Integrating Deep Learning and Statistical Physics for Template Agnostic Novel Protein Structure Prediction

**Abstract**

The protein structure prediction problem is a longstanding challenge within the scientiﬁc community with the potential to revolutionize the healthcare and drug development industries. Structural knowledge of target proteins allows scientists to evaluate how a particular protein may interact with drug candidates and other molecules. Various methods have been employed to generate probable native structures given available amino acid sequences with varying results. Here, we present an approach that employs deep learning methodology alongside statistical physics with the aim to further protein structure predictive capabilities, focused on membrane protein. Sequence-based deep learning algorithms are powerful and accurate predictions for localized structural attributes, such as bond angles and atomic distances, while statistical physics allows for large-scale conformation generation and subsequent samplings, evaluations and optimizations that result in viable, structurally sound predicted native protein folds. To maintain consideration for biological processes, the sequential procedures utilized in this approach were deliberately designed to resemble processes under protein folding theory. Results indicate that initial deep learning predictive processes followed by the exploration of a broad conformational space may be utilized to generate viable native protein structures for scientiﬁc research applications.

**Keyword:**

**Introduction**

Proteins are complex molecules which are an essential part of all living organisms. They originate from the synthesis of sequences of amino acids which then interact with one another and the surrounding environment to ultimately transition into their native, lowest energy structural conﬁgurations. The order of amino acids is vitally important for the protein to function properly in its biological role. Proteins fold into stable three-dimensional shapes, or conformations, that are determined by their amino acid sequence. The function of protein depends on the protein fold or conformation (Schmid *et al.* 2020). Protein misfolding and loss of function leads to several lethal diseases (Chaudhuri et al. 2006). Therefore, protein structure play an important role in structure-based drug discovery (Nero *et al.* 2018). In the case of Covid-19, for example, the spike proteins on the surface of the virus are able to bind to and subsequently hijack a human’s healthy cells, thereby resulting in illness. Knowledge of protein sequences and their structures allows for scientists to explore ways to prevent or interrupt this kind of disease progression ( Huang *et al.* 2020).

Protein structures are determined by X-ray crystallography, nuclear magnetic resonance spectroscopy, or more recently cryo-electron microscopy experiment. As of April 2018, there are close to 140,000 protein structures in the database of biological macromolecular structures ( [http://www.rcsb.org](http://www.rcsb.org/)). However, this number drops to about 50,000 if one keeps sequences with less than 90% sequence identity. This number is obviously small compared with the number of existing proteins. ere were more than 557,000 protein sequences deposited in SwissProt-Uniprot version 2018-03. From the past several decades, computational methods are trying to fill the gap (Delarue *et al.* 2018). Recently protein structure prediction has made significant progress (Laine et al. 2021), due to the increasing number of protein structures in the PDB, the explosion of genome sequencing, and rapid advance in deep learning. In the 14th Critical Assessment of Protein Structure Prediction (CASP14), protein structures predicted by AlphaFold2 can achieve accuracy at the atomic level, based on a new model of neural networks (Jumper *et al.* 2021). However, most proteins have only been predicted in a single conformation. Besides, AlphaFold2 is at a loss when confronted by disordered regions (Porta-Pardo et al. 2022). This observation stimulated the question of whether AlphaFold2 explores membrane protein's large conformational landscape and predicts multiple conformations. For example, in the case of Sec61 translocon complex (membrane protein), the plug-domain of the Sec61 channel is displaced from the channel pore in the open-state conformation; however, the plug occupies the pore in the closed state (Bhadra *et al*. 2021, Lang *et al.* 2017). It has been noticed that AlphaFold2 provide the closed-state conformation of the plug region. Recently, Del Alamo et al. demonstrated that modification of the AlphaFold2 algorithm is needed to identify the multiple conformations of GPCR (Del Alamo et al. 2022). Furthermore, protein structures shift and slide in the presence of small-molecule ligands, sometimes subtly and sometimes dramatically, but AlphaFold2 is not yet equipped to predict these changes. Recently, Computer simulations provide an increasingly realistic picture of large-scale conformational change of proteins; however these methods are computationally expensive (Kayanak *et al.* 2022, Kandathil *et al.* 2019).

Here, we propose a system of predictive processes that integrates deep learning methodologies with statistical physics to provide an elegant solution to the short comings of relation-based protein structure prediction. Deliberate deep learning algorithms may be relied upon for general predictions on protein structure and amino acid arrangements while statistical processes may be utilized to evaluate immense conformational space and select highly probable conﬁgurations for ﬁne-tuning. Furthermore, data engineering and selection measures may be utilized to carry out prediction processes considering only the most reliable and relevant features available.

Our study focuses on Vitamin K epoxide reductase (VKOR). It is an endoplasmic reticulum membrane enzyme protein that sustains blood coagulation through the vitamin K cycle. VKOR is the target of the most widely prescribed oral anticoagulant, Warfarin. Warfarin, a well-known Vitamin K antagonist (VKA), is among the most commonly used drugs worldwide. VKAs are oral anticoagulants used to treat and prevent thromboembolic diseases, including myocardial infarction and stroke, the two leading causes of human death and disability. Like other membrane proteins, structure-and-function studies of VKOR have been challenging, and the results of these studies are often controversial. The human VKOR (HsVKOR) was previously proposed to contain three TMs, a controversial model with both supporting and opposing biochemical evidence (Tie et al. 2012, Cao et al. 2016). However, recently published crystal structures of HsVKOR clearly show that it is a four-TM protein in a bound state (with Warfarin (Liu et al. 2021). Active site of HsVKOR is surrounded by a four-TM bundle and covered by a cap domain, which is made of a short helix (cap helix) followed by a loop (cap loop). Besides Shen et al. found that mutation of HsVKOR increases the size and changes the substrate-binding pocket's shape (Shen et al.). However, the conformation of the HsVKOR at other states and the cap loop conformational dynamics are not clear yet. Here, we discuss the conformational dynamics of HsVKOR using our structure prediction algorithm.

**Method**

**Overview of the structure prediction pipeline**

The protein structure prediction pipeline comprises five modules, namely, (1) baseline structure prediction, (2) random omega generation, (3) protein folding simulation, (4) structure refinement using MD simulation and (5) distance-based corroboration (see Figure 1). In the first module (baseline structure prediction), an unknown target protein's protein backbone geometry has been predicted from the amino acid sequence using the bidirectional LSTM (long short-term memory) model (Hochreiter *et al.* 1997). In addition, omega angle values were obtained through random generation based on a probability distribution from the whole population of protein structures in the second module (random omega generation). In the third step (protein folding simulation), folds with perceived minimal potential energy values with the highest structural integrity were selected using the angle-based NERF (Natural Extension Reference Frame) (ALQuraishi *et al.* 2019) algorithm and MUFold-CL (Zhang *et al.* 2013). Then the highest-ranking folds were selected for structural optimization through Molecular Dynamics (MD) simulation in the fourth model (structure refinement). Finally, in the last module (distance-based corroboration), the predicted inter-residue interactions, which were calculated using DeepMetaPSICOV (Kandathil et al. 2019), were used to evaluate the final folded structure of the target protein. The pipeline is for the query protein with sequence similarity < 80% from the database of know protein structure; otherwise, the template-based modelling algorithm HHpred (Hildebrand *et al* 2009) was used to predict the protein structure.

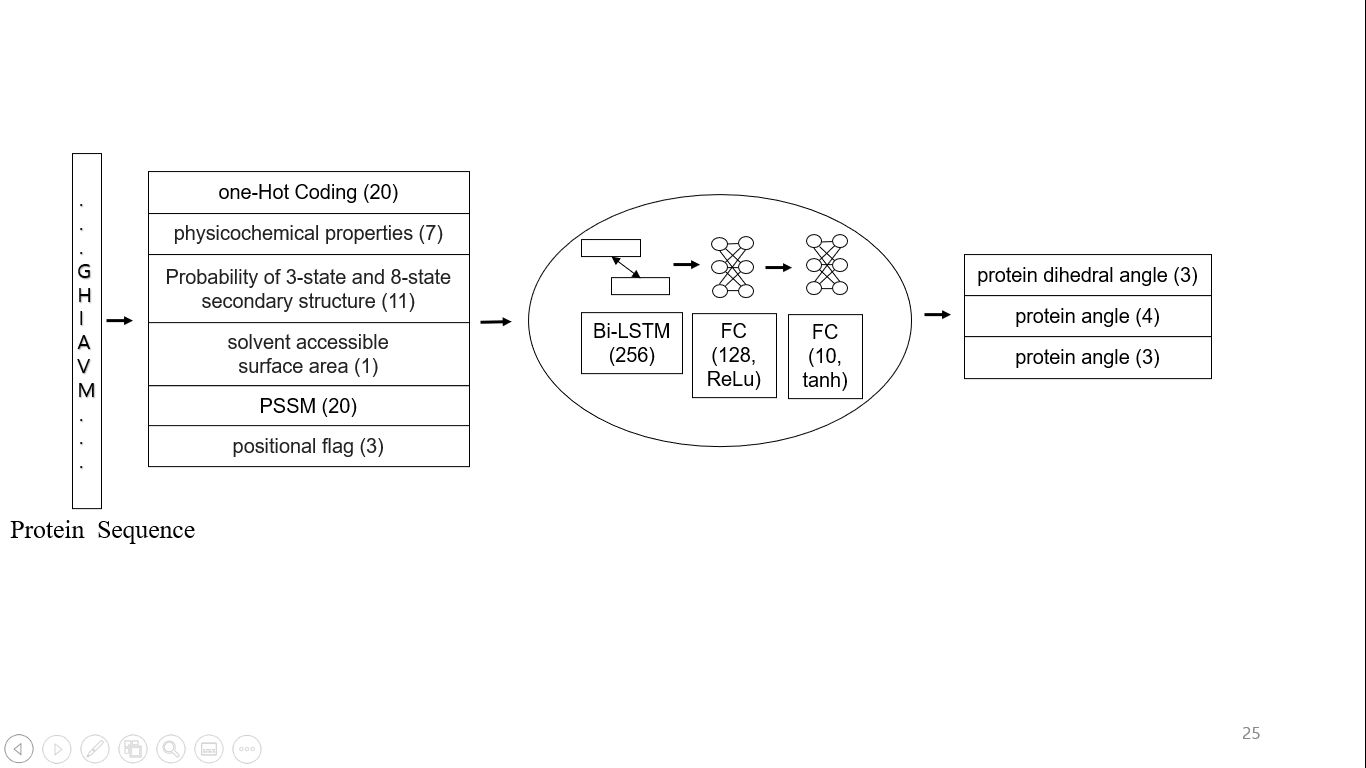
Diagram

Description automatically generated

**Figure 1:** Workflow for the protein structure prediction pipeline given a single query sequence

**Baseline structure prediction**

Prediction of protein backbone geometry is the initial step of prediction of the 3D structure of a protein from the given sequence. In the baseline structure prediction model, we predicted protein backbone dihedral angles, angles and distance using bidirectional LSTM architecture (Figure 2) from a given protein sequence. The baseline model was trained on a nonredundant dataset from PDB consisting of 15000+ proteins with high resolution and R-value constraints.

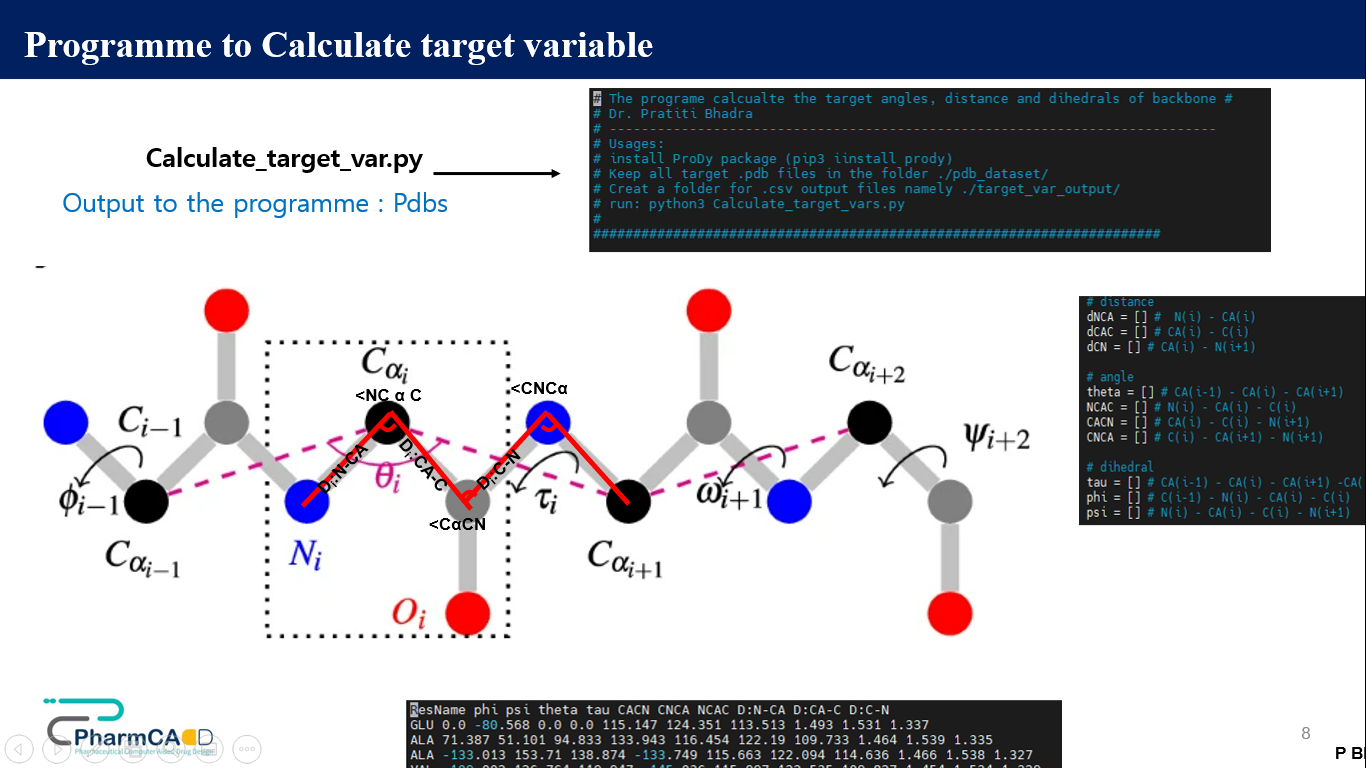


**Figure 2**: The baseline structure prediction

The features of each protein are extracted from its sequence and corresponding multiple sequence alignment (MSA). The basic features include one hot encoding of amino acids (20-dimensional) (Lin *et al.* 2002), seven physicochemical properties (PP) (Meiler et al. 2001), including steric parameter, polarizability, normalized van der Waals volume, hydrophobicity, isoelectric point, helix probability, and sheet probability (7-dimensional), solvent accessible surface area (1 dimenational), probability of 3-state and 8-state secondary structure from SPOT-1D (Hanson *et al.* 2019) (11-dimensional) and 20 dimensional position-specific substitution matrix (PSSM) generated by PSI-BLAST (Altschul et al. 1997). Besides we used a 3-bit positional flag to identify the terminal of the protein sequence. The total dimension of the feature vector is 62.

The one-hot encoding is used to represent an amino acid as a vector of bits. A 20-bit binary vector represents each amino acid. It produces a 20-column vector for each position that only contains a 1 where the letter corresponds to that amino acid, while other bits are 0 (Lin *et al.* 2002). As example, alanine (A) and cysteine (C) are represented by {1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0} and {1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0}, respectively. The sequence order information of amino acids in protein sequence is preserved by this encoding method. The second set of features comes from seven physicochemical properties of amino acid and solvent accessible surface area. These properties of amino acids have been an essential factor for protein fold (Huang et al., 2013).   Protein secondary structure can be regarded as an information bridge that links the primary sequence and tertiary structure. The third set of features represents the probability of secondary structure of each residue of the protein sequence. Protein secondary structures are traditionally characterized as 3 general states: helix (H), strand (E) and coil (C). However, more precise structure-based properties are captured by finer characterization of the secondary structure, namely 8-states: 310 helix (G), α-helix (H), π-helix (I), β-stand (E), bridge (B), turn (T), bend (S), and others (C). The probability value of 3-state and 8-state secondary structures have been calculated using the SPOT-1D method (Hanson et al. 2019).  The last set of features comes from the protein profiles generated using PSI-BLAST (Altschul et al. 1997). In our experiments, the PSI-BLAST parameters were set to (evalue: 0.001; num\_iterations: 3; db: UniRef90) to generate PSSM. Each amino acid in the protein sequence is represented as a vector of 21 real values ranging from 0 to 1, representing the 20 amino acids' PSSM values plus a NoSeq label in the last. The NoSeq label does not include in the feature vector. The features of the PSI-BLAST correspond to the 20 amino acids: "A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V". The PSI-BLAST generates a PSSM or a position specific scoring matrix that represents evolutionary profiles derived from protein multiple sequence alignment.

In our baseline model, we considered three outputs for protein dihedral angles, one for each phi (*ϕ*), psi (*ψ*) and tau (*τ*) angles. Each ϕ and ψ can be associated with exactly one residue or Cα. Psi (*ψ*) is the *N(i)-Cα(i)-C(i)-N(i+1)* dihedral angle for residue *i* and Phi (*ϕ*) is the *C(i-1)-N(i)-Cα(i)-N(i+1)*. Dihedral angle for residue *i.* (see Figure 3). The ranges of phi-psi angle are -180 to 180 degrees. Tau (τ) dihedral angle involving Cα(i-1), Cα(i), Cα(i+1), Cα(i+2) is associated with Cα(i). Besides, we include four angles, namely theta (θ), NCαC, CαCN and CN Cα. Like tau (τ). theta (θ) angle involving Cα(i-1), Cα(i), Cα(i+1) is associated with Cα(i). Other three angles are angles for residue *i.* NCαC, CαCN and CNCα represent the angle between *N(i)-Cα(i)-C(i)*, *Cα(i)-C(i)-N(i+1)* and *C(i-1)-N(i)- Cα(i),* respectively, for residue *i*. Furthermore, another three-distance metrics for a residue also predicted by the baseline model. These distance metrics are the distance between *N(i)-Cα(i), Cα(i)-C(i)* and *C(i)-N(i+1)*.



**Figure 3:** Protein backbone angles: It illustrate protein dihedral (phi, psi and tau) angle, protein backbone angle (<NCαC, <CNCα and <CαCN).

**Result**

**Table 1:** The performance of the baseline model from 5 fold cross validation (mean square error; MAE). Measurement of angles and dihedral angles are in degree. The measurement of distance is in Å.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Phi | Psi | Tau | Theta | CαCN | CNCα | NCαC | CN | NCα | CαC |
| 17.7 | 21.1 | 6.2 | 25.7 | 0.8 | 1.45 | 2.06 | 0.008 | 0.008 | 0.011 |

**Reference**

AlQuraishi, M., 2019. Parallelized natural extension reference frame: parallelized conversion from internal to Cartesian coordinates. *Journal of computational chemistry*, *40*(7), pp.885-892.

Altschul S et al., “Gapped BLAST and PSI-BLAST: a new generation of protein database search programs,” Nucleic Acids Research, vol. 25, no. 17, pp. 3389–3402, January 1997.

Bhadra, P. and Helms, V., 2021. Molecular modeling of signal peptide recognition by eukaryotic Sec complexes. *International Journal of Molecular Sciences*, *22*(19), p.10705.

Cao Z, van Lith M, Mitchell LJ, Pringle MA, Inaba K, Bulleid NJ. The membrane topology of vitamin K epoxide reductase is conserved between human isoforms and the bacterial enzyme. *Biochem J*. 2016;**473**:851–8.

Chaudhuri, T.K. and Paul, S., 2006. Protein‐misfolding diseases and chaperone‐based therapeutic approaches. *The FEBS journal*, *273*(7), pp.1331-1349.

Del Alamo, D., Sala, D., Mchaourab, H.S. and Meiler, J., 2022. Sampling alternative conformational states of transporters and receptors with AlphaFold2. *Elife*, *11*, p.e75751.

Delarue, M. and Koehl, P., 2018. Combined approaches from physics, statistics, and computer science for ab initio protein structure prediction: ex unitate vires (unity is strength)?. *F1000Research*, *7*.

Hanson, J., Paliwal, K., Litfin, T., Yang, Y. and Zhou, Y., 2019. Improving prediction of protein secondary structure, backbone angles, solvent accessibility and contact numbers by using predicted contact maps and an ensemble of recurrent and residual convolutional neural networks. *Bioinformatics*, *35*(14), pp.2403-2410.

Hildebrand, A., Remmert, M., Biegert, A. and Söding, J., 2009. Fast and accurate automatic structure prediction with HHpred. *Proteins: Structure, Function, and Bioinformatics*, *77*(S9), pp.128-132.

Hochreiter, S. and Schmidhuber, J., 1997. Long short-term memory. *Neural computation*, *9*(8), pp.1735-1780.

Huang, Y.F. and Chen, S.Y., 2013. Extracting physicochemical features to predict protein secondary structure. *The Scientific World Journal*, *2013*.

Huang, Y., Yang, C., Xu, X.F., Xu, W. and Liu, S.W., 2020. Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19. *Acta Pharmacologica Sinica*, *41*(9), pp.1141-1149.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A. and Bridgland, A., 2021. Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), pp.583-589.

Kandathil, S.M., Greener, J.G. and Jones, D.T., 2019. Prediction of interresidue contacts with DeepMetaPSICOV in CASP13. *Proteins: Structure, Function, and Bioinformatics*, *87*(12), pp.1092-1099.

Kaynak, B.T., Krieger, J.M., Dudas, B., Dahmani, Z.L., Costa, M.G., Balog, E., Scott, A.L., Doruker, P., Perahia, D. and Bahar, I., 2022. Sampling of Protein Conformational Space Using Hybrid Simulations: A Critical Assessment of Recent Methods. *Frontiers in molecular biosciences*, *9*.

Laine, E., Eismann, S., Elofsson, A. and Grudinin, S., 2021. Protein sequence‐to‐structure learning: Is this the end (‐to‐end revolution)?. *Proteins: Structure, Function, and Bioinformatics*, *89*(12), pp.1770-1786.

Lang, S., Pfeffer, S., Lee, P.H., Cavalié, A., Helms, V., Förster, F. and Zimmermann, R., 2017. An update on Sec61 channel functions, mechanisms, and related diseases. *Frontiers in Physiology*, p.887.

Lin, K., May, A.C. and Taylor, W.R., 2002. Amino acid encoding schemes from protein structure alignments: Multi-dimensional vectors to describe residue types. *Journal of theoretical biology*, *216*(3), pp.361-365.

Liu, S., Li, S., Shen, G., Sukumar, N., Krezel, A.M. and Li, W., 2021. Structural basis of antagonizing the vitamin K catalytic cycle for anticoagulation. *Science*, *371*(6524), p.eabc5667.

Meiler, J., Müller, M., Zeidler, A. & Schmäschke, F. Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks. *Mol. Model. Annu.* **7**, 360–369 (2001)

Nero, T.L., Parker, M.W. and Morton, C.J., 2018. Protein structure and computational drug discovery. *Biochemical Society Transactions*, *46*(5), pp.1367-1379.

Porta-Pardo, E., Ruiz-Serra, V., Valentini, S. and Valencia, A., 2022. The structural coverage of the human proteome before and after AlphaFold. *PLoS computational biology*, *18*(1), p.e1009818.

Schmid, S. and Hugel, T., 2020. Controlling protein function by fine-tuning conformational flexibility. *Elife*, *9*, p.e57180.

Shen, G., Li, C., Cao, Q., Megta, A.K., Li, S., Gao, M., Liu, H., Shen, Y., Chen, Y., Yu, H. and Li, S., 2022. Structural features determining the vitamin K epoxide reduction activity in the VKOR family of membrane oxidoreductases. *The FEBS Journal*.

Tie J-K, Jin D-Y, Stafford DW. Human vitamin k epoxide reductase and its bacterial homologue have different membrane topologies and reaction mechanisms. *J Biol Chem*. 2012;**287**:33945–55.

Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Žídek, A., Bridgland, A., Cowie, A., Meyer, C., Laydon, A. and Velankar, S., 2021. Highly accurate protein structure prediction for the human proteome. *Nature*, *596*(7873), pp.590-596.

Whittle, P.J. and Blundell, T.L., 1994. Protein structure-based drug design. *Annual review of biophysics and biomolecular structure*, *23*(1), pp.349-375.

Zhang, J. and Xu, D., 2013. Fast algorithm for population‐based protein structural model analysis. *Proteomics*, *13*(2), pp.221-229.