TopSuite:
Collection of Deep Learning-Based Metamethods to Predict and Evaluate Protein Structures and Properties

Filip Koenig[1]**, Daniel Mulnaes**[1]**, Karel van der Weg**[2]**, and Holger Gohlke**[1,2]

[1] Institute of Pharmaceutical and Medicinal Chemistry, Heinrich Heine University Düsseldorf,
40225 Düsseldorf, Germany
E-mail: {filip.koenig, daniel.mulnaes, gohlke}@hhu.de

[2] John von Neumann Institute for Computing (NIC), Jülich Supercomputing Centre (JSC),
Institute of Biological Information Processing (IBI-7: Structural Biochemistry),
and Institute of Bio- and Geosciences (IBG-4: Bioinformatics),
Forschungszentrum Jülich, 52425 Jülich, Germany
E-mail: {k.van.der.weg, h.gohlke}@fz-juelich.de

To understand the function of proteins, knowledge of various properties is necessary. While for many properties experimental determination protocols are available, these may be time-consuming and laborious. An alternative is given by computational predictions, building on recent advances in machine learning technology. Here, we present TopSuite, a collection of deep learning-based predictors. TopSuite contains programs for protein model quality assessment, template-based protein structure prediction, domain boundary prediction, secondary structure and membrane topology prediction, and protein-protein interface contact prediction.

1 Introduction

To develop new pharmaceutical and biotechnological products, it is of utmost importance to understand how proteins as targets or enzymes work. Various experimental procedures are available to determine different properties of proteins, with their three-dimensional structure being the most crucial property to gaining insight into protein function. While the experimental determination of protein structure allows for atomic resolution information, it is expensive, time-consuming, and comes without a guarantee of success in all cases[1]. An alternative to experimental determination is computational protein structure prediction. Having been a major scientific problem in computational biology for decades[2], recent advances in deep learning technology led to the development of tools like AlphaFold[3] and RoseTTAfold[4], that show drastic improvements in the accuracy of predicted protein structures, providing structural models of even complex proteins that can match experimental structures within experimental uncertainty.

Apart from pure structure prediction, a variety of additional properties are relevant for understanding the biological context of proteins, e.g., predicting the borders of protein domains or the interaction and orientation between proteins and membranes, i.e., membrane topology. Similar to structure prediction, these tasks have benefited from deep learning techniques[5,6].

A collection of deep learning-based metamethods to predict and evaluate protein structures and properties has been developed over the past seven years in our group, bundled under the name TopSuite. TopSuite contains software for protein model quality assessment (TopScore), template-based protein structure prediction (TopModel), domain boundary prediction (TopDomain), secondary structure and membrane topology prediction (TopProperty), and, under development, protein-protein interface contact prediction (TopInterface)[7–10]. To make the methods available to the public, web interfaces to all programs were developed[11] [a].

2 Methods

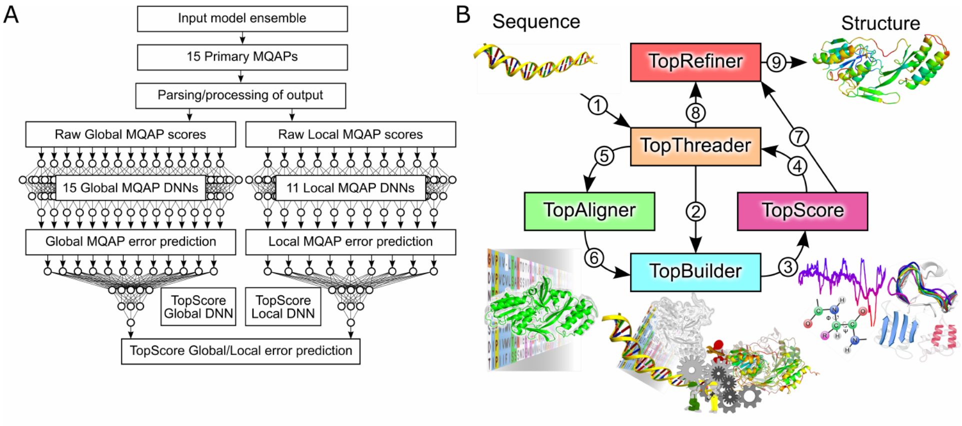

Figure 1. (A) The architecture of TopScore to assess protein model quality. 15 primary predictors are run on the input model ensemble. MQAP stands for Model Quality Assessment Program. The output is processed and fed to individual DNNs for each primary predictor, one set for global and one set for local scores. The prediction output is then used as input for the final DNNs, which predict a global or residue-wise TopScore value. (B) Workflow of TopModel for protein structure prediction. The user provides an input sequence to TopThreader (1). The resulting threader alignments are used by TopBuilder to build models (2), which are subsequently scored with TopScore (3). TopThreader then clusters and ranks the templates to remove false positives (4). TopAligner is used to build ensembles of aligned templates and subsequently consensus alignments (5), which are fed to TopBuilder (6). The models are again scored with TopScore (3). Subsequently, TopThreader takes these models along with primary threader scores and creates a ranking for templates according to the predicted similarity to the native structure (4). After the selection of the templates, TopAligner is used again to build an ensemble of multitemplate and pairwise alignments (5), which TopBuilder builds models from (3). The resulting models are scored with TopScore (4). Models from the multi-template ensemble (7) and single-template models (8) are chosen and used by TopRefiner to create a final model by combination and refinement of good structural parts (9). Figures were taken from Refs. 7 and 8.

All methods in TopSuite are metamethods and have a similar basic structure. In each case, a varying number of primary predictors is run on the user input to produce the raw data for the prediction. These predictions are normalised and set up as input features, potentially combined with other features extracted directly from the input data. One or multiple consecutive stages of deep neural networks (DNN) are then used to predict the respective properties. This kind of architecture is illustrated in Fig. 1A with the DNNs of TopScore as an example. Using multiple DNNs connected in series has the advantage

[a] https://cpclab.uni-duesseldorf.de/topsuite/

that one can present features in different forms to separate stages, allowing the network to learn specific patterns instead of overwhelming it by presenting all available information at once. The final output of the last-stage neural network is processed and, in case our web interfaces are used, is made available as a results webpage.

TopModel has a series of additional workflow modules, as it is not only predicting or classifying properties but builds complete structural models. A scheme of the architecture of TopModel is shown in Fig. 1B. The workflow is split into five separate modules: TopThreader, TopAligner, TopBuilder, TopScore, and TopRefiner. TopThreader uses twelve threading programs in combination with DNNs in a top-down consensus approach to select optimal templates. TopAligner uses eight alignment programs to build an ensemble of pairwise template-template alignments to ensure similar quality and fold of templates for the model building. TopBuilder combines output from Modeller9[12] and Rosetta[13] to build structural models based on alignments with template structures. TopScore is used at various points in the workflow to score models. TopRefiner selects and combines different models to replace regions with low TopScore values.

All programs of TopSuite are programmed with the Python programming language. TopScore and TopModel use a multilayer perceptron (MLP) regression model from SciKit-learn, while TopDomain, TopProperty, and TopInterface use Tensorflow and Keras in combination with ResNet[14–17]. For TopDomain, TopProperty, and TopInterface, it is relevant to encode the local protein neighbourhood to let the DNN capture the local context. Therefore, a sliding window approach is used in which residues surrounding the target residue are represented as the peripheral pixels in an image. This is realised by using convolutional neural networks (CNN), which are well established in the field of image recognition.

3 Results

3.1 TopScore

TopScore predicts a derivative of the lDDT score[18], a superposition-independent target score, called the “lDDT error”, which is defined as 1-lDDT. It provides a global score for the model as a whole, as well as a local score for every residue of a model. To provide an evaluation scheme that is free of clustering information, a second metric besides TopScore is computed, TopScoreSingle. This is done because clustering-free methods are better in selecting the best model, especially if the model ensemble is heterogeneous[7]. In comparison, AlphaFold provides an intrinsic model accuracy estimate called pLDDT[3], calculated directly from the final stage of the single representation. While it also uses an MLP to calculate the score, its input features cannot be interpreted as obviously as the input values to TopScore, making a direct comparison difficult.

TopScore was evaluated with different metrics to benchmark its performance in four distinct tasks: The ability to separate good from wrong models, the accuracy of the predicted global lDDT error, the accuracy in ranking model ensembles, and how good it is at finding the best model. Overall, TopScore and TopScoreSingle significantly outperform all primary predictors in all categories. When calculating correlation scores with Pearson’s $R_{\mathrm{All}}{}^2$, TopScore achieves a value of 0.93 for global and 0.78 for local error prediction. Fig. 2E illustrates the connection between global TopScore and the true lDDT error on a test system, indicating that the error prediction is more precise on very good and very wrong models.

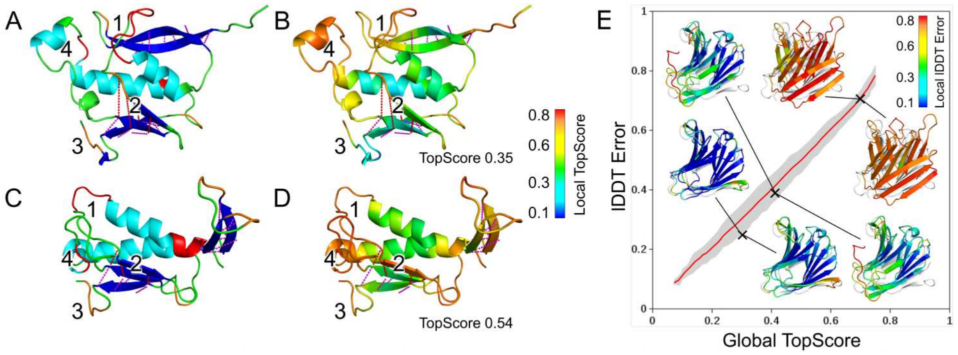

Figure 2. A-D Evaluation of the predicted structure of lipoprotein LipoP from *C. difficile* with experimental results from NMR. (A) Model predicted by TopModel. The model is coloured according to the agreement with NMR results. Blue: Residues of β-strands that agree with the NMR data. Orange: Residues that were found to be in β-strand conformation according to the NMR data but not in the TopModel prediction. Cyan: Residues of α-helices that agree with the NMR data. Red: Residues that were found to be in α-helical conformation according to the NMR data but not in the TopModel prediction. Magenta lines: NOE restraints of β-sheets that agree with the model. Red lines: NOE restraints of β-sheets that disagree with the model. (B) The TopModel model is coloured according to residue-wise TopScore. (C) Model structure predicted by the program dPPAS2 from the LOMETS server. The colour scheme is the same as in panel (A). (D) The model from dPPAS2 is coloured according to residue-wise TopScore. (E) Performance of TopScore. From the 3DRobot dataset[20], three random models from PDB entry 4BMB were selected and coloured according to TopScore (lower triangle) and lDDT error (upper triangle). Figures were taken from Refs. 7 and 8.

3.2 TopModel

TopModel uses five submodules with numerous deep neural networks to build accurate alignments that are used in homology modelling of protein structures. TopModel depends on identifying high-quality templates; therefore, the module TopThreader uses a multi-step DNN-assisted selection procedure to find optimal templates. AlphaFold, on the other hand, can use template information, but such information plays only a minor role in its performance[3]. By contrast, it mainly exploits information extracted from deep sequence alignments. We view the two methods as complementary to each other, with TopModel being a good option for providing high-confidence structural predictions if high-quality templates are available and AlphaFold being a valuable tool in the case no templates are available.

To validate the predictive power of TopModel, nuclear magnetic resonance (NMR) experiments were performed on the lipoprotein LipoP from *C. difficile*. The model from TopModel and the model from the best primary predictor, the dPPSA2 program of LOMETS[19], are compared with experimental results from NMR (Fig. 2A-D). The model from TopModel shows good agreement with the experimental assignments from NMR, with Matthews correlation coefficients (MCC) of 0.81 for β-strands, 0.68 for α-helices, and 0.66 for coils. The good model quality is also corroborated by a TopScore value of 0.35.

Additionally, the performance of TopModel and AlphaFold were compared in terms of TopScore values computed for models generated by either method (Fig. 3). AlphaFold predicts models with a better TopScore than TopModel, in the case of good models, whereas TopModel produces models with a better TopScore in the high-TopScore region.

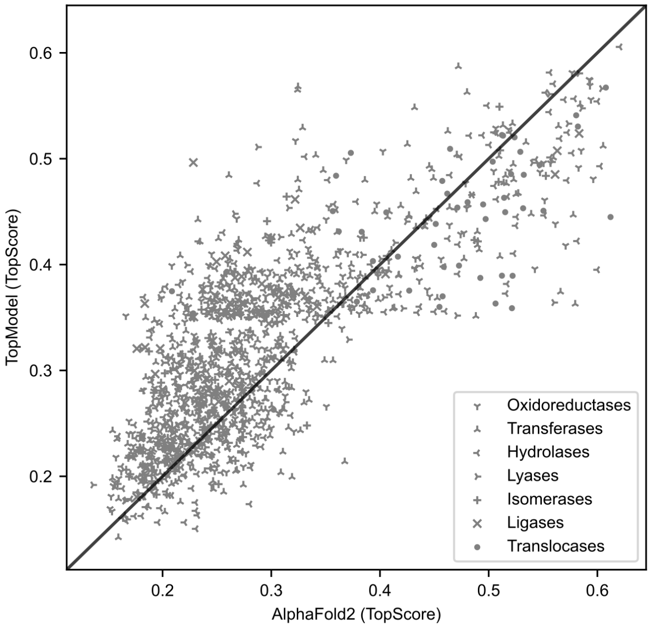

Figure 3. Comparison of TopModel and AlphaFold predictions scored by TopScore each. Lower TopScore values indicate better structural models.

3.3 TopProperty

TopProperty uses eleven primary predictors of secondary structure (SS) and solvent accessibility (SA) and sixteen primary predictors of transmembrane topology (TMT) and membrane exposure (ME) to produce features that are fed to two separate 1D CNNs with four respectively five differently sized sliding windows to perform a meta-prediction on these properties. While it is possible to use predicted structures from AlphaFold to deduce the TMT or ME directly, TopProperty can predict these properties from sequence alone and, therefore, is independent of structure availability.

TopProperty is benchmarked on the NOUMENON dataset[21] for evaluating the performance of SS and SA predictions and on the OPM dataset[22], consisting of transmembrane proteins, for evaluating the performance of SS and SA as well as TMT and ME predictions. TopProperty shows overall superior performance for predicting SS and SA compared to the respective primary predictors when the Q3 metric was used, both for the NOUMENON and the OPM dataset. The Q3 score is the percentage of residues with a correctly predicted 3-state secondary structure[23]. For TMT, TopProperty was benchmarked against methods that specifically predict either transmembrane α-helical bundles (TMH) or transmembrane β-barrels (TMB). TopProperty significantly outperforms all other primary predictors on the Q3 score. For ME, no competitor method provides any predictions for TMHs, therefore, no comparison to another method was possible. For TMBs, only one competitor method provides predictions but outperformed TopProperty on the MCC metric.

3.4 TopDomain

TopDomain applies 53 primary predictors of domain boundaries and feeds the output of these combined with homology- and sequence-based features into two stages of 1D CNNs with varying sliding window sizes. TopDomain contains workflows for sequence- (TopDomain, TopDomain$_{\text{Seq}}$) as well as structure-based (TopDomain$_{\text{Parse}}$) domain boundary predictions. Additionally, a binary predictor (TopDomain$_{\text{TMC}}$) predicts if domain parsing is necessary, e.g., to facilitate subsequent protein structure prediction. TopDomain is benchmarked on two datasets, the manually annotated TopDomain dataset and the CASP dataset based on annotations from the CASP organisers on CASP11-13 proteins. For both datasets, TopDomain, TopDomain$_{\text{Seq}}$, and TopDomain$_{\text{Parse}}$ significantly outperform their respective best primary predictor. TopDomain achieves an F1-score of 73.8 % for domain boundary prediction within ± 10 residues for the TopDomain dataset and a value of 42.8 % for the CASP dataset. Although, in general, tools such as AlphaFold and RoseTTAfold can predict structures for full-length sequences, often no fold can be predicted for parts of the structure. In these cases, it can be fruitful to split the respective sequence into separate domains and try to predict structures for the domains separately.

3.5 TopInterface

TopInterface uses four primary predictors of coevolutionary couplings together with structural, energetic, and sequence-based features as input to three stages of CNNs, leveraging 2D sliding windows to represent the local neighbourhood context. In comparison to AlphaFold, it uses additional ways of building paired multiple sequence alignments that are necessary for the calculation of coevolutionary couplings of protein pairs. Especially, TopInterface has a dedicated workflow to build paired alignments for protein pairs coming from two different species, allowing the modelling of inter-species protein-protein interactions (PPI). TopInterface provides a distance prediction together with an upper and

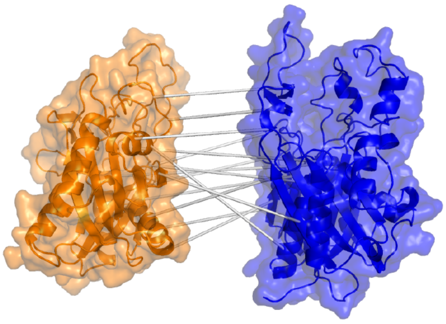

Figure 4. TopInterface distance restraints. TopInterface takes two protein structures as input, extracts various structural, energetic, and sequence-based features, and performs three consecutive stages of deep learning to predict distance restraints with upper and lower bounds for every inter-protein residue-residue pair. These restraints can be used in a docking protocol to create a structural model of the protein-protein complex. The distance restraints are illustrated as grey lines between proteins of PDB ID 5DWZ.

lower bound for every inter-protein residue-residue pair, which can be used in a distance-restrained docking, as presented in Fig. 4. It has workflows for homo- and heterodimers, as well as domain-domain interactions and inter-species PPIs.

4 Conclusion

TopSuite, a collection of metamethods for predicting protein structures and properties, was developed in our group. It consists of the programs TopScore, TopModel, TopDomain, TopProperty, and TopInterface. All programs utilise DNNs to perform their respective tasks, using carefully chosen sets of features and DNN architectures. As new protein structure prediction tools such as AlphaFold and RoseTTAfold have been developed, the accuracy of template-free structure predictions has reached new levels. While the TopSuite programs cannot provide predictions with similar quality without the use of templates, we view TopSuite as a complementary approach for properties that cannot be directly derived from predicted models from AlphaFold or RoseTTAfold, e.g., domain boundary predictions in the case of failing to build a multidomain protein directly from an end-to-end approach. Therefore, we expect TopSuite to be a valuable tool in providing such property predictions. In future developments, we seek to combine the power of sophisticated alignment pairing workflows in TopInterface with AlphaFold for building protein complexes of higher stochiometry.

Acknowledgements

This work was supported in part by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Projektnummer 267205415 - SFB 1208 (project A03 to HG), and by the Bundesministerium für Bildung und Forschung (BMBF) - Förderkennzeichen 031L0182 - InCelluloProtStruct to HG, and by the Helmholtz Association of German Research Centres (Helmholtz School for Data Science in Life, Earth and Energy (HDS-LEE) to HG). We are grateful for computational support and infrastructure provided by the "Zentrum für Informations- und Medientechnologie" (ZIM) at the Heinrich Heine University Düsseldorf and the computing time provided by the John von Neumann Institute for Computing (NIC) to HG on the supercomputer JUWELS at Jülich Supercomputing Centre (JSC) (user ID: HKF7, VSK33). Figs. 1 and 2 reprinted (adapted) with permission from Refs. 7 and 8; Copyright (2018, 2020), American Chemical Society.

References

1. H. Deng, Y. Jia, and Y. Zhang, *Protein structure prediction*, Int. J. Mod. Phys. B **32**, 18, 2018.
2. D. Petrey and B. Honig, *Protein structure prediction: inroads to biology*, Mol. Cell **20(6)**, 811-819, 2005.
3. J. Jumper et al., *Highly accurate protein structure prediction with AlphaFold*, Nature **596(7873)**, 583-589, 2021.
4. M. Baek et al., *Accurate prediction of protein structures and interactions using a three-track neural network*, Science **373(6557)**, 871-876, 2021.

5. Q. Shi et al., *DNN-Dom: predicting protein domain boundary from sequence alone by deep neural network*, Bioinformatics **35(24)**, 5128-5136, 2019.
6. Z. Liu, Y. Gong, Y. Bao, Y. Guo, H. Wang, and G.N. Lin, *TMPSS: A Deep Learning-Based Predictor for Secondary Structure and Topology Structure Prediction of Alpha-Helical Transmembrane Proteins*, Front. Bioeng. Biotechnol. **8**, 629937, 2020.
7. D. Mulnaes and H. Gohlke, *TopScore: Using Deep Neural Networks and Large Diverse Data Sets for Accurate Protein Model Quality Assessment*, J. Chem. Theory Comput. **14(11)**, 6117-6126, 2018.
8. D. Mulnaes, et al., *TopModel: template-based protein structure prediction at low sequence identity using top-down consensus and deep neural networks*, J. Chem. Theory Comput. **16(3)**, 1953-1967, 2020.
9. D. Mulnaes, P. Golchin, F. Koenig, and H. Gohlke, *TopDomain: Exhaustive Protein Domain Boundary Metaprediction Combining Multisource Information and Deep Learning*, J. Chem. Theory Comput. **17(7)**, 4599-4613, 2021.
10. D. Mulnaes, S. Schott-Verdugo, F. Koenig, and H. Gohlke, *TopProperty: Robust Metaprediction of Transmembrane and Globular Protein Features Using Deep Neural Networks*, J. Chem. Theory Comput. **17(11)**, 7281-7289, 2021.
11. D. Mulnaes, F. Koenig, and H. Gohlke, *TopSuite Web Server: A Meta-Suite for Deep-Learning-Based Protein Structure and Quality Prediction*, J. Chem. Inf. Model. **61(2)**, 548-553, 2021.
12. N. Eswar et al., *Comparative protein structure modeling using MODELLER*, Curr. Protoc. Bioinform. **15(1)**, 5-6, 2006.
13. C. A. Rohl et al., *Protein structure prediction using Rosetta*, Meth. Enzymol. **383**, 66-93, 2004.
14. F. Pedregosa et al., *Scikit-learn: Machine learning in Python*, J. Mach. Learn. Res. **12**, 2825-2830, 2011.
15. M. Abadi et al., *Tensorflow: A system for large-scale machine learning*, 12th USENIX Symposium on operating Systems Design and Implementation **OSDI 16**, 265-283, 2016.
16. F. Chollet, *Keras*, 2015.
17. K. He, X. Zhang, S. Ren, and J. Sun, *Deep residual learning for image recognition*, Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 770-778, 2016.
18. V. Mariani, M. Biasini, A Barbato, and T Schwede, *lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests*, Bioinformatics **29(21)**, 2722-2728, 2013.
19. S. Wu and Y. Zhang., *LOMETS: a local meta-threading-server for protein structure prediction*, Nucleic Acids Res. **35(10)**, 3375-3382, 2007.
20. H. Deng, Y. Jia, and Y. Zhang, *3DRobot: automated generation of diverse and well-packed protein structure decoys*, Bioinformatics **32(3)**, 378-387, 2016.
21. G. Orlando, D. Raimondi, and W. F. Vranken, *Observation selection bias in contact prediction and its implications for structural bioinformatics*, Sci. Rep. **6**, 36679, 2016.
22. M. A. Lomize et al., *OPM: orientations of proteins in membranes database*, Bioinformatics **22(5)**, 623-625, 2006.
23. M. Spencer, J. Eickholt, and C. Jianlin, *A Deep Learning Network Approach to ab initio Protein Secondary Structure Prediction*, IEEE/ACM Trans. Comput. Biol. Bioinform. **12(1)**, 103-112, 2014.