JCTC Journal of Chemical Theory and Computation

pubs.acs.org/JCTC Article

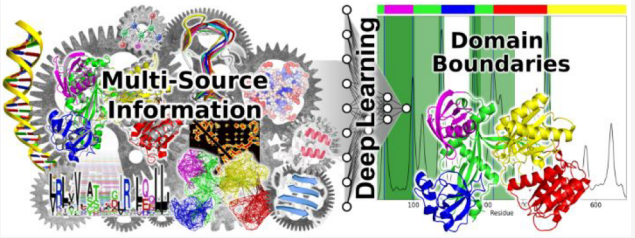
TopDomain: Exhaustive Protein Domain Boundary Metaprediction Combining Multisource Information and Deep Learning

Daniel Mulnaes, Pegah Golchin, Filip Koenig, and Holger Gohlke*

Cite This: *J. Chem. Theory Comput.* 2021, 17, 4599–4613 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Protein domains are independent, functional, and stable structural units of proteins. Accurate protein domain boundary prediction plays an important role in understanding protein structure and evolution, as well as for protein structure prediction. Current domain boundary prediction methods differ in terms of boundary definition, methodology, and training databases resulting in disparate performance for different proteins. We developed TopDomain, an exhaustive metapredictor, that uses deep neural networks to combine multisource information from sequence- and homology-based features of over 50 primary predictors. For this purpose, we developed a new domain boundary data set termed the TopDomain data set, in which the true annotations are informed by SCOPe annotations, structural domain parsers, human inspection, and deep learning. We benchmark TopDomain against 2484 targets with 3354 boundaries from the TopDomain test set and achieve F1 scores of 78.4% and 73.8% for multidomain boundary prediction within ±20 residues and ±10 residues of the true boundary, respectively. When examined on targets from CASP11-13 competitions, TopDomain achieves F1 scores of 47.5% and 42.8% for multidomain proteins. TopDomain significantly outperforms 15 widely used, state-of-the-art *ab initio* and homology-based domain boundary predictors. Finally, we implemented $\text{TopDomain}_{\text{TMC}}$, which accurately predicts whether domain parsing is necessary for the target protein.



INTRODUCTION

Knowing the 3D structure of a protein is important to understand its function[1] and modify its interactions[2] with small molecules or other proteins.[3] Consequently, protein 3D structure determination is a key part of molecular biology and drug discovery. To resolve protein 3D structures, the most commonly used experimental methods are X-ray crystallography,[4] nuclear magnetic resonance spectroscopy (NMR),[5] and cryogenic electron microscopy.[6] However, these methods are time-consuming and costly, and especially large, multidomain, flexible, or transmembrane proteins are not easy to resolve.[7−10] Computational structure prediction is faster and cheaper than experiments, but since it uses experimental data as a foundation, it faces similar problems for predicting large multidomain or transmembrane proteins.[11,12]

Protein domains are independent units that function, evolve, and fold independently and are structurally stable.[13] An estimated 70% of all proteins are multidomain.[14] Multidomain proteins are more abundant in eukaryotes than prokaryotes[15,16] because multiple domains give proteins an evolutionary advantage for folding reliability,[17] structural stability, and new complex functions.[18,19] Thus, studying multidomain proteins is needed to understand and potentially modulate complex biological processes, such as signal transduction[20] and host−pathogen interactions.[21] Furthermore, experimental deletion of domains can reduce protein flexibility and improve protein solubility, making it easier to experimentally determine protein 3D structure if domain knowledge is available.[22] Finally, cutting multidomain proteins into domains makes it faster and more accurate to search through template databases and predict protein structures by computational methods.[23] Knowing the domain boundaries of a protein is therefore highly demanded.

However, even when structural information is available, determining boundaries between protein domains may be difficult, a problem known as domain parsing. The difficulties arise partly due to algorithmic limitations and partly due to the diverse evolutionary processes that lead to different domain architectures, including exon shuffling,[24] uneven crossover during sexual reproduction,[25] and gene copying.[26] These processes give rise to domain patterns such as sequential, inserted, and repeated domains. Accordingly, protein domain annotation databases such as CATH[27] and SCOPe[28] use a combination of human manual annotation, Hidden Markov Model (HMM) comparison, and structure-based methods to annotate domains. HMM-comparison methods[29−37] identify boundaries by domain conservation, but they are highly

Received: February 4, 2021
Published: June 23, 2021

https://doi.org/10.1021/acs.jctc.1c00129
J. Chem. Theory Comput. 2021, 17, 4599–4613

Table 1. Overview of TopDomain Methods[a]

method	input	sequence features	homology features	structure features	output	competitor
TopDomain	sequence	yes	yes	yes*	domain boundaries	ThreaDom
$TopDomain_{Seq}$	sequence	yes	no	no	domain boundaries	ConDo
$TopDomain_{Parse}$	structure	no	no	yes**	domain boundaries	DDOMAIN
$TopDomain_{TMC}$	sequence	yes	yes	no	parsing decision	–

[a]TopDomain takes a sequence as input and uses both sequence- and homology-based features, as well as structure-based features (* from templates), to predict domain boundaries. $TopDomain_{Seq}$ takes a sequence as input and uses only sequence-based features to predict domain boundaries. $TopDomain_{Parse}$ takes a structure as input and uses only structure-based features (** from the input structure) to predict domain boundaries. $TopDomain_{TMC}$ takes a sequence as input and uses templates and boundaries identified by TopDomain to predict whether parsing the sequence into domains is required based on the coverage of each interboundary sequence segment by a given template.

database-dependent, and not all boundaries appear between conserved sequence segments. Three-dimensional structure-based methods[38−40] use 3D coordinates to derive boundary predictions by measuring atomic connectivity between residue regions but rely on general rules fitting for most domains and in turn lack contextual evolutionary knowledge of the protein. Furthermore, since they use a static input structure, they cannot account for structural flexibility.

Historically, domain boundary predictors (DBPs) can be categorized as homology-based or *ab initio* methods and generally use CATH or SCOPe annotations or both as the target boundary labels to predict. For homology-based DBPs,[41−45] the principle is to search through known protein structure- and family databases using one or more HMM-comparison methods, PSI-BLAST searches, or threading algorithms. The boundary information is then mapped from the template to the target sequence using the resulting alignments. Although homology-based DBPs can perform well, their performance depends on the availability of structural homologues and the ability to accurately identify them. *Ab initio* DBPs fall into two groups: statistical methods and machine learning-based methods. Statistical DBPs[46−48] infer domain boundaries using statistical regularities of features such as domain size, residue propensity, and hydrophobicity distribution. Machine learning-based DBPs[49−54] predict boundaries by training models such as support vector machines, random forests, or neural networks on combinations of residue physiochemical properties, predicted structural properties such secondary structure, solvent accessibility, or residue contacts, and position-specific scoring matrices (PSSMs) generated by PSI-BLAST.[55] Although *ab initio* DBPs do not need to find a matching structure in a database to predict domain boundaries, their accuracy is generally lower than homology-based DBPs due to the lack of detailed structural information.

In the last few decades, many DBPs have been developed that vary in terms of domain boundary definition, methodology, machine learning algorithms, and training databases. Therefore, different methods perform differently for different kinds of proteins and domain boundaries. However, this diversity provides an ideal foundation for a metapredictor,[41,45] which, if correctly designed, can result in more accurate and stable predictions than any of its constituent DBPs. Here, we present TopDomain, an exhaustive metapredictor that combines over 50 different primary predictors to provide accurate domain boundary predictions. To train TopDomain, we developed a new domain boundary data set termed the TopDomain data set, which was annotated by structure- and evolution-based automatic methods and, additionally, extensive iterated manual annotation guided by deep learning. In Stage 1 of TopDomain, we extract multisource information from over 50 diverse primary predictors. In Stage 2, to reduce the feature space, we classify each residue based on its distance to a domain boundary using multiple deep neural networks (DNNs). In Stage 3, the reduced informative features of Stage 2 are used for a final DNN to estimate a boundary score using a Gaussian kernel, which is then smoothened and turned into binary predictions using peak detection. We benchmark our TopDomain methods against all widely used and available stand-alone DBPs and find far superior performance for all quality metrics. Hence, TopDomain can aid experimentalists and computational biologists in resolving or predicting large multidomain protein 3D structures by accurately cutting them into domains.

METHODS AND IMPLEMENTATION

Overview. There are four different TopDomain methods with different goals, inputs, and outputs. These methods are summarized in Table 1.

Domain Parsing Necessity. When presented with a new protein sequence for structure elucidation, the first question arising is "Is it necessary to parse the sequence into domains, or is there a good template available that covers all domains?" To answer this question, we made a predictor termed $TopDomain_{TMC}$ (TMC = Template Modeling Coverage). This predictor is built on the rule that parsing the sequence into domains is not required if a template is found that (1) covers at least 80% of every interboundary sequence segment and (2) has a predicted TM-Score,[56] a measure of similarity between a template and a reference structure, above 0.5, indicating correct global fold and domain orientation.[56] $TopDomain_{TMC}$ requires three components: (I) prediction of domain boundaries by TopDomain (see below) to evaluate the coverage of each interboundary sequence segment by a given template; (II) prediction of the TM-Score of each of the templates identified by each of the primary methods of TopDomain; (III) a decision method that returns 0 if cutting the sequence into domains is not required and 1 if it is required, based on the results of components I and II and the rule defined above. A detailed description of $TopDomain_{TMC}$ can be found in the Supporting Information (SI, text T1).

$TopDomain_{Parse}$. To have a fast prediction method for a known protein structure, we devised a fast and accurate parser for predicting domain boundaries if only the native structure is used as an input. For example, this would be useful if further simulation or calculations on a known structure is needed, but only for a specific domain. To achieve this, we trained a set of DNNs that predicts domain boundaries based on the output from DDOMAIN,[38] DomainParser2,[39] and SWORD,[40] three structure-based primary DBPs, which take the native structure as an input. These DNNs were trained in the same way as the other TopDomain methods (see next section), and we call the

https://doi.org/10.1021/acs.jctc.1c00129
J. Chem. Theory Comput. 2021, 17, 4599−4613

resulting predictor TopDomain$_{Parse}$. Since TopDomain$_{Parse}$ uses the native structure as an input, it cannot be compared to other DBPs since their input is the native sequence. Therefore, we compare TopDomain$_{Parse}$ to DNNs trained in the same way as DNNs for other primary predictors (see section DNN Input in Features) on the output of only DDOMAIN, DomainParser2, or SWORD.

TopDomain Architecture. TopDomain consists of three stages (Figure 1): (1) multisource feature generation from primary predictors (Stage 1); (2) multiclass classification of domain boundary distance using three different DNNs with different sliding window sizes of ±10, ±20, and ±40 residues (Stage 2); (3) boundary score regression using DNNs with a sliding window size of ±40 residues and binary boundary assignment using peak detection (Stage 3). Using two stages of DNNs allows us to (1) reduce the vast input feature space to its most informative features, (2) use different sliding window sizes to capture different types of boundary signals, and (3) improve boundary detection by oversampling the less common boundary classes during classification.

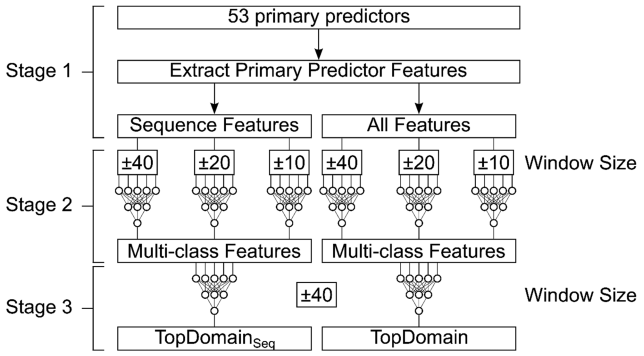


Figure 1. TopDomain architecture. In Stage 1, the target sequence is used as input for 53 primary predictors (see DNN Input section and SI T3 and T4) for multisource feature calculation. In Stage 2, the features are used as input for six DNNs with different sliding window sizes; three DNNs use only sequence-based features (TopDomain$_{Seq}$) and three DNNs use both sequence- and homology-based features (TopDomain). These DNNs each predict the residue distance to a boundary in terms of distance bins. In Stage 3, the output of the Stage 2 DNNs is used to train two regression DNNs; one uses the multiclass distance predictions from sequence-based features (TopDomain$_{Seq}$) and one uses the multiclass distance predictions from both sequence- and homology-based features (TopDomain). The boundary probabilities from TopDomain$_{Seq}$ and TopDomain are smoothened and used to assign discrete boundaries in the target sequence by peak detection.

To evaluate the impact of using homology-based information (i.e., templates and DBPs that search structure-derived databases) in addition to using sequence-based information alone, we trained two different predictors: TopDomain and TopDomain$_{Seq}$. TopDomain$_{Seq}$ uses only input features calculated from the target sequence and from methods that use the target sequence and search sequence databases, whereas TopDomain uses all available information (sequence- and homology-based). We expect that TopDomain will perform better than TopDomain$_{Seq}$, since this method utilizes all the available information, but TopDomain$_{Seq}$ may yield better predictions in cases where no reliable template information is available.

Data Sets. *TopDomain Data Set.* Because of the large set of primary predictors used in TopDomain, the data set for training and validation has to be of limited size to save computational resources during feature calculation but of high quality. Here, a brief outline of the data set generation is given; for a detailed description, see SI Text T2. The TopDomain data set is available at http://dx.doi.org/10.25838/d5p-19. We used the Astral SCOPe data set[28] as a starting point and clustered it to 20% identity for single-domain proteins and 70% identity for multidomain proteins. First, we carefully inspected structure-based predictions from DomainParser2[39] and DDOMAIN[38] and the original SCOPe annotations and used these as a basis for manually annotating all domain boundaries. Using these initial annotations as a target, a preliminary classification DNN was trained. Its predictions were used to inform the human manual re-evaluation and careful reannotation of the data set to rectify mistakes. This DNN was only used to highlight potential errors in the initial manual annotation, and not to annotate the data; all annotations were carried out manually by human inspection. The DNN was subsequently discarded and not used for any of the TopDomain methods to prevent any potential bias. This resulted in 3105 multidomain and 1035 single-domain proteins, which are manually annotated based on structure-based prediction, SCOPe annotations, and human inspection informed by DNN predictions. Finally, we used an in-house algorithm to split the data set into training (80%) and test (20%) parts, termed the TopDomain training set and TopDomain test set, respectively. No two proteins share more than 20% sequence identity between these two splits, but the data sets are otherwise similar in protein size, the number of boundaries, and prediction difficulty. During training, the TopDomain training set is split into 80% used for weight-adjustment of the DNNs and 20% used as validation to prevent overfitting using early stopping (see TopDomain Stage 2 section).

To emulate different prediction difficulties in terms of availability and quality of template information, homology-based DNNs (for TopDomain and primary predictors that rely on homology information) were trained on the TopDomain training data set using three upper cut-offs for sequence identity of homology-based information: 90%, 60%, and 30%. As such, the homology-based DNNs see different amounts of template information during training and learn to balance the use of sequence- and homology-based information. Each homology-based DNN thus uses the information from all three different sequence identity cut-offs (i.e., their training targets were three times more than for sequence-based DNNs). They cannot, therefore, learn to only rely on high-quality homology-based information (high identity template information) but have to balance sequence-based and homology-based features to get optimal performance across a wide range of template qualities.

For homology-based predictors, the final TopDomain training data set contains 9936 targets and 13425 boundaries: For each of the three different sequence identity cut-offs, there are 2486 multidomain proteins containing 4475 boundaries (75%) and 826 single-domain proteins (25%). For sequence-based predictors, the final TopDomain training data set contains 3312 targets and 4475 boundaries, since no sequence identity cutoff is imposed.

For homology-based predictors, the final TopDomain test data set contains 2484 targets and 3354 boundaries: For each of the three different sequence identity cut-offs, there are 619 multidomain proteins containing 1118 boundaries (75%) and 209 single-domain proteins (25%). For sequence-based predictors, the final TopDomain test data set contains 828 targets and 1118 boundaries, since no sequence identity cutoff is imposed.

https://doi.org/10.1021/acs.jctc.1c00129
J. Chem. Theory Comput. 2021, 17, 4599−4613

CASP Data Set. Furthermore, to validate TopDomain, $TopDomain_{Seq}$, and $TopDomain_{Parse}$ on an external data set, in which boundary definitions are not based on our careful manual curation, we used the CASP11, CASP12, and CASP13 proteins and the domain boundaries provided by the CASP organizers as the true domain boundary annotations. To emulate the availability of template information at the time of the CASP competitions, we impose a 30% sequence identity cutoff when running TopDomain, $TopDomain_{Seq}$, and $TopDomain_{Parse}$ on this data set. The data set is termed the CASP domain data set. Because of the more limited homology-based information, the higher number of single-domain proteins compared to the TopDomain data sets (47% and 25% respectively), and the greater difficulty of CASP targets, we expect the performance of all predictors on this data set to be lower than on the TopDomain test data set.

Features. To account for the large diversity of proteins and domain architectures, primary predictors are used for TopDomain that generate diverse information from multiple sources, which can nevertheless be classified into two overall types of input features. The first feature type consists of boundary predictions. These include homology-based predictions from ThreaDom,[41] InterProScan[57] (including all of its component primary predictors), DOMPRED,[42] and FIEFDom,[43] which are used by TopDomain but not $TopDomain_{Seq}$. Furthermore, they include *ab initio* predictions from ScoobyDomain,[47] PPRODO,[54] DOBO,[51] DROP,[49] DomPro,[52] DOMCUT,[46] ConDo,[50] and DeepDom,[53] used by TopDomain and $TopDomain_{Seq}$. The second feature type consists of predictions that can aid boundary prediction. These predictions include homology-based features and sequence-based features.

Homology-Based Features. One important feature type is derived from template information obtained by running various threading methods[58−63] on domain databases and structure databases.[28,44,64] The templates are then analyzed with DSSP[65] and the structure-based domain parsers DDOMAIN, DomainParser2, and SWORD.[38−40] These differ from homology-based primary predictors (e.g., ThreaDom) since the predictions from structure-based domain parsers show boundary locations in identified templates rather than the target protein. The accuracy of these features thus depends on the accuracy of the domain parser as well as the correctness of the identified template and the mapping between template and target. The resulting features (boundaries, secondary structure, solvent accessibility, and dihedral angles) are mapped back to the target sequence using the threading alignments. These features are not used by $TopDomain_{Seq}$. A detailed description of the homology-based features is given in SI Text T3.

Sequence-Based Features. In addition to homology-based features, sequence-based features calculated from the target sequence are used. The rationale behind the use of these features is that domain boundaries and protein domains are highly diverse. Therefore, information about the target protein will aid in the prediction of domain boundaries. The sequence-based features include PSSM features and position-specific gap propensities as well as predictions of generic structural features, such as secondary structure,[66−68] solvent accessibility,[67,69] dihedral angles,[67] residue disorder,[48,70−72] and residue−residue contacts.[73−79] They also include specific structural features such as transmembrane topology,[80−82] presence of signal peptides,[83] domain repeats and solenoid repeats,[84−86] and coiled-coil regions.[87,88] To account for inaccuracies in predicted features from a single method, multiple methods are used for each type of feature to allow the DNNs to learn from a diverse set of predictions. These features are used by both TopDomain and $TopDomain_{Seq}$. A detailed description of the sequence-based features is given in the SI Text T4.

DNN Input. In total, the DNN input comprises 208 features for TopDomain, 156 features for $TopDomain_{Seq}$, and seven features for $TopDomain_{Parse}$ derived from 24 homology-based, 29 sequence-based, and three structure-based domain parser DBPs. A detailed description of the conversion of features for DNN input is given in SI Text T5.

To compare the performance of TopDomain, $TopDomain_{Seq}$, and $TopDomain_{Parse}$ with the performance of the primary DBPs in a fair manner, DNNs were also trained for each primary DBP on the TopDomain training set. That way, the performance of a primary DBP becomes comparable regardless of the boundary definitions or data sets used when each respective DBP was created because the optimization in terms of DNN architecture, DNN training, cutoff estimation, and peak detection is identical for all DBPs. The only difference is the input features, which derive either from the output of a specific DBP or a combination of features (TopDomain, $TopDomain_{Seq}$ and $TopDomain_{Parse}$). We expect good primary DBPs to give informative outputs, which, when used as features for the DNNs, result in accurate predictions, and *vice versa*.

Deep Neural Networks. *TopDomain Stage 2.* There are two main goals of TopDomain Stage 2: First, it captures different types of boundary peaks using sliding window sizes of ±10, ±20, and ±40 residues. The different window sizes are motivated because some boundary signals may span many residues, for example, in long flexible or disordered linkers, but others may comprise only a few residues, for example, in a tight hinge between domains. The former can be well captured by a large sliding window and the latter by a small one. Second, TopDomain Stage 2 assigns each residue the probability to be in one of six classes based on the residue distance to a true boundary and minimizes the penalty of misclassification. Using multiclass classification is motivated by the improvement in residue contact prediction obtained by classifying distance bins rather than binary contact cut-offs.[89] The six boundary classes are boundary residues (distance of 0 residues), residues near boundaries (distance of 1−5, 6−10, 11−15, or 16−20 residues, respectively), and nonboundary residues (distance >20 residues). In total, this results in 18 output probabilities per residue from 3 window sizes and 6 boundary classes.

For TopDomain Stage 2, we use a Residual Network (ResNet), a type of deep convolutional neural network for image recognition.[90] In traditional deep neural networks, backpropagation requires the multiplication of many small partial derivatives, one for each layer. This effect causes the gradients of the loss function to shrink to zero for networks of many layers, causing them to stop learning. To resolve this vanishing gradient problem,[91] the ResNet architecture provides residual connections straight to earlier layers. Traditionally, an image is composed of an ordered array of pixels with different color channels. In TopDomain, each "image" is a sliding window of the target sequence, with each residue in the window serving as a pixel in this 1D image. For each residue, each one of the different features calculated by the primary predictors is stored as a "color channel".

To learn the probability of each boundary class, a ResNet with 18 layers, categorical cross-entropy loss function,[92] and a softmax activation function[92] of the last layer is used. The input shapes of the DNNs vary based on the DBPs and the window

https://doi.org/10.1021/acs.jctc.1c00129
J. Chem. Theory Comput. 2021, 17, 4599−4613

size. The hyperparameters, including the oversampling scheme, learning-rate scheduler, and early stopping criteria, are identical for all primary DBP networks as well as for TopDomain, $\text{TopDomain}_{\text{Seq}}$, and $\text{TopDomain}_{\text{Parse}}$. For further details on TopDomain Stage 2, see SI Text T6 and Table S2.

TopDomain Stage 3. The main goal of TopDomain Stage 3 is to detect boundary residues by prediction of a single boundary score for each residue. This is done by using the 18 output probabilities of Stage 2 as input features along with a 19th feature called the "Stage 3 filtering score". This score is a binary feature designed to separate putative boundary residues (distance ≤20 residues) from nonboundary residues (distance >20 residues) based on the nonboundary probability predicted by the Stage 2 DNNs. The score is based on cut-offs for each DNN (Table S1), calculated by maximizing the harmonic mean of the fraction of nonboundary residues above the cutoff and the fraction of boundary residues below it (eq 1).

$$\text{Goodness} = \left(\frac{N_{\text{Boundary}}}{N_{\text{Boundary-below-cutoff}}} + \frac{N_{\text{Nonboundary}}}{N_{\text{Nonboundary-above-cutoff}}} \right)^{-1} \quad (1)$$

The filtering score is one if the nonboundary probabilities of all three Stage 2 DNNs are below their respective cut-offs and zero otherwise. That way, the Stage 3 filtering score separates regions of the protein with significant Stage 2 boundary signals from regions without.

The Stage 3 DNNs are ResNets with 50 layers and a window size of ±40 residues to predict a single boundary score by regression.[93] For each residue, there are 1539 input features from the 18 + 1 features mentioned above times 81 residues in the window. The input feature space is mapped to one value for identifying a boundary and diminishing false-positive signals. This boundary score is defined by eq 2.

$$\text{Score}(r_i) = e^{-(D_i/k)^2} \quad (2)$$

$\text{Score}(r_i)$ is the boundary score of residue r_i in the sequence, D_i is the distance of that residue to the nearest true boundary, and k is the width of the kernel. We use a kernel size of 10 to avoid penalizing residues close to a boundary for being predicted as boundaries but still giving the DNNs an incentive to predict boundaries as close to the true position as possible. This differs from previous methods, where multiple residues near the boundary are considered equally probable (e.g., all residues within ±20 positions of a boundary or all linker residues between domains).[51] The Stage 3 DNNs were trained on 80% of the TopDomain data set as training data and 20% as validation data to prevent overfitting. All hyperparameters, including the oversampling scheme, learning-rate scheduler, and early stopping criteria, are identical for all primary DBPs as well as for TopDomain, $\text{TopDomain}_{\text{Seq}}$, and $\text{TopDomain}_{\text{Parse}}$. For further details on TopDomain Stage 3, see SI Text T7 and Table S2.

In order to smoothen the raw DNN output, we use Gaussian kernel smoothing with the same functional form as eq 1 and a kernel size of 3 residues, which is equivalent to a 1D convolution operation. To assign specific positions as boundaries, peak detection is used on the smoothened boundary score. The peak detection parameters selected are peak height and peak prominence (the distance between a peak and the nearest other peak). These parameter were optimized by a grid-search to maximize the F1 score of each predictor using the strict quality criterion (see section Quality Criteria) (Table S3). The boundary region is predicted as a function of peak height. This prediction is based on a logistic function fitted on the 1σ (68%) confidence interval of the distance between predicted and true boundary (Figure S2). For further details on peak detection and boundary region prediction, see SI Text T8. The peak detection and boundary region prediction is optimized identically for each primary DBP as well as TopDomain, $\text{TopDomain}_{\text{Seq}}$, and $\text{TopDomain}_{\text{Parse}}$.

$\text{TopDomain}_{\text{TMC}}$. The main goal of $\text{TopDomain}_{\text{TMC}}$ is to inform the user if parsing the input sequence into domains before structure prediction is required for accurate template-based structure prediction. This task requires two additional DNNs, the first of which predicts the TM-Score of a given template with respect to the native structure and the second of which uses the predicted TM-Score and the template coverage of the predicted interboundary segments based on the boundary prediction from TopDomain to predict whether parsing the sequence into domains is necessary.

The first task is accomplished by a multilayer perceptron regression network[92] since only a single score is needed for each template and the number of input features is limited. This network considers 11 input features for each template (sequence coverage, identity, and similarity between template and target and agreement between structural features measured in the template and predicted for the target, e.g. secondary structure, solvent accessibility, dihedral angles, and residue contacts). The target value is the TM-score,[56] which measures structural similarity between the template and the target structure ranging from 0 (mismatch) to 1 (perfect match). The network is designed with eight layers and uses the sigmoid activation function to scale the output between 0 and 1. As for other TopDomain DNNs, this network was trained on the templates from 80% of the TopDomain training data set and validated on the templates from the remaining 20% to prevent overfitting.

The second task is accomplished by a multilayer perceptron classification network with six layers and uses the binary cross-entropy loss function to classify the target as either requiring domain parsing or not. The final prediction is that domain parsing is not needed for accurate structure prediction if at least one template is classified as good enough. The templates that informed this decision are reported. Otherwise, the prediction is that domain parsing is required.

Method Availability. A stand-alone package for TopDomain alongside the installation of our other TopSuite server packages on which TopDomain depends is available at https://cpclab.uni-duesseldorf.de/topsuite/standalone_download.html.

Quality Criteria. The most common quality criterion for domain boundary prediction in literature is to classify predicted boundaries within 20 residues of the true boundary as correct.[41,51,53] Given this criterion, termed the literature criterion, the precision, recall, and F1 score for the boundary prediction is calculated. However, the literature criterion could potentially classify predicted boundaries as correct even if these boundaries were to place entire secondary structure elements on the wrong side of the boundary or cut secondary structure elements into pieces. Furthermore, this criterion can leave a large fraction of the protein as an acceptable boundary.

To estimate the impact of the quality criterion on the predictor performance, we designed a simple random boundary predictor called RanDom. This predictor only uses the sequence length as input. It predicts random boundaries in the sequence following only two rules: (1) Starting at 80 residues, place one boundary for every 80 residues in the protein up to a maximum

https://doi.org/10.1021/acs.jctc.1c00129
J. Chem. Theory Comput. 2021, 17, 4599−4613

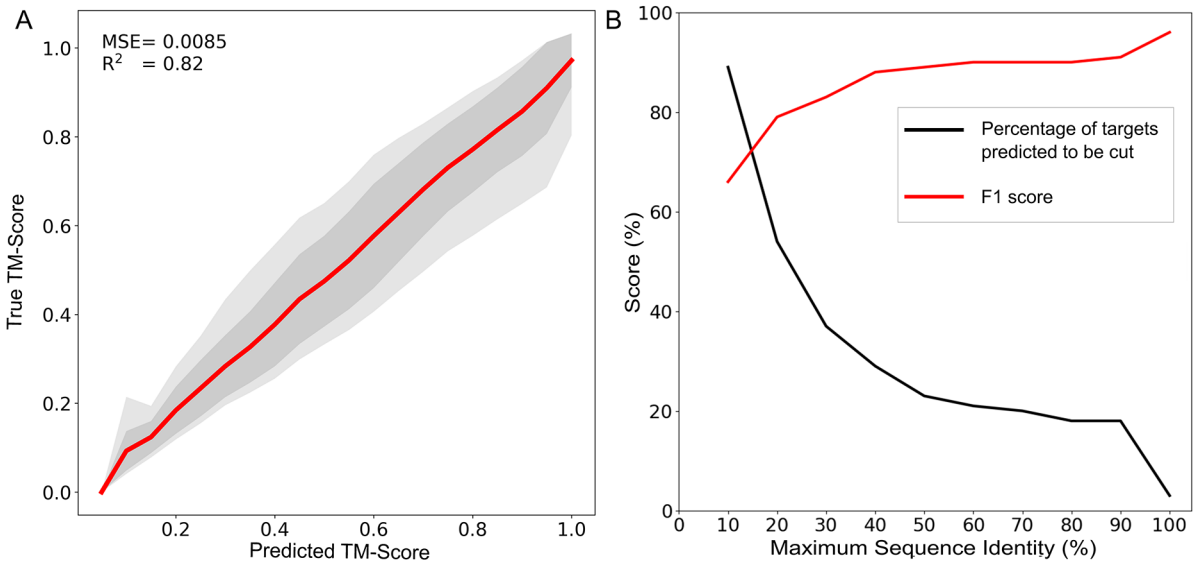


Figure 2. TopDomain$_{TMC}$ performance. (A) Correlation between predicted and true TM-Score for templates identified by primary threaders for proteins of the TopDomain test set. Predictions are binned, and the mean (red line) and asymmetric standard deviations (gray shaded areas) are calculated for each bin. One and two standard deviations above and below the mean are indicated in dark and light gray, respectively. (B) F1 score of TopDomain$_{TMC}$ decision function (red line) and percentage of targets predicted to require domain cutting (black line) plotted as a function of maximum sequence identity permitted for identified templates. TopDomain$_{TMC}$ constantly shows high performance for classifying whether the target sequence needs to be parsed into domains. As the identity of the identified templates declines, the likelihood increases that domain parsing is required. Both panels A and B are based on the TopDomain test set (2484 targets, 268214 templates).

of ten boundaries. (2) Do not place a boundary closer than 40 residues to an existing boundary or the termini of the sequence. We evaluated the performance of RanDom for different distance cut-offs on the TopDomain data set. Surprisingly, for the literature criterion, RanDom obtains a precision, recall, and F1 score of 19.8%, 44.4%, and 27.4%, respectively, which is better than some primary DBPs (Table S4).

Therefore, in addition to the literature criterion, we use the strict criterion. Here, a boundary is considered correctly predicted if placed within ±10 residues of the true boundary. At this cutoff, RanDom obtains a precision, recall, and F1 score of 10.2%, 22.8%, and 14.1%, respectively, which is again better than some primary DBPs (Table S5).

Finally, we use the residue distance between predicted and true boundary as a criterion for prediction quality. This is done to examine how close to the true boundary predictions from different DBPs can be expected to be, since boundaries closer to the true boundary minimize the chances of accidentally assigning secondary structure elements to the wrong domain or cutting secondary structure elements in pieces.

RESULTS

TopDomain Data Set. The TopDomain data set consists of two data sets, the TopDomain training data set (3312 proteins with 4475 boundaries) and the TopDomain test data set (828 proteins with 1118 boundaries). The relationship between the number of proteins and number of boundaries for different sequence lengths across the whole TopDomain data set can be seen in Figure S1. No two proteins in the TopDomain test data set or the TopDomain training data set share more than 20% sequence identity, but otherwise the two data sets are similar in terms of protein size distribution, distribution of number of protein boundaries, and prediction difficulty.

TopDomain$_{TMC}$ Performance. To evaluate the performance of TopDomain$_{TMC}$ to estimate the TM-Score of the templates identified by primary threaders, we calculate the Pearson's coefficient of determination (Pearson's R^2) and the mean-squared error (MSE) between the predicted and true TM-Scores of each template identified by primary threaders for the TopDomain test set (Figure 2A). The results indicate good performance for estimating the template TM-Score.

To evaluate the performance of the TopDomain$_{TMC}$, we evaluate the decision to cut the protein at the predicted domain boundaries or not (based on predicted boundaries and predicted template TM-Scores) compared to the true decision (based on true boundaries and true template TM-Scores). Most proteins in the data set do not require domain parsing (Figure 2B) since at least one reliable template with decent coverage and correct domain orientation tends to be identified. Thus, we used the F1 score to focus on the True Positive class (Parse) and weight precision and recall equally. We expect that as sequence identity gets lower and the templates become more distantly related to the target sequence, the decision performance will decline. The results of this analysis are shown in Figure 2B. These results show a stable and high performance for TopDomain$_{TMC}$, which declines only slowly with a dropping identity of the identified templates. Furthermore, as the identity of the identified templates declines, the likelihood increases that TopDomain$_{TMC}$ predicts that domain parsing is required. Finally, as expected, when no cutoff is imposed (100% sequence identity permitted), domain parsing is predicted for virtually no targets.

Stage 2 Performance. To evaluate the performance of the Stage 2 DNNs, we calculated the ability of the Stage 3 filtering score (a feature calculated from the Stage 2 DNN output) to focus on putative boundary residues (boundary distance ≤20 residues) compared to nonboundary residues (boundary distance >20 residues) and, thus, function as an informative feature for Stage 3. For each DBP, we calculated the retention (percentage of true boundaries with a filtering score of 1) and the reduction (percentage of residues with a filtering score of 0).

https://doi.org/10.1021/acs.jctc.1c00129
J. Chem. Theory Comput. 2021, 17, 4599−4613

The retention shows the recall a DBP would obtain if the filtering score were the decider of boundary prediction, while reduction shows how much the filtering score can focus the attention of the DNN to a small subset of the data. An informative feature will limit the boundary search space (high reduction) while maintaining a high potential recall (high retention). The results are shown in Figure 3. They demonstrate

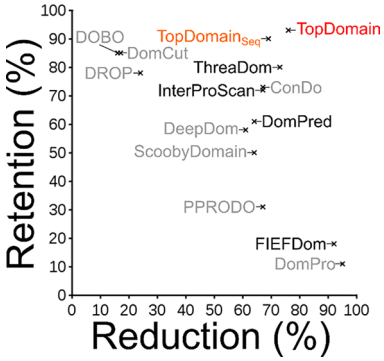


Figure 3. Stage 2 performance for different DBP on the TopDomain test data set. The *x*-axis shows the reduction (percentage of residues with a Stage 3 filtering score of 0) and, hence, the ability of the filtering score to focus on a small subset of the input data. The *y*-axis shows the retention (number of true boundaries with a filtering score of 1) and, hence, how well the filtering score identifies all true boundaries. Structure-based DBPs are indicated in black and sequence-based DBPs in gray. TopDomain is highlighted in red and TopDomain$_{Seq}$ in orange.

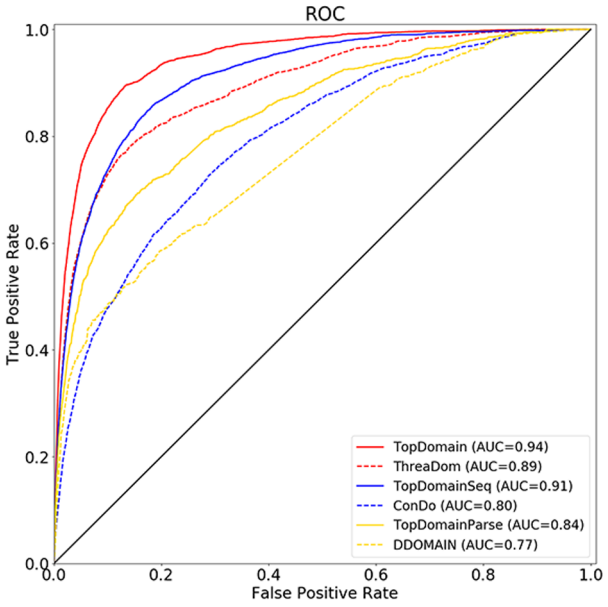


Figure 4. ROC of TopDomain and Primary DBP scores. This figure shows the receiver-operator characteristic curves for the best primary DBPs and TopDomain methods and the area under the curve (AUC) for each predictor. These scores reflect the ability of each DBP score to separate nonboundary residues from boundary residues. They do not, however, reflect how boundaries are assigned, since boundary assignment depends not only on the score of an individual residue but also on the height and prominence of the entire boundary peak. Homology-based predictors are shown in red, sequence-based predictors are shown in blue, and structural-based domain parsers are shown in yellow. The black diagonal line reflects a random boundary score, which has equal probability of assigning a residue as boundary and nonboundary. Performance is calculated for the TopDomain test set (1857 multidomain targets and 627 single-domain targets, 3354 boundaries).

that, for Stage 2, TopDomain has the overall best retention and a reduction smaller by only ∼10% than FIEFDom and DomPro; however, the latter two DBPs show the worst retention of all DBPs. TopDomain$_{Seq}$'s retention is almost as high as TopDomain's, with ∼10% smaller reduction.

Stage 3 Performance. To evaluate the performance of Stage 3 DNN boundary scores, we calculated the area under the receiver−operator characteristic curves of the boundary scores of TopDomain methods as well as primary DPBs (Figure 4; see Figure S3 for a plot of all DBPs). This analysis shows that scores of TopDomain methods outperform the best primary DBPs in terms of their ability to separate boundaries from nonboundaries. TopDomain$_{Seq}$ even slightly outperforms ThreaDom despite not using any homology-based information.

To evaluate the performance of the Stage 3 boundary prediction (which takes both the boundary score and the peak-detection performance into consideration), we calculated the Matthews correlation coefficient (MCC) for classifying a protein as single- or multidomain and precision, recall, and F1 score for boundary prediction in multidomain proteins. Proteins with any boundary predicted are classified as multidomain; all others are classified as single-domain. The scores are calculated for all primary DBPs and TopDomain predictors (TopDomain, TopDomain$_{Seq}$, and TopDomain$_{Parse}$). We compare the performance of each TopDomain predictor to the best primary DBPs for three categories: homology-based, sequence-based, and structure-based prediction (Figure 5, see Tables S4 and S5 for numerical values for all DBPs). The scores are calculated based on the literature criterion (Figure 5A, Table S4) and the strict criterion (Figure 5B, Table S5). When using the strict quality criterion, the performance gaps between TopDomain methods and primary DBPs are even more pronounced for boundary prediction in multidomain proteins (Figure 5B), indicating that TopDomain methods predict boundaries closer to the true position than other DBPs. Performances are evaluated both on the TopDomain test data set and the CASP data set.

As to boundary prediction in multidomain proteins, TopDomain significantly outperforms the best homology-based DBP ThreaDom, TopDomain$_{Seq}$ significantly outperforms the best sequence-based DBP ConDo, and TopDomain$_{Parse}$ significantly outperforms the best structure-based DBP DDOMAIN on the TopDomain test data set. TopDomain$_{Seq}$ even outperforms homology-based DBPs such as DOMPRED, InterProScan, and FIEFDom (Tables S4 and S5). It is also clear that the template information used in TopDomain significantly ($p < 0.0001$, McNemar's test[94]) improves performance compared to TopDomain$_{Seq}$. Both TopDomain and TopDomain$_{Seq}$ are more accurate than TopDomain$_{Parse}$, which may initially seem surprising considering that these methods do not have access to the 3D information in the native structure. However, TopDomain$_{Parse}$ lacks evolutionary information, which is apparently more severe since evolutionary processes cause domain boundaries.

On the CASP data set, performance drops for all predictors compared to the TopDomain test set. This is likely due to the limitations on homology-based information (template identity limited to 30%) and sequence information (CASP targets are difficult to predict also for *ab initio* methods due to fewer sequence homologues) for this data set.

https://doi.org/10.1021/acs.jctc.1c00129
J. Chem. Theory Comput. 2021, 17, 4599−4613

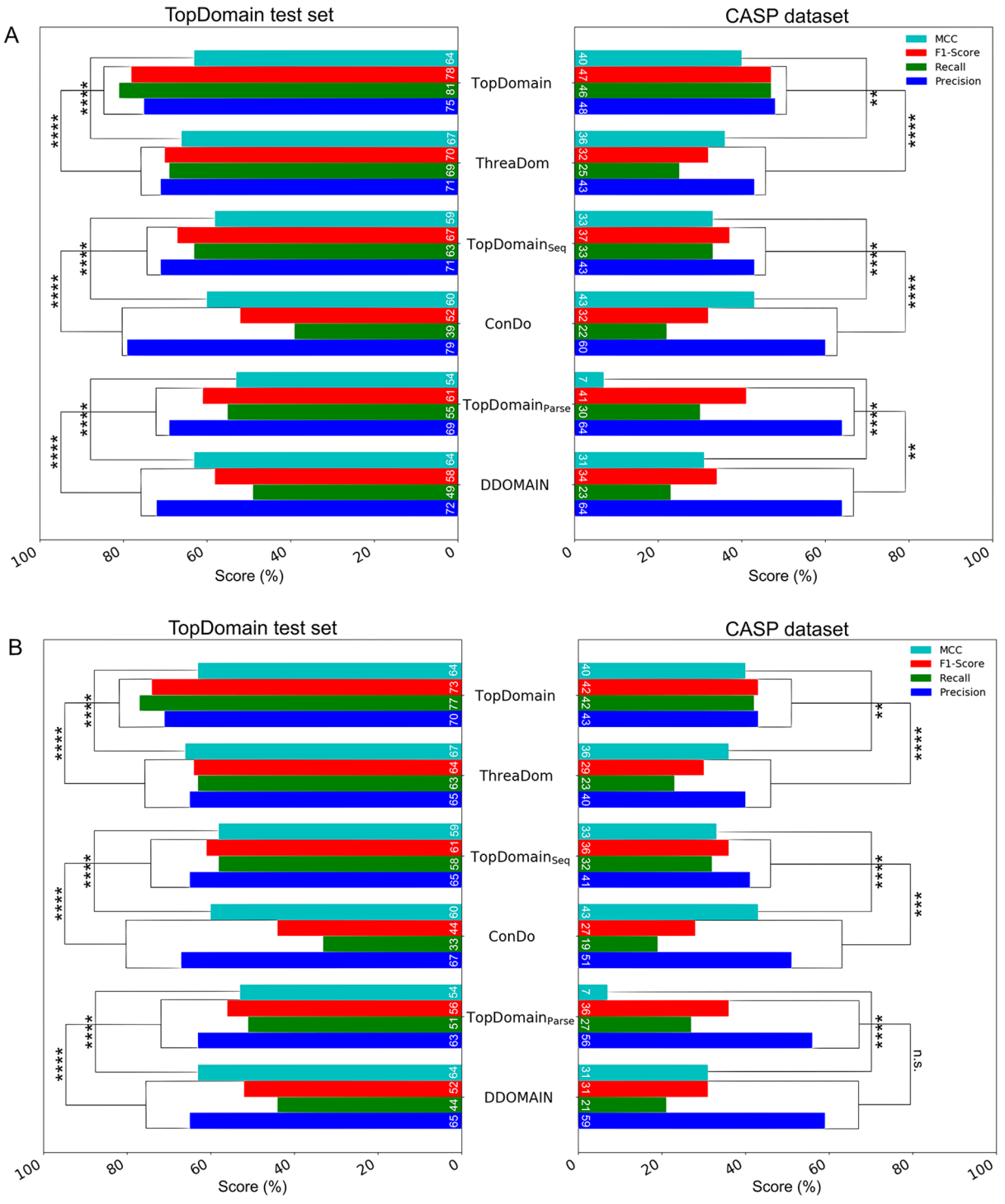


Figure 5. TopDomain Stage 3 performance classifying proteins as single- or multidomain and for boundary prediction in multidomain proteins. The performance is calculated using the literature criterion (boundary distance ≤20 residues) (A) and the strict criterion (boundary distance ≤10 residues) (B). The ability to classify single- vs multidomain proteins is evaluated in terms of MCC and shown in cyan bars. The boundary prediction performance is evaluated in terms of precision, recall, and F1 scores, shown in blue, green, and red bars, respectively. Performance is compared to the best primary DBPs (in terms of F1 score) for each category: TopDomain and ThreaDom (homology-based), TopDomain$_{Seq}$ and ConDo (sequence-based), and TopDomain$_{Parse}$ and DDOMAIN (structure-based). Performance is calculated for the TopDomain test set (1857 multidomain targets, 3354 boundaries) and the CASP domain data set (82 multidomain targets, 304 boundaries). Significant differences between TopDomain methods (TopDomain, TopDomain$_{Seq}$, TopDomain$_{Parse}$) and primary DBPs are calculated using McNemar's test[94] and indicated with brackets (*p < 0.05, **p < 0.01, ***p < 0.001, ****p < 0.0001).

For classifying single- and multidomain proteins, the TopDomain predictors are overall slightly (1−10 MCC points) worse than the best primary DBPs on the TopDomain test set, which applies especially for TopDomain$_{Parse}$. This result reflects that the TopDomain training data set contains mostly multidomain proteins (75% multidomain) to emulate the

https://doi.org/10.1021/acs.jctc.1c00129
J. Chem. Theory Comput. 2021, 17, 4599−4613

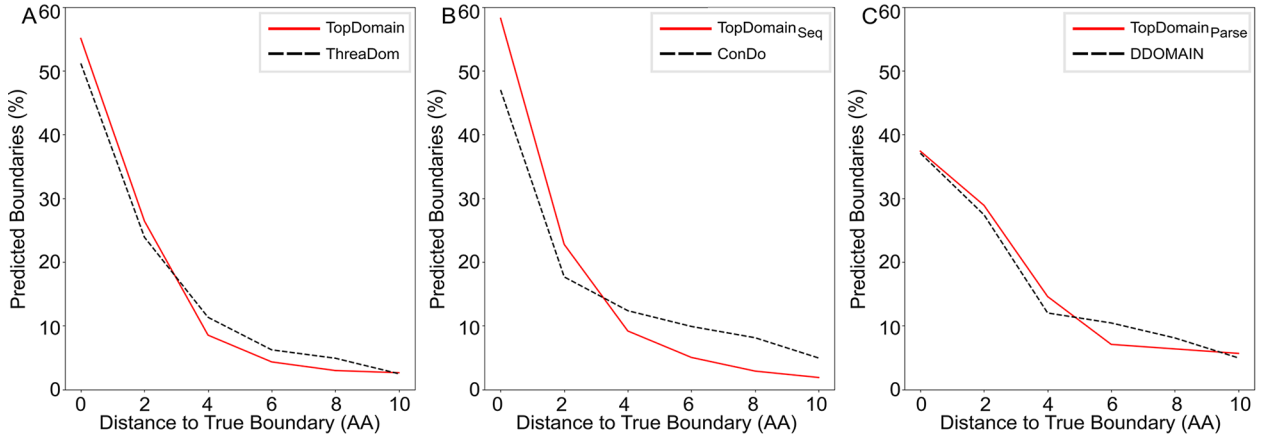


Figure 6. Distributions of the distance between predicted and true boundaries on the TopDomain test set. Homology-based DBPs (A), sequence-based DBPs (B), and structure-based DBPs (C). TopDomain, TopDomain$_{Seq}$, and TopDomain$_{Parse}$ show a shift toward lower distances compared to the best primary DBPs in each category. The number of true boundaries in the CASP data set was too few to give proper sampling in each distance bin for an informative analysis.

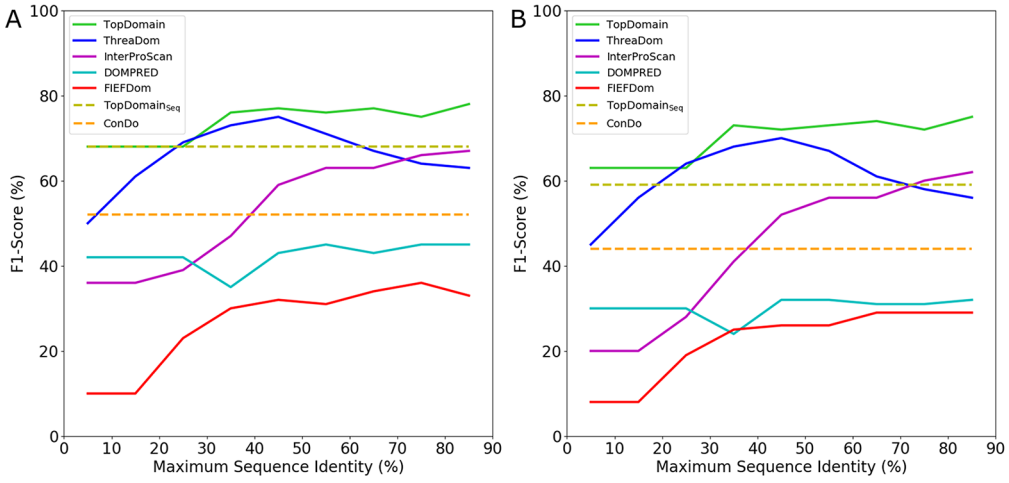


Figure 7. F1 score as a function of the highest sequence identity between a homologue and the target sequence as evaluated on the TopDomain data set. The scores are calculated based on the literature criterion (A) and the strict criterion (B). The F1 score of homology-based DBPs generally decreases as the maximum sequence identity drops. The performance of ConDo and TopDomain$_{Seq}$ are indicated as references, as these are independent of template information. TopDomain shows a much smaller decline than other DBPs, indicating a more robust performance regardless of the availability of homology-based information.

proportion of single- and multidomain proteins found in real life. The CASP data set (53% multidomain), on the other hand, and training data sets used by other primary DBPs are highly biased toward single domain proteins.[50,53,95] This difference in database compositions slightly biases TopDomain methods to predict boundaries in single-domain proteins. By contrast, DBPs like ConDo and DDOMAIN are highly biased toward not predicting any boundaries and classifying most proteins as single-domain, resulting in low recall values (Figure 5, Tables S4 and S5). Finally, as to applications in template-based structure prediction, given that TopDomain$_{TMC}$ confidently tells the user whether cutting a protein at the predicted boundaries is necessary (Figure 2), slightly overpredicting boundaries in single-domain proteins is much less of a concern than under-predicting boundaries in multidomain proteins.

For the primary DBPs, the high performance of ThreaDom is likely because it uses more sophisticated homologue detection (multiple threading algorithms) than InterProScan (HMM-Comparison and PSI-BLAST), DOMPRED (PSI-BLAST), and FIEFDom (PSI-BLAST). Similarly, the high performance of ConDo is likely because it is the only one of all the tested primary DBPs using coevolutionary information. There is a notable difference between boundary detection precision and recall for several primary DBPs. ConDo, DOMpro, and FIEFDom have very low recall compared to precision, indicating a high bias toward predicting single-domain proteins. In contrast, DOMCUT and PPRODO have low precision compared to recall, indicating a high bias toward overpredicting boundaries and classifying proteins as multidomain (Tables S4 and S5).

Distance to the True Boundary. Precision, recall, and F1 score give good estimates of the global quality of the boundary prediction. To evaluate the local quality of the predictions, we examined histograms of the distance between predicted and true boundaries (Figure 6). The results reveal not only that TopDomain, TopDomain$_{Seq}$, and TopDomain$_{Parse}$ perform globally better than the best primary DBPs but that boundaries predicted by them are also closer to the true boundaries.

https://doi.org/10.1021/acs.jctc.1c00129
J. Chem. Theory Comput. 2021, 17, 4599−4613

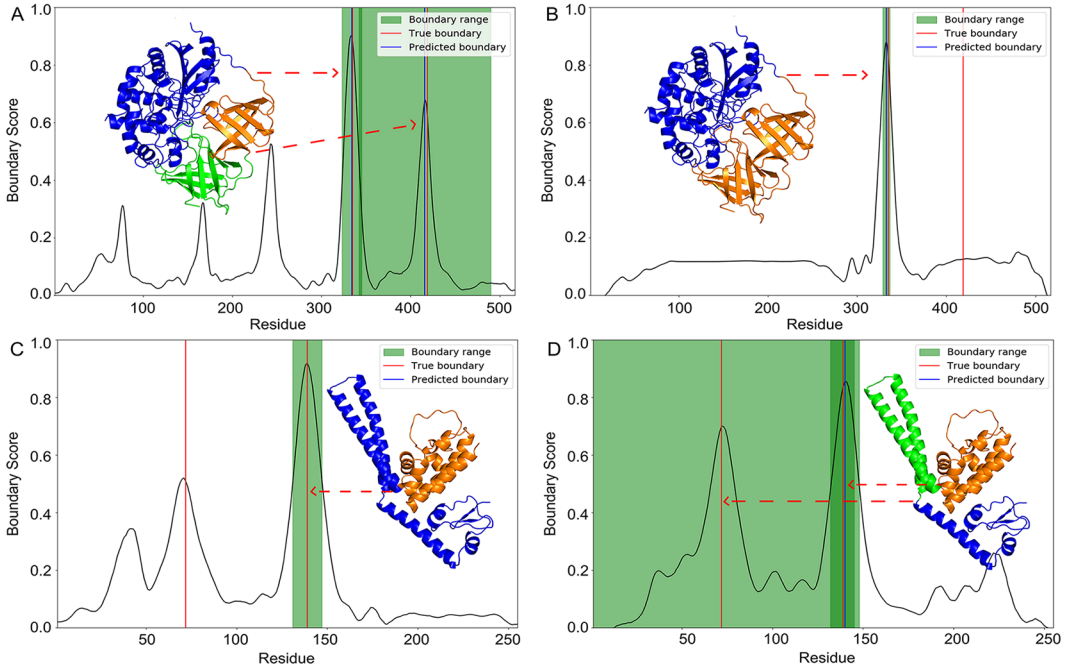


Figure 8. TopDomain and ThreaDom comparison. (A, B) 1EI5_A from the TopDomain test data set has three domains, including two repeat domains. TopDomain accurately identifies both boundaries due to two pronounced peaks (A). However, the second boundary has a large confidence range due to its low peak height. ThreaDom completely fails to detect this second boundary between the repeat domains (B). The boundary score of ThreaDom shows no peak for the missed boundary, indicating that the failure does not stem from failed peak detection. (C, D) 1NT2_B from the TopDomain test set has three domains separated by two boundaries. TopDomain is only able to identify one boundary since peak detection fails to detect the first peak (C). This failure is probably because this boundary is not found by any other primary DBP except ThreaDom. ThreaDom correctly identifies both boundaries since the peaks are pronounced enough for both (D). However, the first boundary has a large confidence range due to its low peak height.

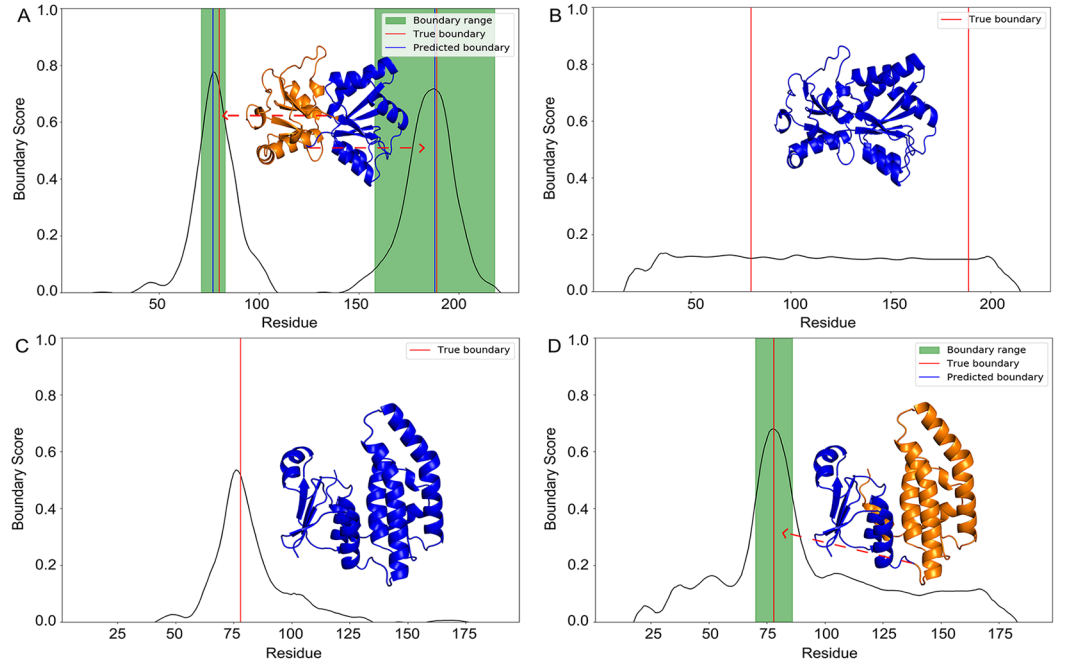


Figure 9. TopDomain$_{Seq}$ and ConDo comparison. (A, B) 1ATG_A from the TopDomain test set has two domains, a discontinuous one (blue) and an inserted one (orange). TopDomain$_{Seq}$ predicts both domains correctly due to two pronounced peaks in the boundary score (A). ConDo fails to predict any boundaries in this protein (B). The boundary score of ConDo shows no peaks for the missed boundaries, indicating that the failure does not stem from failed peak detection. (C, D) 5H5L_B from the TopDomain test set has two domains separated by one boundary. TopDomain$_{Seq}$ fails to identify this boundary since peak detection fails to detect this peak (C). This failure is probably because this boundary is not found by any other primary DBP except ConDo. ConDo correctly predicts this boundary since the boundary score peak is pronounced enough (D).

https://doi.org/10.1021/acs.jctc.1c00129
J. Chem. Theory Comput. 2021, 17, 4599−4613

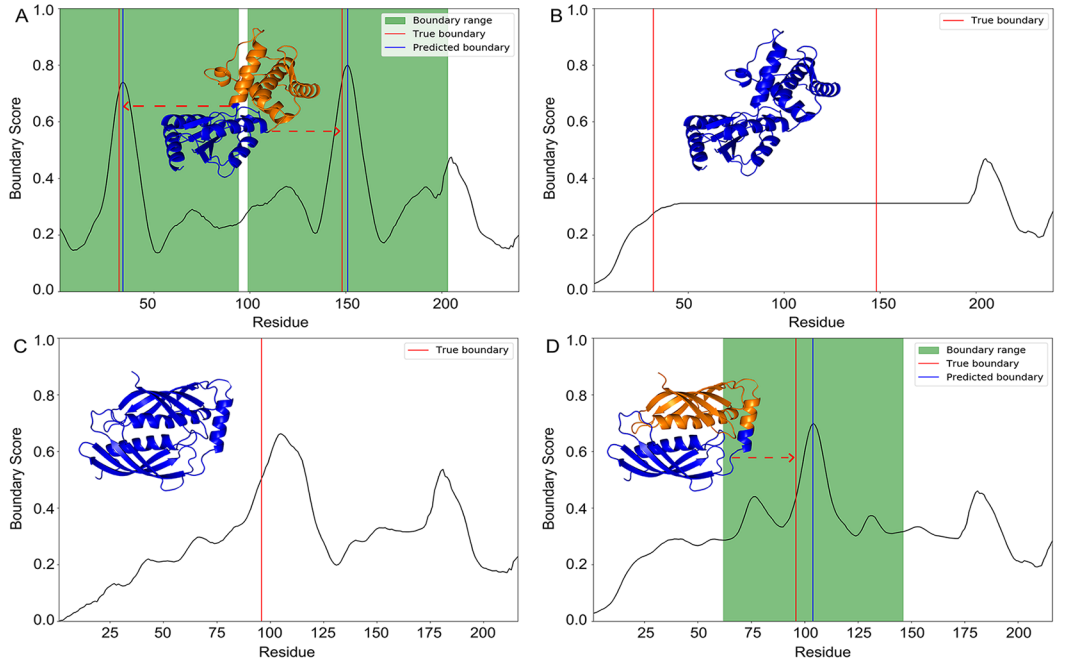


Figure 10. TopDomain$_{Parse}$ and DDOMAIN comparison. (A, B) 1DEK_A from the TopDomain test set has two domains, including a discontinuous one (blue) and an inserted one (orange). TopDomain$_{Parse}$ predicts both domains correctly. However, both boundary peaks have a large confidence range due to their low peak heights (A). DDOMAIN fails to predict any boundaries in this protein (B). The boundary score of DDOMAIN shows no peaks for the missed boundaries, indicating that the failure does not stem from failed peak detection. (C, D) 1WNH_A from the TopDomain test set has two domains separated by a single boundary. TopDomain$_{Parse}$ fails to identify this boundary since peak detection fails to detect this peak (C). This failure is probably because this boundary is not found by any other primary DBP except DDOMAIN. DDOMAIN correctly predicts this boundary since the boundary score peak is pronounced enough (D). However, the boundary has a large confidence range due to its low peak height.

Dependence on Structural Information. We expect that as the sequence identity to the closest template decreases and the detection of true templates becomes more difficult, the performance of homology-based predictors will decrease. To quantify this trend, we binned the predictions on the TopDomain data set according to the highest sequence identity obtained for each homology-based DBP. We then calculated the F1 score for homology-based DBPs for each bin and compared their performance to the F1 scores of the best sequence-based DBPs, TopDomain$_{Seq}$ and ConDo, which are independent of homology-based information. The results are shown in Figure 7.

TopDomain shows a much smaller decline than other DBPs, indicating a more robust performance regardless of the availability of homology-based information and the used quality criterion. ThreaDom has better performance for medium sequence identity ranges (30−50%) than for high ones, possibly resulting from originally being trained in the absence of highly homologous templates. Other homology-based DBPs show a decrease in performance as sequence identity declines, most notably for InterProScan.

Because TopDomain predicts boundaries at the local level, using a sliding window approach, the prediction performance is calculated at the per-boundary level rather than at the protein level and is independent of protein size or boundary number. We expect certain protein classes to be harder to predict than others (such as transmembrane proteins and intrinsically disordered proteins), likely because these protein classes are experimentally more difficult to resolve and thus more sparsely populated in structure databases. This, in turn, limits training data and makes homology detection more difficult.

Example Cases. To improve our understanding of the underlying reasons for the performance differences between TopDomain methods and the different primary DBPs, we examined many targets for which TopDomain predictions differ from the best primary DBPs. In particular, we examined the performance of TopDomain compared to ThreaDom (Figure 8), TopDomain$_{Seq}$ compared to ConDo (Figure 9), TopDomain$_{Parse}$ compared to DDOMAIN (Figure 10), and TopDomain compared to ThreaDom on CASP targets (Figure 11).

Overall, the examples illustrated in Figures 8−11 indicate the difficulty of boundary prediction. When TopDomain methods fail to identify boundaries, it is generally not because of a missing boundary score signal but because the peak is too weak for peak detection to assign boundaries given the optimized parameters (Table S3; Figures 8C, 9C, and 10C). This is often the case when a boundary is found only by a single primary DBP. On the other hand, TopDomain methods show superior performance precisely because the multisource information from the large number of features for TopDomain allows it to detect boundaries that are completely missed by some DBPs (Figures 8A, 9A, and 10A). This illustrates the power of integrating information from multiple sources using DNNs.

Furthermore, the detailed evaluation of several CASP targets suggests that the lower performance of all DBPs on the CASP data set results not only from more challenging cases. Instead, upon human inspection, the true annotations by CASP can be questioned for several targets (Figures 11). This is likely because CASP domain annotations are assigned with a different goal than domain boundary prediction. CASP domain annotations are used to decide which parts of the target should be evaluated

https://doi.org/10.1021/acs.jctc.1c00129
J. Chem. Theory Comput. 2021, 17, 4599−4613

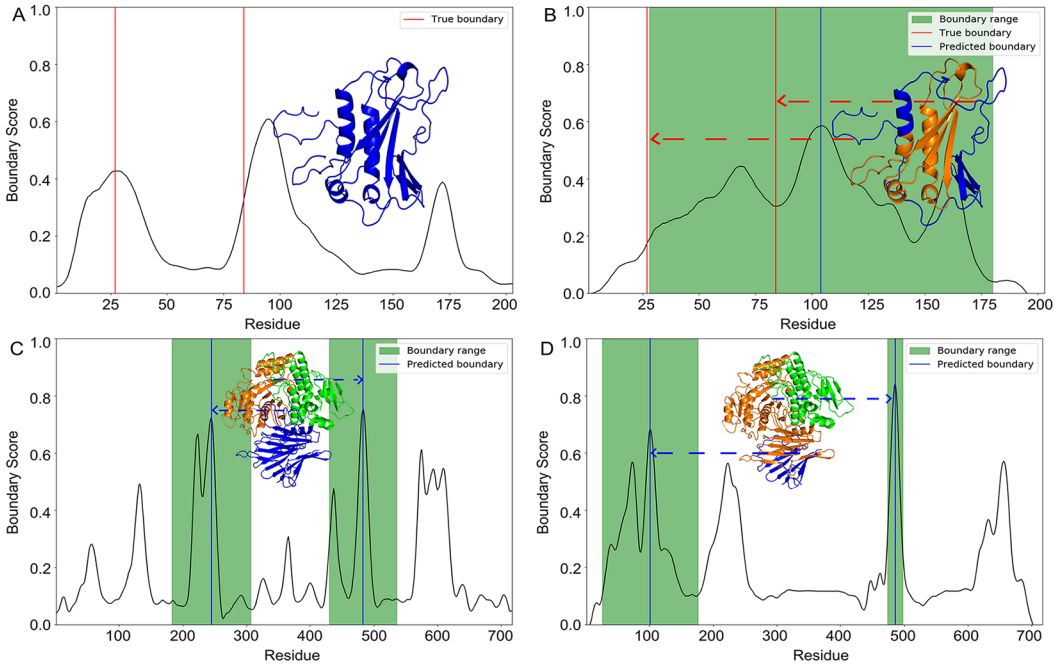


Figure 11. TopDomain and ThreaDom comparison on CASP. (A,B) T0771 from CASP 11 is considered a two-domain protein with two boundaries by CASP. The first boundary separates a small disordered region at the N-terminus. TopDomain considers this protein as single-domain since neither of the boundary score peaks is pronounced enough for boundary assignment (A). This looks plausible based on human inspection. ThreaDom predicts one of the CASP boundaries (B). However, the boundary is placed in a helix and has a confidence range spanning almost the entire protein due to its low peak height. (C, D) T1009 from CASP 13 is considered a single-domain protein by CASP. TopDomain predicts two confident boundaries for this protein, which upon human inspection look very plausible (C). ThreaDom predicts one of these boundaries but also predicts a second boundary in a different position, which upon human inspection looks implausible as it is located inside of a beta-strand of a beta-sandwich domain (D).

together during CASP quality assessment. Thus, it is likely that for easier CASP targets, some boundaries are omitted because the annotators deem that these targets should be evaluated in full and not on a per-domain basis. Similarly, it is likely that for difficult targets, some boundaries might be assigned to reward predictors for getting just parts of the model correct, even if these parts do not constitute independently folding domains in the evolutionary sense.

CONCLUDING REMARKS

Protein domain prediction is often the first step in protein structure prediction and has a large impact on downstream predictions such as template identification and prediction of residue contacts and distances. This is caused in part by the decreased ability of search algorithms such as PSI-BLAST,[55] HMMER,[29] and HHBLITS[96] to detect correct sequence matches for long sequences with low sequence identity due to limited coverage and in part by the sparse representation of multidomain structures in structure databases like the PDB in general.

A large multidomain protein may have only a few homologous sequences or templates available with the same domain architecture (i.e., a high coverage), but many matches for each domain individually. This can limit the number of effective sequences[79] for multidomain protein multiple sequence alignments, which are the basis for protein property predictions and template identification. Thus, protein property and structure prediction may be greatly improved by accurately cutting the target sequence into domains at the domain boundaries and subsequently predicting each domain separately. Furthermore, protein folding simulations with methods such as ITASSER[97] or ROSETTA[98] become exceedingly expensive as the size of the folded sequence increases, making it a feasible option to divide a protein into domains and fold these individually before combining them into a full-length model.

We presented two metapredictors of protein domain boundaries, TopDomain and $\text{TopDomain}_{\text{Seq}}$, and a protein domain parser, $\text{TopDomain}_{\text{Parse}}$, each of which are the most accurate to date for boundary prediction in multidomain proteins compared to any other of many tested state-of-the-art DBPs. Furthermore, we presented a predictor, $\text{TopDomain}_{\text{TMC}}$, that accurately predicts whether parsing the protein into domains before structure prediction is needed based on the available templates and their domain coverage and orientation. Finally, we developed a simple rule-based random predictor, RanDom, as a baseline, which only uses the sequence length and two generic rules to predict boundaries. Unexpectedly, some DBPs from literature performed on par with or worse than RanDom for several quality metrics (Tables S4 and S5).

As expected, we found ThreaDom to be the best homology-based primary DBP, followed by the commonly used InterProScan. This is likely due to these DBPs more sophisticated methods for template detection. We found that the best *ab initio* primary DBP is ConDo due to its sophisticated use of coevolution information and deep learning.

Different primary DPBs offer different advantages. Homology-based methods are often the most accurate but slower to compute, and their performance depends both on the availability of template structures and the ability to detect them correctly. Sequence-based methods are often less accurate but faster to compute and perform better when sequence information is available but structure information is not. *Ab initio*

https://doi.org/10.1021/acs.jctc.1c00129
J. Chem. Theory Comput. 2021, 17, 4599−4613

methods that rely only on the primary sequence are the fastest to use but the least accurate. Metamethods such as TopDomain give the best performance regardless of information availability, because they combine information from multiple sources at the cost of a slower computation time.

TopDomain currently predicts only the boundaries between domains and not the domains themselves. As such, it does not pair pieces of discontinuous domains into full domains. We plan to extend TopDomain with a metapredictor that uses the predicted boundaries to reconstruct discontinuous domains. TopDomain is available free of charge as a Web server at https://cpclab.uni-duesseldorf.de/topsuite/topdomain.php.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.1c00129.

Detailed descriptions of TopDomain$_{TMC}$, the TopDomain data set generation, TopDomain homology-based features, TopDomain sequence-based features, TopDomain feature conversion for DNNs, TopDomain training procedures for Stage 2 and Stage 3, parameters for boundary peak detection and confidence estimation, and prediction performance for all DBPs benchmarked in this paper (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Holger Gohlke − *Institut für Pharmazeutische und Medizinische Chemie, Heinrich-Heine-Universität Düsseldorf, 40225 Düsseldorf, Germany; John von Neumann Institute for Computing (NIC), Jülich Supercomputing Centre (JSC), Institute of Biological Information Processing (IBI-7: Structural Biochemistry) & Institute of Bio- and Geosciences (IBG-4: Bioinformatics), 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*

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

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

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.1c00129

Author Contributions

D.M. developed the method, implemented primary predictors and input features, curated data sets, performed data analysis, and wrote the manuscript. P.G. implemented primary predictors, curated the CASP data set, designed and trained DNNs, implemented postprocessing, performed benchmark calculations and analysis, and wrote the manuscript. D.M. and P.G. manually inspected and curated domain boundaries in the TopDomain data set. F.K. implemented the TopDomain web server. H.G. conceived the study, supervised and managed the project, secured funding and resources for the project, and revised the manuscript. 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. This work was supported in part by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), Projektnummer 267205415, SFB 1208 (project A03 to H.G.), and by the Bundesministerium für Bildung und Forschung (BMBF), Förderkennzeichen 031L0182, InCellulo-ProtStruct, to H.G. We are grateful for computational support and infrastructure provided by the "Zentrum für Informations- und Medientechnologie" (ZIM) at the Heinrich Heine University Düsseldorf and the computing time provided by the John von Neumann Institute for Computing (NIC) to H.G. on the supercomputer JUWELS at Jülich Supercomputing Centre (JSC) (user ID HKF7, VSK33).

■ REFERENCES

(1) Widderich, N.; Pittelkow, M.; Höppner, A.; Mulnaes, D.; Buckel, W.; Gohlke, H.; Smits, S. H.; 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.

(2) Jones, S.; Thornton, J. M. Principles of Protein-Protein Interactions. *Proc. Natl. Acad. Sci. U. S. A.* **1996**, *93*, 13−20.

(3) Scott, D. E.; Bayly, A. R.; Abell, C.; Skidmore, J. Small Molecules, Big Targets: Drug Discovery Faces the Protein−Protein Interaction Challenge. *Nat. Rev. Drug Discovery* **2016**, *15*, 533.

(4) Kendrew, J. C.; Bodo, G.; Dintzis, H. M.; Parrish, R.; Wyckoff, H.; Phillips, D. C. A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis. *Nature* **1958**, *181*, 662−666.

(5) Huang, C.; Kalodimos, C. G. Structures of Large Protein Complexes Determined by Nuclear Magnetic Resonance Spectroscopy. *Annu. Rev. Biophys.* **2017**, *46*, 317−336.

(6) Egelman, E. H. The Current Revolution in Cryo-Em. *Biophys. J.* **2016**, *110*, 1008−1012.

(7) Chruszcz, M.; Wlodawer, A.; Minor, W. Determination of Protein Structures - a Series of Fortunate Events. *Biophys. J.* **2008**, *95*, 1−9.

(8) Frueh, D. P.; Goodrich, A. C.; Mishra, S. H.; Nichols, S. R. Nmr Methods for Structural Studies of Large Monomeric and Multimeric Proteins. *Curr. Opin. Struct. Biol.* **2013**, *23*, 734−739.

(9) Rawson, S.; Iadanza, M.; Ranson, N.; Muench, S. Methods to Account for Movement and Flexibility in Cryo-Em Data Processing. *Methods* **2016**, *100*, 35−41.

(10) Campbell, M. G.; Cheng, A.; Brilot, A. F.; Moeller, A.; Lyumkis, D.; Veesler, D.; Pan, J.; Harrison, S. C.; Potter, C. S.; Carragher, B.; Grigorieff, N. Movies of Ice-Embedded Particles Enhance Resolution in Electron Cryo-Microscopy. *Structure* **2012**, *20*, 1823−1828.

(11) Fiser, A. Template-Based Protein Structure Modeling. *Computational Biology*; Springer: 2010; pp 73−94.

(12) Xu, D.; Zhang, Y. Ab Initio Protein Structure Assembly Using Continuous Structure Fragments and Optimized Knowledge-Based Force Field. *Proteins: Struct., Funct., Genet.* **2012**, *80*, 1715−1735.

(13) Ponting, C. P.; Russell, R. R. The Natural History of Protein Domains. *Annu. Rev. Biophys. Biomol. Struct.* **2002**, *31*, 45−71.

(14) Han, J.-H.; Batey, S.; Nickson, A. A.; Teichmann, S. A.; Clarke, J. The Folding and Evolution of Multidomain Proteins. *Nat. Rev. Mol. Cell Biol.* **2007**, *8*, 319−330.

(15) Ekman, D.; Björklund, Å. K.; Frey-Skött, J.; Elofsson, A. Multi-Domain Proteins in the Three Kingdoms of Life: Orphan Domains and Other Unassigned Regions. *J. Mol. Biol.* **2005**, *348*, 231−243.

(16) Apic, G.; Gough, J.; Teichmann, S. A. Domain Combinations in Archaeal, Eubacterial and Eukaryotic Proteomes. *J. Mol. Biol.* **2001**, *310*, 311−325.
(17) Brockwell, D. J.; Smith, D. A.; Radford, S. E. Protein Folding Mechanisms: New Methods and Emerging Ideas. *Curr. Opin. Struct. Biol.* **2000**, *10*, 16−25.
(18) Bhaskara, R. M.; Srinivasan, N. Stability of Domain Structures in Multi-Domain Proteins. *Sci. Rep.* **2011**, *1*, 40.
(19) Vishwanath, S.; de Brevern, A. G.; Srinivasan, N. Same but Not Alike: Structure, Flexibility and Energetics of Domains in Multi-Domain Proteins Are Influenced by the Presence of Other Domains. *PLoS Comput. Biol.* **2018**, *14*, e1006008.
(20) Schroeder, J. I.; Allen, G. J.; Hugouvieux, V.; Kwak, J. M.; Waner, D. Guard Cell Signal Transduction. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **2001**, *52*, 627−658.
(21) Rescigno, M.; Borrow, P. The Host-Pathogen Interaction: New Themes from Dendritic Cell Biology. *Cell* **2001**, *106*, 267−270.
(22) Dale, G. E.; Oefner, C.; D'Arcy, A. The Protein as a Variable in Protein Crystallization. *J. Struct. Biol.* **2003**, *142*, 88−97.
(23) Kim, D. E.; Chivian, D.; Malmström, L.; Baker, D. Automated Prediction of Domain Boundaries in Casp6 Targets Using Ginzu and Rosettadom. *Proteins: Struct., Funct., Genet.* **2005**, *61*, 193−200.
(24) Kolkman, J. A.; Stemmer, W. P. Directed Evolution of Proteins by Exon Shuffling. *Nat. Biotechnol.* **2001**, *19*, 423−428.
(25) Kondrashov, A. S. Deleterious Mutations and the Evolution of Sexual Reproduction. *Nature* **1988**, *336*, 435−440.
(26) Koonin, E. V.; Fedorova, N. D.; Jackson, J. D.; Jacobs, A. R.; Krylov, D. M.; Makarova, K. S.; Mazumder, R.; Mekhedov, S. L.; Nikolskaya, A. N.; Rao, B. S.; et al. A Comprehensive Evolutionary Classification of Proteins Encoded in Complete Eukaryotic Genomes. *Genome Biol.* **2004**, *5*, R7.
(27) Pearl, F. M.; Bennett, C.; Bray, J. E.; Harrison, A. P.; Martin, N.; Shepherd, A.; Sillitoe, I.; Thornton, J.; Orengo, C. A. The Cath Database: An Extended Protein Family Resource for Structural and Functional Genomics. *Nucleic Acids Res.* **2003**, *31*, 452−455.
(28) Fox, N. K.; Brenner, S. E.; Chandonia, J.-M. SCOPe: Structural Classification of Proteins—Extended, Integrating Scop and Astral Data and Classification of New Structures. *Nucleic Acids Res.* **2014**, *42*, D304−D309.
(29) Eddy, S. R. Accelerated Profile Hmm Searches. *PLoS Comput. Biol.* **2011**, *7*, e1002195.
(30) Bateman, A.; Birney, E.; Cerruti, L.; Durbin, R.; Etwiller, L.; Eddy, S. R.; Griffiths-Jones, S.; Howe, K. L.; Marshall, M.; Sonnhammer, E. L. The Pfam Protein Families Database. *Nucleic Acids Res.* **2002**, *30*, 276−280.
(31) Yeats, C.; Lees, J.; Reid, A.; Kellam, P.; Martin, N.; Liu, X.; Orengo, C. Gene3d: Comprehensive Structural and Functional Annotation of Genomes. *Nucleic Acids Res.* **2007**, *36*, D414−D418.
(32) Thomas, P. D.; Campbell, M. J.; Kejariwal, A.; Mi, H.; Karlak, B.; Daverman, R.; Diemer, K.; Muruganujan, A.; Narechania, A. Panther: A Library of Protein Families and Subfamilies Indexed by Function. *Genome Res.* **2003**, *13*, 2129−2141.
(33) Wu, C. H.; Nikolskaya, A.; Huang, H.; Yeh, L. S. L.; Natale, D. A.; Vinayaka, C. R.; Hu, Z. Z.; Mazumder, R.; Kumar, S.; Kourtesis, P.; et al. Pirsf: Family Classification System at the Protein Information Resource. *Nucleic Acids Res.* **2004**, *32*, D112−D114.
(34) Haft, D. H.; Loftus, B. J.; Richardson, D. L.; Yang, F.; Eisen, J. A.; Paulsen, I. T.; White, O. Tigrfams: A Protein Family Resource for the Functional Identification of Proteins. *Nucleic Acids Res.* **2001**, *29*, 41−43.
(35) Wilson, D.; Madera, M.; Vogel, C.; Chothia, C.; Gough, J. The Superfamily Database in 2007: Families and Functions. *Nucleic Acids Res.* **2007**, *35*, D308−D313.
(36) Haft, D. H.; Selengut, J. D.; White, O. The Tigrfams Database of Protein Families. *Nucleic Acids Res.* **2003**, *31*, 371−373.
(37) Letunic, I.; Doerks, T.; Bork, P. Smart 7: Recent Updates to the Protein Domain Annotation Resource. *Nucleic Acids Res.* **2012**, *40*, D302−D305.
(38) Zhou, H.; Xue, B.; Zhou, Y. Ddomain: Dividing Structures into Domains Using a Normalized Domain−Domain Interaction Profile. *Protein Sci.* **2007**, *16*, 947−955.
(39) Xu, Y.; Xu, D.; Gabow, H. N. Protein Domain Decomposition Using a Graph-Theoretic Approach. *Bioinformatics* **2000**, *16*, 1091−1104.
(40) Postic, G.; Ghouzam, Y.; Chebrek, R.; Gelly, J.-C. An Ambiguity Principle for Assigning Protein Structural Domains. *Science Advances* **2017**, *3*, e1600552.
(41) Xue, Z.; Xu, D.; Wang, Y.; Zhang, Y. Threadom: Extracting Protein Domain Boundary Information from Multiple Threading Alignments. *Bioinformatics* **2013**, *29*, i247−i256.
(42) Bryson, K.; Cozzetto, D.; Jones, D. T. Computer-Assisted Protein Domain Boundary Prediction Using the Dom-Pred Server. *Curr. Protein Pept. Sci.* **2007**, *8*, 181−188.
(43) Bondugula, R.; Lee, M. S.; Wallqvist, A. Fiefdom: A Transparent Domain Boundary Recognition System Using a Fuzzy Mean Operator. *Nucleic Acids Res.* **2008**, *37*, 452−462.
(44) Marchler-Bauer, A.; Lu, S.; Anderson, J. B.; Chitsaz, F.; Derbyshire, M. K.; DeWeese-Scott, C.; Fong, J. H.; Geer, L. Y.; Geer, R. C.; Gonzales, N. R.; et al. Cdd: A Conserved Domain Database for the Functional Annotation of Proteins. *Nucleic Acids Res.* **2011**, *39*, D225−D229.
(45) Quevillon, E.; Silventoinen, V.; Pillai, S.; Harte, N.; Mulder, N.; Apweiler, R.; Lopez, R. Interproscan: Protein Domains Identifier. *Nucleic Acids Res.* **2005**, *33*, W116−W120.
(46) Suyama, M.; Ohara, O. Domcut: Prediction of Inter-Domain Linker Regions in Amino Acid Sequences. *Bioinformatics* **2003**, *19*, 673−674.
(47) George, R. A.; Lin, K.; Heringa, J. Scooby-Domain: Prediction of Globular Domains in Protein Sequence. *Nucleic Acids Res.* **2005**, *33*, W160−W163.
(48) Linding, R.; Russell, R. B.; Neduva, V.; Gibson, T. J. Globplot: Exploring Protein Sequences for Globularity and Disorder. *Nucleic Acids Res.* **2003**, *31*, 3701−3708.
(49) Ebina, T.; Toh, H.; Kuroda, Y. Drop: An Svm Domain Linker Predictor Trained with Optimal Features Selected by Random Forest. *Bioinformatics* **2011**, *27*, 487−494.
(50) Hong, S. H.; Joo, K.; Lee, J. Condo: Protein Domain Boundary Prediction Using Coevolutionary Information. *Bioinformatics* **2019**, *35*, 2411−2417.
(51) Eickholt, J.; Deng, X.; Cheng, J. Dobo: Protein Domain Boundary Prediction by Integrating Evolutionary Signals and Machine Learning. *BMC Bioinf.* **2011**, *12*, 43.
(52) Cheng, J.; Sweredoski, M. J.; Baldi, P. Dompro: Protein Domain Prediction Using Profiles, Secondary Structure, Relative Solvent Accessibility, and Recursive Neural Networks. *Data Min. Knowl. Discovery* **2006**, *13*, 1−10.
(53) Jiang, Y.; Wang, D.; Xu, D. Deepdom: Predicting Protein Domain Boundary from Sequence Alone Using Stacked Bidirectional Lstm. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*; World Scientific: 2019; Vol. *24*, pp 66−75.
(54) Sim, J.; Kim, S. Y.; Lee, J. Pprodo: Prediction of Protein Domain Boundaries Using Neural Networks. *Proteins: Struct., Funct., Genet.* **2005**, *59*, 627−632.
(55) Altschul, S. F.; 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.
(56) Zhang, Y.; Skolnick, J. Tm-Align: A Protein Structure Alignment Algorithm Based on the Tm-Score. *Nucleic Acids Res.* **2005**, *33*, 2302−2309.
(57) Jones, P.; Binns, D.; Chang, H.-Y.; Fraser, M.; Li, W.; McAnulla, C.; McWilliam, H.; Maslen, J.; Mitchell, A.; Nuka, G.; et al. Interproscan 5: Genome-Scale Protein Function Classification. *Bioinformatics* **2014**, *30*, 1236−1240.
(58) 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.

https://doi.org/10.1021/acs.jctc.1c00129
J. Chem. Theory Comput. 2021, 17, 4599−4613

(59) 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.
(60) Jaroszewski, L.; Rychlewski, L.; Li, Z.; Li, W.; Godzik, A. Ffas03: A Server for Profile−Profile Sequence Alignments. *Nucleic Acids Res.* **2005**, *33*, W284−W288.
(61) Peng, J.; Xu, J. Raptorx: Exploiting Structure Information for Protein Alignment by Statistical Inference. *Proteins: Struct., Funct., Genet.* **2011**, *79*, 161−171.
(62) 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.
(63) 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.
(64) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(65) Kabsch, W.; Sander, C. Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers* **1983**, *22*, 2577−2637.
(66) Fang, C.; Shang, Y.; Xu, D. Mufold-Ss: New Deep Inception-inside-Inception Networks for Protein Secondary Structure Prediction. *Proteins: Struct., Funct., Genet.* **2018**, *86*, 592−598.
(67) Heffernan, R.; Yang, Y.; Paliwal, K.; Zhou, Y. Capturing Non-Local Interactions by Long Short-Term Memory Bidirectional Recurrent Neural Networks for Improving Prediction of Protein Secondary Structure, Backbone Angles, Contact Numbers and Solvent Accessibility. *Bioinformatics* **2017**, *33*, 2842−2849.
(68) Wang, S.; Peng, J.; Ma, J.; Xu, J. Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields. *Sci. Rep.* **2016**, *6*, 18962.
(69) Joo, K.; Lee, S. J.; Lee, J. Sann: Solvent Accessibility Prediction of Proteins by Nearest Neighbor Method. *Proteins: Struct., Funct., Genet.* **2012**, *80*, 1791−1797.
(70) Ward, J. J.; McGuffin, L. J.; Bryson, K.; Buxton, B. F.; Jones, D. T. The Disopred Server for the Prediction of Protein Disorder. *Bioinformatics* **2004**, *20*, 2138−2139.
(71) Wang, S.; Weng, S.; Ma, J.; Tang, Q. Deepcnf-D: Predicting Protein Order/Disorder Regions by Weighted Deep Convolutional Neural Fields. *Int. J. Mol. Sci.* **2015**, *16*, 17315−17330.
(72) Necci, M.; Piovesan, D.; Dosztányi, Z.; Tosatto, S. C. Mobidb-Lite: Fast and Highly Specific Consensus Prediction of Intrinsic Disorder in Proteins. *Bioinformatics* **2017**, *33*, 1402−1404.
(73) Jones, D. T.; Singh, T.; Kosciolek, T.; Tetchner, S. Metapsicov: Combining Coevolution Methods for Accurate Prediction of Contacts and Long Range Hydrogen Bonding in Proteins. *Bioinformatics* **2015**, *31*, 999−1006.
(74) Jones, D. T.; Buchan, D. W.; Cozzetto, D.; Pontil, M. Psicov: Precise Structural Contact Prediction Using Sparse Inverse Covariance Estimation on Large Multiple Sequence Alignments. *Bioinformatics* **2012**, *28*, 184−190.
(75) Seemayer, S.; Gruber, M.; Söding, J. Ccmpred—Fast and Precise Prediction of Protein Residue−Residue Contacts from Correlated Mutations. *Bioinformatics* **2014**, *30*, 3128−3130.
(76) Marks, D. S.; Hopf, T. A.; Sander, C. Protein Structure Prediction from Sequence Variation. *Nat. Biotechnol.* **2012**, *30*, 1072.
(77) Adhikari, B.; Hou, J.; Cheng, J. Dncon2: Improved Protein Contact Prediction Using Two-Level Deep Convolutional Neural Networks. *Bioinformatics* **2018**, *34*, 1466−1472.
(78) Jones, D. T.; Kandathil, S. M. High Precision in Protein Contact Prediction Using Fully Convolutional Neural Networks and Minimal Sequence Features. *Bioinformatics* **2018**, *34*, 3308.
(79) Michel, M.; Menéndez Hurtado, D.; Elofsson, A. Pconsc4: Fast, Accurate and Hassle-Free Contact Predictions. *Bioinformatics* **2019**, *35*, 2677.
(80) Käll, L.; Krogh, A.; Sonnhammer, E. L. A Combined Transmembrane Topology and Signal Peptide Prediction Method. *J. Mol. Biol.* **2004**, *338*, 1027−1036.
(81) Krogh, A.; Larsson, B.; Von Heijne, G.; Sonnhammer, E. L. Predicting Transmembrane Protein Topology with a Hidden Markov Model: Application to Complete Genomes. *J. Mol. Biol.* **2001**, *305*, 567−580.
(82) Hayat, S.; Elofsson, A. Boctopus: Improved Topology Prediction of Transmembrane B Barrel Proteins. *Bioinformatics* **2012**, *28*, 516−522.
(83) Petersen, T. N.; Brunak, S.; Von Heijne, G.; Nielsen, H. Signalp 4.0: Discriminating Signal Peptides from Transmembrane Regions. *Nat. Methods* **2011**, *8*, 785.
(84) Szklarczyk, R.; Heringa, J. Tracking Repeats Using Significance and Transitivity. *Bioinformatics* **2004**, *20*, i311−i317.
(85) Jorda, J.; Kajava, A. V. T-Reks: Identification of Tandem Repeats in Sequences with a K-Means Based Algorithm. *Bioinformatics* **2009**, *25*, 2632−2638.
(86) Söding, J.; Remmert, M.; Biegert, A. Hhrep: De Novo Protein Repeat Detection and the Origin of Tim Barrels. *Nucleic Acids Res.* **2006**, *34*, W137−W142.
(87) Lupas, A. Predicting Coiled-Coil Regions in Proteins. *Curr. Opin. Struct. Biol.* **1997**, *7*, 388−393.
(88) Ludwiczak, J.; Winski, A.; Szczepaniak, K.; Alva, V.; Dunin-Horkawicz, S. Deepcoil—a Fast and Accurate Prediction of Coiled-Coil Domains in Protein Sequences. *Bioinformatics* **2019**, *35*, 2790.
(89) AlQuraishi, M. Alphafold at Casp13. *Bioinformatics* **2019**, *35*, 4862−4865.
(90) Veit, A.; Wilber, M. J.; Belongie, S. Residual Networks Behave Like Ensembles of Relatively Shallow Networks. *Proc. 30th Int. Conf. Neural Inf. Process. Syst.* **2016**, 550−558.
(91) Hochreiter, S. The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* **1998**, *6*, 107−116.
(92) Bishop, C. M. *Pattern Recognition and Machine Learning*; Springer: 2006.
(93) He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*; IEEE: 2016; pp 770−778.
(94) McNemar, Q. Note on the Sampling Error of the Difference between Correlated Proportions or Percentages. *Psychometrika* **1947**, *12*, 153−157.
(95) Zhou, X.; Hu, J.; Zhang, C.; Zhang, G.; Zhang, Y. Assembling Multidomain Protein Structures through Analogous Global Structural Alignments. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116*, 15930−15938.
(96) 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.
(97) Zhang, Y. I-Tasser Server for Protein 3d Structure Prediction. *BMC Bioinf.* **2008**, *9*, 40.
(98) 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.

https://doi.org/10.1021/acs.jctc.1c00129
J. Chem. Theory Comput. 2021, 17, 4599−4613