRMISKPVVAIVSFRNVPAGOMQIRREFRGK-AEIKVVKNTLLEALDALG----GDYL	75
RLA0_METKA	MAVKAKGQPPSGYEPKVAEWRREVEKELKELMDEYENVGLVDLEGIPAPQLOEIRAKLRERDTIIRMSRNTLMRIALEEKLDER--PELE	88
RLA0_METTH	-----MAHVAEWKKEVQELHDLIKGYEVVGIANLADIPARQLOKMRQTLRDS-ALIRMSKKTLLISLALAKAGREL--ENVD	74
RLA0_METTL	-----MITAESEHKIAPWKIEEVNKLKELLKNGQIVALVDMMEVPAQLOEIRDKIR-GTMTLKMSRNTLIERAIKEVAEETGNPEFA	82
RLA0_METVA	-----MIDAKSEHKIAPWKIEEVNALKELLSANVIALIDMMEVPAQLOEIRDKIR-DQMTLKMSRNTLIKRAVEEVAEETGNPEFA	82
RLA0_METJA	-----METKVKAHVAPWKIEEVKTLKGLIKSKPVVAIVDMMDVPAPQLOEIRDKIR-DKVKLRMSRNTLIERALKEAAEELNPKLA	81
RLA0_PYRAB	-----MAHVAEWKKEVEELANLIKSPVIALVDVSSMPAYPLSQMRRLIRENGLLRVSNTLIELAIKKAQELGKPELE	77
RLA0_PYRHO	-----MAHVAEWKKEVEELAKLIKSPVIALVDVSSMPAYPLSQMRRLIRENGLLRVSNTLIELAIKKAQELGKPELE	77
RLA0_PYRFU	-----MAHVAEWKKEVEELANLIKSPVIALVDVSSMPAYPLSQMRRLIRENGLLRVSNTLIELAIKKAQELGKPELE	77
RLA0_PYRKO	-----MAHVAEWKKEVEELANLIKSPVIALVDVAGVPAYPLSKMRDKLR-GKALLRVSNTLIELAIKRAAQELGQPELE	76
RLA0_HALMA	-----MSAESERKTETIPWQEEVDATVEMIESYESVGVVNIAGIPSRQLODMRRDLHGT-AELRVSNTLLEALDDVD----DGLE	79
RLA0_HALVO	-----MSESEVRQTEVIPWQKREEDVDFIESYESVGVVGVAGIPSRQLOSMRRELHGS-AAVRMSRNTLVNRALEVN----DGFE	79
RLA0_HALSA	-----MSAEEQRTTEVPWKRQEVAVLDLETYDSVGVVNVVTGIPSKQLODMRRGLHGQ-AAVRMSRNTLLVRALEEAG----DGLD	79
RLA0_THEAC	-----MKEVSQKKELVNEITRIKASRSVAIVDTAGIRTRIQIDIRGKNRGK-INLKVIKKTLLFKALENLGD----EKLS	72
RLA0_THEVO	-----MRKINPKKKEIVSELAQDITKSKAVAVDIKGVTRIQMODIRAKNRDK-VKIKVVKKTLLFKALDSIND----EKLT	72
RLA0_PICTO	-----MTEPAQWKIDFVKNLENEINSRKVAAIVSIKGLRNNEFQKIRNSIRDK-ARIKVSRRARLLRLAIENTGK----NNIV	72
ruler	1.....10.....20.....30.....40.....50.....60.....70.....80.....90	

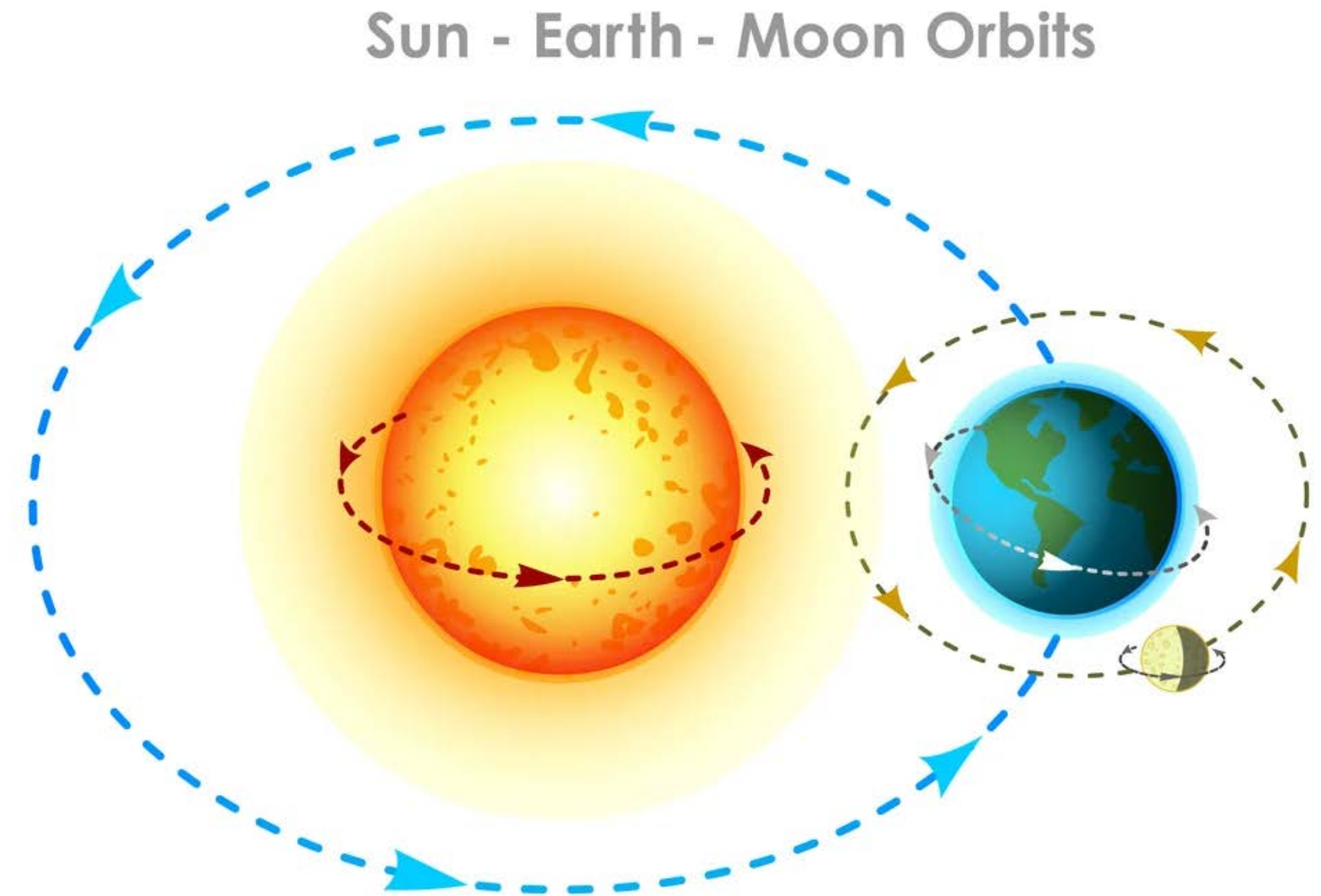
Structure prediction methods

- 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.



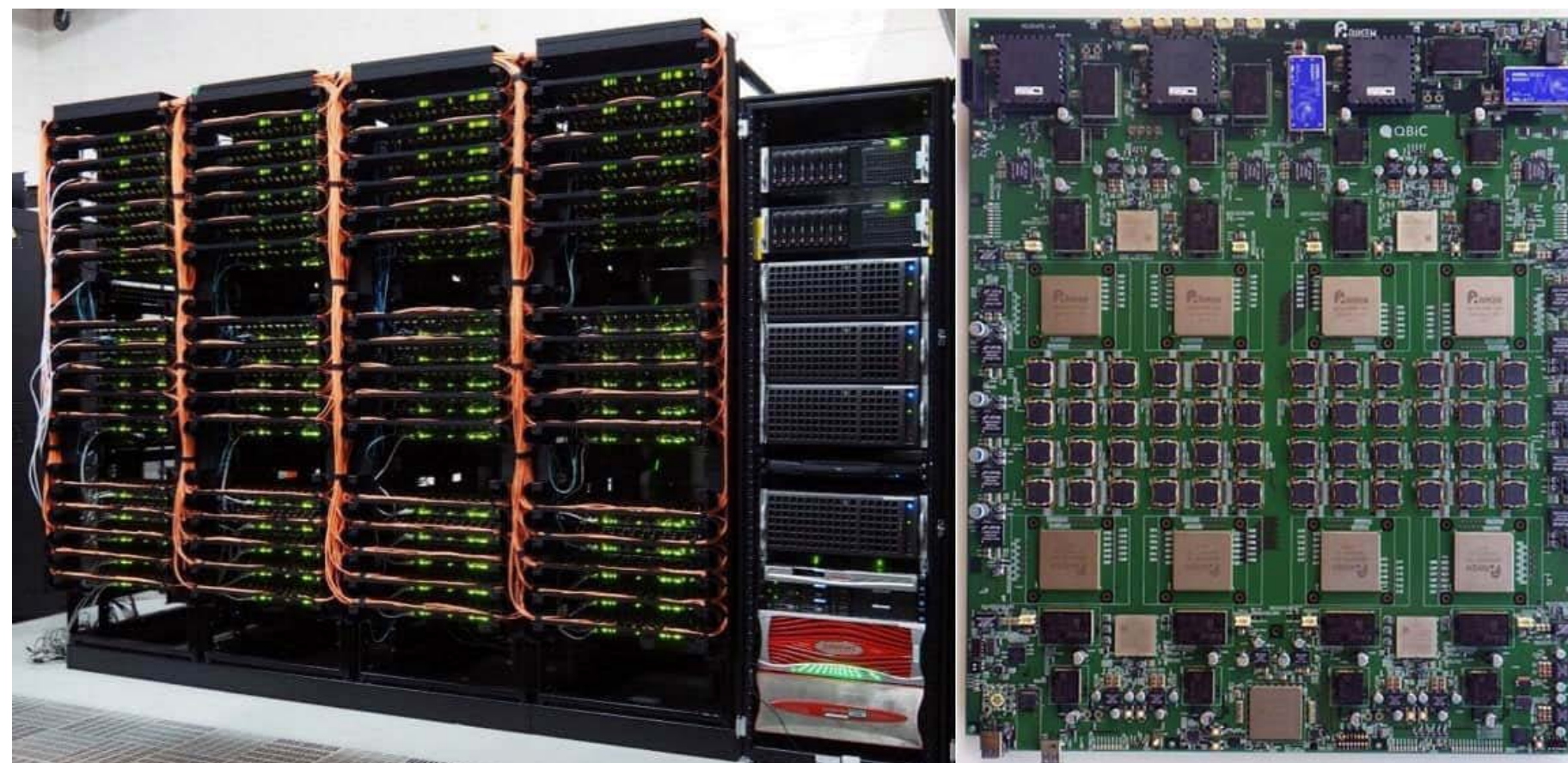
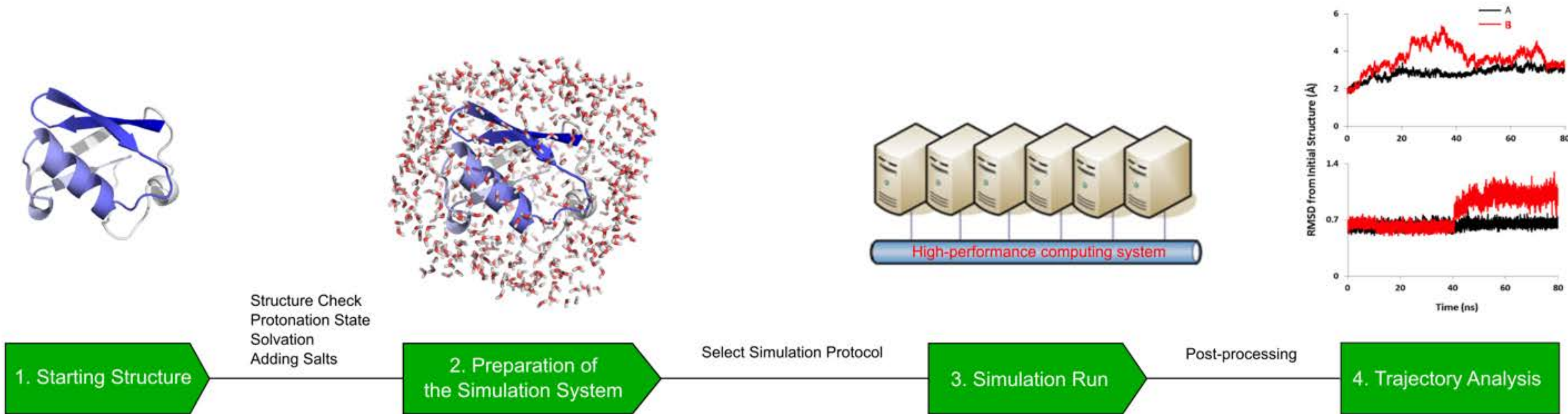
Three-body problem

- 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



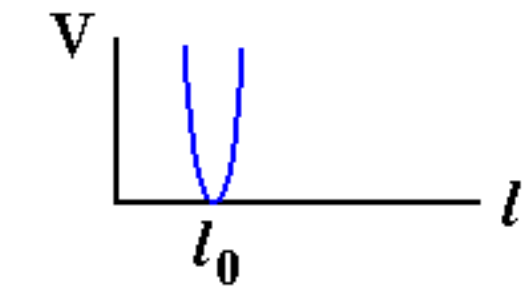
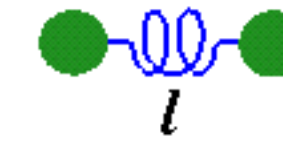
Molecular dynamics

Physical potential

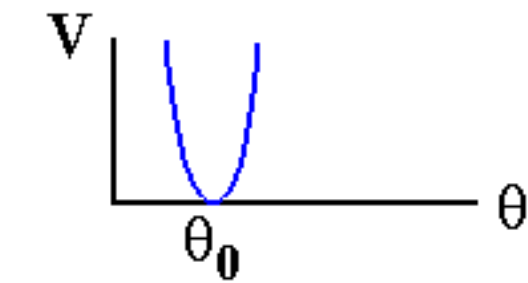
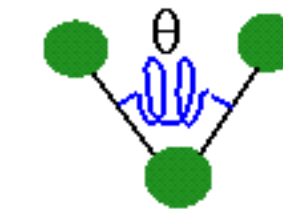
- Compute new positions for particles based on forces acting on them
- Take a lot of small steps
- Can we get a good approximation first?
- Why follow physics when we have statistics?

Empirical Potential Energy Function

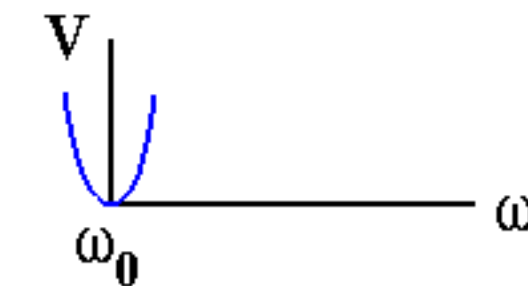
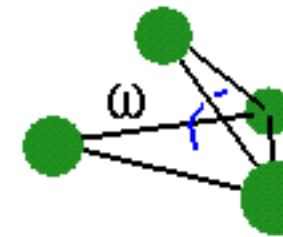
Bonds



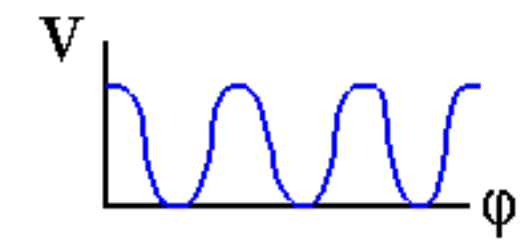
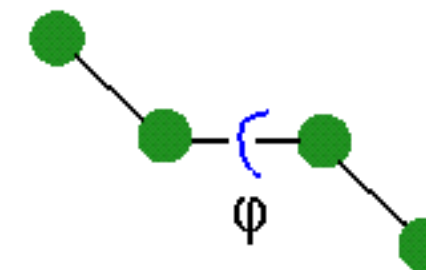
Angles



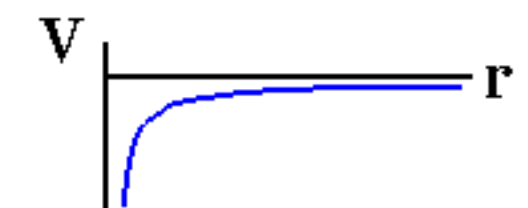
Improper
Dihedrals



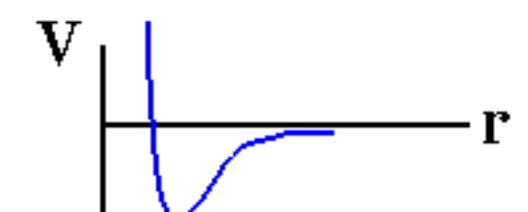
Torsions



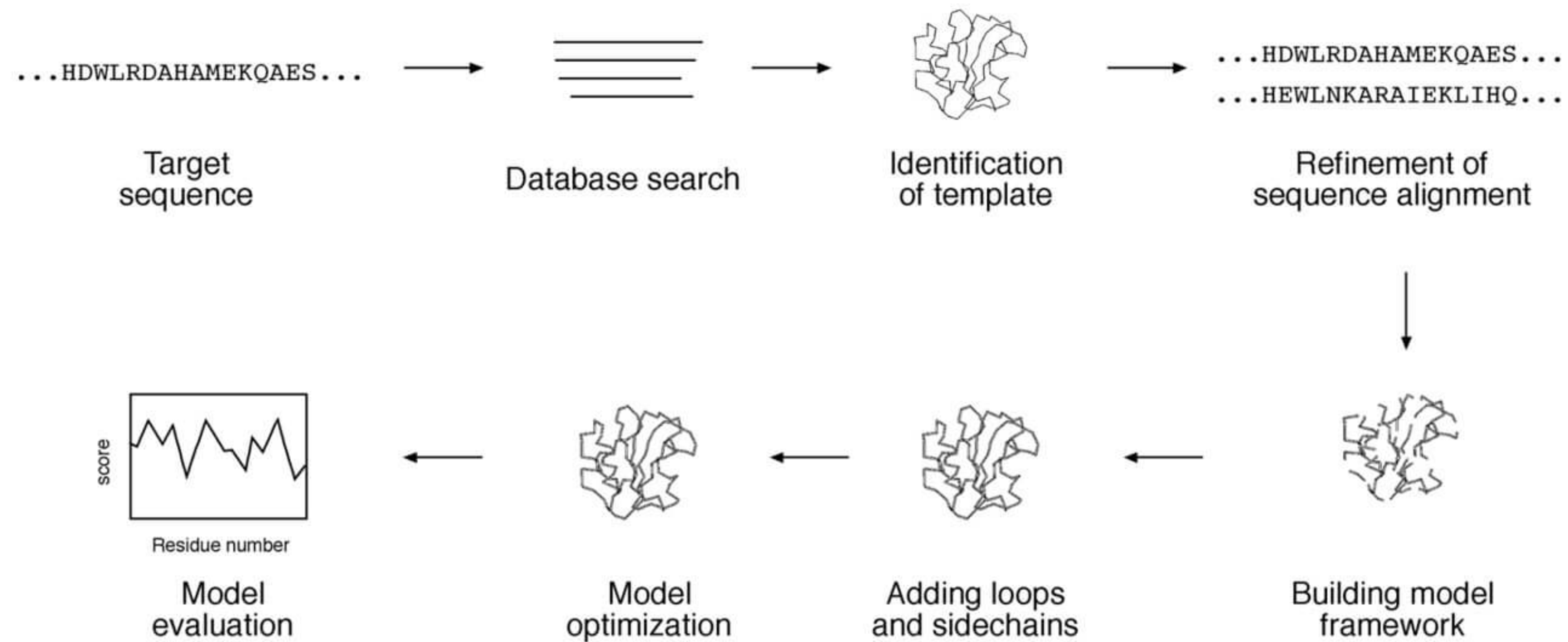
Electrostatics



van der Waals



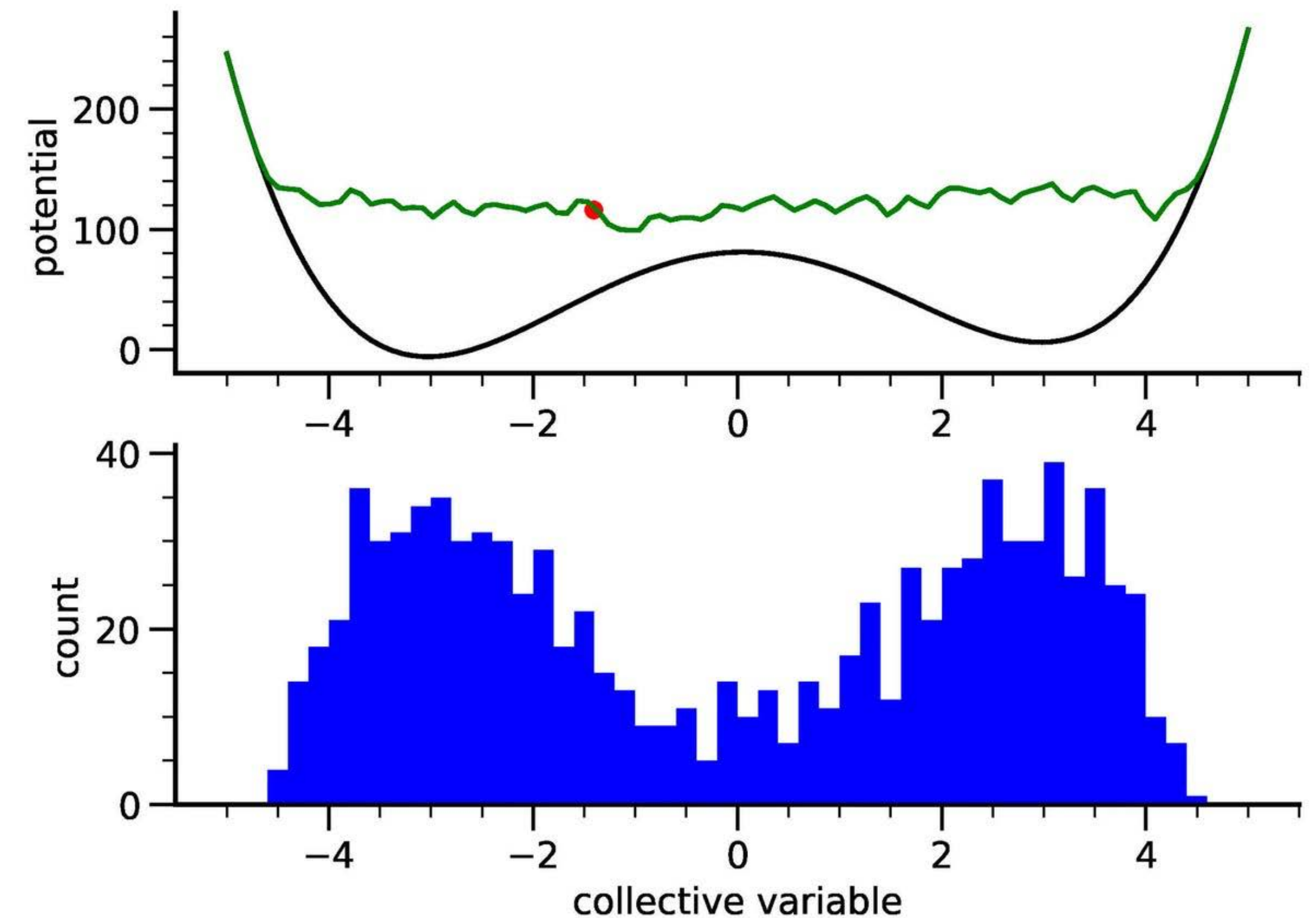
Homology modeling



Molecular dynamics

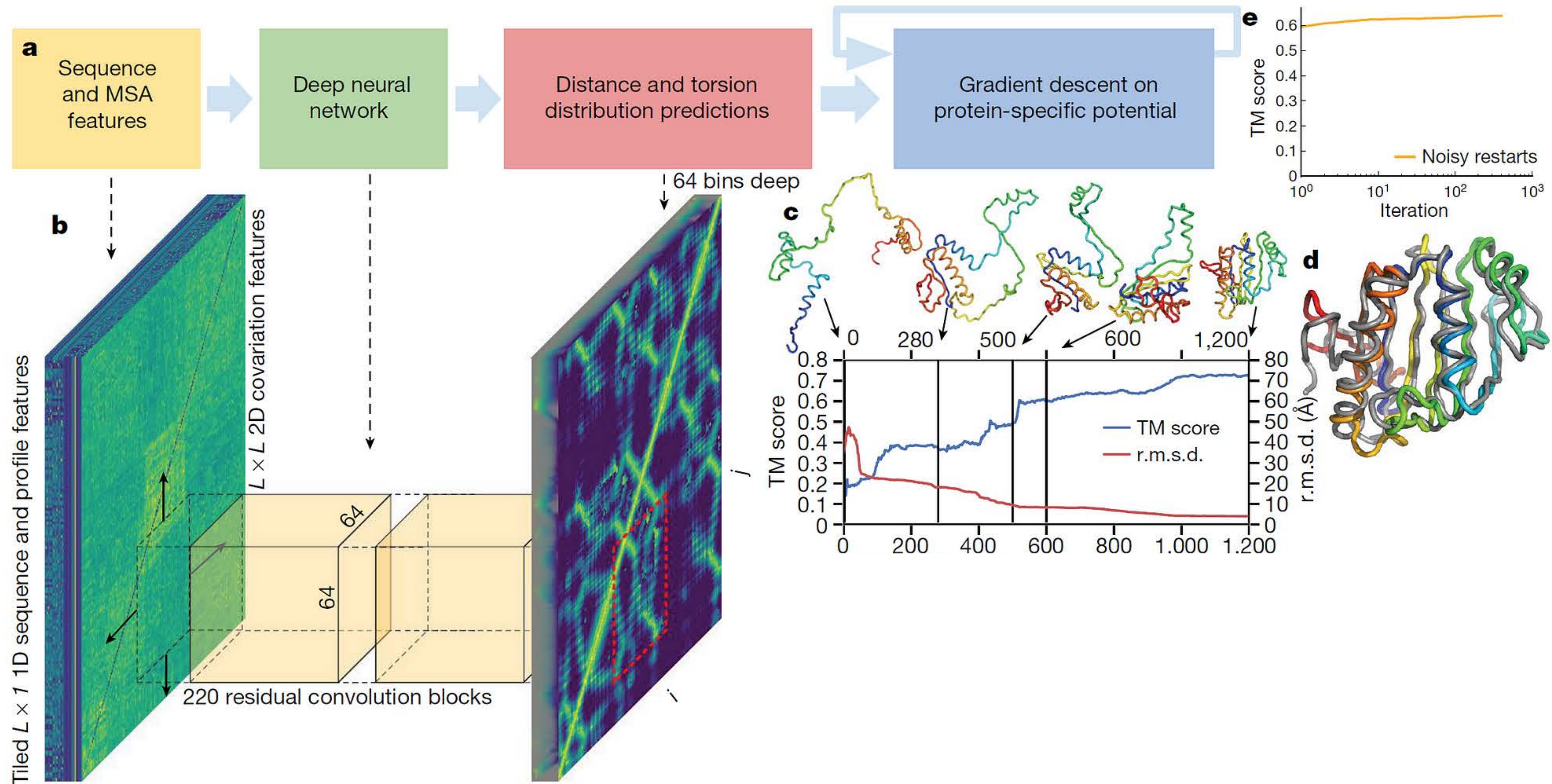
Statistical potential

- 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



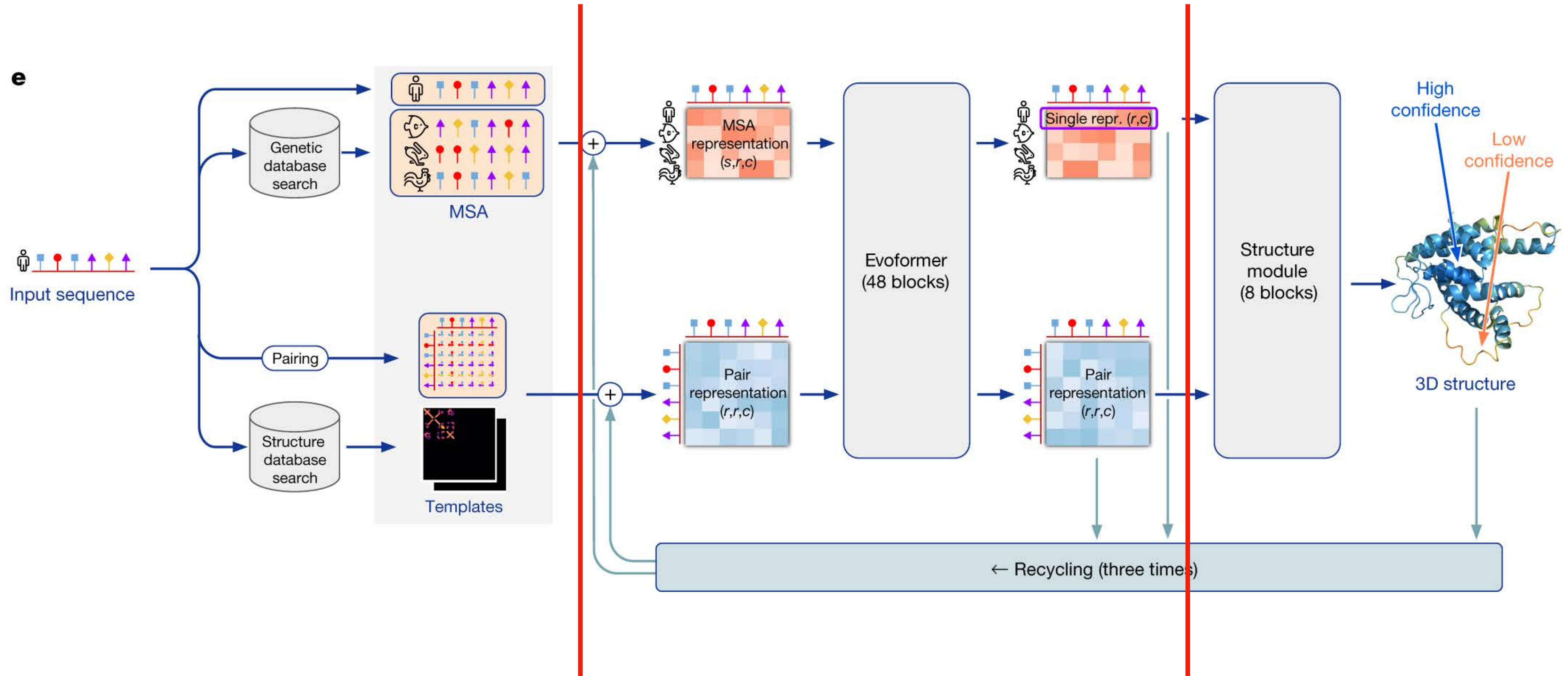
AlphaFold 1

Predict the potential, convolution, MSA features



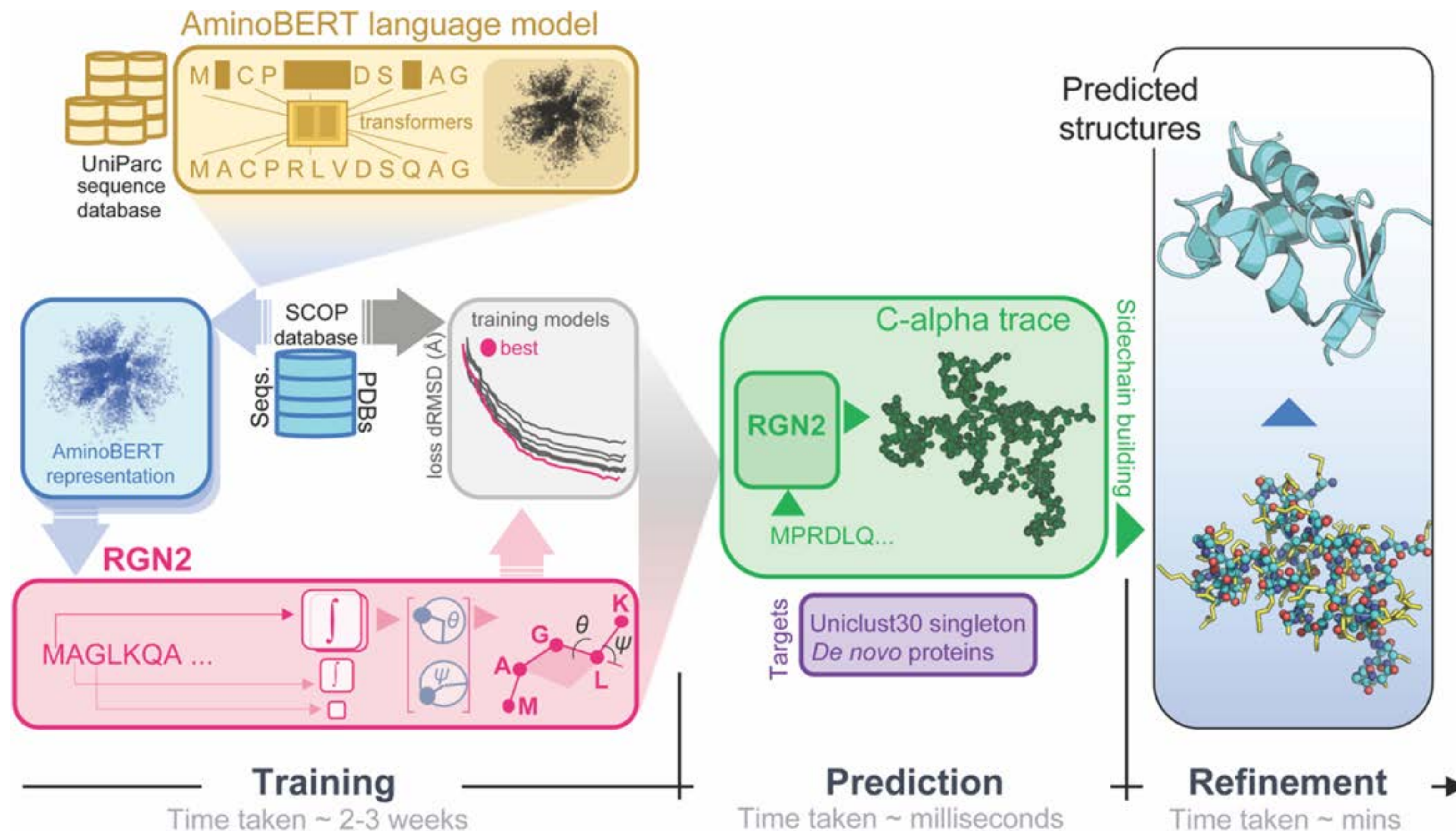
AlphaFold 2

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



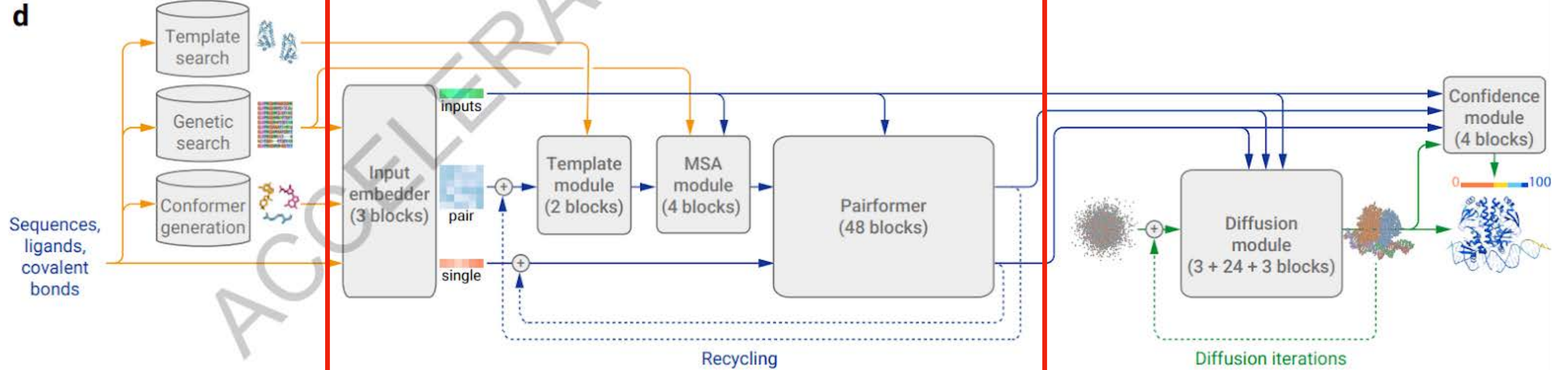
Single Seq

- 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