JCTC Journal of Chemical Theory and Computation

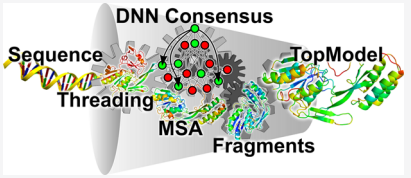
pubs.acs.org/JCTC Article

TopModel: Template-Based Protein Structure Prediction at Low Sequence Identity Using Top-Down Consensus and Deep Neural Networks

Daniel Mulnaes, Nicola Porta, Rebecca Clemens, Irina Apanasenko, Jens Reiners, Lothar Gremer, Philipp Neudecker, Sander H. J. Smits, and Holger Gohlke*

Cite This: *J. Chem. Theory Comput.* 2020, 16, 1953−1967 Read Online

ABSTRACT: Knowledge of protein structures is essential to understand proteins' functions, evolution, dynamics, stabilities, and interactions and for data-driven protein- or drug design. Yet, experimental structure determination rates are far exceeded by that of next-generation sequencing, resulting in less than 1/1000th of proteins having an experimentally known 3D structure. Computational structure prediction seeks to alleviate this problem, and the Critical Assessment of Protein Structure Prediction (CASP) has shown the value of consensus and meta-methods that utilize complementary algorithms. However, traditionally, such methods employ majority voting during template selection and model averaging during refinement, which can drive the model away from the native fold if it is underrepresented in the ensemble. Here, we present TopModel, a fully automated meta-method for protein structure prediction. In contrast to traditional consensus and meta-methods, TopModel uses top-down consensus and deep neural networks to select templates and identify and correct wrongly modeled regions. TopModel combines a broad range of state-of-the-art methods for threading, alignment, and model quality estimation and provides a versatile workflow and toolbox for template-based structure prediction. TopModel shows a superior template selection, alignment accuracy, and model quality for template-based structure prediction on the CASP10−12 datasets compared to 12 state-of-the-art stand-alone primary predictors. TopModel was validated by prospective predictions of the nisin resistance protein (NSR) protein from *Streptococcus agalactiae* and LipoP from *Clostridium difficile*, showing far better agreement with experimental data than any of its constituent primary predictors. These results, in general, demonstrate the utility of TopModel for protein structure prediction and, in particular, show how combining computational structure prediction with sparse or low-resolution experimental data can improve the final model.



INTRODUCTION

Knowing the 3D structure of a protein is important to understand its stability,[1] dynamics, function,[2] structural evolution,[3] and interactions with ligands[4,5] or other proteins.[6] Consequently, protein structure prediction is an essential part of knowledge-based protein engineering,[7] drug design and -discovery,[8] and function assignment.[9,10] At present, X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy are the dominating experimental methods for structure determination, but both are too time-consuming to keep up with current high-throughput genome sequencing information. Computational structure prediction has sought to alleviate this problem, and in the last decades, many approaches have been developed, raising the question of which method to use for a given sequence of interest. The biological information that can be derived from a structure prediction depends on its accuracy: high-confidence models based on homologous templates are generally suitable for computational ligand docking and virtual compound screening, while models with medium confidence can be useful for the identification of functionally important sites and disease-associated mutations.[11]

The field of computational structure prediction has driven many advances in structural bioinformatics, the most important being the development of threading algorithms that seek to identify a template structure most similar to the native structure of a target sequence of interest. These developments include fast and sensitive alignment methods such as iterated search methods,[12] position-specific scoring matrices,[13,14] sequence profile alignment,[15] profile−profile alignment,[16,17] and hidden Markov models.[18−22] The accuracy of threading algorithms has been further improved by adding structural features such as predicted secondary structure, residue contacts, solvent accessibility,[23] residue depth,[24] and backbone dihedral angles[25] to the alignment and scoring functions. Additionally, probabilistic modeling,[26,27] depth-dependent alignment of structure fragments,[28] multiple template and structure alignment,[29] normalized *Z*-scores,[16,23]

Received: August 15, 2019
Published: January 22, 2020

https://dx.doi.org/10.1021/acs.jctc.9b00825
J. Chem. Theory Comput. 2020, 16, 1953−1967

and sequence-based solvation potentials[17] have been employed to increase performance. Advances in multiple structure/sequence alignment methods, model building, clustering, and quality estimation have also had a large impact in the field.[30,31] Meta-approaches have proven to be one of the major advances,[32] as evident by the consistent high ranking of the Zhang meta-server[33] in the blind Critical Assessment of Protein Structure Prediction (CASP) experiments. The meta-server methodology is to produce structure predictions using information from multiple different algorithms[33,34] and either rerank or combine their output to produce better predictions than any of their component predictors. Considering the diversity of optimization procedures, training sets and quality measures, it is not surprising that meta-methods provide more robust results with a higher overall quality.

Here, we present a meta-approach to template-based structure prediction, which uses a top-down consensus approach rather than traditional majority-voting consensus termed TopModel. The development of TopModel was inspired by the success of meta-approaches in CASP experiments.[35] The CASP experiments, however, are undertaken on a working group rather than an algorithmic level, and competing groups use different algorithms not only for threading but also for alignment, model construction, model refinement, model evaluation, and model selection. It is therefore difficult to assign the differences in model quality from different groups to improvement of a specific step of the structure prediction workflow.

The aim of TopModel is therefore to individually optimize four steps of the structure prediction pipeline: template selection, target−template(s) alignment, model selection, and model combination and refinement. By focusing on each step individually, we aim to improve the final quality of models produced by TopModel. TopModel aims to provide a versatile and accurate toolbox for template-based protein structure prediction, expand the applicability of existing algorithms for threading, alignment, model quality estimation, and refinement via an automated integration of all methods, and yield high-quality structure predictions even for low sequence identities that are in agreement with experimental data.

Ab initio folding methods have in recent years seen a large increase in model accuracy because of a revolution in using image recognition deep neural networks (DNNs) for predicting residue−residue contacts and distances.[36] The aim of TopModel, however, is to establish an automated workflow for template-based modeling (TBM) in order to explore how deep learning can improve template selection and how well the use of structural information from multiple templates and alternate alignments can improve model quality compared to single-template models. In parallel to the development of TopModel, we are working on an ab initio folding pipeline that builds on the recent advances in prediction of residue−residue distances, which we aim to combine with the template-based folding in TopModel for improved performance.

METHODS AND IMPLEMENTATION

TopModel. TopModel is a protein structure prediction workflow with five modules that are executed consecutively or can be used individually. A simplified depiction of the interaction between the different TopModel modules can be seen in Figure 1; a detailed description of each module is given below.

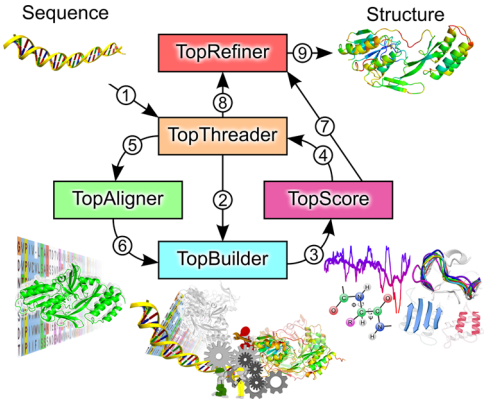


Figure 1. Simplified interaction between TopModel modules. The target sequence is given as an input to TopThreader (1), which searches for templates using different primary threaders. TopThreader uses TopBuilder (2) to build models from the primary threader alignments, template structures, and target sequence, which are scored with TopScore (3) and used by TopThreader (4) together with primary threader scores to rank and cluster templates and remove false positives. TopThreader then uses TopAligner (5) to align templates and construct consensus alignments, which are built with TopBuilder (6), scored with TopScore (3), and used together with primary threader scores in TopThreader (4) to rank templates by predicted similarity to the native structure. After template selection, TopAligner (5) is used to generate a large ensemble of pairwise and multitemplate alignments from which models are built with TopBuilder (6) and scored with TopScore (3). Models are selected from the multitemplate ensemble (7) and the single-template models (8) by TopRefiner, which combines and refines the models to produce a final model (9).

1. TopThreader. TopThreader identifies template structures from a target sequence based on predictions from 12 different primary threading programs using a top-down consensus approach instead of traditional majority voting.
2. TopAligner. TopAligner makes an ensemble of alignments between the target sequence and the provided templates based on template−template alignments from eight different primary alignment programs and template−target alignments from TopThreader.
3. TopBuilder. TopBuilder makes models of the target sequence based on alignments from TopAligner or TopThreader and templates from TopThreader using Modeller9[37] and Rosetta.[38]
4. TopScore. TopScore and TopScoreSingle[39] predict the global and local error of models based on predictions from 15 primary model quality assessment programs. TopScoreSingle is similar to TopScore but does not include clustering information and is therefore suitable when the best model is not part of a cluster.
5. TopRefiner. TopRefiner selects, combines, and refines models made by TopBuilder based on predicted global and local errors from TopScore and TopScoreSingle.

TopThreader. The threading process is the first and most critical step of template-based protein structure prediction.[33] It has three main goals: (1) identification of correct template structures for a target sequence, also known as fold recognition or threading, (2) target−template alignment, and (3) ranking of templates according to their similarity to the native structure. The TopModel threading module TopThreader uses a combination of DNNs, model quality prediction by

https://dx.doi.org/10.1021/acs.jctc.9b00825
J. Chem. Theory Comput. 2020, 16, 1953−1967

TopScore and TopScoreSingle,[39] and sequence/structure alignments to predict the TM-Score[40] between each template and the native structure, remove false-positive templates, calculate consensus alignments, and rank templates by their predicted TM-Score. The TM-Score is a robust measure of structural similarity between two proteins, which is independent of the protein sizes.

Prediction of template quality is similar to protein model quality assessment but not identical. First, template similarity to the native structure differs from model quality because of different possible target−template alignments, which is one of the main determinants of template-based model quality. In other words, a template may be similar to the native structure, but if the target−template alignment is wrong, the resulting model can have a low quality. Consequently, while a template has just one TM-Score to the native structure, models built from different alignments between the target and the template may have different model qualities, which can obscure the detection of the best template. Second, template similarity to the native structure is based on comparison between structures with different sequences and sizes, while model quality is based on comparison between structures of the same size and sequence as the native structure. Thus, while a small partially matching template may have the right fold for a given part of the target sequence, a model based on such a template alone could have a poor quality because of low coverage. These differences are important, especially for hard cases, in which threaders may prefer a wrong template with a large coverage over a short template with a correct fold but poor coverage. As such, the prediction of template similarity to the native structure is a challenging task.

TopThreader has eight steps outlined here. In the Supporting Information, a detailed description of the TopThreader workflow (Text T1 and Figure S1), the DNN training (Text T1, Figure S2, and Table S2), and the primary threading programs (Text T2 and Table S4) can be found.

1. Primary Threaders. TopThreader uses 20 primary threading algorithms from 12 primary threaders and selects the top 5 templates from each threader (Table S1). All threaders are run with default settings following the provided instructions by their respective authors.
2. Prefiltering. Prefiltering allows the user to discard templates according to cutoffs with respect to, for example, sequence identity, coverage, experimental method, or submission date. By default, templates with less than 30% coverage and artificially designed proteins are removed.
3. Alignment Fitting. TopThreader fits all pairwise threading alignments to the template structures and target sequence to ensure that residues match exactly.
4. Score templates using DNNs. TopThreader initially predicts a target−template TM-Score (Initial Score) using DNNs. DNNs' input features include primary threader scores and values calculated from threading alignments such as sequence identity and target coverage.
5. Redundancy clustering. TopThreader clusters templates at 90% sequence identity and pairwise TM-Score of 0.9, selecting the cluster centroid with the highest Initial Score. Alignments from other threaders/templates in the cluster are transferred to the centroid by superimposing their target−template alignments to the (nearly identical) centroid while minimizing changes to the alignment.
6. False positive removal. Removal of false positives is critical to ensure correct fold recognition. TopThreader first clusters templates structurally to remove bias toward folds with many templates. For each cluster, DNNs are used to predict the centroid TM-Score (Filtering Score). Templates are then structurally aligned to the best centroid based on Filtering Score and TopScoreSingle of a model built from the template. Using a top-down consensus approach, models are discarded if they are dissimilar (TM-Score < 0.4) to the best centroid.
7. Consensus. TopThreader uses local and global quality scores of models from different pairwise threading alignments combined with a structural alignment of all templates to calculate consensus alignments for each template.
8. Ranking. The final template ranking is based on the predicted TM-Score from a DNN with input features from all previous steps. This score, the TopThreader Score, has a Pearson's R^2 of 0.77 with the true TM-Score of the template.

A key difference between TopThreader and consensus methods such as the MULTICOM[41] or Zhang servers[42] is that consensus in TopThreader is calculated based on DNN-predicted template similarity (TM-Score) to the native structure and top-down structural comparison to the highest scoring template. This contrasts with traditional consensus approaches similar to those mentioned above, in which the frequency with which a fold is identified is the driving factor of the consensus decision. TopThreader therefore has the advantage that even if the majority of identified templates or alignments are wrong, it can find true templates and good alignments if the highest scoring template is correct. This selection scheme is a key advantage in cases where correlated threading results produce a bias toward the same false-positive templates or wrong alignments, as seen for the CASP target T0742 as well as for prospective modeling of the nisin resistance protein (NSR) from *Streptococcus agalactiae* (*Sa*NSR; see the Experimental Validation section). An analogous situation is found in protein model quality assessment, in which clustering methods (which determine the quality based on consensus between models) perform worse at selecting the best model; if this model does not belong to a cluster, a task single-model and quasi-single-model methods handle better.[43] In turn, the top-down approach is at a disadvantage if the highest scoring template does not have the correct fold, in which case a potentially correct fold could be discarded when being compared to the highest scoring template.

TopAligner. The use of information from multiple templates can improve model quality by increasing total target coverage or improving pairwise alignments between templates and the target by matching structural elements of different templates.[26] This improvement depends heavily on the quality of the templates and their similarity to each other, however. If the quality difference between the best template and other identified templates is large, including sub-par information from bad templates may decrease model quality or distort multiple alignments. Therefore, the TopAligner module calculates an ensemble of pairwise and multiple alignments using every possible combination of the top five compatible

https://dx.doi.org/10.1021/acs.jctc.9b00825
J. Chem. Theory Comput. 2020, 16, 1953−1967

(pairwise TM-Score > 0.5) templates. TopAligner uses eight different state-of-the-art programs for template−template alignment (Table S1) and all primary threader and consensus alignments from TopThreader for template−target alignment. Each pairwise template−target alignment is weighted both globally and locally according to the weights calculated by TopThreader from model quality assessment with TopScore, residue-wise lDDT to the best scoring pairwise-alignment model, and residue-wise sequence similarity between the target and template. A detailed description of TopAligner and its primary alignment programs can be found in the Supporting Information Texts T3 and T4.

TopBuilder. All alignments from TopAligner are modeled using the TopBuilder module, which is also used at the initial modeling stages of TopThreader. TopBuilder uses Modeller9[37] and the partial thread function of Rosetta[38] to construct models based on alignments and template structures. It includes algorithms for knot detection and elimination, multiple types of loop refinement selected automatically based on loop size, and four methods for model refinement.[44−47] A detailed description of TopBuilder can be found in the Supporting Information Text T5. By default, model refinement is done by side-chain repacking with RASP.[45]

TopScore. The ensemble of models generated by TopBuilder is evaluated using TopScore and TopScoreSingle.[39] Because TopAligner produces more alignments based on multiple templates, model selection with TopScore is, due to the use of clustering information, biased toward selecting a multitemplate model. As mentioned (see the TopAligner section), this bias can in some cases lead to worse models because of inclusion of information from worse templates. Therefore, it is key to consider both TopScore and TopScoreSingle when selecting models for refinement and model combination (see the TopRefiner section).

TopRefiner. Previous work[34,41,48] has shown that combining different templates or models can improve the accuracy of the final model. Previous work has focused on combining pairwise alignments,[41] extracting consensus restraints from templates,[42] or averaging models.[49] The TopRefiner module refines models using model quality assessment, model fragmenting, fragment recombination, template/model hybridization, and fragment-guided MD refinement in order to remove regions with predicted errors and combine good fragments into full-length models. Models are first selected from the TopAligner (top-ranked model for each template combination) and TopThreader (top five primary threader models and top five consensus models according to TopScore and TopScoreSingle) model ensembles. From these models, regions predicted to contain errors by TopScore or TopScoreSingle are removed, and the resulting fragments are recombined into improved models. After fragment recombination, the models are used to construct new structural alignments to all identified templates, from which hybrid models are built using Rosetta.[38] Finally, the best models from each of the previous steps of the refinement are selected and refined with Modrefiner,[46] followed by a second round of model fragmenting and recombination. The final model is selected as the highest ranked model in the largest cluster according to TopScore. A detailed description of TopRefiner can be found in the Supporting Information Text T6 and Figure S3.

DATASETS

Screening. To train the DNNs of TopThreader on a set of diverse structures and difficulties (with respect to low sequence identity), a screening protocol is used, in which a set of known structures are repredicted while removing templates with a sequence identity above a given cutoff. The sequence identity cutoffs were chosen as 90, 60, and 30%, respectively, to simulate trivial, easy, and difficult modeling situations. A detailed description of the screening can be found in the Supporting Information.

CASP Dataset. To evaluate how TopModel performs when compared to other automated methods in the field, the conditions of the CASP10, CASP11, and CASP12 experiments were approximated. By turning on the PDB submission date filter in TopThreader, templates submitted on the day of or after the submission of a CASP target are removed, a procedure similar in nature to the CAMEO experiments.[50] A CASP target was kept if it fulfills three criteria: (1) the target native structure must be submitted to the PDB while writing this article, allowing for comparison between the model and native structure, (2) the target must not have been canceled during the CASP competition by the organizers, and (3) the sequence identity between the sequence released for prediction and the resolved native structure must be at least 50%. Applying these filtering criteria leaves 140 template-based targets and 46 free modeling targets (Table S3). It is important to note that this approximation will not yield the exact same results as if TopModel was run at the time of each CASP competition. Because threader and sequence databases have been updated since the respective competitions, quality scores (such as *e*-values and *Z*-scores) calculated by primary threaders, as well as primary feature predictions (such as secondary structure), will differ from what they would have been at the time of the competition. This can lead to hits that would have been identified with scores above significance cutoffs at the time of CASP competition but now have scores below the cutoffs for the updated databases. This effect is compounded by database clustering, in particular, for threaders that only return a fixed number of hits, of which a significant portion may be released too recently and thus removed by the filter. However, despite these approximations, it can serve as a useful indicator of structure prediction performance. None of the CASP targets were considered for the training of the TopThreader DNNs. This dataset will be referred to as the CASP dataset and is used as external evaluation of TopModel performance.

Experimental Validation. To evaluate the performance of TopModel on two de novo cases, we modeled the *Sa*NSR protein from the nisin operon of *S. agalactiae* (Uniprot ID A0A140UHB6)[51] before its release to the PDB and the LipoP from *Clostridium difficile* (Uniprot ID Q18BL3). These structures were then experimentally validated by crystallization[51] or by agreement with small-angle X-ray scattering (SAXS) and NMR data (see Experimental Validation).

RESULTS AND DISCUSSION

Evaluation of TopThreader. The aim of template selection with TopThreader is to retrieve a set of templates ranked according to their similarity (according to TM-Score) to the native structure. To evaluate how well this goal is achieved, we calculate the following: for each target in the CASP dataset (Table S3), the highest TM-Score between the

https://dx.doi.org/10.1021/acs.jctc.9b00825
J. Chem. Theory Comput. 2020, 16, 1953−1967

native structure and any template identified in the top five templates of any primary threader is calculated to find the best obtainable TM-Score given the primary threader results if template selection by TM-Score is perfect. Then, for TopThreader and each primary threader, the highest TM score of the top five ranked templates is compared to this best obtainable score. From this comparison, we calculate $\Delta TM_{100} = 100 \cdot (\max[TM_{\text{all templates}}] - \max[TM_{\text{top5 templates}}])$. Based on this ΔTM_{100} score for a given target, we define three categories: (I) the best template is found ($\Delta TM_{100} < 5$), (II) an adequate template is found (ΔTM_{100} [5−15]), and (III) no adequate template is found ($\Delta TM_{100} > 15$). We count the frequency of each category for each primary threader and for TopThreader for three subsets of the CASP dataset: (1) cases assigned by CASP organizers as TBM targets, (2) cases assigned as free modeling (FM) targets, and (3) all (TBM + FM) targets. The results are presented in Figure 2 (see Table S5 for numerical values).

The Ghent implementation of the Freeman−Halton exact test for 3 × 3 contingency tables[52] was used to determine the significance between the categorization of TopThreader and each primary threader in terms of three categories (I, II, and III) described above (see Table S6 for summarized normalized tables). Accordingly, all differences are highly significant ($p < 0.01$) for all cases showing a large and significant benefit to selecting templates with TopThreader over any of the tested stand-alone primary threaders.

The results in Figure 2 show that for the CASP subsets [(A) TBM, (B) FM, (C) TBM + FM], TopThreader identifies the best template (category I, blue) as one of the top five templates in 92, 56, and 83% of the cases, respectively. Furthermore, an adequate template (category II, yellow) is found in 4, 24, and 9% of the cases. An (at least) adequate template (according to TM-Score) is not identified in only 4, 20, and 8% of the cases (category III, red). It also becomes clear that for FM targets, it is more difficult to select the template with the best TM-Score (Figure 2B) because all primary threaders and TopThreader fail to identify the best template for ∼20% of targets. It is important to note, however, that for FM targets, most TM-Scores are close to or below 0.4 even for the best template and as such poorly reflect structural similarity in the first place, as two random structures will have a TM-Score of 0.17 when aligned.[40]

In addition to evaluating absolute performance for all top five templates, we evaluated the difference in template TM-Score of each of the top five ranked templates by normalizing the TM-Score of a template with a given rank to the template with that rank if the templates had been ranked according to true (rather than predicted) TM-Score. These normalized scores were then averaged, resulting in values closer to 1 corresponding to a ranking similar on average to a perfect ranking by true TM-Score rather than predicted TM-Score. The full results can be found in Table S7 and show that, in terms of ranking, TopThreader has a significantly better performance compared to the best primary threaders for TBM targets, with an average increase of 2% across all top five template ranks. For FM targets, a large improvement is seen for the top-ranked model (7%) and lower performance than primary threaders for subsequent ranks. This is surprising considering that templates for FM targets are close to or below the 0.4 TM-Score limit used by TopThreader to distinguish true from false templates, and because of CASP organizers, these targets should have no templates available. This suggests

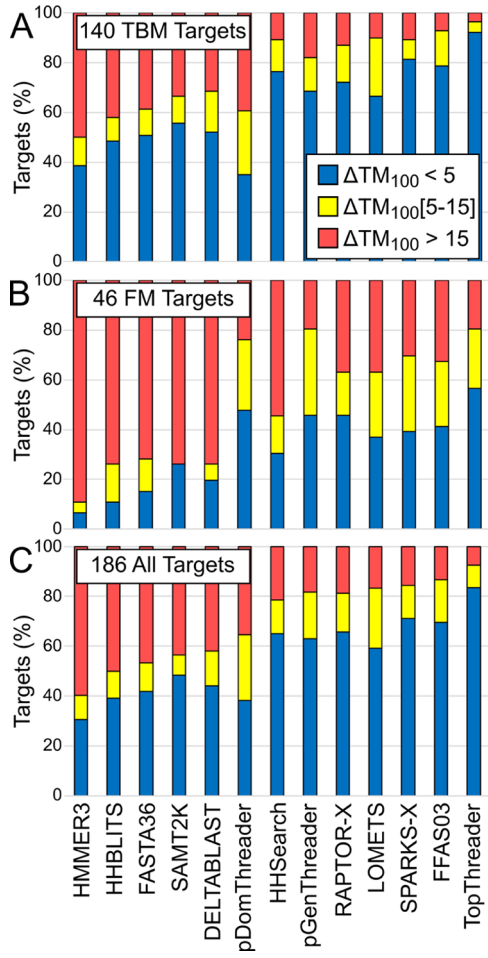


Figure 2. Template enrichment by TopThreader compared to primary threaders. Comparison of template selection performance on the CASP dataset. Performance is evaluated based on the ΔTM_{100} score, which evaluates the difference between the best of the top five ranked templates of a given threader and the best template found by any threader. For each target, three categories are selected: (I) the best template is found ($\Delta TM_{100} < 5$), (II) an adequate template is found (ΔTM_{100} [5−15]), and (III) no adequate template is found ($\Delta TM_{100} > 15$). The values represent percentages of targets in the CASP dataset for TBM (A), FM (B), and all (C) targets. Differences between TopThreader and the best primary threader for each subset are highly significant ($p < 0.01$) according to the Ghent implementation of the Freeman−Halton exact test for 3 × 3 contingency tables.[52] For numerical values, see Tables S5 and S6.

that even for extremely remote structural similarities, TopThreader is able to distinguish between low-quality templates and a random match to some degree, as is also shown in Figure 2B. The lower ranking performance for FM targets for ranks other than the top-ranked template is an effect of TopThreader requiring structural consensus between selected templates. Primary threaders do not require consensus and can therefore rank multiple incompatible folds highly. This gives a higher chance that one of the lower ranked templates is the best, while TopModel only finds the best template if it is either ranked at the top or is structurally similar to the top-ranked template.

Evaluation of TopAligner. To evaluate the effect of using TopAligner to sample alignments with different template combinations and alignment programs, we compared models built from primary threading alignments (TopThreader step 7) with models from TopRefiner stage 1, which are selected from

https://dx.doi.org/10.1021/acs.jctc.9b00825
J. Chem. Theory Comput. 2020, 16, 1953−1967

the TopAligner and TopThreader ensembles but without modifying the models themselves. Model quality is evaluated in terms of GDT_TS score,[53] which is used in CASP to evaluate model quality by comparing a model to the native structure and evaluates intermodel C_α atom distance conservation given different distance thresholds. We calculated the change in GDT_TS score between the two alignment ensembles. However, as we are interested in the relative change in model quality, we calculate the percentage-wise difference denoted as ΔGDT_TS. All models are built with TopBuilder and selected either with TopScoreSingle or according to the true GDT_TS score, and thus, only the alignment ensemble used to generate the models differ. There is no bias from the composition or size of the model ensemble because neither TopScoreSingle nor the true GDT_TS score depends on composition or size of the model ensemble. This allows us to compare the use of an ensemble of multitemplate and single-template alignments to the use of an ensemble of only single-template threading alignments. The results are shown in Figure 3A.

These findings indicate that sampling different alignments and combinations of templates using TopAligner in the majority of cases (56 and 82% of TBM and FM targets, respectively, if selected with TopScoreSingle) leads to little change in GDT_TS score. This result is expected, as for most targets, the different templates cover similar residues or are so similar that model quality is comparable. Furthermore, FM targets rarely have many similar templates identified by TopThreader because TopThreader requires all identified templates to have the same fold as the top-ranked template, which is rarely the case for FM targets. For TBM targets, using multiple templates leads to a decrease in GDT_TS score in 9 and 5% of cases if selected by TopScoreSingle or by best GDT_TS, respectively. This indicates that in a small number of cases, model quality decreases by using multiple templates, usually because of introduction of alignment errors when aligning poor templates with good ones. More importantly, however, for 22% of TBM targets, the GDT_TS score improves by 5−20%, and for 9% of targets, it improves by >20%. This shows an over 3 times higher chance that using TopAligner to sample different multitemplate alignments will increase model quality. These findings are in line with previous work, showing that using multiple templates and sampling alternate alignments can improve model accuracy.[41]

Evaluation of TopRefiner. TopRefiner has three aims: (1) selection of a small ensemble of good models built by TopThreader and TopAligner to be used for model combination and refinement, (2) combination of selected models to generate an ensemble of models converging on the correct fold, and (3) selecting the best possible model as the final TopModel prediction.

To evaluate the achievement of the first goal, we calculate ΔGDT_TS between the best model from the stage 1 ensemble and the best model achieved at any previous step of the TopModel workflow from any alignment of any template. We find that in just 6% of TBM targets, this distance is more than 5 GDT_TS units (26% for FM targets). The cases in which this distance is large are primarily those in which template selection with TopThreader fails to select the best template. This confirms that the models selected for refinement and model combination represent good models compared to those generated at earlier steps in the pipeline.

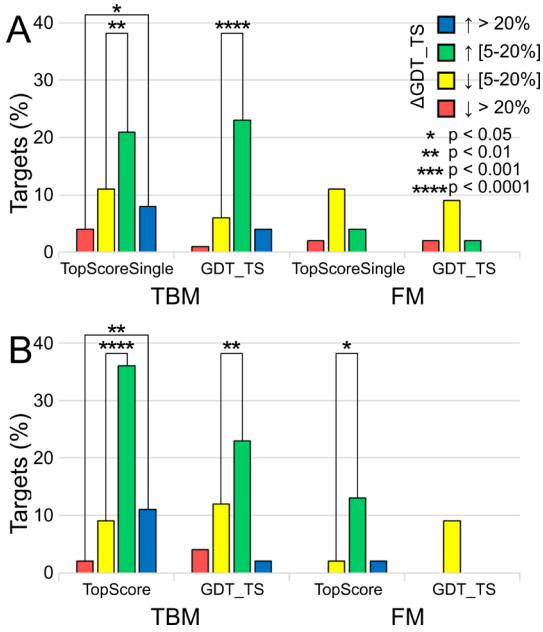


Figure 3. Impact of using TopAligner and TopRefiner on model quality. The relative change in GDT_TS score (ΔGDT_TS) is calculated by comparing a model selected before and after running TopAligner (A) or TopRefiner (B), respectively. (A) Difference in model quality when selected from a multi-/single-template model ensemble from TopAligner/TopThreader compared to selection from a single-template pairwise primary threader model ensemble. (B) Difference in model quality when selected from the first stage of TopRefiner (before refinement) compared to selection from the last stage of TopRefiner (after refinement). The models are selected either by true GDT_TS or by TopScoreSingle (A) or TopScore (B). Five categories are defined based on ΔGDT_TS: no change (ΔGDT_TS < 5%), small increase/decrease (ΔGDT_TS ↑/↓ [5−20%]; green/yellow), and large increase/decrease (ΔGDT_TS ↑/↓ > 20%; blue/red). The "no change" category is the most abundant and is not shown as it reflects no significant change in model quality. Significance is calculated using a one-tailed *t*-test between corresponding increase/decrease categories (blue-red and green-yellow, respectively). The null hypothesis is that the probability of model quality increase of a given amount (5−20 or >20% ΔGDT_TS) is the same as the probability of quality decrease by the same amount. Pairwise comparisons where this hypothesis can be rejected are indicated with brackets and corresponding p-values (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$, and ****: $p < 0.0001$). The number of samples used is the number of CASP targets in the TBM (140) and FM (46) categories.

To see how well the second goal is achieved, the models from TopRefiner stage 1, which are not refined but simply selected from the TopThreader/TopAligner model ensembles, are compared with models from TopRefiner stage 4, which is after refinement. If TopRefiner is successful, significantly more targets should see an increase in GDT_TS compared to those with a decrease. The result of this comparison is shown in Figure 3B and demonstrates that in 42% of TBM targets (92% for FM targets) ΔGDT_TS is <5%, indicating that in these cases, no significant change in GDT_TS score is observed, either because the starting models are too far from the true fold to be refined (most FM targets) or because the starting models are so close to the true structure that no improvement is seen (most TBM targets). However, for TBM targets, we find that there is a significant advantage of refinement, with over 2 times as many systems showing an increase in GDT_TS rather than

https://dx.doi.org/10.1021/acs.jctc.9b00825
J. Chem. Theory Comput. 2020, 16, 1953−1967

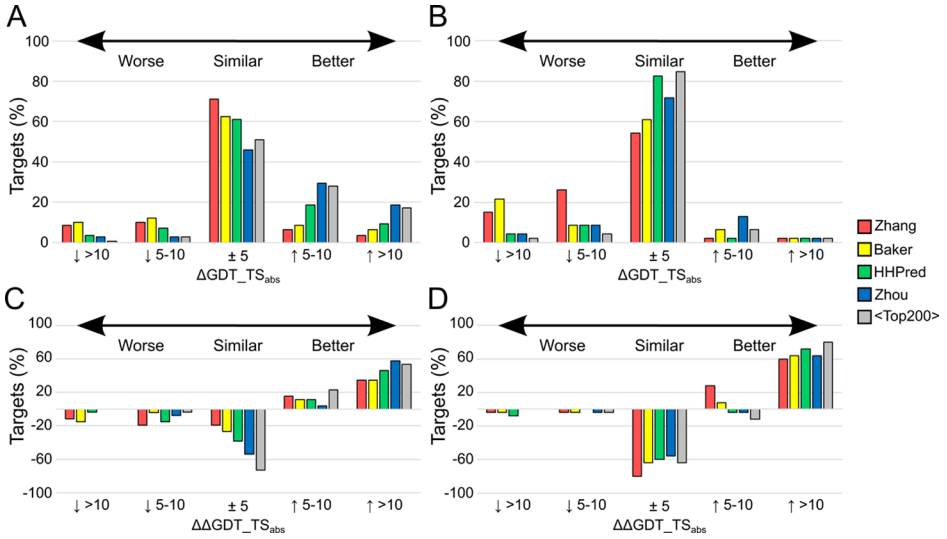


Figure 4. GDT_TS comparisons between TopModel and CASP servers. The bars represent comparison between TopModel and one of the four established CASP servers [the Zhang server (red), the Baker Server (yellow), the HHPred server (green), and the Zhou server (blue)] as well as the average of the top 200 server submissions for each target (gray). The Zhang server and Baker server both make use of ab initio folding and domain parsing, putting them at an advantage over TopModel. (A) ΔGDT_TS$_{abs}$ for CASP TBM targets indicates for how many of CASP TBM targets TopModel shows similar, worse, or better model quality than other established servers. (B) ΔGDT_TS$_{abs}$ for CASP FM targets indicates for how many of CASP FM targets TopModel shows similar, worse, or better model quality than other established servers. (C) $\Delta\Delta$GDT_TS$_{abs}$ for multidomain TBM CASP targets shows the change in the number of targets for which TopModel performs worse, similar, or better than established servers, if domain parsing, domain-wise modeling, and domain recombination were used. A large shift from worse/similar model qualities to better model qualities is seen. (D) $\Delta\Delta$GDT_TS$_{abs}$ for multidomain FM CASP targets shows the change in the number of targets for which TopModel performs worse, similar, or better than established servers, if domain parsing, domain-wise modeling, and domain recombination were used. A large shift from worse/similar model qualities to better model qualities is seen.

a decrease. It is interesting to see that model selection with TopScore shows a larger improvement than according to true GDT_TS. This shows that part of the benefit of refinement is an improved ability to select the best model, and not only an improvement of the models themselves, indicating that for many targets convergence to the native fold is a key part of refinement.

Comparison to CASP Stage 2 Models. To evaluate the performance of the entire TopModel pipeline, the final TopModel models from the CASP datasets are compared with the highest ranked CASP stage 2 models (CASP stage 2 consists of the top 200 automated server models for each target) from four established CASP servers: the Zhang[42] server (the best automated server in CASP8-13) and Baker server,[38] both of which use domain parsing and ab initio folding as part of their pipeline, and the HHPred[54] and Zhou[27] servers, which do not. Because TopModel has no ab initio folding module and does not parse the target sequence into domains, servers that include such methods are expected to be at an advantage. To evaluate the performance based on the part of the target structure that was solved experimentally, rather than the sequence submitted for prediction, only experimentally resolved residues were evaluated. For each target, the GDT_TS score was calculated for the final model produced by TopModel and the top-ranked model from each of the servers mentioned above, as well as the distribution of all server submissions in the stage 2 dataset. As we are interested in the absolute difference in model quality, rather than the relative increase in model quality, we classify each CASP target based on the difference in GDT_TS score (ΔGDT_TS$_{abs}$) between the final model from TopModel and the top-ranked model from each server. The results can be found in Figure 4A,B for TBM and FM targets, respectively.

As of now, TopModel has no domain parsing module to cut the input sequence into domains before modeling. Therefore, in the cases where multiple domains have good templates but no template covers the whole sequence, TopModel will match the best (often largest) domain template, leaving the other domains without a template. Therefore, TopModel is at a disadvantage for large multidomain targets for which no template is found that covers all domains. We expect this to be particularly detrimental for FM targets, most of which have multiple domains. To estimate the hypothetical performance that TopModel could achieve if multidomain targets were modeled domain-wise and combined in the correct way, the CASP domain annotations (released after the end of each competition) were used to parse the sequences of multidomain targets into their respective domains. Each domain was then submitted to TopModel separately, given the same restrictions as for regular targets to emulate previous CASP rounds. For each target, a weighted average (by the number of residues) of the GDT_TS scores of the respective domains is calculated as the hypothetical accuracy if domain parsing and combination was used. In the same way, the best ranked models of the servers used for comparison were parsed into domains, and the weighted average GDT_TS was calculated. This eliminates the relative orientation of domains as a factor for GDT_TS, and the difference in scores thus stems from the ability to fold the domains (largely determined by the identification of good templates). It is important to stress here that this analysis is done using a posteriori domain assignments, which were determined by experts with access to the known structures. Therefore, the performance increase of TopModel depicted in Figure 4C,D is merely an estimate of the potential upper bound of accuracy that could be obtained if perfect domain predictions and combination of domains into full-length

https://dx.doi.org/10.1021/acs.jctc.9b00825
J. Chem. Theory Comput. 2020, 16, 1953−1967

models were available. Given the difficulty of domain boundary prediction a priori, the boundary predictions of the other servers are unlikely to be perfect; as a result, they are at a disadvantage in this comparison.

Then we compare the GDT_TS score of models built from the CASP sequence released for prediction with this weighted average and evaluate the change in ΔGDT_TS_{abs} ($\Delta\Delta GDT\_TS_{abs}$). If $\Delta\Delta GDT\_TS_{abs}$ is positive, domain parsing improves model quality relative to other servers, and if negative, it deteriorates model quality. The results are depicted in Figure 4C,D for TBM and FM targets, respectively.

Our findings show that despite being at a disadvantage compared to the Zhang and Baker servers because of lack of domain parsing and ab initio folding, 71 and 63% of TBM target models from TopModel are of comparable quality to the Zhang and Baker servers, respectively, while 10 and 15% of TBM target models have a higher quality and 19 and 22% have a lower quality, respectively (Figure 4A). Compared to pure template-based servers such as HHPred and Zhou servers, on the other hand, TopModel has a clear advantage, with 28 and 48% of TBM targets having higher quality and 11 and 6% having lower quality, respectively. For FM targets, despite having no ab initio module, TopModel shows comparable accuracy to the Zhang and Baker servers for 54 and 61% of targets, respectively (Figure 4B) but a lower accuracy for 41 and 30% of targets, which is not surprising given the lack of ab initio folding and domain prediction in TopModel (most FM targets are multidomain targets). In terms of binary classification ($\Delta GDT\_TS_{abs} > 0$), TopModel produces better models for TBM targets than the Zhang and Baker servers in 46% of the cases for both and better models than the HHPred and Zhou servers in 60, and 71% of the cases, respectively. For FM targets, the corresponding values are 32 and 39% for Zhang and Baker servers and 46 and 51% for HHPred and Zhou servers, respectively. The trend where TopModel performs better than template-based servers HHPred and Zhou but worse than the Zhang and Baker servers due to their use of domain prediction and ab initio folding is also seen when dividing the CASP targets according to the highest sequence identity identified by TopModel (Table S8).

The results in Figure 4C,D show that a large improvement is possible for multidomain targets if the sequence is parsed into domains, predicted separately, and combined into a full-chain model. When compared to the Zhang server, for example, for TBM targets, the percentage of multidomain targets for which TopModel is worse than the Zhang server drops by 31 points, while the percentage of targets for which TopModel is better than the Zhang server increases by 51 points. For FM targets, the same trend is seen, with the percentage of worse models dropping by 8 points and the percentage of better models increasing by 88 points. Similar trends are observed for the other three investigated servers. This indicates that correctly parsing the input sequence into domains has a large impact on the quality of multidomain models, in particular, for FM targets. Note again, though, that because of the nature of the analysis using a posteriori-determined domain boundaries, these results should be considered merely as an outlook for the potential benefit of accurate domain parsing and not used for comparing TopModel performance with that of the other servers.

We speculate that the reason behind this is that accurately identifying a partially matching template for a large multidomain protein is difficult, especially for methods that have been trained to identify templates for single domains. As such, many FM targets may have been classified as such because of a failure to detect templates using the full sequence as a query and not because of an actual lack of templates. These results show that when properly parsed into domains and searching for each domain, template detection is easier and distant structural homologues become detectable for many targets that would traditionally be considered without templates. Thus, a large model quality improvement is achievable by predicting domain boundaries and combining the domains into a final model. However, in order to achieve such accuracy on prospective targets, accurate domain prediction and domain combination are required, which is therefore the focus of our future work.

The results in Figure 4A,B show that TopModel has comparable or better performance than the average server submission (gray) for the majority of targets (97% for TBM, 93% for FM) and performs significantly better than template-based servers without ab initio folding such as the Zhou and HHPred servers. TopModel even shows comparable or better performance than the Baker and Zhang servers for 82 and 78% of TBM targets and 59 and 70% of FM targets, respectively. These data show the benefit of using a top-down consensus rather than majority voting and the benefit of combining threading scores, model quality, and structural alignment using DNNs for ranking and selecting templates.

It is interesting to examine a case such as T0742 from CASP10. For this target, the vast majority of predictions from CASP servers, including the consensus-based MULTICOM server, fail to identify the best template (PDB ID 3TZG, identity = 14%, coverage = 70%, GDT_TS = 0.31) and instead predict a fold based on the wrong template identified by the majority of threaders. TopModel, however, identifies PDB ID 3TZG as the best template, a direct effect of its ability to discard wrong templates even when the consensus is indicating that they should be correct. A similar effect is seen for the prospective modeling of *Sa*NSR (see below).

Hard Cases. Although TopModel correctly folds most CASP TBM targets and has better template selection and alignment than any of its primary threaders (Figures 1−3), there are cases where it fails to predict the best template when comparing GDT_TS scores to those of other competing servers. Aside from the issues of simulating previous CASP rounds mentioned earlier, manual inspection indicated three main types of such cases where TopModel is at a disadvantage compared to servers such as those of Zhang and Baker.

First, for several targets, no template is found that covers all domains of the target. Based on CASP annotations released after the competitions, 39% of targets in the CASP dataset are multidomain targets (18% of TBM, 98% of FM). Additionally, there are several targets (including T0721, T0737, and T0755) that are nonconsecutive multidomain targets annotated as single domain by the CASP organizers, for which TopModel either is only able to match one domain (such as for T0755) or finds a slightly different and more favorable (better score from TopScore) conformation, resulting in a lower GDT_TS score (such as for T0922 and T0833).

Second, there are many targets (in particular, FM targets), for which the sequence submitted for prediction differs significantly from that of the resolved native structure. In most such cases, the native structure covers only a small fraction of the residues submitted for prediction. This makes structure prediction much more difficult because threading

https://dx.doi.org/10.1021/acs.jctc.9b00825
J. Chem. Theory Comput. 2020, 16, 1953−1967

Table 1. Inspection of Hard TopThreader TBM Targets[a]

ID	template	threaders	identity↑ (%)	coverage↑ (%)	Initial Score↑	TopScore single↓	GDT_TS↑	TM↑
T0700	2OVR*	HHSearch	17	79	0.64	0.59	0.49	0.51
	3V7D (I)	HHBlits HHSearch	21	52	0.66	0.59	0.35	0.47
	1EZJ (F)	pDomThreader	21	60	0.46	0.49	0.21	0.22
T0812	4BQ2*	LOMETS	10	87	0.45	0.60	0.28	0.45
	1H6Y (I)	RAPTORX	12	60	0.51	0.60	0.12	0.20
	3LY6 (F)	RAPTOR-X	18	59	0.48	0.60	0.14	0.33
T0818	4HYZ*	HHBlits HHSearch	15	55	0.64	0.72	0.25	0.37
	3H51 (I)	SPARKSX FFAS03	13	80	0.68	0.69	0.14	0.34
	4CE4 (F)	LOMETS	13	91	0.56	0.62	0.15	0.25

[a]Summary of scores for the CASP TBM targets for which TopThreader fails to select the best templates. * indicates the best template according to the lowest GDT_TS for a model built from that template. (I) indicates the highest ranked template according to the Initial Score, which is a prediction of template TM-Score using only sequence-derived features from primary threaders. (F) indicates the highest ranked template according to the Filtering Score, which is a prediction of template TM-Score using both sequence-derived features from primary threaders and the predicted error in the resulting model according to TopScoreSingle. The GDT_TS and TM columns indicate structural similarity between the best model from a given template and the native structure (not the TM-Score of the template). The arrows ↑↓ indicate if a score gets better with increasing or decreasing values, respectively.

algorithms focus on templates that cover as much of the target sequence as possible, when in fact only a small fraction of it can be resolved. For these targets, servers that use domain prediction have an advantage as they mitigate the inherent threader bias toward high target coverage by cutting the sequence into predicted domains.

Third, there are a few cases in which TopThreader discards the best template for TBM targets as a false positive. Three such cases (T0678, T0700, and T0818) were identified. To examine these cases, the best template, the highest ranked template by the Initial Score, and the highest ranked template identified by the Filtering Score are compared to the native structure in terms of GDT_TS and TM scores. The results are shown in Table 1.

T0700. For this target, the best template (PDB ID 2OVR) is discarded because it scores much worse by TopScoreSingle, which lowers the Filtering Score. This shows that despite higher coverage, models built from such a template will not always exhibit a better model quality, and as such, selection by model quality alone does not guarantee that the best template is found.

T0812. For this target, the best template (PDB ID 4BQ2) has a lower Initial Score than both the false-positive templates PDB ID 1H6Y and PDB ID 3LY6. All three templates result in models with identical scores from TopScoreSingle. This shows that using scores from primary threaders alone does not guarantee that the best template is found.

T0818. For this target, the best template (PDB ID 4HYZ) has lower coverage and consequently also worse TopScoreSingle score than both the best ranked template according to the Initial Score (PDB ID 3H51) and Filtering Score (PDB ID 4CE4), leading to a false-positive template being selected because of higher coverage. This is similar to T0700 in the sense that a higher weight on the Initial Score would have led to a better model, but in this case, the template with lower coverage is better, unlike for T0700 and T0812 where the higher coverage templates are better.

Analyzing the few TBM cases for which TopThreader does not select the best template shows that template selection is a complex task and that no single feature is likely to result in a flawless prediction for every target. However, the performance of TopThreader (Figure 2) shows that taking features from both primary threaders and model quality into account using DNNs significantly improved template selection. We expect that using predicted residue−residue contacts can further improve the template selection to resolve such issues.

Prospective Prediction and Experimental Validation of *Sa*NSR. Because TopModel uses a different consensus methodology than other methods, it can potentially go against the majority of threading results and give a prediction better than any of its constituent predictors. To illustrate the effect of this kind of consensus, we prospectively predicted the structure of *Sa*NSR (PDB ID 4Y68) prior to experimental structure determination and submission to the PDB. *Sa*NSR is a member of the S41 protease family, which degrades the lantibiotic nisin, and thus contributes to the congenital resistance against nisin of *S. agalactiae*.[51]

We then compared the model from TopModel to the distribution of primary threader models in terms of how close each model is to the experimental structure (measured by GDT_TS) (Figure 5). The results reveal that models based on most of the primary threader alignments are of poor quality, with a median GDT_TS score of 38, while the model from TopModel is much more accurate, with a GDT_TS score of 55 and a C_α atom root-mean-square deviation (rmsd) of 3.1 Å. The main reason for the failing of primary threaders in this case is that in most available templates, there are one or more large domain insertions. This causes the majority (82%) of threaders to thread the N-terminal helix bundle sequence onto the wrong domain (see the SPARKSX example in Figure 5) because of low (≤16%) sequence identity, incorrectly folding it into a β-sheet domain. However, because this β-sheet domain is scored poorly by TopScore and TopScoreSingle, the helix bundle is recovered in the model from TopModel. There is a minority of primary predictor models (18%) that show a correctly aligned helix bundle N-terminal domain. However, these contain significant differences in other parts of the model and are with traditional majority voting consensus far outweighed by the incorrect alignments, showing the benefit of using a top-down consensus approach rather than majority voting.

Prospective Prediction and Experimental Validation of LipoP from *C. difficile*. Building on the previous successes, TopModel was used to prospectively predict the structure of the lipoprotein LipoP from *C. difficile* (*C. difficile*). The lipoprotein is encoded as the gene CD1348 in the genome of

https://dx.doi.org/10.1021/acs.jctc.9b00825
J. Chem. Theory Comput. 2020, 16, 1953−1967

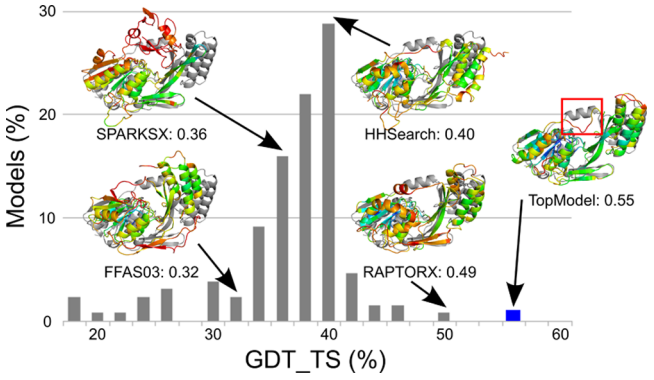


Figure 5. Prospective modeling of the NSR protein from *S. agalactiae*. The model quality distribution (in terms of GDT_TS score) of primary threader models for the NSR protein from *S. agalactiae* for prospective modeling before the release of the native structure (gray) to the PDB. The vast majority (82%) of models show an incorrectly threaded N-terminal domain (see the SPARKSX model). A minority of models (18%) show a correctly threaded helical domain (HHSearch, RAPTORX, and FFAS03) on a few templates, often with large errors elsewhere in the model (such as β-sheets shown in red). Because TopModel does not use majority voting, the model produced (blue box) is of far better quality (GDT_TS = 55) than those produced by primary threaders (median GDT_TS = 38), while majority voting consensus would produce a model in the middle of the distribution at a GDT_TS of ~38. Model examples from the different bins are colored according to the residue-wise lDDT score[55] to the native structure, with red showing incorrectly modeled regions and blue showing perfect agreement with the crystal. The largest error in the TopModel model is the fact that the residues linking the helical bundle with the catalytic core of the protein do not fold into an α-helix (red box). This is because no model from any of the primary predictors correctly folds these residues into a helix, and as such, TopRefiner has no fragment that it can select during model fragmenting and refinement, which would produce a helix for these residues; the secondary structure prediction by PSIPRED[56] also fails to identify these residues as helical.

C. difficile directly in front of the CprABC operon, which confers resistance against antimicrobial compounds such as the lantibiotic nisin.[57] Similar to the NSR protein, LipoP is likely involved in lantibiotic resistance, although its exact function remains unresolved to date. Database searches revealed that this protein is only present in the genus of *Clostridia*.

The templates identified by TopThreader (sequence identity in parenthesis; chain after the "_") are PDB IDs 5J7R_A (11%), 6GZ8_A (18%), 2JNV_A (18%), 5O5J_C (9%), and 3GKU_A (8%). Interestingly, the top-ranked structure PDB ID 5J7R is a putative lipoprotein from *Clostridium perfringens*, and as such, it is suggested to share the biological function with the homologue from *C. difficile*, despite the sequence identity being far below the 30% sequence identity limit generally considered the twilight zone for template-based structure prediction.[31] The final model quality predicted by TopScore was 0.35, indicating about 35% error in the model. This shows that the model may not be highly confident, which is expected given the low sequence identity and the fact that the first 43 residues (28% of the protein) of the N-terminus (termed the tail region) are unstructured and therefore highly mobile (a description of the tail region is available in the Supporting Information Text T7). To validate the model and identify errors, NMR experiments were therefore carried out to determine the secondary structure and β-strand pairing, and SAXS experiments were performed to estimate the shape and radius of gyration (R_G). A detailed description of the NMR and SAXS experiments can be found in the Supporting Information Text T7.

The initial model from TopModel has a good agreement with the secondary structure assignment and matches two out of three NOE β-strand pairings (strand 1/2 and strand 4/5) from NMR. The Matthews correlation coefficient (MCC) between the DSSP[58] secondary structure of the model from TopModel and the experimental assignment from NMR is 0.81 for β-strands, 0.68 for α-helices, and 0.66 for coil. However, there are still discrepancies between the predicted model and the experimental data. Four differences can be identified (Figure 6A,B): (1) α-helix 1 is eight residues shorter in the model than indicated by NMR, which is also indicated by TopScore showing the loop between α-helix 1 and β-strand 3 to have a high error (>50% residue-wise error). This is also the reason that random coil and helical MCCs are much lower than those for β-strands. (2) β-Strand 3 is indicated by NMR NOE measurements to be shifted by two residues, which produces a longer loop between α-helix 1 and β-strand 3, the loop indicated by TopScore to contain high errors. Because the

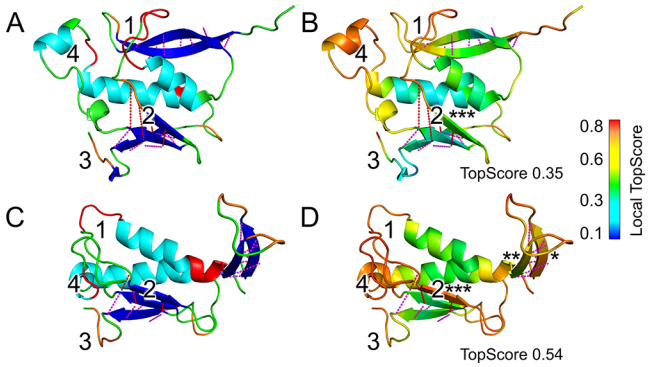


Figure 6. Prospective modeling of LipoP from *C. difficile* (disordered tail not shown for clarity). (A) Agreement of the TopModel model with secondary structure assignments and NOE restraints from NMR. Blue: β-sheet residues showing agreement between the model and NMR data. Orange: residues identified to be in a β-strand in NMR but not found so in the model. Cyan: α-helical residues showing agreement between the model and NMR data. Red: residues identified to be α-helical in NMR but not found so in the model. Magenta lines: experimental β-sheet NOE restraints showing agreement with the model. Red lines: experimental β-sheet NOE restraints showing a shift of two residue positions of β-strand 3 (**). (B) Model colored according to residue-wise TopScore. Yellow/red regions indicate regions with a high residue-wise error (>50%). C. Best model (according to TopScore) from primary predictors (dPPAS2 from the LOMETS server). The coloring scheme is the same as in (A). (D) Best model (according to TopScore) from primary predictors (dPPAS2 from the LOMETS server). The coloring scheme is the same as in (B) The shift of β-strand 3 (***) is only one residue in this model, placing two hydrophobic valines on the wrong side of the sheet and exposing them to the solvent. Furthermore, β-strands 1 (*) and 2 (**) are exposed to the solvent, exposing five hydrophobic isoleucines and one leucine to the solvent, all of which are buried in the TopModel prediction. Numbers 1−4 relate to the location of the errors described in the main text for panels (A,B) and corresponding locations in the best primary threader model in panels (C,D). In panels (A,B), these errors are caused by the fact that no template-based model from any primary predictor folds these regions correctly, which leaves TopRefiner unable to select a correctly folded fragment for these residues.

https://dx.doi.org/10.1021/acs.jctc.9b00825
J. Chem. Theory Comput. 2020, 16, 1953−1967

shift is of two residues, the hydrophobic valines in this sheet are still buried. (3) According to NMR, the C-terminus of the protein is folded into a β-strand, which most likely pairs up with the previous stand (β-strand 6), which in the model is too short by three residues. (4) α-Helix 2 scores poorly according to TopScore because of a difference in the helix length of one residue and its proximity to errors 1 and 2.

All of the differences in the LipoP model are due to the fact that TopModel is a template-based structure prediction method, which does not use ab initio folding. When none of the initial template-based models from pairwise or multi-template alignments produce correct structural fragments, TopRefiner is unable to select a fragment with the correct fold for such residues. A comparison of the final model from TopModel and the two highest ranked templates is shown in Figure S4 and illustrates this point. To correct differences such as these, ab initio folding is required in order to supplement the template-based model ensemble with models from ab initio folding and enable TopRefiner to select folded fragments not present in the templates. The Zhang server[42] had the same issue, which was remedied by the inclusion of ab initio models from QUARK.[59] It is important to note, however, that without the use of TopModel, the highest scoring model from primary threading alignments, generated by dPPAS2 from the LOMETS server, is of much lower quality (Figure 6C,D).

Because TopModel does not include any ab initio folding as of yet, we carried out molecular dynamics (MD) simulations for a total of 600 ns starting from the TopModel model, either using only the folded domain or the full-length sequence including the disordered tail, in an attempt to improve agreement with the available experimental data in terms of NMR secondary structure assignment and radius of gyration (R_g) from SAXS. The best snapshot from the globular domain simulations and the best snapshot from the full length model simulations were selected according to agreement with NMR and SAXS data, respectively. These two snapshots were combined with TopBuilder and energy-minimized to create a final full-length refined model (Figure 7). A detailed description of the MD simulations, the selection protocol, and the structural refinement can be found in the Supporting Information Text T7. The final refined model shows a secondary structure MCC of 0.81 for β-sheets, 0.88 for α-helices, and 0.78 for random coils. The propensity for each residue to be helical or β-sheet across all simulations can be found in Figure S5. The initial shape agreement with SAXS (see Table S9 and Figure S7 for experimental data) has a χ^2 of 49.8 (Figure S7A,B), which is high but not surprising given the highly mobile disordered tail. The model shows a radius of gyration of 26.7 Å, which compares favorably to the experimentally determined value of 24.3 Å (Table S9). Most interestingly, in the MD simulations, the loop between α-helix 1 and β-strand 3 shows some α-helix formation (see Figure S6 for a normalized distribution of secondary structure agreement with NMR across the MD simulations). After combining the two models agreeing best with NMR and SAXS using TopBuilder, we find that error (1) (Figure 7A,B) has been mostly corrected, in that α-helix 1 has been extended to nearly the same length indicated by NMR. None of the other errors were significantly impacted by the MD refinement; however, one cannot expect MD simulations to be able to fix alignment errors on the time scales applied (20 × 30 ns).

To explore if the high χ^2 with respect to SAXS data is caused by the disordered tail, a truncated version of the protein was

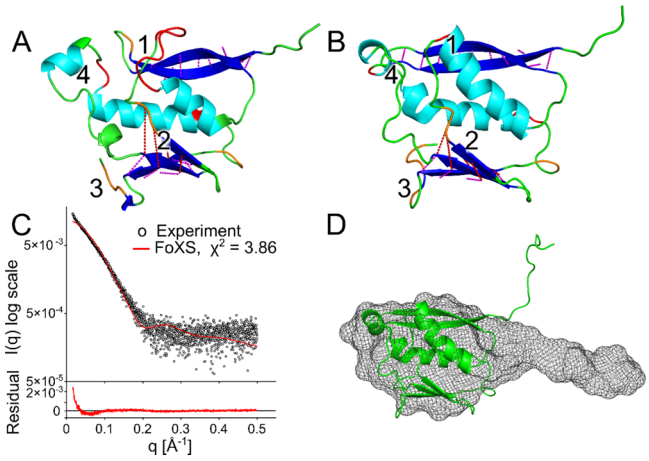


Figure 7. Model of LipoP from *C. difficile* after MD refinement and selection according to agreement with sparse experimental structural data. (A) Agreement of the TopModel model with secondary structure assignments and NOE restraints from NMR. The numbers indicate the location of errors, as described in the text previously. This panel is as in Figure 6A and again shown for ease of comparison here. Blue: β-sheet residues showing agreement between model and NMR data. Orange: residues identified to be in a β-strand in NMR but not found so in the model. Cyan: α-helical residues showing agreement between the model and NMR data. Red: residues identified to be α-helical in NMR but not found so in the model. Magenta lines: experimental β-sheet NOE restraints showing agreement with the model. Red lines: experimental β-sheet NOE restraints showing a shift of two residue positions of β-strand 3. (B) Agreement between the model after MD refinement, selection according to the agreement with experimental NMR and SAXS data, and combination with TopBuilder (see the main text and Supporting Information Text T7 for details) and experimental NMR data; colors are following the same scheme as in panel (A). The extension of α-helix 1 is seen. (C) Agreement between the experimental scattering data from SAXS (black) and simulated scattering curve of the MD model (red); FoXS[60,61] was used for simulating the scattering curve. The fit plots depict log-intensity vs q (Å^{-1}), and the residual plot shows the difference between experimental and computed intensity vs q (Å^{-1}). (D) Volumetric envelope of LipoP, as calculated from the scattering data using GASBOR,[62] is shown in gray mesh. The MD model of LipoP (green) was docked into the volumetric envelope using SUPCOMB.[63] Disagreement with SAXS is found mainly for the disordered tail of LipoP.

expressed, in which the first 30 of the 43 disordered tail residues were removed. When SAXS measurements of the truncated protein are compared to the full-length model after MD refinement and combination with TopBuilder (Figure 7A) from which the same tail residues were removed, the shape agreement increases markedly, as indicated by a drop in χ^2 from 49.8 to 3.9 (Figures 7C,D and S7), confirming that the initial disagreement with the full-length SAXS data is indeed caused by the high mobility of the disordered tail and that the shape of the folded domain shows a high agreement with the experiment.

In short, the modeling of LipoP from *C. difficile* clearly demonstrates the value of close interplay between computational structure prediction with TopModel and the use of sparse experimental structural data not only to validate and improve the predicted model but also to identify structural parts that still lack accuracy.

Preliminary Competition in CASP13. Despite TopModel development not being finished at the time of the

https://dx.doi.org/10.1021/acs.jctc.9b00825
J. Chem. Theory Comput. 2020, 16, 1953−1967

CASP13 competition, in particular, lacking most of the TopRefiner module, we decided to compete as a human server. The CASP12 and CASP13 competitions saw a huge impact of recent developments in ab initio folding in terms of highly accurate residue−residue contact and distance predictions, which also have had a large impact in template selection to remove false-positive templates. CASP13 also had the highest number of multidomain targets of any CASP competition to date, with some targets having more than 1000 residues. As such, we did not expect TopModel to rank very well compared to servers that utilize these tools such as the Zhang and A7D servers. TopModel showed a very good performance for several targets, however, most notably targets T1016-D1 (Figure 8A), T1014-D2 (Figure 8B), and T0964-D1 (Figure 8C), for which the models produced by TopModel were ranked second. Overall, our findings in CASP13 confirm our conclusions from our own benchmarking on the CASP dataset, in that while deep learning does improve template-based structure prediction, ab initio folding and domain prediction are required for folding large multidomain structures and structures without known templates.

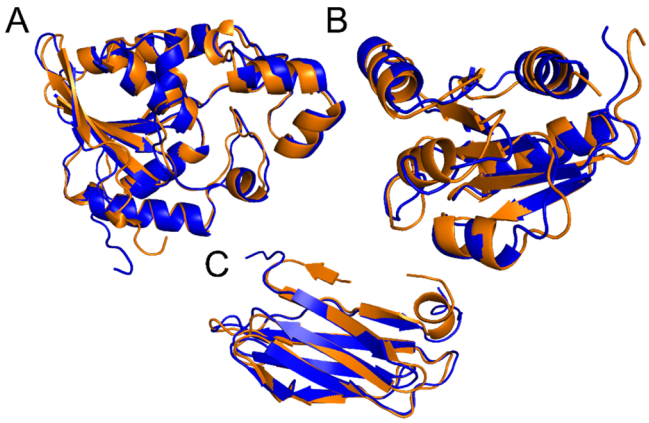


Figure 8. Examples of highly accurate structure predictions from TopModel in CASP13. (A) T1016-D1: A7D predicted the best model (blue) and TopModel predicted the second best one (orange) ($GDT\_TS_{TopModel}$ = 81.9, GDT_TS_{Best} = 85.4, C_α-$rmsd_{TopModel\_to\_Best}$ = 1.1 Å). (B) T1014-D2: McGuffin predicted the best model (blue) and TopModel predicted the second best one (orange) ($GDT\_TS_{TopModel}$ = 76.4, GDT_TS_{Best} = 76.7, C_α-$rmsd_{TopModel\_to\_Best}$ = 1.5 Å). (C) T0964-D1: MESHI predicted the best model (blue) and TopModel predicted the second best one (orange) ($GDT\_TS_{TopModel}$ = 78.7, GDT_TS_{Best} = 80.0, C_α-$rmsd_{TopModel\_to\_Best}$ = 1.6 Å). rmsd was calculated using the align function in PyMol.[64] The native structures were not released while writing this article.

CONCLUDING REMARKS

In this study, we introduced TopModel, a fully automated meta-method for protein structure prediction, which improves template-based threading beyond any of the 12 evaluated primary predictors. Instead of using majority voting during template selection and model averaging during refinement as other approaches,[41,42] TopModel uses top-down consensus and DNNs to select templates and identify and correct wrongly modeled regions. TopModel builds on numerous well-founded approaches to template-based structure prediction in terms of primary programs used for threading, alignment, model building, refinement, and model quality estimation. Yet, aside from the aspect of automation, TopModel offers several advantages over using these programs individually: we demonstrate a significant improvement for template selection and alignment accuracy because of sophisticated template selection with TopThreader, use of multiple alternate alignments between different combinations of templates with TopAligner, and model refinement with TopRefiner using TopScore and TopScoreSingle to detect wrongly modeled regions of the protein.

By applying TopModel to our CASP dataset, which includes targets from CASP10, 11, and 12 with released structures, we showed that TopModel consistently performs better than the average competing server and outperforms established template-based servers such as the Zhou and HHPred servers. Yet, we identified two areas in which TopModel currently falls short of state-of-the art predictors, mainly in terms of ab initio structure prediction and domain prediction for multidomain targets. As seen for top ranking servers in CASP12 and CASP13, such methods are required to be competitive for multidomain targets for which no template is available that covers all domains or for targets for which a correct template structure cannot be detected by threading.

Because of its meta-server composite nature, running TopModel usually takes 24−48 h on four cores for medium-sized proteins of 100−200 residues but may take up to a week for larger proteins. Parsing the input sequence into domains will clearly be beneficial here, as each domain can then be predicted in parallel, which will significantly decrease the total runtime. Proceeding that way is comparable to other folding methods such as the Zhang or Baker servers.

TopModel currently treats all proteins the same without special regard for specific classes such as transmembrane proteins, intrinsically disordered proteins, solenoid proteins, or coiled-coil proteins. These special classes exhibit traits, however, that can be predicted, such as transmembrane topology, intrinsic disorder, or coiled-coil regions. We intend to implement a meta-predictor of such protein characteristics to optimize the template selection performance of TopModel.

Early versions of TopModel have been applied to several systems, including enzymes,[2,4] ethylene receptors,[65] and restriction factors,[66] and yielded good predictions that agreed with experimental results and/or allowed for guiding of biochemical experiments. Here, we applied TopModel to predict the structure of the *Sa*NSR protein de novo; subsequent experimental structure determination by X-ray crystallography showed that TopModel predicted the correct fold even when the vast majority of primary threaders produced incorrect alignments and models. Finally, we used TopModel to predict the structure of LipoP, which showed good agreement with data from NMR spectroscopy and SAXS. The modeling of LipoP highlights the utility of the method and shows how the close interplay between computational structure prediction and sparse or low-resolution experimental data can be used synergistically to improve the final model.

Overall, we have shown that TopModel outperforms other stand-alone methods in the field with regard to template selection, template−target alignment, and model quality. However, TopModel is at a disadvantage when compared to black-box automated online servers, which utilize recent developments in residue−residue contact predictions, ab initio folding, and domain predictions. Therefore, we are focusing future work on contact prediction, ab initio folding, and domain prediction to improve the performance of TopModel for such targets. The TopModel suite is available as stand-

https://dx.doi.org/10.1021/acs.jctc.9b00825
J. Chem. Theory Comput. 2020, 16, 1953−1967

alone program from https://cpclab.uni-duesseldorf.de/index.php/Software. See Supporting Information Text T8 with respect to a description of the TopSuite content and disk space requirements.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.9b00825.

Detailed descriptions of TopThreader, TopAligner, TopBuilder, and TopRefiner; brief descriptions of the primary methods used as part of TopThreader and TopAligner; detailed description of the training and test datasets and the training of the DNNs and DNN features and architectures for TopThreader; targets in the CASP evaluation dataset; numerical data for the evaluation of TopThreader performance on this dataset in terms of template selection and template ranking compared to primary predictors; and detailed descriptions of the experimental validation, data collection, and MD simulation-based refinement for LipoP from *C. difficile* (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Holger Gohlke − *Institut für Pharmazeutische und Medizinische Chemie, Heinrich-Heine-Universität Düsseldorf, 40225 Düsseldorf, Germany; Institute of Biological Information Processing (IBI-7: Structural Biochemistry) & JuStruct and John von Neumann Institute for Computing (NIC) & Jülich Supercomputing Centre (JSC), Forschungszentrum Jülich GmbH, 52425 Jülich, Germany;* orcid.org/0000-0001-8613-1447; Phone: (+49) 211 81 13662; Email: gohlke@uni-duesseldorf.de; Fax: (+49) 211 81 13847

Authors

Daniel Mulnaes − *Institut für Pharmazeutische und Medizinische Chemie, Heinrich-Heine-Universität Düsseldorf, 40225 Düsseldorf, Germany*

Nicola Porta − *Institut für Pharmazeutische und Medizinische Chemie, Heinrich-Heine-Universität Düsseldorf, 40225 Düsseldorf, Germany*

Rebecca Clemens − *Institute für Biochemie, Heinrich-Heine-Universität Düsseldorf, 40225 Düsseldorf, Germany*

Irina Apanasenko − *Institut für Physikalische Biologie, Heinrich-Heine-Universität Düsseldorf, 40225 Düsseldorf, Germany; Institute of Biological Information Processing (IBI-7: Structural Biochemistry) & JuStruct, Forschungszentrum Jülich GmbH, 52425 Jülich, Germany*

Jens Reiners − *Institute für Biochemie, Heinrich-Heine-Universität Düsseldorf, 40225 Düsseldorf, Germany; Center for Structural Studies Heinrich-Heine-Universität Düsseldorf, 40225 Düsseldorf, Germany*

Lothar Gremer − *Institut für Physikalische Biologie, Heinrich-Heine-Universität Düsseldorf, 40225 Düsseldorf, Germany; Institute of Biological Information Processing (IBI-7: Structural Biochemistry) & JuStruct, Forschungszentrum Jülich GmbH, 52425 Jülich, Germany*

Philipp Neudecker − *Institut für Physikalische Biologie, Heinrich-Heine-Universität Düsseldorf, 40225 Düsseldorf, Germany; Institute of Biological Information Processing (IBI-7: Structural Biochemistry) & JuStruct, Forschungszentrum Jülich GmbH, 52425 Jülich, Germany;* orcid.org/0000-0002-0557-966X

Sander H. J. Smits − *Institute für Biochemie, Heinrich-Heine-Universität Düsseldorf, 40225 Düsseldorf, Germany; Center for Structural Studies Heinrich-Heine-Universität Düsseldorf, 40225 Düsseldorf, Germany*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.9b00825

Author Contributions

H.G. and D.M. jointly conceived the study. D.M. developed the method, performed computations, analyzed the results, and wrote the manuscript. N.P. performed the molecular dynamics simulations and analyzed the disordered tail. H.G. supervised and managed the project, secured funding and resources for the project, and revised the manuscript. R.C. prepared samples for SAXS. J.R. performed SAXS measurement and data analysis. R.C., L.G., and S.S. prepared the NMR samples. P.N. recorded the NMR experiments. I.A. and P.N. analyzed the NMR spectra. All authors reviewed and approved the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We are grateful to the developers of all primary programs used in this work for making their methods available as stand-alone to the scientific community. In particular, we are thankful to the developers of Phyrestorm for providing their clustering tree and the developers of MergeAlign2 for providing matrices used for multiple alignment with MergeAlign2. We acknowledge access to the Jülich-Düsseldorf Biomolecular NMR Center. This work was supported in part by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) − Projektnummer 267205415/SFB 1208 (project A03 to HG and B03 to PN) and Projektnummer 270650915 / Research Training Group GRK 2158 (project TP4a to HG and SS)-, and by the Bundesministerium für Bildung und Forschung (BMBF) − Förderkennzeichen 031L0182 / InCelluloProtStruct to H.G. We are grateful for the 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 H.G. on the supercomputer JUWELS at Jülich Supercomputing Centre (JSC) (user ID: HKF7). The Center for Structural Studies is funded by the Deutsche Forschungsgemeinschaft (DFG grant numbers 417919780 and INST 208/761-1 FUGG).

■ REFERENCES

(1) Rathi, P. C.; Höffken, H. W.; Gohlke, H. Quality matters: Extension of clusters of residues with good hydrophobic contacts stabilize (hyper) thermophilic proteins. *J. Chem. Inf. Model.* **2014**, *54*, 355−361.

(2) Widderich, N.; Pittelkow, M.; Höppner, A.; Mulnaes, D.; Buckel, W.; Gohlke, H.; Smits, S. H. J.; Bremer, E. Molecular dynamics simulations and structure-guided mutagenesis provide insight into the architecture of the catalytic core of the ectoine hydroxylase. *J. Mol. Biol.* **2014**, *426*, 586−600.

(3) Ingles-Prieto, A.; Ibarra-Molero, B.; Delgado-Delgado, A.; Perez-Jimenez, R.; Fernandez, J. M.; Gaucher, E. A.; Sanchez-Ruiz, J. M.;

Gavira, J. A. Conservation of protein structure over four billion years. *Structure* **2013**, *21*, 1690−1697.
(4) Gohlke, H.; Hergert, U.; Meyer, T.; Mulnaes, D.; Grieshaber, M. K.; Smits, S. H. J.; Schmitt, L. Binding region of alanopine dehydrogenase predicted by unbiased molecular dynamics simulations of ligand diffusion. *J. Chem. Inf. Model.* **2013**, *53*, 2493−2498.
(5) Yang, J.; Roy, A.; Zhang, Y. Protein−ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. *Bioinformatics* **2013**, *29*, 2588.
(6) Janin, J. Assessing predictions of protein−protein interaction: the CAPRI experiment. *Protein Sci.* **2005**, *14*, 278−283.
(7) Aehle, W.; Sobek, H.; Amory, A.; Vetter, R.; Wilke, D.; Schomburg, D. Rational protein engineering and industrial application: Structure prediction by homology and rational design of protein-variants with improved "washing performance"—the alkaline protease from Bacillus alcalophilus. *J. Biotechnol.* **1993**, *28*, 31−40.
(8) Cavasotto, C. N.; Phatak, S. S. Homology modeling in drug discovery: current trends and applications. *Drug discovery today* **2009**, *14*, 676−683.
(9) Roy, A.; Yang, J.; Zhang, Y. COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. *Nucleic Acids Res.* **2012**, *40*, W471.
(10) Roche, D. B.; Buenavista, M. T.; Mcguffin, L. J. The FunFOLD2 server for the prediction of protein−ligand interactions. *Nucleic Acids Res.* **2013**, *41*, W303−W307.
(11) Zhang, Y. Protein structure prediction: when is it useful? *Curr. Opin. Struct. Biol.* **2009**, *19*, 145−155.
(12) Altschul, S.; Madden, T. L.; Schäffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389−3402.
(13) Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **1999**, *292*, 195−202.
(14) Boratyn, G. M.; Schäffer, A. A.; Agarwala, R.; Altschul, S. F.; Lipman, D. J.; Madden, T. L. Domain enhanced lookup time accelerated BLAST. *Biol. Direct* **2012**, *7*, 12.
(15) Panchenko, A. R. Finding weak similarities between proteins by sequence profile comparison. *Nucleic Acids Res.* **2003**, *31*, 683−689.
(16) Rychlewski, L.; Li, W.; Jaroszewski, L.; Godzik, A. Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Sci.* **2000**, *9*, 232−241.
(17) Lobley, A.; Sadowski, M. I.; Jones, D. T. pGenTHREADER and pDomTHREADER: new methods for improved protein fold recognition and superfamily discrimination. *Bioinformatics* **2009**, *25*, 1761−1767.
(18) Soding, J. Protein homology detection by HMM−HMM comparison. *Bioinformatics* **2005**, *21*, 951−960.
(19) Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **2011**, *7*, No. e1002195.
(20) Remmert, M.; Biegert, A.; Hauser, A.; Söding, J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods* **2012**, *9*, 173−175.
(21) Madera, M. Profile Comparer: a program for scoring and aligning profile hidden Markov models. *Bioinformatics* **2008**, *24*, 2630−2631.
(22) Karplus, K. SAM-T08, HMM-based protein structure prediction. *Nucleic Acids Res.* **2009**, *37*, W492.
(23) Xu, D.; Jaroszewski, L.; Li, Z.; Godzik, A. FFAS-3D: improving fold recognition by including optimized structural features and template re-ranking. *Bioinformatics* **2013**, *30*, 660.
(24) Chakravarty, S.; Varadarajan, R. Residue depth: a novel parameter for the analysis of protein structure and stability. *Structure* **1999**, *7*, 723−732.
(25) Wu, S.; Zhang, Y. MUSTER: improving protein sequence profile−profile alignments by using multiple sources of structure information. *Proteins: Struct., Funct., Bioinf.* **2008**, *72*, 547−556.
(26) Peng, J.; Xu, J. RaptorX: exploiting structure information for protein alignment by statistical inference. *Proteins: Struct., Funct., Bioinf.* **2011**, *79*, 161−171.
(27) Yang, Y.; Faraggi, E.; Zhao, H.; Zhou, Y. Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates. *Bioinformatics* **2011**, *27*, 2076−2082.
(28) Zhou, H.; Zhou, Y. Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. *Proteins: Struct., Funct., Bioinf.* **2005**, *58*, 321−328.
(29) Fernandez-Fuentes, N.; Rai, B. K.; Madrid-Aliste, C. J.; Eduardo Fajardo, J.; Fiser, A. Comparative protein structure modeling by combining multiple templates and optimizing sequence-to-structure alignments. *Bioinformatics* **2007**, *23*, 2558−2565.
(30) Moult, J. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Curr. Opin. Struct. Biol.* **2005**, *15*, 285−289.
(31) Floudas, C. A.; Fung, H. K.; Mcallister, S. R.; Mönnigmann, M.; Rajgaria, R. Advances in protein structure prediction and de novo protein design: A review. *Chem. Eng. Sci.* **2006**, *61*, 966−988.
(32) Rychlewski, L.; Fischer, D. LiveBench-8: The large-scale, continuous assessment of automated protein structure prediction. *Protein Sci.* **2009**, *14*, 240−245.
(33) Wu, S.; Zhang, Y. LOMETS: a local meta-threading-server for protein structure prediction. *Nucleic Acids Res.* **2007**, *35*, 3375−3382.
(34) Wang, Z.; Eickholt, J.; Cheng, J. MULTICOM: a multi-level combination approach to protein structure prediction and its assessments in CASP8. *Bioinformatics* **2010**, *26*, 882−888.
(35) Moult, J.; Fidelis, K.; Kryshtafovych, A.; Schwede, T.; Tramontano, A. Critical assessment of methods of protein structure prediction (CASP)—round x. *Proteins: Struct., Funct., Bioinf.* **2014**, *82*, 1−6.
(36) Schaarschmidt, J.; Monastyrskyy, B.; Kryshtafovych, A.; Bonvin, A. M. J. J. Assessment of contact predictions in CASP12: Co-evolution and deep learning coming of age. *Proteins: Struct., Funct., Bioinf.* **2018**, *86*, 51−66.
(37) Webb, B.; Sali, A. Comparative protein structure modeling using Modeller. *Curr. Protoc. Bioinf.* **2014**, *47*, 5.6.1−5.6.32.
(38) Rohl, C. A.; Strauss, C. E.; Misura, K. M.; Baker, D. Protein structure prediction using Rosetta. *Methods in Enzymology*; Elsevier, 2004; Vol. *383*, pp 66−93.
(39) Mulnaes, D.; Gohlke, H. TopScore: Using Deep Neural Networks and Large Diverse Data Sets for Accurate Protein Model Quality Assessment. *J. Chem. Theory Comput.* **2018**, *14*, 6117.
(40) Zhang, Y.; Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **2005**, *33*, 2302−2309.
(41) Li, J.; Deng, X.; Eickholt, J.; Cheng, J. Designing and benchmarking the MULTICOM protein structure prediction system. *BMC Struct. Biol.* **2013**, *13*, 2.
(42) Zhang, Y. I-TASSER server for protein 3D structure prediction. *BMC Bioinf.* **2008**, *9*, 40.
(43) Kryshtafovych, A.; Barbato, A.; Fidelis, K.; Monastyrskyy, B.; Schwede, T.; Tramontano, A. Assessment of the assessment: evaluation of the model quality estimates in CASP10. *Proteins: Struct., Funct., Bioinf.* **2014**, *82*, 112−126.
(44) Wang, Q.; Canutescu, A. A.; Dunbrack, R. L., Jr SCWRL and MolIDE: computer programs for side-chain conformation prediction and homology modeling. *Nat. Protoc.* **2008**, *3*, 1832.
(45) Miao, Z.; Cao, Y.; Jiang, T. RASP: rapid modeling of protein side chain conformations. *Bioinformatics* **2011**, *27*, 3117−3122.
(46) Xu, D.; Zhang, Y. Improving the physical realism and structural accuracy of protein models by a two-step atomic-level energy minimization. *Biophys. J.* **2011**, *101*, 2525−2534.
(47) Bhattacharya, D.; Nowotny, J.; Cao, R.; Cheng, J. 3Drefine: an interactive web server for efficient protein structure refinement. *Nucleic Acids Res.* **2016**, *44*, W406−W409.

https://dx.doi.org/10.1021/acs.jctc.9b00825
J. Chem. Theory Comput. 2020, 16, 1953−1967

(48) Cheng, J. A multi-template combination algorithm for protein comparative modeling. *BMC Struct. Biol.* **2008**, *8*, 18.
(49) Wallner, B.; Larsson, P.; Elofsson, A. Pcons. net: protein structure prediction meta server. *Nucleic Acids Res.* **2007**, *35*, W369−W374.
(50) Haas, J.; Barbato, A.; Behringer, D.; Studer, G.; Roth, S.; Bertoni, M.; Mostaguir, K.; Gumienny, R.; Schwede, T. Continuous Automated Model EvaluatiOn (CAMEO) complementing the critical assessment of structure prediction in CASP12. *Proteins: Struct., Funct., Bioinf.* **2018**, *86*, 387−398.
(51) Khosa, S.; Frieg, B.; Mulnaes, D.; Kleinschrodt, D.; Hoeppner, A.; Gohlke, H.; Smits, S. H. Structural basis of lantibiotic recognition by the nisin resistance protein from Streptococcus agalactiae. *Sci. Rep.* **2016**, *6*, 18679.
(52) Ghent, A. W. A method for exact testing of 2X2, 2X3, 3X3, and other contingency tables, employing binomial coefficients. *Am. Midl. Nat.* **1972**, *88*, 15−27.
(53) Zemla, A.; Venclovas, Č.; Moult, J.; Fidelis, K. Processing and analysis of CASP3 protein structure predictions. *Proteins: Struct., Funct., Bioinf.* **1999**, *37*, 22−29.
(54) Söding, J.; Biegert, A.; Lupas, A. N. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.* **2005**, *33*, W244−W248.
(55) Mariani, V.; Biasini, M.; Barbato, A.; Schwede, T. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics* **2013**, *29*, 2722−2728.
(56) Mcguffin, L. J.; Bryson, K.; Jones, D. T. The PSIPRED protein structure prediction server. *Bioinformatics* **2000**, *16*, 404−405.
(57) Clemens, R.; Zaschke-Kriesche, J.; Khosa, S.; Smits, S. H. Insight into two ABC transporter families involved in lantibiotic resistance. *Front. Mol. Biosci.* **2018**, *4*, 91.
(58) Kabsch, W.; Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**, *22*, 2577−2637.
(59) Zhang, W.; Yang, J.; He, B.; Walker, S. E.; Zhang, H.; Govindarajoo, B.; Virtanen, J.; Xue, Z.; Shen, H.-B.; Zhang, Y. Integration of QUARK and I-TASSER for Ab Initio Protein Structure Prediction in CASP11. *Proteins: Struct., Funct., Bioinf.* **2016**, *84*, 76−86.
(60) Schneidman-Duhovny, D.; Hammel, M.; Tainer, J. A.; Sali, A. Accurate SAXS profile computation and its assessment by contrast variation experiments. *Biophys. J.* **2013**, *105*, 962−974.
(61) Schneidman-Duhovny, D.; Hammel, M.; Tainer, J. A.; Sali, A. FoXS, FoXSDock and MultiFoXS: Single-state and multi-state structural modeling of proteins and their complexes based on SAXS profiles. *Nucleic Acids Res.* **2016**, *44*, W424−W429.
(62) Svergun, D. I.; Petoukhov, M. V.; Koch, M. H. J. Determination of domain structure of proteins from X-ray solution scattering. *Biophys. J.* **2001**, *80*, 2946−2953.
(63) Kozin, M. B.; Svergun, D. I. Automated matching of high- and low-resolution structural models. *J. Appl. Crystallogr.* **2001**, *34*, 33−41.
(64) Delano, W. L. *PyMOL*, 2002.
(65) Milić, D.; Dick, M.; Mulnaes, D.; Pfleger, C.; Kinnen, A.; Gohlke, H.; Groth, G. Recognition motif and mechanism of ripening inhibitory peptides in plant hormone receptor ETR1. *Sci. Rep.* **2018**, *8*, 3890.
(66) Zhang, Z.; Gu, Q.; Vasudevan, A. a. J.; Hain, A.; Kloke, B.-P.; Hasheminasab, S.; Mulnaes, D.; Sato, K.; Cichutek, K.; Häussinger, D. Determinants of FIV and HIV Vif sensitivity of feline APOBEC3 restriction factors. *Retrovirology* **2016**, *13*, 46.

https://dx.doi.org/10.1021/acs.jctc.9b00825
J. Chem. Theory Comput. 2020, 16, 1953−1967