ROPIUS0: A deep learning-based protocol for protein structure prediction and model selection and its performance in CASP14

Mindaugas Margelevičius*

Institute of Biotechnology, Life Sciences Center, Vilnius University, Vilnius, 10257, Lithuania

Abstract A protocol ROPIUS0 for protein structure prediction and model selection is presented. At the core of the ROPIUS0 protocol is the deep learning module developed for the selection of protein structural models. It is shown that the direct use of predicted inter-residue distances may be sufficient to discriminate between correct and incorrect protein folds, considering only a small fraction of predicted distances. Having finished the latest CASP14 prediction season, a ROPIUS0 variant based on model selection ranks 13th in the category of tertiary structure prediction. Its performance is on par with top-performing automated prediction servers when tested on the CASP13 dataset. The results suggest ways to improve searching for structurally similar and homologous proteins without considerably increasing speed.

1 Introduction

The problem of protein three-dimensional (3D) structure prediction is fundamental to Bioinformatics, and structure prediction is particularly important in biomedical research. Knowledge of the structure and thermodynamics of the protein can lead to a much deeper understanding of biological processes. It would greatly facilitate vaccine and drug design (Maximova *et al.*, 2016).

The structure prediction problem can be addressed in various ways. The common classification distinguishes between two basic approaches (Abriata *et al.*, 2019). One is based on modeling by homology to solved structures (Croll *et al.*, 2019). The sequence of the protein of interest, the target, is searched for similarities in a database of sequences with known 3D structure or structures using homology search (Margelevičius, 2020a; Steinegger *et al.*, 2019a) or threading tools (Zhu *et al.*, 2018; Zheng *et al.*, 2019b). A significant similarity to a protein from the database suggests that the target protein can adopt a similar structure, and the identified protein can represent a good candidate for using it as a template to model the structure of the target protein. The accuracy of a structural model strongly depends on the extent of the template's structural similarity to the target protein. Not less important factor is the alignment accuracy between the template and the target. Inaccurate alignment deteriorates the quality of a structural model.

*Correspondence: mindaugas.margelevicius@bti.vu.lt

The other approach is de novo structure prediction, which predicts the structure without using structural templates. De novo structure prediction has improved dramatically over the past decade (Kryshtafovych *et al.*, 2019), with the most successful methods being those that use predicted distances or contacts between residue pairs as restraints (Senior *et al.*, 2020; Zheng *et al.*, 2019a; Hou *et al.*, 2019). Restraints are typically generated using artificial neural networks (NNs). A standard choice for the NN backbone architecture has been a residual network (ResNet) (He *et al.*, 2016) due to its good learning and performance characteristics and convolutional nature, which fits well with learning on variable-length proteins. Given the type of NN architecture, the set of input data to the NN, the NN depth expressed in hidden layers, training policy and the size of the training dataset are among the most influential factors that affect the accuracy of predictions.

Inseparable from protein modeling is model accuracy estimation or model selection. Its goal is to select from the set of models generated during the modeling process models structurally most similar to the target protein. More generally, given the target sequence and a set of structural models, a model selection algorithm aims to select the most accurate model, and, therefore, it is as important as discriminating between correct and incorrect protein folds. In the context of independent model accuracy estimation (Olechnovič and Venclovas, 2017; Studer *et al.*, 2020), the progress in deep learning has fostered the application of deep NNs to predicting model accuracy locally for each residue (Hurtado *et al.*,

2018; Shuvo *et al.*, 2020).

The present study introduces a protocol ROPIUS0 (an acronym for "Restraint-Oriented Protocol for Inference and Understanding of protein Structures") for protein structure prediction and model selection. The rationale behind the ROPIUS0 protocol is to apply homology modeling when structural templates can be identified. Otherwise, it is aimed at selecting the most accurate models from a set of independently generated structural models.

Model selection is expected to be accurate if it produces estimates that correlate with discrepancies between the model and the target. According to the ROPIUS0 protocol, predicted distances between residues are contrasted with the distances observed in the model, and the differences are used to calculate model quality scores. This way, the accuracy of predictions determines the accuracy of model selection. Given the great prediction power of deep learning models, the deep learning module is at the core of the ROPIUS0 protocol.

ROPIUS0 employs a convolutional encoder-decoder NN architecture for predicting the distribution of distances between residues. The encoder part encodes the input into a latent space capturing dependencies in every input receptive field patch. The decoder part produces distance predictions by operating on the information encoded by the encoder. Predictions are therefore a result of considering the input representing the entire target.

The main difference from existing methods that use predicted distances between residues to predict model quality scores (Jing and Xu, 2020) is that ROPIUS0 uses distances predicted by the deep learning module directly for model accuracy estimation and selection. We show that the direct use of predicted distances may be sufficient to achieve satisfactory results even in the presence of some limitations, which we discuss in Section 3. We registered and tested ROPIUS0 (ropius0) and its variant ROPIUS0QA (ropius0QA) in the latest CASP14 season of critical assessment of methods of protein structure prediction (CASP) (Kryshtafovych *et al.*, 2019). ROPIUS0QA was officially ranked 13 (and among the top 10 independent groups) in the category of tertiary structure prediction. When tested on the CASP13 dataset, ROPIUS0QA was found to perform on par with top-performing automated prediction servers.

In this study, we focus on the model selection algorithm since the model selection-based variant ROPIUS0QA produced better results but describe the complete ROPIUS0 protocol for consistency. We further discuss the implications of this algorithm for improving homology search and modeling.

2 Results

In this section, we first describe the workflow of the ROPIUS0 protocol and define how ROPIUS0QA differs from it. We then proceed with the results of the protocols.

2.1 ROPIUS0 workflow

The goal of the ROPIUS0 protocol is to predict the structure of a given target protein sequence (Fig. 1). The target is modeled by homology modeling, or its structure is selected from independently generated models using the deep learning module (the feature prediction part in Fig. 1). The independently generated models represent CASP-hosted server predictions, which are available for (human) groups assigned three weeks to build models for a target. In general, they represent structural models generated by one or more methods, e.g., web servers.

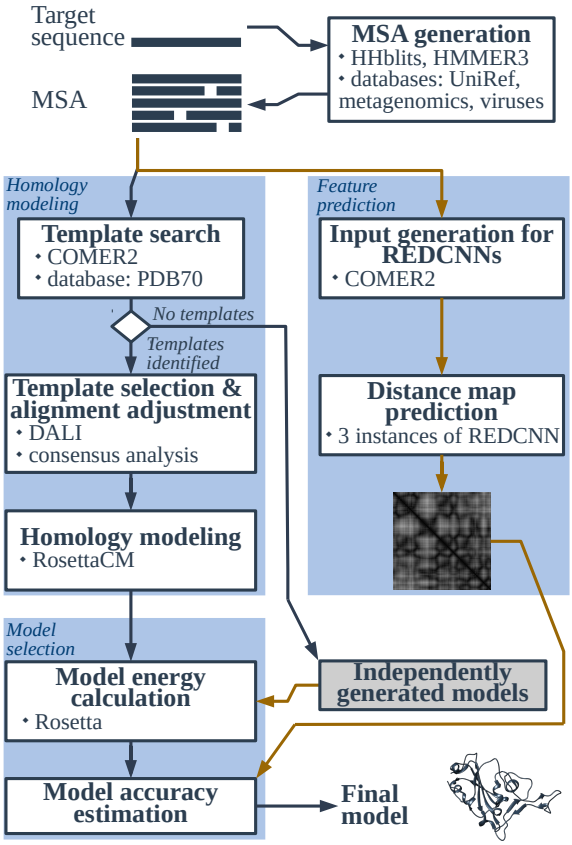

Figure 1. Workflow of the ROPIUS0 protocol. White boxes represent processes. The external and specifically developed tools and resources used in a process are listed at the bottom of the box. The path highlighted in yellow denotes the part of the workflow executed by ROPIUS0QA.

The approach of modeling is determined by the availability of structural templates for homology modeling. Available templates lead to invoking a structure-based procedure for template selection and alignment adjustment, followed by homology modeling. Concurrently, structural features of the target are predicted that are used as restraints for model ranking and selection. The restraints on the distances between the CB atoms of the target protein are generated using residual encoder-decoder convolutional neural networks (REDCNNs).

Structural models generated by either of the modeling approaches are ranked by the model accuracy estimation module, which calculates the energy of each model structure and evaluates the closeness of the distances between CB atoms (CA for Gly) in a model to those predicted by the REDCNNs. The model with the highest estimated accuracy is considered the most probable. The ROPIUS0 components are described in more detail in Section 4.

ROPIUS0QA is a special case of the ROPIUS0 protocol. It is designed to always perform model selection on a set of independently generated models. The difference between ROPIUS0 and ROPIUS0QA lies, therefore, in homology modeling and refinement (Section 4.4). ROPIUS0 generates models by homology and ranks them using the REDCNNs when templates can be identified, whereas ROPIUS0QA always selects among independently generated models. The ROPIUS0QA path in the ROPIUS0 workflow is highlighted in yellow in Figure 1.

2.2 Performance in CASP14

CASP is an initiative to rigorously evaluate predictions submitted by registered prediction groups (Kryshtafovych *et al.*, 2019). CASP prediction seasons are organized biennially. The latest CASP14 season for tertiary structure prediction started on 18th May and lasted until 21st August 2020. The sequences of unpublished target structures were issued every workday with a time frame of three days and three weeks for the server and human groups, respectively, to submit their predictions. The human groups were allowed to use the models submitted by the server groups.

At the end of CASP14, the independent assessors evaluated the predictions on a protein domain basis so that the orientation of the domains in a model did not influence the evaluation results. The evaluation takes into account the complexity of each target and is based on calculating the z-score for the GDT_TS score (Zemla, 2003) of the superposition between a model and the target domain. The total z-score accumulated over all domains determines a group's position in the final ranking (Abriata *et al.*, 2019).

We registered ROPIUS0 and ROPIUS0QA for the CASP14 season as two human prediction groups (group IDs 254 and 039, respectively). The results are shown in Figure 2. The official ranking for tertiary structure prediction is available in Ref. Protein Structure Prediction Center (2020).

The first observation is that one group, AlphaFold2 (427), stands out above the rest, with most of the targets modeled with high accuracy. The group's deep learning approach (DeepMind, 2020) is likely to change the direction of protein structure prediction.

ROPIUS0 and ROPIUS0QA rank 48th and 13th, respectively, among 146 groups. The ROPIUS0QA approach of using model selection from the server predictions without any further refinement was more accurate than the ROPIUS0 approach including homology modeling.

We see three reasons for this effect. The first is that protein structure prediction methods have improved substantially over recent years. Deep learning frameworks are able to accurately predict structural characteristics for using them as restraints in protein structure optimization and lead to high-quality models without considering templates. Second, the top-performing groups do not use homology modeling directly (except for group 183, tFold-CaT, among the top 50 groups). A typical approach instead converts structural information of identified templates to additional restraints for optimization or uses them as additional input for a more confident prediction of target inter-residue distances and orientations. Its advantage in modeling insertions in the target sequence is evident. Yet, it can also be seen from the perspective of processing deletions. An attempt to remove a fragment in the template structure by homology modeling can end up (to the extent depending on the modeling tool) with the distorted geometry of aligned fragments, even though the alignment can be correct. The third reason is model selection. The model accuracy estimation module was mainly tested through the ROPIUS0QA path. This module for ranking homology models was used by the ROPIUS0 protocol only in the second half of the CASP14 season (Section 4.4). The results show that using it during the entire CASP14 season would have been advantageous.

Judging from the available description of the groups' methods (CASP, 2020), ROPIUS0QA ranks best among the groups that made selections from the server predictions and submitted models without refinement (Fig. 2). Although ROPIUS0QA ranks models by providing accuracy estimates representing errors in Angstroms, no efforts were made for these estimates to correlate well with GDT_TS scores. The primary goal, instead, was to select the best available model (Section 4.8).

ROPIUS0 was also registered in the model accuracy

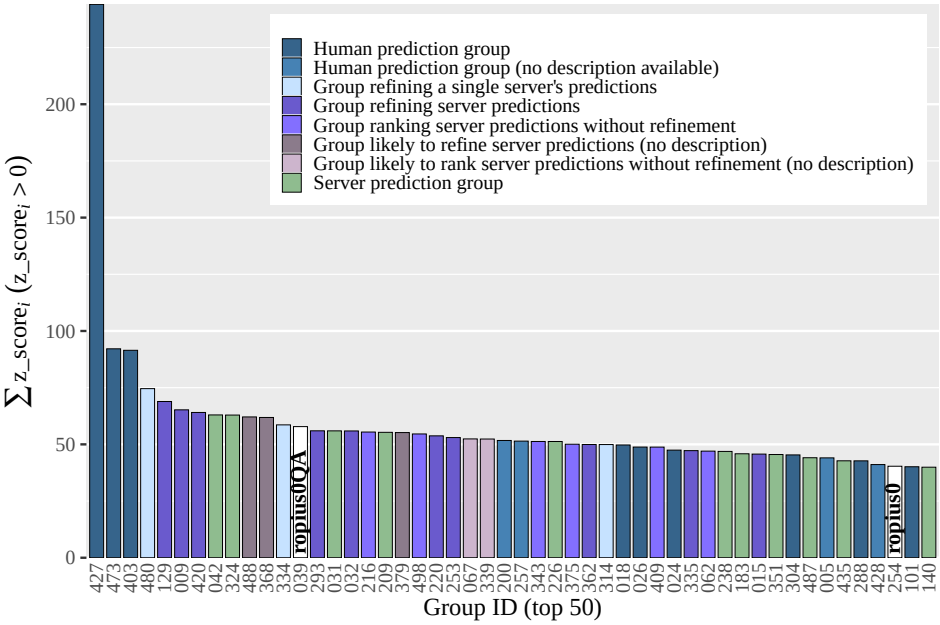

Figure 2. Ranking of the groups registered for the CASP14 season in the tertiary structure prediction category. The ranking is based on the sum of the z-scores found to be positive for the structural models designated by the group under consideration as best (1st model) (Protein Structure Prediction Center, 2020). The top 50 groups are shown. Server prediction groups are represented in green.

estimation (EMA) category. Model ranking in this category differed from the ranking by ROPIUS0QA to make selected models more consistent with the ROPIUS0 predictions (Section 4.4).

The rest of the Results section, therefore, focuses on the analysis of the ROPIUS0QA method and model selection, while Section 4 and Supplementary Sections S1.1–S1.3 present and provide results of a structure-based sequence alignment algorithm for improving alignment accuracy.

2.3 Test results on the CASP13 targets

To further explore the ROPIUS0QA's ability to correctly select structural models, we tested ROPIUS0QA on the CASP13 dataset. The test was performed in the same manner as that used during the CASP14 season. ROPIUS0QA was applied to the full targets, and the models selected from the CASP-hosted server predictions were evaluated on a target domain basis, following the (same) z-score-based official procedure for ranking groups (Protein Structure Prediction Center, 2018).

Normally, ROPIUS0QA uses two sets of thresholds for two application modes that depend on target difficulty (Section 4.8). Figure 3 shows that both ROPIUS0QA and its single-mode version (ropius0QA_SM), which simplify application, achieve similar results, which are on par with those of the top-performing server groups. The conclusion does not change (Supplementary Fig. S5) if the analysis excludes ten targets used to determine the REDCNN application thresholds and five domains having a HMMER (Eddy, 2011) hit at a significance threshold of E-value = 0.01 to a member of the REDCNN training and validation datasets. (A significance threshold of E-value = 0.01 for searching the sequences of the training and validation datasets corresponds to E-value $\approx$ 330 for searching a UniRef50 sequence database.)

Table 1 and 2 report ROPIUS0QA's performance contrasted with DeepAccNet (Hiranuma1 *et al.*, 2021) and two of the best single-model accuracy prediction methods in CASP13 (Won *et al.*, 2019): ProQ3D (Uziela *et al.*, 2017) and QMEANDisCo (FaeNNz) (Studer *et al.*, 2020) evaluated using latest sequence databases in the same way as ROPIUS0QA (see Supplementary Section S1.4 for details). Importantly, ROPIUS0QA compares favorably even when multiple sequence alignments (MSAs) for generating input to the REDCNNs originated from searching a database (Uniclust30 (Mirdita *et al.*, 2017) 10/2017) with a tool (HHblits (Remmert *et al.*, 2012) v3.0.0), both released before the start of the CASP13 season. The result of this test, which brings ROPIUS0QA closer to the conditions at the time of CASP13, implies that sequence

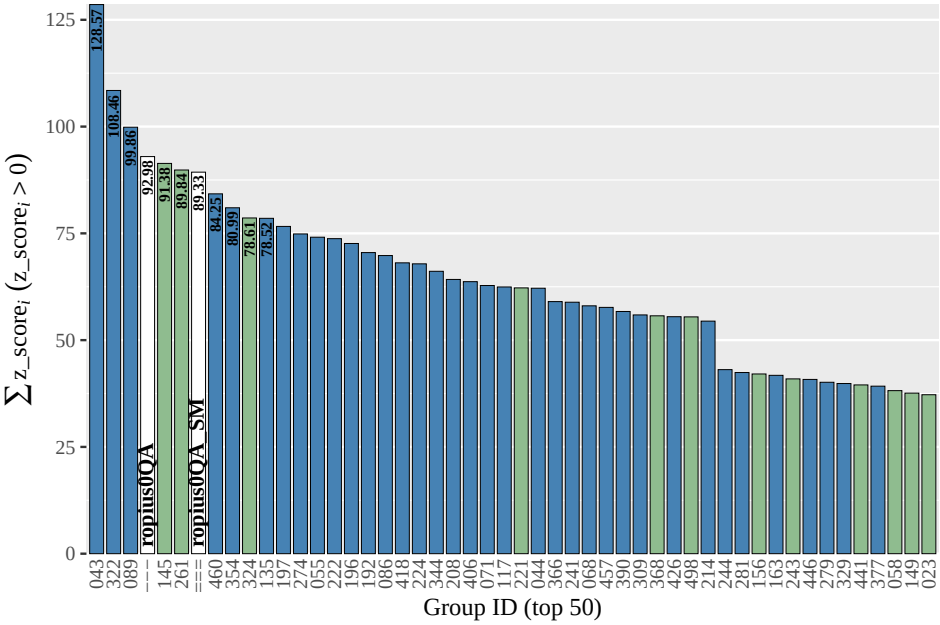

Figure 3. Results of the test on the CASP13 targets (104 domains) in the tertiary structure prediction category. ROPIUS0QA and ROPIUS0QA_SM were not registered for the CASP13 season. The ranking is based on the sum of the z-scores found to be positive for the structural models designated by the group under consideration as best (1st model) (Protein Structure Prediction Center, 2018). The top 50 groups are shown. Server prediction groups are represented in green.

quantity alone does not determine MSA quality. The accuracy and relevance of sequence alignments are just as important.

This is illustrated with two examples (Table 3). ROPIUS0QA's performance was measured for different variants of input MSAs constructed for the same target. Appending the initial MSA with aligned sequences simulated using the same MSA model (Margelevičius, 2019) does not change the result. In contrast, sequences simulated with noise deteriorate the quality of the MSA and, consequently, the result. The same effect is observed when the number of HHblits iterations for T1005 increases from 2 to 6. The MSA obtained after 6 iterations inflated with weakly related sequences, where the sequences at the significance limit at iteration 2 became highly significant, and the statistical significance of the highly significant sequences at iteration 2 considerably decreased.

The dependence of ROPIUS0QA's performance on the amount and diversity of sequences in an MSA for the CASP13 targets is shown in Figure 4. A weak correlation between the effective number of sequences in the MSA (Supplementary Section S2.2) and the distance from the highest GDT_TS score (Pearson's $r = -0.13$ for Neff, and $r = -0.22$ for log(Neff)) demonstrates ROPIUS0QA's ability to differentiate between accurate and inaccurate models for a considerable part of the shallow MSAs.

2.4 Deep learning performance justification

The REDCNNs in the ROPIUS0 protocol (Fig. 1) were trained on a small set of training data (Supplementary Section S2.1). Yet, their application to model selection has been shown to produce competitive results. An explanation for this phenomenon lies in the application of REDCNN predictions.

Inaccurate distance predictions diminish the ability to select the most accurate structural models. In application, the probability of using inaccurate predictions made by the REDCNNs is reduced, involving only confident predictions. Distance predictions made with probability at least equal to some threshold are retained, while those of lower probability are ignored (Section 4.8).

Figure 5 shows the dependence of the accuracy of distances, predicted by combining the output of the three REDCNNs, on the threshold probability for the CASP14 targets for which the structure is available. (A similar result obtained for the CASP13 targets is shown in Supplementary Fig. S7.) We observe that increasing the threshold probability leads to more accurate inter-residue distance predictions, independently of how far

Method	$\sum_{z_\text{score}_i>0} z_\text{score}_i$	Runtime (s), (s.d.)
ROPIUS0QA	92.98	133.35 (128.75)
ROPIUS0QA_SM	89.33	—"—
ROPIUS0QA*	90.87	—"—
ROPIUS0QA_SM*	88.47	—"—
QMEANDisCo	65.36	58.27 (47.62)
ProQ3D	60.67	6104.60 (6635.30)

Table 1. Evaluation results on the CASP13 targets. An asterisk (*) denotes a ROPIUS0QA version provided only with the Uniclust30 database released before the CASP13 season. The average runtimes per target (Runtime) exclude the time spent on sequence search. The runtimes for each CASP13 target are plotted in Supplementary Figure S6. The CASP13 dataset of the available DeepAccNet predictions (Hiranuma1 *et al.*, 2021) lacks data for three targets T0954, T0999, and T1005. The sum of the z-scores for three DeepAccNet versions (Standard, Bert, and MSA) excluding these three targets (six domains) is 57.52, 61.19, and 77.97, respectively. s.d., standard deviation.

Method	$\sum_{z_\text{score}_i>0} z_\text{score}_i$
ROPIUS0QA	69.77
ROPIUS0QA_SM	68.26
ROPIUS0QA*	67.14
ROPIUS0QA_SM*	66.31
DeepAccNet-MSA	63.25
DeepAccNet-Standard	48.69
QMEANDisCo	48.49
ProQ3D	47.69
DeepAccNet-Bert	47.11

Table 2. Evaluation results on 76 CASP13 target domains. Ten targets used to determine thresholds for the REDCNNs are excluded. Five more targets/domains (T0954, T0966, T0983, T0999, and T1017s1-D1) having sequence similarity to a member of the training and validation sets are excluded too. An asterisk (*) denotes a ROPIUS0QA version using the Uniclust30 database released before the CASP13 season. The results for three DeepAccNet versions (Standard, Bert, and MSA) were obtained from the available predictions on the CASP13 dataset (Hiranuma1 *et al.*, 2021).

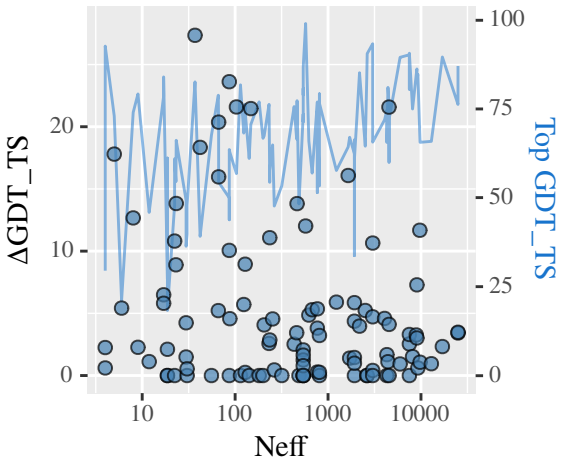

Figure 4. Difference (ΔGDT_TS) between the highest GDT_TS among the server groups (Top GDT_TS, line) and the ROPIUS0QA GDT_TS plotted against the effective number of sequences (Neff) for the groups' first models of the CASP13 targets.

apart residues are in the sequence. Yet, the number of distances predicted with at least the threshold probability decreases with increasing threshold. A small number of predictions may be insufficient to capture key fold-determining inter-residues properties. However, the average number of predictions approximately corresponds to the target length with a mean error slightly greater than 1Å at a relatively high threshold probability of 0.2 (Fig. 5), suggesting an optimal configuration.

The reader is referred to Supplementary Section S1.5 for a comparison of prediction accuracy with trRosetta (Yang *et al.*, 2020).

3 Discussion

One of the goals of developing ROPIUS0 was to build a system for evaluating the feasibility of improving homology detection. The ROPIUS0 protocol consists of two mutually connected parts. The first is for homology modeling. The other is formed by the deep learning module designed for the selection of protein structural models. We registered ROPIUS0 as a prediction group and tested it during the CASP14 season. Its variant ROPIUS0QA aimed at unconditionally carrying out model selection was also tested. Model selection on its own is important for fold recognition, protein structure and function prediction, and modeling of protein complexes. How model selection is related to homology detection will be discussed below.

ROPIUS0QA showed satisfactory results according to CASP14's ranking in the category of tertiary structure prediction and based on an additional test on the CASP13 targets. The architecture of the REDCNN, although being straightforward, was considered adequate for sequence-based only model selection. Its function to

	T1005				T1015s1		
	MSA_n2	+rndmsa_r0	+rndmsa_r0.1	MSA_n6	MSA_n4	+rndmsa_r0	+rndmsa_r0.1
Neff	201.36	230.36	230.36	1992.63	51.48	100.48	100.48
z_score	0.73	0.73	0.06	−0.07	1.73	1.73	−0.72

Table 3. Example of the importance of alignment quality rather than quantity alone. The effective number of sequences (Neff) in the MSA and the resultant z_score for different variants of MSAs provided as input to ROPIUS0QA are reported for two CASP13 targets, T1005 and T1015s1. MSA_ni represents the MSA obtained after i iterations of an HHblits search against the Uniclust30 database. +rndmsa_r0 and +rndmsa_r0.1 represent the original MSA (MSA_n2 or MSA_n4) appended with aligned sequences simulated (Margelevičius, 2019) using the original MSA as a model without noise (r0) and with noise (r0.1), respectively.

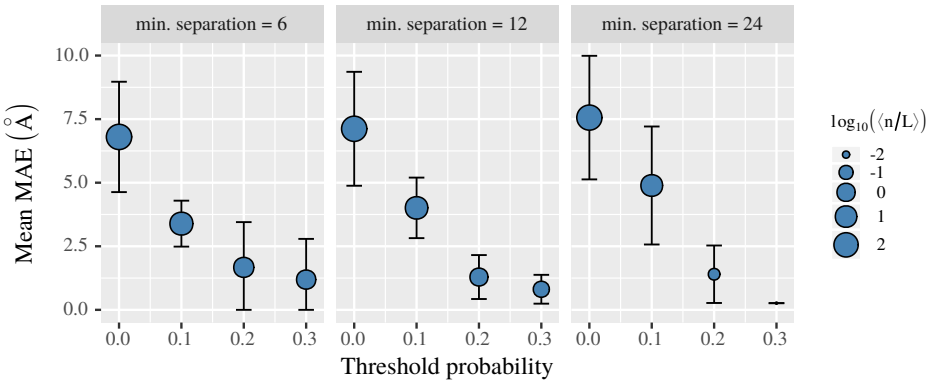

Figure 5. Mean of the mean absolute error (Mean MAE) between inter-residue distances predicted by the combination of the three REDCNNs and those observed in the CASP14 target structures. The mean MAE is calculated for a set of distances predicted with probability greater than the threshold probability by considering distances between any two residues separated by at least 6, 12, or 24 residues in the sequence. The point size is linearly related to the logarithm of the average number of predictions, expressed in the target length L, $\langle n/L \rangle$. ($\log_{10}(\langle n/L \rangle) = 0$ implies $\langle n/L \rangle = 1$.) Error bars represent the standard error of the mean MAE.

predict distances between residues is employed to contrast predictions with the distances observed in structural models. Differences in distances are then used to estimate the accuracy of models. Such an approach provides a reduction in computational complexity with respect to methods that use a two-level prediction system based on predicting distances first and local (residue-wise) model accuracy afterwards.

Although primarily designed for global model accuracy estimation, the REDCNN architecture can be extended to allow for multiple heads for the simultaneous prediction of distances between residues and position-specific differences in distances. Such modification would require augmenting the input format with structural models.

In general, the ROPIUS0 protocol can be transformed into an end-to-end differential framework unifying the present homology modeling and de novo modeling with model accuracy estimation. It would generate structural models from inter-residue distances and backbone dihedral angles predicted based on input data augmented with restraints transferred from identified templates (if any). Meanwhile, the discriminator component in the context of a generative adversarial network would produce model accuracy estimates for selection among the generated models.

The accuracy of modeling by homology in both a unified framework and the present study depends on the accuracy of the alignment between the target sequence and a template. ROPIUS0 applied a TSA3 algorithm (Section 4) to template selection and target-template alignment adjustment. Its effectiveness in the process of building structural models of higher quality is demonstrated in Supplementary Sections S1.1–S1.3. TSA3's usefulness lies in general applicability, i.e., it can be applied to alignments produced by any method.

One of the main limitations of the present study arises with limited training data. While more advanced attention-based NN architectures (Devlin *et al.*, 2019) may lead to better performance, limited training data hinders learning using any NN model. This problem was overcome by focusing only on confident inter-residue distance predictions, which showed high accuracy (Figure 5). Increasing the amount of training data

should lead to better model selection performance, but using a subset of predictions also has important implications for threading and homology search in general.

We intentionally used the COMER2 software (Margelevičius, 2020a) to generate sequence-based input to the REDCNN. Testing model selection using sequence-based inputs allowed us to evaluate both the accuracy of REDCNN predictions and how these predictions correlate with recognizing the correct structural fold. The result of including confident predictions offers prospects for improving the sensitivity of homology search without sacrificing speed.

4 Methods

This section provides details on the components of the ROPIUS0 protocol (Fig. 1).

4.1 Profile construction

The availability of templates is determined and templates for homology modeling are searched for by profile-profile alignment. The sensitivity of the search largely depends on the quality of the profile constructed for the target sequence. To collect reliably aligned homologous sequences for a multiple sequence alignment (MSA) from which a profile is constructed, the procedure shown in Figure 6 is used.

The UniRef30 database of MSAs (Mirdita *et al.*, 2017), filtered to 30% sequence identity, is searched with the target sequence using HHblits (Remmert *et al.*, 2012) for two to six iterations. If the result contains a number of sequences sufficient to construct an informative COMER2 (Margelevičius, 2020a) profile, no further searches are conducted. The information content of the resulting COMER2 profile is measured by the effective number of observations (ENO) (Margelevičius, 2019). Its value rarely exceeds 14, and a profile is considered to be informative if its ENO is greater than or equal to 4.

If insufficient sequences appear in the search results, HHblits is used to search the BFD metagenomic protein database of MSAs (Steinegger *et al.*, 2019b), filtered to 30% sequence identity. Concurrently, two viral sequence databases RVDB-prot (Bigot *et al.*, 2020) and IMG/VR (Paez-Espino *et al.*, 2017) are searched using the jackhmmer (first iteration) and hmmsearch (subsequent iterations) programs from the HMMER3 software package (Eddy, 2011) for two to four iterations if the target is a viral protein (Fig. 6).

If the overall search results do not imply an informative COMER2 profile, the final round of searches in the UniRef50 (Suzek *et al.*, 2015) and the MGnify metagenomic (Mitchell *et al.*, 2020) sequence database

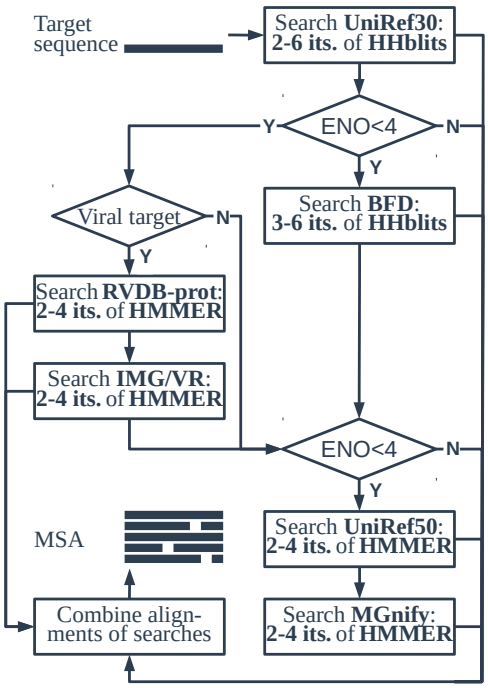

Figure 6. Flowchart of the procedure for generating an MSA for the target sequence.

is performed using the same settings for HMMER3 as described above.

All searches are configured with an E-value threshold for output and sequence inclusion in subsequent iterations of 0.01. The progression of iterations stops as soon as the statistical significance of a match in the previous iteration drastically decreases over subsequent iterations.

Before constructing a profile, the significant alignments from all searches conducted are combined into a single MSA. A profile is constructed from the resulting MSA using the makepro program from the COMER2 software package.

4.2 Template search

For identifying structural templates, the PDB70 database (Margelevičius, 2020b) of COMER2 profiles corresponding to PDB (Berman *et al.*, 2000) sequences filtered to 70% identity is searched using COMER2 (Margelevičius, 2020a) with a query profile constructed for the target sequence. The output of the search is controlled by the option MINPP of a posterior probability threshold for alignment extension. MINPP values of 0.28 (default) and 0.1 were used, and alignments produced using a value of 0.1 were selected if the aligned query coverage was not much larger than that obtained

using the default value. A MINPP value of 0 was not used, as semi-global alignments produced in this setting can include non-homologous, structurally divergent terminal regions.

4.3 Template selection and alignment adjustment algorithm TSA3

An algorithm for template selection and alignment adjustment TSA3 has been developed (Fig. 7). The rationale behind the algorithm is that statistically significantly identified templates with the most extensive alignment coverage considered reliable represent the most appropriate candidates for protein homology modeling. The regions considered reliably aligned between a template and the target result from applying the TSA3 algorithm to that template. The algorithm applies to every template of interest individually.

Let tmpl_s denote the template whose alignment with the target sequence is to be adjusted by applying the TSA3 algorithm. Let T be the set of *transition* templates tmpl_i homologous to both tmpl_s and the target. According to the algorithm, reliably aligned regions are aligned consistently across the alignment variants between tmpl_s and the target. The tmpl_s-target alignment variants are obtained by first aligning the transition templates tmpl_i with both the target and tmpl_s. For each tmpl_$i \in T$, two pairwise alignments, tmpl_i-tmpl_s and tmpl_i-target alignments, are merged to form the MSA using tmpl_i as a reference sequence. Leaving tmpl_i out then produces one tmpl_s-target alignment variant (Fig. 7).

The accuracy of the TSA3 algorithm depends on pairwise alignment quality. Since one set of pairwise alignments includes template sequences for which the structure is known, structure comparison, considered the most accurate, is used to produce the alignments between tmpl_s and all the templates tmpl_$i \in T$. We use DALI Holm *et al.* (2008).

A schematic of the TSA3 algorithm is shown in Figure 7. In addition to indicating reliably aligned regions, alignment variants sorted by proximity to the consensus alignment offer alternatives for adjusting the alignment between the target and tmpl_s to be used for homology modeling. The TSA3 algorithm is most helpful when the set T includes structurally similar templates.

4.4 Homology modeling

Homology modeling is used to predict the structure of the target sequence when structural templates are available (Section 4.2). Modeling is performed with RosettaCM Song *et al.* (2013) version 2019.35.60890 using the ref2015 score function Park *et al.* (2016), which is applied to modeling every target domain using optionally multiple templates and to combining domain structural models into a complete structure of the target. Target domain boundaries are identified using COMER2 and from REDCNN predictions (see Section 4.5). One or more templates for modeling are selected and their alignments to the target are adjusted using the TSA3 algorithm (Section 4.3) if multiple templates are available. Five to eight selected structural models (method described in Section 4.8 and used in the second half of the CASP14 season) of each identified target domain are refined using the Rosetta relax Conway *et al.* (2014) computer program. Relaxed models were not considered if they were more divergent from the template(s) than the selected starting models. The number of decoys generated by RosettaCM and relax was 200. The same refinement using relax is applied to selected CASP-hosted server models when templates are unavailable.

4.5 REDCNN

A REDCNN is used to predict a probability distribution of distances between the CB atoms of each pair of the target residues. Its fully convolutional architecture, shown in Figure 8, is based on SegNet (Badrinarayanan *et al.*, 2017). It consists of the encoder and decoder parts. The encoder contains a series of groups of blocks, each followed by a maxpool layer. The maxpooling operation applied with a window of size 2×2 and a stride of the same size halves the dimensions of each two-dimensional (2D) input feature map. Each block in a group preceding a maxpool layer contains a convolution layer followed by batch normalization and ReLU activation.

The decoder has an upscaling structure. In the decoder, a convolution transpose layer, batch normalization and ReLU activation precede each group of blocks organized similarly as in the encoder. A stride of size 2×2 used for transposed convolution ensures the dimensions of an output feature map agree with those of a feature map fed into the maxpool layer at the corresponding level in the encoder. The units of each feature map produced by a maxpool layer of the encoder are connected by skip connections to the units of the corresponding feature map passed to a convolution transpose layer of the decoder (Fig. 8).

A probability distribution of distances is predicted by the softmax layer. The output for a length-L target protein or its fragment is an $L \times L$ map whereby each cell is a probability distribution over 64 classes, which represent distances discretized into intervals of length 1Å $([0,2),[2,3),\ldots,[63,64),[64,\infty))$. Clusters of relatively short (mean) distances in local regions of the predicted map can provide additional indication of domain boundaries for the target protein.

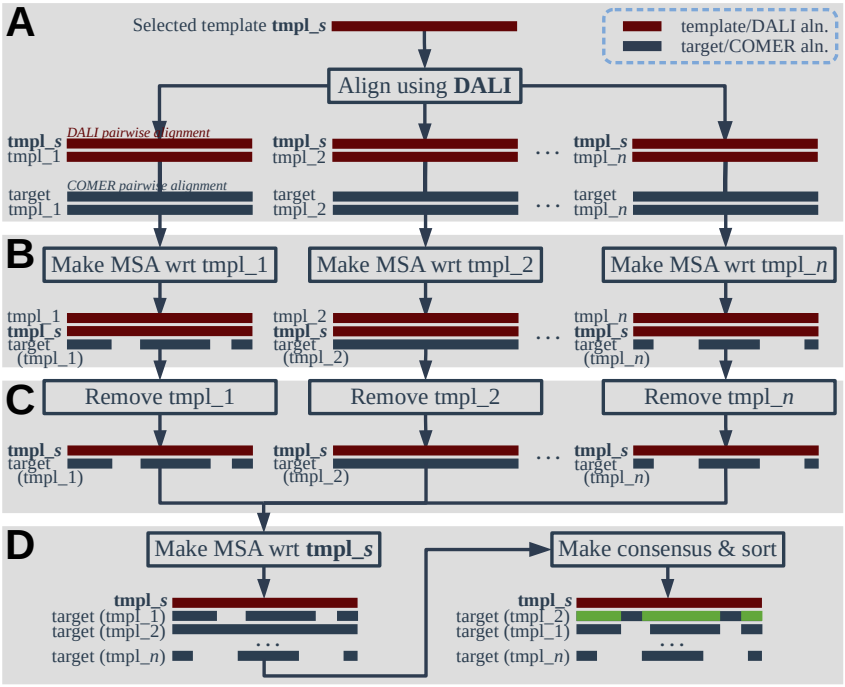

Figure 7. TSA3 algorithm for template selection and alignment adjustment. Its goal is to identify reliably aligned regions for each template tmpl_s of interest. The algorithm starts with producing the pairwise structural alignments between tmpl_s and all members of the template set $\{\text{tmpl_}i\}_{i=1}^{n}$ using DALI (A). Next, for each template tmpl_i ($i = 1, \ldots, n$), the MSA is derived from the DALI pairwise alignment between tmpl_s and tmpl_i and the COMER pairwise alignment between the target and tmpl_i (target (tmpl_i); B). Template tmpl_i is removed from the MSA to leave tmpl_s and the target associated through it in the pairwise alignment (C). Consensus analysis of the target alignment variants in the MSA derived from the indirect pairwise alignments between tmpl_s and the target reveals reliably aligned regions (green; D). The TSA3 algorithm applied independently to multiple templates tmpl_s suggests for protein modeling candidates with extensive reliably aligned regions.

4.6 REDCNN input

Multiple $L \times L$ attribute matrices stacked one on top of another constitute the input to a REDCNN. Each matrix represents a specific type of data for a target or its fragment of length L. A matrix entry indexed by i and j contains a value that characterizes a relation between residues at positions i and j. The number of matrices is called the number of channels in the input.

There are two groups of attribute matrices. The first group is obtained from a COMER profile constructed from an MSA representing the family of the target sequence (Section 4.1). Profile data extends along the target sequence. Therefore, the value of a profile attribute at position i is assigned to row i and replicated across all columns of an attribute matrix. Accordingly, the same value at profile position i is additionally associated with column i and replicated across all rows to obtain a 2D representation of profile data.

Profile attributes at position i and the number of channels required are the following: 1 channel of sequence position index; 1 channel of residue code; 20 channels of amino acid substitution probabilities (probability distribution $\boldsymbol{f}_i$) (Margelevičius, 2016); 7 channels of transition probabilities from and to match (M), insert (I), and delete (D) states; 3 channels of effective numbers of observations; 1 channel of the sum of the five largest posterior cluster membership probabilities and the probability of a new cluster for $\boldsymbol{f}_i$ under the hierarchical Dirichlet process (HDP) mixture model (Margelevičius, 2016); 2 channels of the largest cluster membership probability under the HDP mixture model and the index of the corresponding cluster; 2 channels of the squared Euclidean norm and the logarithm of the prior probability of the context vector (Margelevičius, 2018); 19 channels of the context vector itself; 3 channels of secondary structure probabilities for coil (C), strand (E), and helix (H) classes (Jones, 1999).

The second group of attribute matrices is 2D by nature and corresponds to the mutual information and the cross-covariance matrix between observations at positions i and j of the MSA (and target sequence) for each i and j.

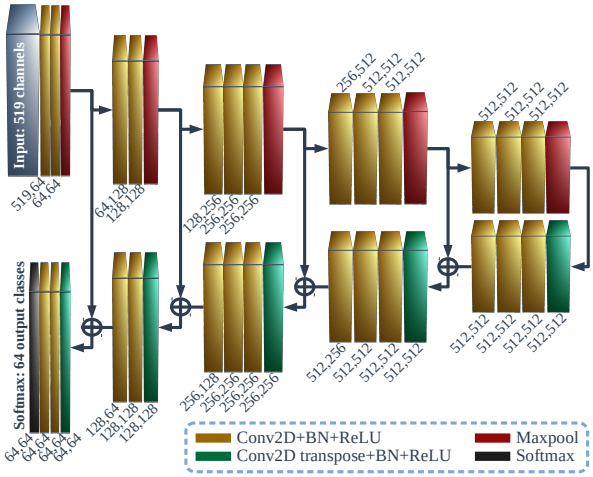

Figure 8. REDCNN architecture. Two numbers below or above blocks indicate the number of input feature maps (channels in the first layer) and the number of filters in the convolution and the convolution transpose layers, respectively. For the softmax layer, the second number is the number of classes. Kernels of size 3×3 are used for convolution and transposed convolution. Input has 519 channels. Output depth corresponds to a probability distribution over 64 classes predicted by the last softmax layer. BN, batch normalization.

The cross-covariance matrix of two random vectors $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$ of dimension 20 is

$$\mathrm{Cov}(\boldsymbol{v}_i, \boldsymbol{v}_j) = \mathrm{E}(\boldsymbol{v}_i \boldsymbol{v}_j^{\mathsf{T}}) - \mathrm{E}(\boldsymbol{v}_i)\mathrm{E}(\boldsymbol{v}_j)^{\mathsf{T}}.$$

We estimate it from a sample of sequences, as found in the MSA, that have residues aligned at positions i and j. An observation $\boldsymbol{v}_{sk}$ of random variable $\boldsymbol{v}_k$ records an occurrence of the residue found in sequence s at position k. The ath element of $\boldsymbol{v}_{sk}$ is $v_{ska} = n_k w_s$, where n_k is the number of sequences at position k in the MSA, and w_s ($\sum_s w_s = 1$) is the weight of sequence s with residue type a at position k (Henikoff and Henikoff, 1994). The total number of channels for the cross-covariance matrix is 400 (20×20).

Mutual information is calculated as the Kullback-Leibler divergence between estimated probability distributions $D_{\mathrm{KL}}(p_{ij} \parallel q_{ij})$, where the joint probabilities of observing residue types a and b at positions i and j are

$$p_{ij}(a,b) = [\mathbf{1}^{\mathsf{T}}(\textstyle\sum_s \boldsymbol{v}_{si}\boldsymbol{v}_{sj}^{\mathsf{T}})\mathbf{1}]^{-1} \textstyle\sum_s v_{sia} v_{sjb} \quad \text{and}$$
$$q_{ij}(a,b) = [(\mathbf{1}^{\mathsf{T}} \textstyle\sum_s \boldsymbol{v}_{si})^{-1} \textstyle\sum_s v_{sia}] \times [(\mathbf{1}^{\mathsf{T}} \textstyle\sum_s \boldsymbol{v}_{sj})^{-1} \textstyle\sum_s v_{sjb}];$$

$\mathbf{1}$ is a vector of ones. Mutual information requires one channel.

With a total of 519 channels, input to a REDCNN is produced using the COMER2 software package. The makepro program constructs a profile from an MSA, whereas the makecov program calculates cross-covariance matrices and mutual information for all pairs of positions i and j ($i \leq j$). We refer to this type of input as a *promage*.

4.7 Combining multiple REDCNN predictions

We apply three trained instances of the REDCNN model (Supplementary Section S2.1) to predicting inter-residue distances and model accuracy estimation (Section 4.8). In the former case, predictions made by three REDCNNs are combined to form a mixture of probability distributions with equal weights for each pair of residues.

4.8 REDCNN-based model accuracy estimation

A trained REDCNN model is used to estimate the accuracy of a protein structural model of length L. Accuracy estimates are produced by evaluating how close the distances between CB atoms in the structural model match those predicted by the REDCNN.

Residue-wise accuracy estimates are calculated for each protein residue and represent errors in Angstroms, as required by CASP. An estimate for residue i ($i \in [1, L]$) is obtained by first calculating the mean absolute difference between distances observed in the structure, and rounded to the nearest integer, and predicted distances over pairs of residues (i, j) ($j = 1, \ldots, i-4, i+4, \ldots, L$ for $4 < i \leq L-4$) such that they are predicted, with probability at least the threshold probability, to be separated by a distance at most the threshold distance. The mean absolute difference obtained for residue i is divided by the expected difference for random structures ($= 1/63 \times 63 \cdot 62/2$) and multiplied by 8 so that the expected error is 8Å. We additionally multiply the estimate by 2 to account for using the thresholds and for a relatively small probability of long distances between residues in protein structures.

The global accuracy score provides an estimate for the entire protein structure. It is calculated as $1 - \frac{1}{8|\mathsf{R}|} \sum_{i \in \mathsf{R}} \hat{e}_i$, where $\hat{e}_i$ is the accuracy estimate calculated for residue i, and R corresponds to the set of residues included in the calculation, whose size is not necessarily L due to the thresholding and the possibility of missing residues in the structural model. The global accuracy estimate ranges from 0 to 1.

The calculation of final global and local estimates is based on averaging the estimates obtained using three trained instances of the REDCNN model (Supplementary Section S2.1). The estimates also depend on the physics of the structural model evaluated using Rosetta (Conway *et al.*, 2014). The final local estimate for

residue i is, therefore, a linear combination of the average estimate and the exponentiated sum of certain Rosetta energy terms for residue i, with weights $1 - r$ and r $(0 < r < 1)$, respectively. The final global estimate is calculated as described above, using the final local estimates.

The set of parameters that applies to calculating the accuracy estimates includes the probability and distance thresholds for each of the three REDCNNs, the list of Rosetta energy terms, and the weight r. We determined their optimal configuration on a small subset of CASP13 targets.

We distinguished between two conditions/modes. These conditions represent the situations of abundance and scarcity of sequences homologous to the sequence of the protein under consideration. They also represent the CASP template-based (TBM) and free modeling (FM) categories for target sequences.

Two groups of five targets served as datasets. The first group of targets for which numerous homologous sequences could be found consisted of the following targets: T0965, T0976, T0996, T0997, and T1014. The targets in the second group were T0950, T0987, T0990, T1005, and T1010. They represented sequences of the FM category.

The optimization criterion was set to achieve the best match between the rankings of the top 50 structural models ranked officially by CASP13 according to GDT_TS (Zemla, 2003) and the rankings obtained by the final global estimate for these structural models. The weight $r = 0.2$ and a list of four Rosetta energy terms, fa_elec, hbond_lr_bb, hbond_sc, and p_aa_pp, were found to be optimal for both groups of targets. The probability and distance threshold values for each of the three REDCNNs were found to be $(0.2, 32)$, $(0.2, 32)$, and $(0.2, 20)$ for the first group of targets. The corresponding values for the second group of targets were $(0.3, 20)$, $(0.1, 10)$, and $(0.1, 8)$.

Data Availability

The ROPIUS0 software is available at https://github.com/minmarg/ropius0 Trained TensorFlow models are deposited to Zenodo at https://doi.org/10.5281/zenodo.4450107 Code, scripts, and data for tests on CASP datasets are deposited to Zenodo at https://doi.org/10.5281/zenodo.4451302 and https://doi.org/10.5281/zenodo.4783276 CASP datasets are available at https://predictioncenter.org.

Acknowledgements

The author is grateful to Albertas Timinskas for help in preparing data for the experiments.

Funding

This work was supported by the European Regional Development Fund [Grant No. 01.2.2-LMT-K-718-01-0028] under a grant agreement with the Research Council of Lithuania (LMTLT).

Competing interests

The author declares no competing interests.

References

Abriata, L., Tamò, G., and Dal Peraro, M. (2019). A further leap of improvement in tertiary structure prediction in CASP13 prompts new routes for future assessments. *Proteins*, **87**(12), 1100–1112.

Badrinarayanan, V., Kendall, A., and Cipolla, R. (2017). SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, **39**(12), 2481–2495.

Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., *et al.* (2000). The Protein Data Bank. *Nucleic Acids Res.*, **28**(1), 235–242.

Bigot, T., Temmam, S., Pérot, P., and Eloit, M. (2020). RVDB-prot, a reference viral protein database and its HMM profiles. *F1000Res*, **8**, 530.

CASP (2020). Abstract book (fourteenth round). https://predictioncenter.org/casp14/doc/CASP14_Abstracts.pdf.

Conway, P., Tyka, M., DiMaio, F., Konerding, D., and Baker, D. (2014). Relaxation of backbone bond geometry improves protein energy landscape modeling. *Protein Sci.*, **23**(1), 47–55.

Croll, T., Sammito, M., Kryshtafovych, A., and Read, R. (2019). Evaluation of template-based modeling in CASP13. *Proteins*, **87**(12), 1113–1127.

DeepMind (2020). AlphaFold: a solution to a 50-year-old grand challenge in biology. https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume*

1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Eddy, S. (2011). Accelerated profile HMM searches. *PLoS Comput. Biol.*, **7**(10), e1002195.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Henikoff, S. and Henikoff, J. G. (1994). Position-based sequence weights. *J. Mol. Biol.*, **243**(4), 574–578.

Hiranuma1, N., Park, H., Baek, M., Ivan Anishchenko, I., Dauparas, J., and Baker, D. (2021). Improved protein structure refinement guided by deep learning based accuracy estimation. *Nat. Commun.*, **12**, 1340.

Holm, L., Kääriäinen, S., Rosenström, P., and Schenkel, A. (2008). Searching protein structure databases with DaliLite v.3. *Bioinformatics*, **24**(23), 2780–2781.

Hou, J., Wu, T., Cao, R., and Cheng, J. (2019). Protein tertiary structure modeling driven by deep learning and contact distance prediction in CASP13. *Proteins*, **87**(12), 1165–1178.

Hurtado, D., Uziela, K., and Elofsson, A. (2018). Deep transfer learning in the assessment of the quality of protein models.

Jing, X. and Xu, J. (2020). Improved protein model quality assessment by integrating sequential and pairwise features using deep learning. *Bioinformatics*.

Jones, D. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **292**(2), 195–202.

Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K., and Moult, J. (2019). Critical assessment of methods of protein structure prediction (CASP)—Round XIII. *Proteins*, **87**(12), 1011–1020.

Margelevičius, M. (2016). Bayesian nonparametrics in protein remote homology search. *Bioinformatics*, **32**(18), 2744–2752.

Margelevičius, M. (2018). A low-complexity add-on score for protein remote homology search with COMER. *Bioinformatics*, **34**(12), 2037–2045.

Margelevičius, M. (2019). Estimating statistical significance of local protein profile-profile alignments. *BMC Bioinformatics*, **20**, 419.

Margelevičius, M. (2020a). COMER2: GPU-accelerated sensitive and specific homology searches. *Bioinformatics*, **36**(11), 3570–3572.

Margelevičius, M. (2020b). COMER2 profile databases version 2.2. https://sourceforge.net/projects/comer2/files/comer2-profile-databases-2.02.

Maximova, T., Moffatt, R., Ma, B., Nussinov, R., and Shehu, A. (2016). Principles and overview of sampling methods for modeling macromolecular structure and dynamics. *PLoS Comput. Biol.*, **12**(4), e1004619.

Mirdita, M., von den Driesch, L., Galiez, C., Martin, M., Söding, J., and Steinegger, M. (2017). Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Res.*, **45**(D1), D170–D176.

Mitchell, A., Almeida, A., Beracochea, M., Boland, M., Burgin, J., Cochrane, G., *et al.* (2020). MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Res.*, **48**(D1), D570–D578.

Olechnovič, K. and Venclovas, Č. (2017). VoroMQA: Assessment of protein structure quality using interatomic contact areas. *Proteins*, **85**(6), 1131–1145.

Paez-Espino, D., Chen, I.-M., Palaniappan, K., Ratner, A., Chu, K., Szeto, E., *et al.* (2017). IMG/VR: a database of cultured and uncultured DNA viruses and retroviruses. *Nucleic Acids Res.*, **45**(D1), D457–D465.

Park, H., Bradley, P., Greisen, P. J., Liu, Y., Mulligan, V., Kim, D., *et al.* (2016). Simultaneous optimization of biomolecular energy functions on features from small molecules and macromolecules. *J. Chem. Theory Comput.*, **12**(12), 6201–6212.

Protein Structure Prediction Center (2018). CASP13: Group performance based on combined z-scores. https://predictioncenter.org/casp13/zscores_final.cgi.

Protein Structure Prediction Center (2020). CASP14: Group performance based on combined z-scores. https://predictioncenter.org/casp14/zscores_final.cgi.

Remmert, M., Biegert, A., Hauser, A., and Söding, J. (2012). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods*, **9**(2), 173–175.

Senior, A., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., *et al.* (2020). Improved protein structure prediction using potentials from deep learning. *Nature*, **577**(7792), 706–710.

Shuvo, M., Bhattacharya, S., and Bhattacharya, D. (2020). QDeep: distance-based protein model quality estimation by residue-level ensemble error classifications using stacked deep residual neural networks. *Bioinformatics*, **36**, i285–i291.

Song, Y., DiMaio, F., Wang, R.-R., Kim, D., Miles, C., Brunette, T., *et al.* (2013). High-resolution comparative modeling with RosettaCM. *Structure*, **21**(10), 1735–1742.

Steinegger, M., Meier, M., Mirdita, M., Vöhringer, H., Haunsberger, S., and Söding, J. (2019a). HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics*, **20**, 473.

Steinegger, M., Mirdita, M., and Söding, J. (2019b). Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nat. Methods*, **16**(7), 603–606.

Studer, G., Rempfer, C., Waterhouse, A., Gumienny, R., Haas, J., and Schwede, T. (2020). QMEANDisCo—distance constraints applied on model quality estimation. *Bioinformatics*, **36**(6), 1765–1771.

Suzek, B., Wang, Y., Huang, H., McGarvey, P., Wu, C., and the UniProt Consortium (2015). UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, **31**(6), 926–932.

Uziela, K., Menéndez Hurtado, D., Shu, N., Wallner, B., and Elofsson, A. (2017). ProQ3D: improved model quality assessments using deep learning. *Bioinformatics*, **33**(10), 1578–1580.

Won, J., Baek, M., Monastyrskyy, B., Kryshtafovych, A., and Seok, C. (2019). Assessment of protein model structure accuracy estimation in CASP13: Challenges in the era of deep learning. *Proteins*, **87**(12), 1351–1360.

Yang, J., Anishchenko, I., Park, H., Peng, Z., Ovchinnikov, S., and Baker, D. (2020). Improved protein structure prediction using predicted interresidue orientations. *Proc. Natl. Acad. Sci. USA*, **117**(3), 1496–1503.

Zemla, A. (2003). LGA: A method for finding 3d similarities in protein structures. *Nucleic Acids Res.*, **31**(13), 3370–3374.

Zheng, W., Li, Y., Zhang, C., Pearce, R., Mortuza, S., and Zhang, Y. (2019a). Deep-learning contact-map guided protein structure prediction in CASP13. *Proteins*, **87**(12), 1149–1164.

Zheng, W., Zhang, C., Wuyun, Q., Pearce, R., Li, Y., and Zhang, Y. (2019b). LOMETS2: improved meta-threading server for fold-recognition and structure-based function annotation for distant-homology proteins. *Nucleic Acids Res.*, **47**(W1), W429–W436.

Zhu, J., Wang, S., Bu, D., and Xu, J. (2018). Protein threading using residue co-variation and deep learning. *Bioinformatics*, **34**(13), i263–i273.