Predicting structures of protein complexes using AlphaFold

Troy Whitfield

Whitehead Institute
Bioinformatics and Research Computing

February 29, 2024

Outline

1. **Introduction**
   - Experiments
   - Predictions
   - Interpretation
2. **Complexes**
3. **Applications**
4. **References**

Protein structure prediction: Nature method of 2022

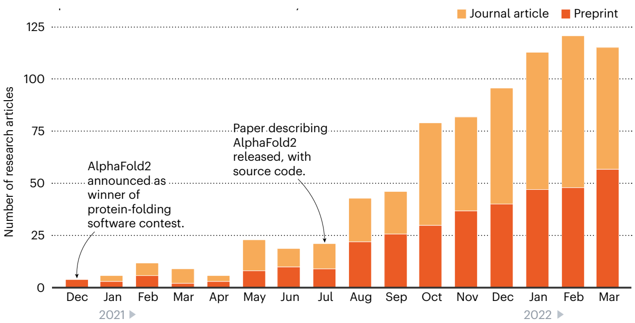


Fig. 1: *The number of research papers and preprints citing the AlphaFold2 AI software has shot up since its source code was released in July 2021 [Callaway, 2022].*

Protein structure informs us about function

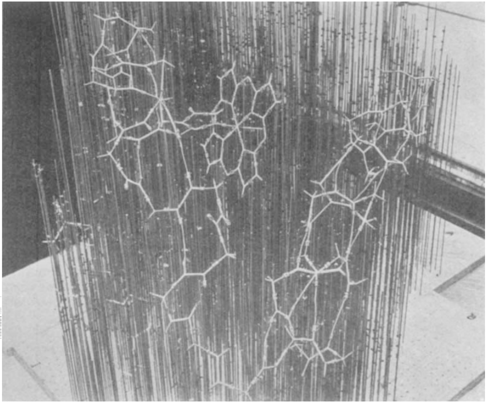


Fig. 2: *Myoglobin structure solved by X-ray crystallography, 1958. Ball and stick model of myoglobin surrounding its heme group, vertical rods and pins outline electron density [Kendrew* et al.*, 1960].*

Why is predicting protein structure so difficult?

- Levinthal's "paradox": Levinthal noted that the number of configurations a protein can sample is vast (e.g. 3^{100} for a 101 residue protein), far too many for the protein to sample on measured folding times (e.g. seconds) [Levinthal, 1969].
- Proteins don't *need* chaperones: experimental *in vitro* denaturation and refolding. From such studies on ribonuclease A, Anfinsen hypothesized that the native protein structure is fully determined by amino acid sequence [Anfinsen, 1973].
- Chemical kinetics help us understand the paradox: unfavorable configurations tend to be avoided and favorable local configurations tend to be stable [Zwanzig *et al.*, 1992].

Approaches to fold proteins on a computer

1. Physics based [Wolynes, 2015, Shaw *et al.*, 2010].
2. Homology modeling [Levitt, 1992, Sali and Blundell, 1993].
3. Evolutionary correlations: co-evolving amino acids imply contacts [Shindyalov *et al.*, 1994, Morcos *et al.*, 2011].

Proteins fold on energy landscapes

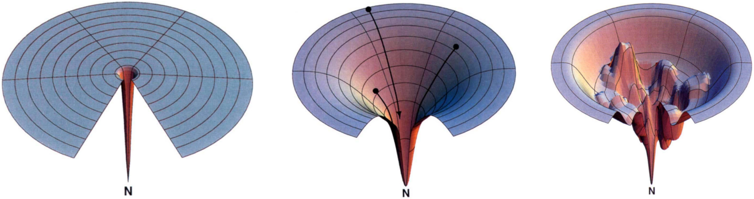


Fig. 3: *Low-dimensional visualizations of different folding landscapes, including a Levinthal "golf-course" landscape (left), idealized folding funnel (center) and rugged landscape (right) [Dill and Chan, 1997].*

Proteins can fold in computer simulations, but...

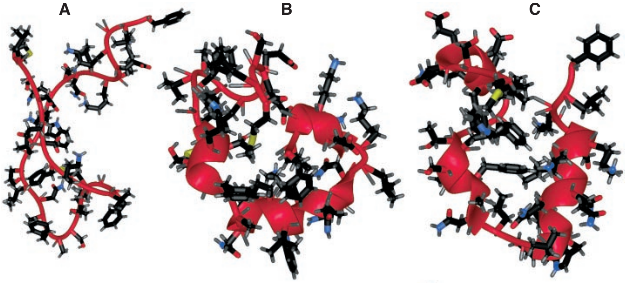


Fig. 4: *Ribbon representations of an unfolded (A), partially folded (at 980 ns, B) and native (C) 36-residue villin headpiece subdomain, HP-36 [Duan and Kollman, 1998].*

Homologous protein domains can serve as templates for predicting structure

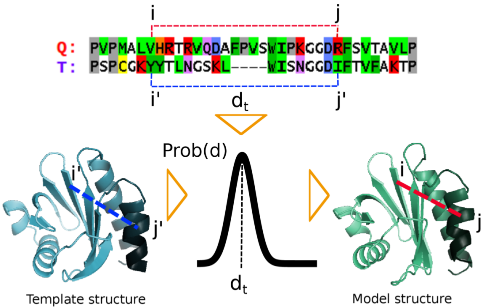
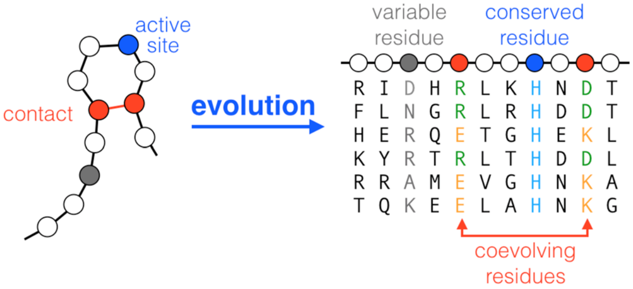

i
j
Q: PVPMALVHRTRVQDAFPVSWIPKGGDRFSVTAVLP
T: PSPCGKYYTLNGSKL----WISNGGDIFTVFAKTP
i'
d_t
j'
Prob(d)
i'
j'
d_t
i
j
Template structure
Model structure

Fig. 5: *MODELLER's [Sali and Blundell, 1993] statistical approach to homology modeling. The unknown distance d between two atoms in residues i and j of the query protein (Q) is described by a probability distribution Prob(d) that is peaked around the distance dt between the corresponding atoms in residues i′ and j′ of the template protein (T) [Meier and Söding, 2015].*

Protein structure computed from evolutionary sequence variation



```
R I D H R L K H N D T
F L N G R L R H D D T
H E R Q E T G H E K L
K Y R T R L T H D D L
R R A M E V G H N K A
T Q K E E L A H N K G
```

Fig. 6: *Evolutionary constraints shaping the variability between homologous sequences: while constraints on individual residues (e.g. active sites) lead to variable levels of amino-acid conservation, the conservation of contacts leads to the coevolution of structurally neighboring residues and therefore to correlations between columns in a multiple-sequence alignment of homologous proteins[Cocco* et al.*, 2018].*

Protein structure computed from evolutionary sequence variation

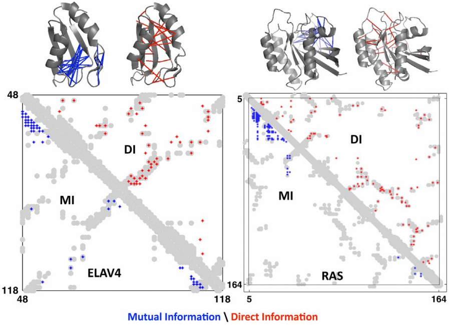


Fig. 7: *Extraction of evolutionary information about residue coupling and predicted contacts from multiple sequence alignments works much better using the global statistical model (direct information, DI) than the local statistical model (mutual information, MI) [Marks* et al.*, 2011].*

AlphaFold’s deep learning architecture

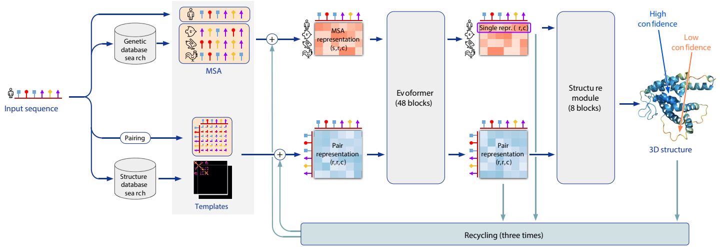


Fig. 8: *Arrows show the information flow among the various algorithmic components. Array shapes are shown in parentheses with s, number of sequences; r, number of residues; c, number of channels. Compared with an earlier deep learning architecture, “key innovations in the Evoformer block are new mechanisms to exchange information within the MSA and pair representations that enable direct reasoning about the spatial and evolutionary relationships.” [Jumper* et al.*, 2021].*

Reminder: neural network training adjusts weights

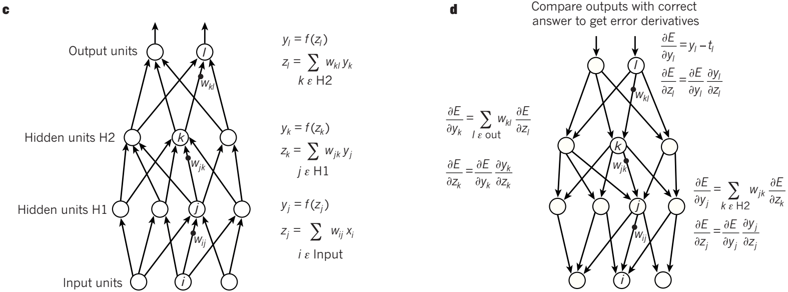

c
Output units
Hidden units H2
Hidden units H1
Input units
l
k
j
i
w_{kl}
w_{jk}
w_{ij}
$y_l = f(z_l)$
$z_l = \sum_{k \in H2} w_{kl} y_k$
$y_k = f(z_k)$
$z_k = \sum_{j \in H1} w_{jk} y_j$
$y_j = f(z_j)$
$z_j = \sum_{i \in Input} w_{ij} x_i$
d
Compare outputs with correct answer to get error derivatives
$\frac{\partial E}{\partial y_l} = y_l - t_l$
$\frac{\partial E}{\partial z_l} = \frac{\partial E}{\partial y_l} \frac{\partial y_l}{\partial z_l}$
$\frac{\partial E}{\partial y_k} = \sum_{l \in out} w_{kl} \frac{\partial E}{\partial z_l}$
$\frac{\partial E}{\partial z_k} = \frac{\partial E}{\partial y_k} \frac{\partial y_k}{\partial z_k}$
$\frac{\partial E}{\partial y_j} = \sum_{k \in H2} w_{jk} \frac{\partial E}{\partial z_k}$
$\frac{\partial E}{\partial z_j} = \frac{\partial E}{\partial y_j} \frac{\partial y_j}{\partial z_j}$

Fig. 9: *To generally illustrate neural network training, consider a multilayer neural network with backpropagation [LeCun* et al.*, 2015]. Training consists in iterating to minimize the loss.*

Success of AlphaFold

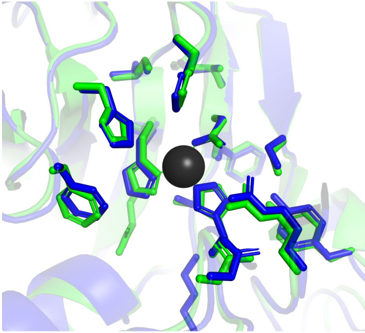
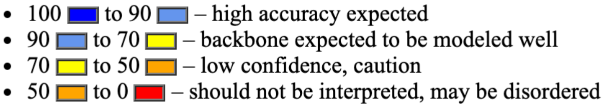


Fig. 10: *An example of a well-predicted zinc binding site. AlphaFold has accurate side chains even though it does not explicitly predict the zinc ion [Jumper* et al., *2021].*

Assessing the quality of predicted structures

- AlphaFold reports *predicted* per-residue Local Distance Difference Test (pLDDT), based on the lDDT-C$_\alpha$ metric [Mariani *et al.*, 2013].



- 100 🟦 to 90 🟦 – high accuracy expected
- 90 🟦 to 70 🟨 – backbone expected to be modeled well
- 70 🟨 to 50 🟧 – low confidence, caution
- 50 🟧 to 0 🟥 – should not be interpreted, may be disordered

- Predicted Alignment Error (PAE) "indicates the expected positional error at residue x if the predicted and actual structures are aligned on residue y". Measured in $\AA$ and capped at 31.75 $\AA$ [Varadi *et al.*, 2022].
- Predicted TM-score is a *global* superposition metric that lies between $(0, 1]$, with better structures having higher pTM [Zhang and Skolnick, 2004].

Monomeric folding example: HuR

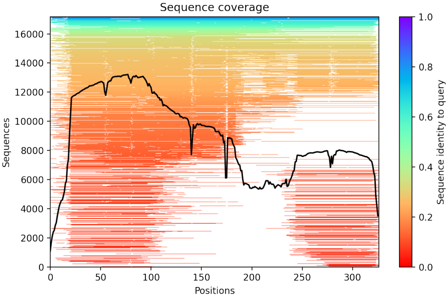


Fig. 11: *Multiple sequence alignment for human antigen R (HuR), a post-transcriptional regulator with three RNA recognition motifs (RRMs). The RRM sequence is clearly more evolutionarily conserved than the linking sequences.*

Monomeric folding example: HuR

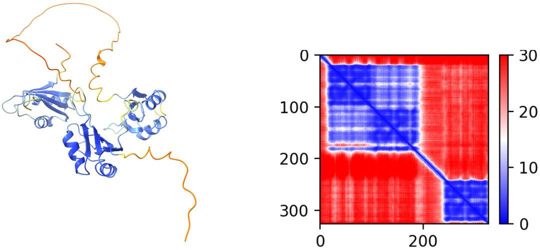
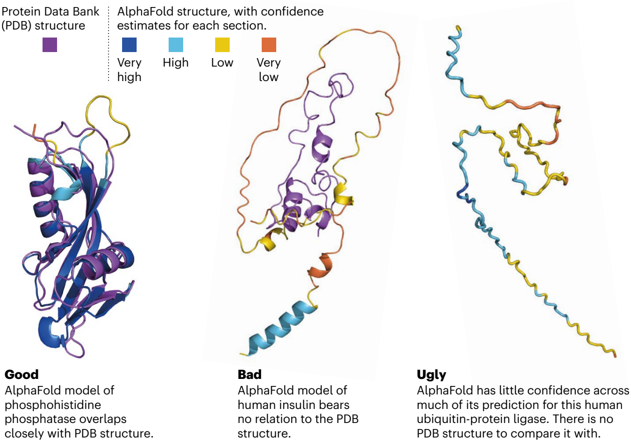


Fig. 12: *Predicted structure for the human antigen R (HuR) protein, a post-transcriptional regulator. The predicted structure has high confidence in the three RRMs (see pLDDT, left), while the PAE (right) indicates low confidence in their relative positions and pTM=0.487.*

Gallery of AlphaFold structural predictions


Protein Data Bank (PDB) structure
AlphaFold structure, with confidence estimates for each section.
Very high
High
Low
Very low

Good
AlphaFold model of phosphohistidine phosphatase overlaps closely with PDB structure.

Bad
AlphaFold model of human insulin bears no relation to the PDB structure.

Ugly
AlphaFold has little confidence across much of its prediction for this human ubiquitin-protein ligase. There is no PDB structure to compare it with.

Fig. 13: *AlphaFold's predictions of a folded protein's structure come with confidence estimates. Superimposing each model on the experimentally determined structure (if available) shows the accuracy of the prediction. [Callaway, 2022].*

Some limitations

- Intrinsic disorder, like that seen in Fig. 13, accounts for approximately 1/3 of the human proteome [Callaway, 2022]. Disordered domains can become structured in the presence of other factors.
- Because of its training, AlphaFold is not sensitive to missense mutations that may disrupt protein structure [Buel and Walters, 2022].
- Sometimes predicted structures with favorable metrics turn out to be incorrect [Callaway, 2022].
- Domains that are predicted to be disordered may sometimes fold using different tools (e.g. roseTTAFold [Baek *et al.*, 2021]).

Application to a multimeric protein complex: actin and profilin-1

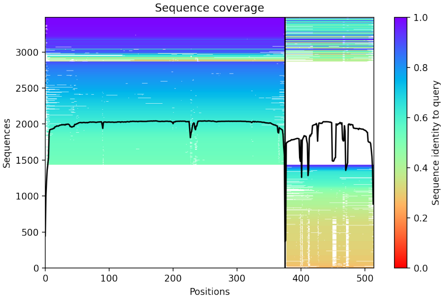


Fig. 14: *Multiple sequence alignment of a heterodimer. Note the species-matched alignments of orthologous proteins: amino acids that co-evolve between the two proteins carry information about the interface.*

Application to a multimeric protein complex: actin and profilin-1

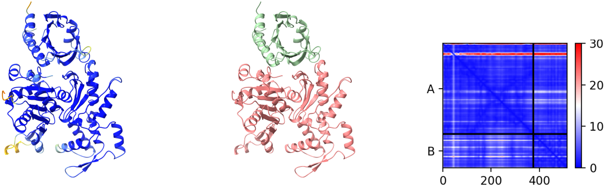


Fig. 15: *Predicted structure for the human actin-PFN1 complex closely agrees with experimental structure, Note the consistently high pLDDT and low PAE.*

Application to a decoy multimeric protein complex: myoglobin and HuR RRM1

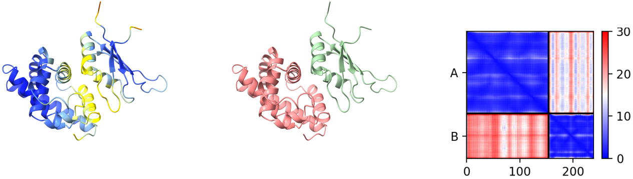


Fig. 16: *Predicted structure for a possible myoglobin-HuR complex appears inplausible: note both the poor interfacial pLDDT and high inter-subunit PAE.*

Ways to access AlphaFold predictions

1. DeepMind/EMBL-EBI online database of predicted structures [Varadi *et al.*, 2022].
2. ColabFold [Mirdita *et al.*, 2022] via ChimeraX [Goddard *et al.*, 2018, Pettersen *et al.*, 2021].
3. Run AlphaFold or AlphaFold Multimer at Whitehead (see BaRC Best Practices webpage).
4. Run AlphaFold at MIT (e.g. SuperCloud).

Considerations for predicting protein complexes

- Predicting protein complexes remains a challenge, although AlphaFold Multimer has shown remarkable success [Evans *et al.*, 2022].
- In a recent benchmark test of protein complex prediction, AlphaFold predicted near-native structures for 43% of the complexes, far greater than the 9% rate for unbound protein-protein docking [Yin *et al.*, 2022].
- AlphaFold does poorly at predicting antigen-antibody complexes, evidently "due to lack of coevolution signal" [Yin *et al.*, 2022].

What’s next?

- AF2Complex [Gao *et al.*, 2022].
- ESMFold [Lin *et al.*, 2023].
- Structures of protein-nucleic acid complexes [Esmaeeli *et al.*, 2023].
- Screening for putative binding partners [Yu *et al.*, 2023].
- ...

Summary

- AlphaFold offers predicted protein structures of unprecedented quality and breadth, most easily accessed via the AlphaFold Protein Structure Database.
- AlphaFold predictions can readily be accessed using ChimeraX, but can also be carried out at Whitehead and elsewhere at MIT.
- Be aware of how AlphaFold works and some of the limitations (e.g. insensitivity to missense mutations) that implies.

References I

- Anfinsen,C.B. (1973) Principles that govern the folding of protein chains.
  Science, **181** (4096), 223–230.

- Baek,M., DiMaio,F., Anishchenko,I., Dauparas,J., Ovchinnikov,S., Lee,G.R., Wang,J., Cong,Q., Kinch,L.N., Schaeffer,R.D., Millán,C., Park,H., Adams,C., Glassman,C.R., DeGiovanni,A., Pereira,J.H., Rodrigues,A.V., van Dijk,A.A., Ebrecht,A.C., Opperman,D.J., Sagmeister,T., Buhlheller,C., Pavkov-Keller,T., Rathinaswamy,M.K., Dalwadi,U., Yip,C.K., Burke,J.E., Garcia,K.C., Grishin,N.V., Adams,P.D., Read,R.J. and Baker,D. (2021) Accurate prediction of protein structures and interactions using a three-track neural network.
  Science, **373** (6557), 871–876.

References II

- Buel,G.R. and Walters,K.J. (2022) Can AlphaFold2 predict the impact of missense mutations on structure?
  Nat. Struct. Mol. Biol., **29** (1), 1–2.

- Callaway,E. (2022) What's next for AlphaFold and the AI protein-folding revolution.
  Nature, **604** (7905), 234–238.

- Cocco,S., Feinauer,C., Figliuzzi,M., Monasson,R. and Weigt,M. (2018) Inverse statistical physics of protein sequences: a key issues review.
  Rep. Prog. Phys., **81** (3), 032601.

- Dill,K.A. and Chan,H.S. (1997) From Levinthal to pathways to funnels.
  Nat. Struct. Biol., **4** (1), 10–19.

References III

- Duan,Y. and Kollman,P.A. (1998) Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution.
  Science, **282** (5389), 740–744.

- Esmaeeli,R., Bauzá,A. and Perez,A. (2023) Structural predictions of protein-DNA binding: MELD-DNA.
  Nucleic Acids Res., **51** (4), 1625–1636.

- Evans,R., O'Neill,M., Pritzel,A., Antropova,N., Senior,A., Green,T., Žídek,A., Bates,R., Blackwell,S., Yim,J., Ronneberger,O., Bodenstein,S., Zielinski,M., Bridgland,A., Potapenko,A., Cowie,A., Tunyasuvunakool,K., Jain,R., Clancy,E., Kohli,P., Jumper,J. and Hassabis,D. (2022) Protein complex prediction with AlphaFold-Multimer.
  bioRxiv, .

References IV

- Gao,M., Nakajima An,D., Parks,J.M. and Skolnick,J. (2022) AF2Complex predicts direct physical interactions in multimeric proteins with deep learning.
  Nat. Commun., **13** (1), 1744–1757.

- Goddard,T.D., Huang,C.C., Meng,E.C., Pettersen,E.F., Couch,G.S., Morris,J.H. and Ferrin,T.E. (2018) UCSF ChimeraX: Meeting modern challenges in visualization and analysis.
  Protein Sci., **27** (1), 14–25.

References V

- Jumper,J., Evans,R., Pritzel,A., Green,T., Figurnov,M., Ronneberger,O., Tunyasuvunakool,K., Bates,R., Žídek,A., Potapenko,A., Bridgland,A., Meyer,C., Kohl,S.A.A., Ballard,A.J., Cowie,A., Romera-Paredes,B., Nikolov,S., Jain,R., Adler,J., Back,T., Petersen,S., Reiman,D., Clancy,E., Zielinski,M., Steinegger,M., Pacholska,M., Berghammer,T., Bodenstein,S., Silver,D., Vinyals,O., Senior,A.W., Kavukcuoglu,K., Kohli,P. and Hassabis,D. (2021) Highly accurate protein structure prediction with AlphaFold.
  Nature, **596**, 583–589.

- Kendrew,J.C., Dickerson,R.E., Strandberg,B.E., Hart,R.G., Davies,D.R., Phillips,D.C. and Shore,V.C. (1960) Structure of myoglobin: A three-dimensional Fourier synthesis at 2 A. resolution.
  Nature, **185** (4711), 422–427.

References VI

- LeCun,Y., Bengio,Y. and Hinton,G. (2015) Deep learning.
  Nature, **521** (7553), 436–444.

- Levinthal,C. (1969) How to fold graciously.
  In *Mossbauer Spectroscopy in Biological Systems, Proceedings of a Meeting held at Allerton House, Monticello, IL*, (Debrunner,P., Tsibris,J.C.M. and M unck,E., eds),. University of Illinois Press Urbana, IL pp. 22–24.

- Levitt,M. (1992) Accurate modeling of protein conformation by automatic segment matching.
  J. Mol. Biol., **226** (2), 507–533.

References VII

- Lin,Z., Akin,H., Rao,R., Hie,B., Zhu,Z., Lu,W., Smetanin,N., Verkuil,R., Kabeli,O., Shmueli,Y., Dos Santos Costa,A., Fazel-Zarandi,M., Sercu,T., Candido,S. and Rives,A. (2023) Evolutionary-scale prediction of atomic-level protein structure with a language model.
  Science, **379** (6637), 1123–1130.

- Mariani,V., Biasini,M., Barbato,A. and Schwede,T. (2013) IDDT: a local superposition-free score for comparing protein structures and models using distance difference tests.
  Bioinformatics, **29** (21), 2722–2728.

- Marks,D.S., Colwell,L.J., Sheridan,R., Hopf,T.A., Pagnani,A., Zecchina,R. and Sander,C. (2011) Protein 3D structure computed from evolutionary sequence variation.
  PLoS one, **6** (12), e28766.

References VIII

Meier,A. and Söding,J. (2015) Automatic Prediction of Protein 3D Structures by Probabilistic Multi-template Homology Modeling.
PLoS Comput. Biol., **11** (10), e1004343.

Mirdita,M., Schütze,K., Moriwaki,Y., Heo,L., Ovchinnikov,S. and Steinegger,M. (2022) ColabFold: making protein folding accessible to all.
Nat. Methods, **19** (6), 679–682.

Morcos,F., Pagnani,A., Lunt,B., Bertolino,A., Marks,D.S., Sander,C., Zecchina,R., Onuchic,J.N., Hwa,T. and Weigt,M. (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families.
Proc. Natl. Acad. Sci. USA, **108** (49), E1293–301.

References IX

- Pettersen,E.F., Goddard,T.D., Huang,C.C., Meng,E.C., Couch,G.S., Croll,T.I., Morris,J.H. and Ferrin,T.E. (2021) UCSF ChimeraX: Structure visualization for researchers, educators, and developers.
  Protein Sci., **30** (1), 70–82.

- Sali,A. and Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints.
  J. Mol. Biol., **234** (3), 779–815.

- Shaw,D.E., Maragakis,P., Lindorff-Larsen,K., Piana,S., Dror,R.O., Eastwood,M.P., Bank,J.A., Jumper,J.M., Salmon,J.K., Shan,Y. and Wriggers,W. (2010) Atomic-level characterization of the structural dynamics of proteins.
  Science, **330** (6002), 341–346.

References X

- Shindyalov,I.N., Kolchanov,N.A. and Sander,C. (1994) Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations?
  Protein Eng., **7** (3), 349–358.

- Varadi,M., Anyango,S., Deshpande,M., Nair,S., Natassia,C., Yordanova,G., Yuan,D., Stroe,O., Wood,G., Laydon,A., Žídek,A., Green,T., Tunyasuvunakool,K., Petersen,S., Jumper,J., Clancy,E., Green,R., Vora,A., Lutfi,M., Figurnov,M., Cowie,A., Hobbs,N., Kohli,P., Kleywegt,G., Birney,E., Hassabis,D. and Velankar,S. (2022) AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models.
  Nucleic Acids Res., **50** (D1), D439–D444.

References XI

- Wolynes,P.G. (2015) Evolution, energy landscapes and the paradoxes of protein folding.
  Biochimie, **119**, 218–230.

- Yin,R., Feng,B.Y., Varshney,A. and Pierce,B.G. (2022) Benchmarking AlphaFold for protein complex modeling reveals accuracy determinants.
  Protein Sci., **31** (8), e4379.

- Yu,D., Chojnowski,G., Rosenthal,M. and Kosinski,J. (2023) AlphaPulldown-a python package for protein-protein interaction screens using AlphaFold-Multimer.
  Bioinformatics, **39** (1).

- Zhang,Y. and Skolnick,J. (2004) Scoring function for automated assessment of protein structure template quality.
  Proteins: Struct. Funct. Genet., **57** (4), 702–710.

References XII

Zwanzig,R., Szabo,A. and Bagchi,B. (1992) Levinthal's paradox.
Proc. Natl. Acad. Sci. USA, **89** (1), 20–22.