JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

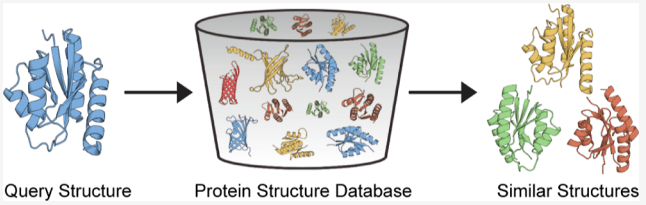
Rapid and Accurate Protein Structure Database Search Using Inverse Folding Model and Contrastive Learning

Qiuyi Lyu, Hong Wei, Shuaishuai Chen, Zhenling Peng,* and Jianyi Yang*

Cite This: https://doi.org/10.1021/acs.jcim.5c02385 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Protein structure database search has become increasingly challenging due to the growing number of experimental and computational structures. We introduce mTM-align2, a novel two-step approach for rapid and accurate protein structure database search. In the first step, protein structures are first transformed into embeddings using a pretrained inverse folding model (ESM-IF) and 3D Zernike polynomials. The ESM-IF embeddings are further optimized through a contrastive learning network, which is trained on ∼7 million structure pairs. Structures with similar embeddings are returned on the fly in this step. The second step employs a rapid structure alignment program to refine top candidates, ensuring high precision and producing high-quality alignments. Extensive benchmarks reveal that mTM-align2 performs competitively compared to other leading methods, completing monomeric structure search in seconds with over 90% precision for the top 10 hits. The t-SNE visualization of the mTM-align2 embeddings for thousands of structures demonstrates that our embeddings are structurally informed, capturing the global structural features. A web server for mTM-align2 is accessible at https://yanglab.qd.sdu.edu.cn/mTM-align/.



INTRODUCTION

The primary objective of protein structure database search is to efficiently identify similar structures within a structure database, such as the Protein Data Bank (PDB).[1] The most accurate approach is based on pairwise structure alignment using tools like TM-align[2] and US-align.[3] However, performing database-wide structure alignments is extremely time-consuming, especially when the database is large. For example, TM-align will require a few weeks to search the PDB database with a single query.[4]

With advancements in protein structure prediction, over 200 million structures predicted by AlphaFold[5] are accessible in the AlphaFold DB (AFDB).[6] Searching against these structures is a great challenge. A straightforward approach to accelerate the structure search is to cluster proteins with similar structures. Dali[7] is one of the most well-recognized methods in this area, which applies a walking strategy to expand hits in clustered structures. mTM-align[8,9] combines both sequence and structure similarity in the clustering and provides fast pairwise and multiple structure alignment. TM-search[10] applies an iterative clustering method to reduce the structure comparison and uses TM-align to perform pairwise structure alignment. Although these methods could significantly speed up the search, it is still challenging for them to handle large databases such as AFDB.

To achieve high-speed search, a few methods were proposed using shape-based descriptors,[11−14] which however are less accurate. 3D-SURFER[14] compares the global shape similarity using the 3D Zernike Moment.[15] 3D-AF-Surf[11] improves the performance through deep learning and supports structure search against AFDB. Nevertheless, like 3D-SURFER, it does not consider the residue-level similarity. BioZernike[12] further enhances retrieval performance through two key innovations. First, it incorporates residue-level information into the moment computation, improving structural representation. Second, it introduces an "alignment descriptor" to facilitate the alignment, leading to a higher precision. In addition to the Zernike moment-based methods, Omakage[13] employs the incremental distance rank profile to represent proteins shape and the Gaussian mixture model to align protein structures.

Recent works show that deep neural networks have great potential in protein structure representation.[16−25] GraSR[24] and Progres[23] utilize a graph neural network to acquire the protein structure embedding, which is further optimized under a contrastive learning framework. DeepFold[18] and FoldExplorer[20] use the convolutional/graph attention neural network to encode protein structure, together with sequence information from protein language model. AlphaFind[25] utilizes the learned metric index[26] approach to generate protein structure embedding, an extremely compressed representation of protein structure, supporting fast structure search against AFDB. Foldclass[21] acquires protein structure similarity by comparing

Received: September 30, 2025
Revised: November 12, 2025
Accepted: November 14, 2025

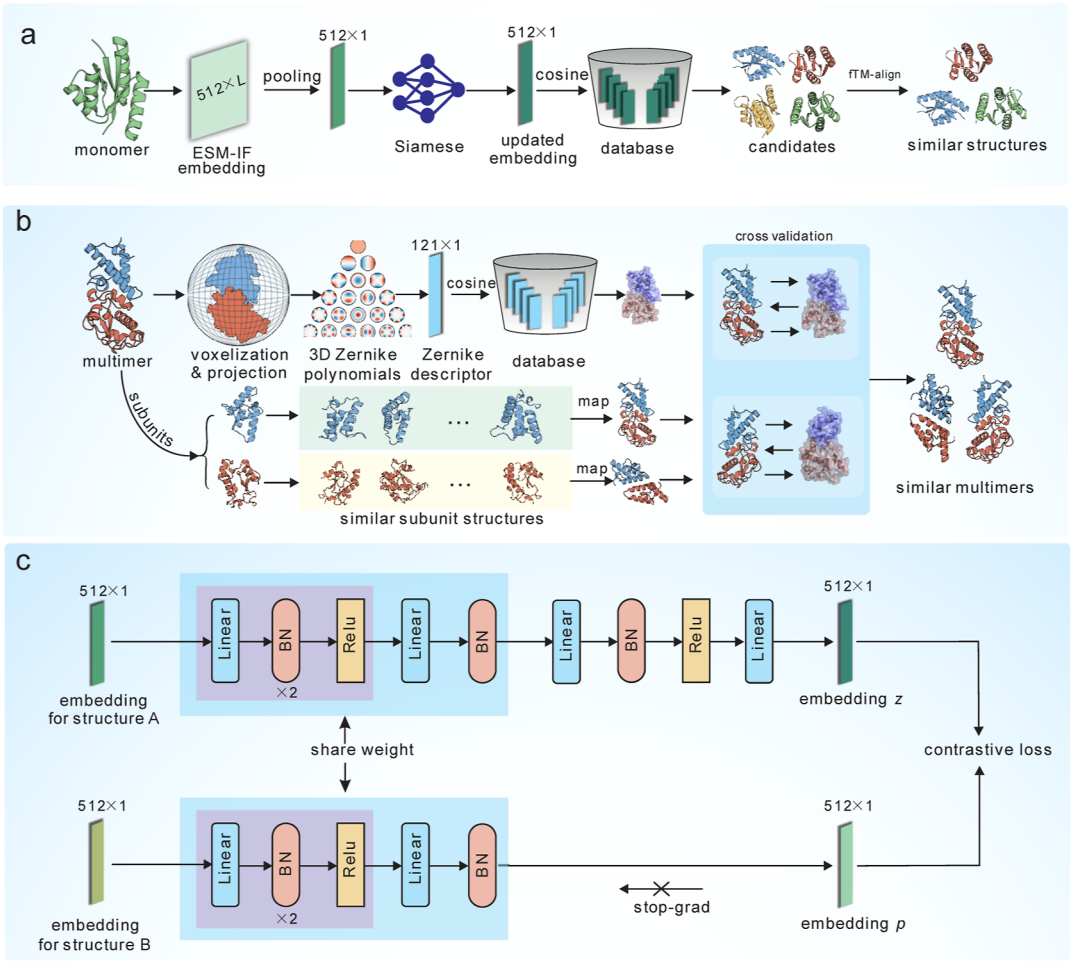


Figure 1. Overall architecture of mTM-align2 for protein structure database search. (a) Algorithm for monomeric structure search. The search process begins with the conversion of the query structure into a 2D embedding using the pretrained inverse folding model (ESM-IF). This 2D embedding is then reduced to a 1D embedding through sum pooling. A contrastive learning network, specifically an asymmetric Siamese network (illustrated in Figure c) is used to optimize the 1D embedding. The similarity between the query and database embeddings is calculated on the fly with cosine function. Finally, the candidates with high similarity to the query are filtered using the structure alignment program fTM-align. (b) Algorithm for multimeric structure search (see Figure S8 for more details). It comprises two key modules. The first module utilizes the 3D Zernike polynomials (ZPM) to convert each structure into a 121-D descriptor vector. The similarity between query and database vectors is also calculated by cosine function. The second module is based on monomeric structure search. The monomeric hits for each subunit structure are mapped to their corresponding multimeric structures. The hits from both modules are combined to generate the final set of multimeric structures. (c) Siamese network with an asymmetric architecture to optimize the ESM-IF embedding. For training, the raw embeddings for two structures are fed into the network twice by exchanging their order. For inference, the embedding z from the upper branch serves as the final embeddings for structures.

their constituent domains, using the program Merizo-search[21] for domain segmentation. TM-VEC,[16] PLMSearch[17] and DHR[22] are representative methods that infer structure similarity from amino acid sequence using protein language models without exact structure comparisons.

Foldseek[4] is a widely used method for searching various structure databases. It converts each protein structure into an artificial amino acid sequence using a Variational Autoencoder (VAE) network, followed by the use of the MMseqs2[27] program for quick detection of similar sequences. It was recently extended to search and align multimeric structures, resulting in the new method Foldseek-Multimer.[28] Reseek[29] aligns protein structures using a mega-alphabet of ∼85 billion states for Cα feature vectors and demonstrates superior sensitivity in homologue detection.

In this study, we present mTM-align2, an enhanced version of mTM-align tailored for rapid protein structure database search. We utilize the inverse folding model (ESM-IF[30]) and 3D Zernike polynomials[12] in conjunction with a contrastive learning network to improve the speed and accuracy. Benchmarks demonstrate that mTM-align2 is competitive with other methods for searching both monomeric and multimeric structures. The t-SNE visualization demonstrates that the mTM-align2 embeddings effectively capture the global features of protein structures, providing valuable insights into the underlying architectural similarities shared by various structures.

RESULTS

Overview of mTM-align2. The major steps of structure search by mTM-align2 are shown in Figure 1. For monomeric structure, the inverse folding model ESM-IF[30] is used to generate a 2D embedding ($512 \times L$, where L is the length of the structure). This embedding is reduced to a raw embedding (512-D vector) by row-wise sum pooling. The raw embedding from the inverse folding space is then transformed into an

https://doi.org/10.1021/acs.jcim.5c02385
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

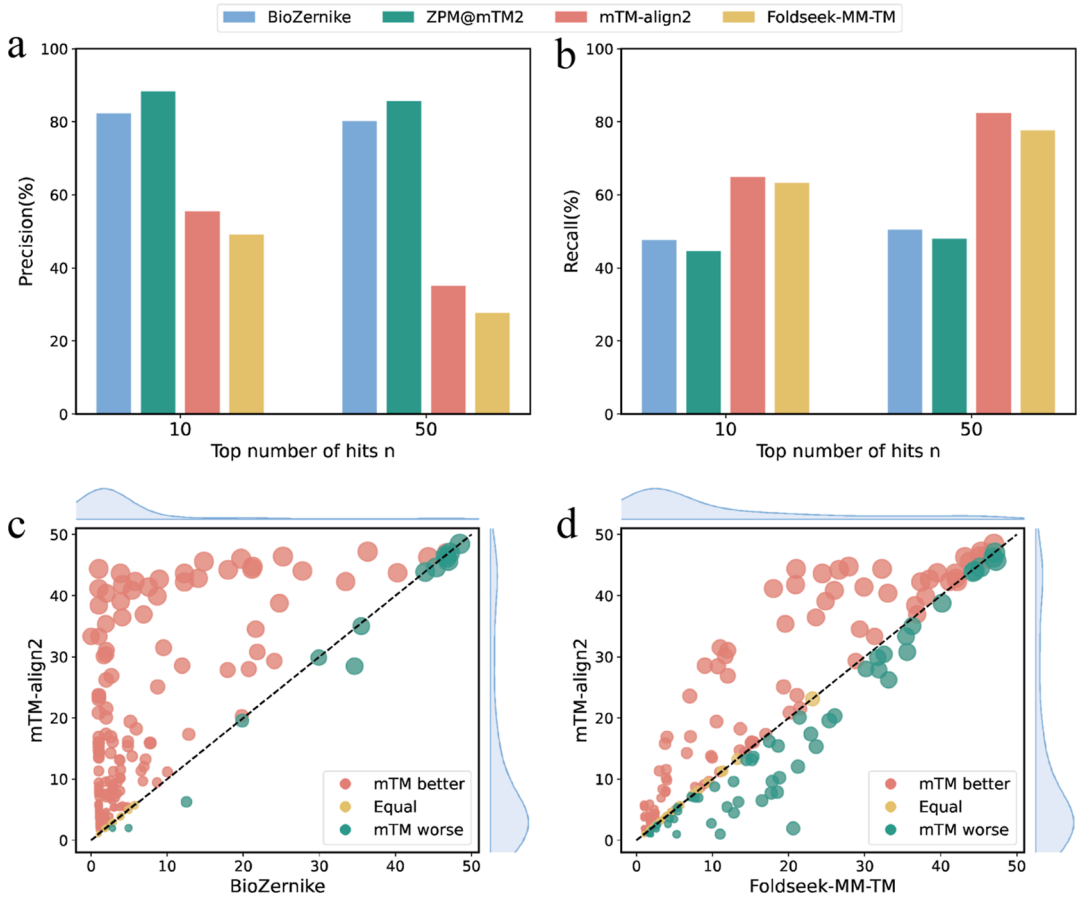


Figure 2. Performance on multimeric structure search. (a) Precision and (b) recall metrics were evaluated for searching 286 structures in the multimer test set, focusing on the top 10 and 50 hits. (c,d) Present the summed TM-scores for the true positives among the top 50 hits. Each point in these graphs represents an individual structure from the test set. The radius of each point is proportional to the number of structures retrieved by US-align. Blue lines along the axes are the kernel density estimation of the data.

updated embedding in the Euclidian space using contrastive learning network (Siamese,[31] see Figure 1c). The similarity between two structures can be calculated on the fly with the cosine function. The top 1000 candidate structures are further filtered and aligned using the fast structure alignment program fTM-align.[2]

For multimeric structures, two modules are utilized to perform the search (Figure 1b). The first module employs the 3D Zernike polynomials (ZPM)[12] to identify structures with similar shapes. Each multimeric structure is represented as a 121-D vector derived from ZPM (refer to Methods). The similarity between two vectors is calculated using the cosine function. A maximum of 1000 hits are returned from this module. The second module is based on the monomeric structure search procedure. All subunit structures are first extracted from the multimeric structure. Each subunit is then processed through the monomeric structure search pipeline. The returned monomers are then mapped back to their corresponding multimers, resulting in a maximum of 1000 multimeric hits. Finally, hits from both modules are combined, resulting in the final set of multimers (see Methods).

mTM-align2 Outperforms Other Methods for Multimeric Structure Search. Here, we compare mTM-align2 with other multimeric structure search methods, including BioZernike,[12] and Foldseek-MM-TM.[28] Foldseek-MM-TM is a variant version of Foldseek-MM[28] that filters hits using TM-align. We also assess the performance of two variants of mTM-align2: ZPM@mTM2, which utilizes only the ZPM method, and IFM@mTM2, which employs only the IFM method. The ground truth is defined according to US-align, where a hit is considered a true positive if the TM-score exceeds 0.65, a threshold employed by Foldseek-MM.

The comparison involves searching a set of 286 multimeric structures against a nonredundant database of approximately 310,000 multimeric structures. As shown in Figure 2a, mTM-align2 outperforms Foldseek-MM-TM in terms of precision. For the top 10 to top 50 hits, mTM-align2 achieves precision rates ranging from 55.52% to 35.08%, compared to 48.67% to 26.17% for Foldseek-MM-TM (see Figure S1). The low precision for both methods (<60%) is due to the diversity of multimer structures. The low precision observed in both methods can be attributed to the limited number of similar structures available in the database for most multimeric structures (refer to the subsequent analysis for Figure 2c). If we consider only the top five predictions, the average precision of mTM-align2 rises to over 90%. A previous study[32] indicates that the distribution of multimeric structures in the 3D Complex database is highly skewed, characterized by a significant number of structure families containing only a few members. This finding aligns with our results.

The Zernike polynomials-based methods (ZPM@mTM2 and BioZernike) demonstrate higher precision than mTM-align2 and Foldseek-MM-TM, at the expense of lower recall. This is because these approaches focus solely on global

https://doi.org/10.1021/acs.jcim.5c02385
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

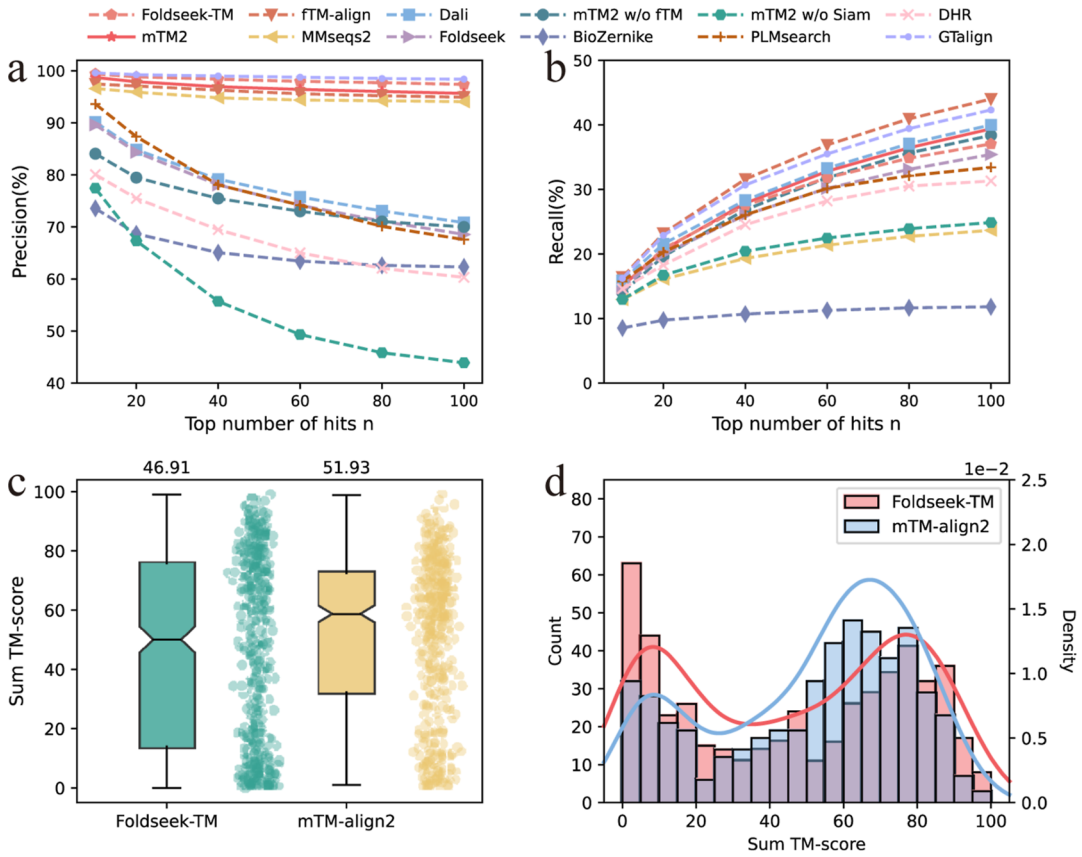


Figure 3. Performance of monomeric structure search. (a) Precision and (b) recall metrics were assessed on the monomer test data set, which includes 500 structures. (c,d) Comparisons of the sum TM-scores for the true positives among the top 100 hits. Each point in (c) represents an individual structure from the test set. The average sum TM-scores are listed on the top of the box. (d) Distributions of the sum TM-score for mTM-align2 and Foldseek-TM, where bars and curves are Count and Density, respectively.

similarity, often overlooking local structural details and returning a limited number of hits. For instance, the average number of hits returned by ZPM@mTM2 and BioZernike is 13 and 15, respectively, leading to lower recall values (see Figure 2b). Two examples are given to illustrate the limitations of relying solely on Zernike polynomials-based descriptors. In the first example, though the two structures (PDB IDs: 1BAR, 2IOY, Figure S3a) have similar shape (Zernike score 0.96), their local structures are very different with a low TM-score of 0.25. On the contrary, though the two structures (PDB IDs: 8DMG, 5DYP, Figure S3b) have different overall shapes (Zernike score 0.94), they show a high structure similarity (an average TM-score of 0.75). To balance precision and recall, we present a precision−recall curve for the methods mTM-align2, Foldseek-MM-TM, and ZPM@mTM2, as shown in Figure S2. At a stricter threshold (>0.98), ZPM@mTM2 achieves high precision of approximately 93%, effectively identifying structures with globally similar shapes. However, as the threshold decreases (<0.90), the number of false positives increases, resulting in a significant drop in precision to below 40%. This observation highlights the importance of integrating Zernike polynomials with complementary descriptors, such as those derived from the IFM, to capture a broader range of structural features.

The recall values for all methods are summarized in Figure 2b. It needs to be noted that up to top 50 hits are considered in this experiment, thus, the average recall rate may be low. To show the upper limit of the recall rate, we use US-align to define the ground truth and present its top 10 to 50 hits recall rate in Figure S1b. Notably, mTM-align2 achieves higher recall than other methods. For the top 10 hits, mTM-align2 secures a recall rate of 64.91%, compared to 63.56% for Foldseek-MM-TM. This difference becomes more pronounced for the top 50 hits, where mTM-align2 achieves a recall of 82.4% versus 77.59% for Foldseek-MM-TM.

We also compare mTM-align2 with BioZernike and Foldseek-MM-TM based on the sum TM-score (denoted by sTM-score) of the true positives among the top 50 hits. Out of the 286 queries, mTM-align2 outperforms BioZernike for 181 queries and performs worse for 14 queries (see Figure 2c). Notably, the improvement over BioZernike is significant for 61 structures, where the difference in sTM-score is more than 10. A substantial number of points (151) are located in the lower left area (sTM-score <10 for mTM-align2 and BioZernike), indicating that both mTM-align2 and BioZernike return a limited number of true positives. This arises from the relatively small number of similar structures available for these targets, averaging only five according to US-align.

In the comparison with Foldseek-MM-TM based on sTM-score (see Figure 2d), mTM-align2 performs better for 112 structures and worse for 52 structures. Notably, mTM-align2 significantly outperforms Foldseek-MM-TM for 22 structures (with sTM-score difference greater than 10). An example is illustrated in Figure S3c (PDB ID: 8F5F), which is a homo dimer. In this case, 44 structures with TM-scores >0.65 are identified by US-align. mTM-align2 successfully found 37 of them, compared to only 2 by Foldseek-MM-TM. Two structures that Foldseek-MM-TM failed to recognize (PDB

https://doi.org/10.1021/acs.jcim.5c02385
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

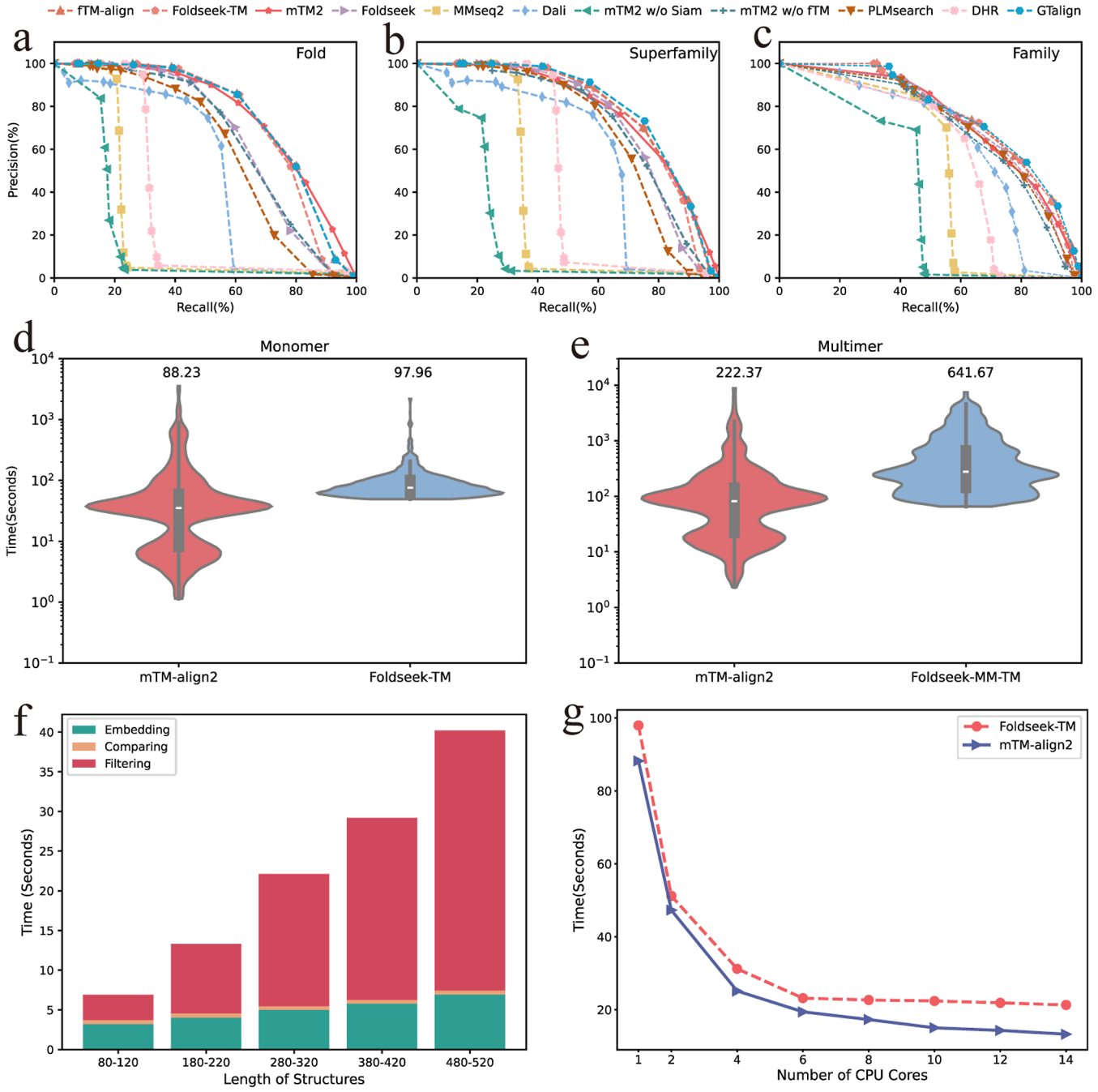


Figure 4. Performance on the SCOPe data set and running time analysis. (a–c) Precision–recall curves on the SCOPe40 test set. These comparisons are made at the fold, superfamily, and family labels, respectively. Hits that share the same fold, superfamily, or family labels as the query are defined as true positives. (d, e) Average running time per structure: the experiments were conducted using monomer and multimer data sets, respectively, with a single CPU core for all methods. (f) Average running time for structures of varying lengths: this evaluation focuses on structures from the monomer data set, also using a single CPU core. (g) Average running time per monomeric structure: this analysis was performed on the monomer data set with different numbers of CPU cores.

IDs: 4MPN and 4MP7) are shown in the figure. This failure may be attributed to misalignment of certain alpha helices and beta sheets (highlighted in the red box), resulting in different local 3D interaction (3Di) state sequences in Foldseek-MM-TM.

mTM-align2 is Competitive with Other Methods for Monomeric Structure Search. We compare mTM-align2 with other methods for monomeric structure search, including Foldseek,[4] Foldseek-TM,[4] fTM-align,[2] BioZernike,[12] DALI,[7] MMseqs2,[27] PLMSearch,[17] DHR,[22] GTalign.[19] Two variants of mTM-align2 are also assessed here: mTM-align2 without fTM-align (mTM2 w/o fTM), and mTM-align2 without the Siamese network (mTM2 w/o Siam). The comparison is based on a search of 500 structures against a monomeric structure database of approximately 730,000 structures. A hit is defined as a true positive if it shares a TM-score >0.5 with the query structure, as calculated by TM-align.

The results are summarized in Figure 3. mTM-align2 achieves comparable precision (>95%) to Foldseek-TM (see Figure 3a). This is anticipated because both methods apply a

https://doi.org/10.1021/acs.jcim.5c02385
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

two-step search strategy. The first step is to quickly find candidate structures that are similar to the query; the second step filters the top hits using the accurate but slow structure alignment. When the filtering is removed, both methods experience reduced precision. For example, the precision of the top 10 hits drops from 98.71% to 84.06% for mTM-align2, compared to a reduction from 99.39% to 89.61% for Foldseek. DALI demonstrates slightly better precision and recall compared to mTM-align2_noTM and Foldseek, but at the cost of being orders of magnitude slower.

Interestingly, the sequence alignment-based method MMseqs2 has similar precision to mTM-align2 and Foldseek-TM. This may be because it only returns hits with similar sequences, which usually implies similar structures. However, MMseqs2 is not able to detect remote homologies, which share similar structure but dissimilar sequence to the query, as evidenced by its low recall values (see Figure 3b). GTalign achieves the highest precision at a high computational cost as its alignment-based retrieval process is significantly more time-consuming than the other methods. We also note that BioZernike performs poorly in terms of both precision and recall, consistent with previous observations.[12]

As shown in Figure 3a,b, the protein language model-based methods, PLMSearch and DHR, exhibit lower accuracy compared to the full versions of mTM-align2 and Foldseek-TM. Notably, PLMSearch performs comparably to mTM-align2 and Foldseek when the structure alignment-based filtering is removed, likely due to its use of PfamScan-based filtering.[17] In contrast, DHR remains less accurate than both mTM-align2 and Foldseek, even if the structure alignment-based filtering is removed. This suggests that the structure-based embeddings utilized in Foldseek (3Di) and mTM-align2 (ESM-IF) are more informative for quantifying structural similarity than the sequence-based embeddings derived from protein language models. This conclusion is further supported by the subsequent ablation analysis.

We note that the recall values for all methods are low. For instance, the recall value for mTM-align2 is 39.38% even when considering the top 100 hits. To understand this data, we calculate the recall values for the structure alignment-based method fTM-align, which also yields a low recall value of 44.01% for the top 100 hits. The low recall values can be attributed to the limited number of top hits considered (a maximum of 100). The average and median number of structures sharing a TM-score >0.5 for the 500 testing structures are 585 and 314, respectively, which is significantly larger than 100 (see Figure S4).

The sTM-scores among the top 100 hits returned by mTM-align2 and Foldseek-TM are presented in Figure 3c,d. mTM-align2 has a higher average sTM-score than Foldseek-TM (51.93 vs 46.91). The distributions in Figure 3d indicate that mTM-align2 outperforms Foldseek-TM when the sTM-score is less than 70. Specifically, mTM-align2 performs better than Foldseek-TM for 216 out of the 319 targets that the sTM-scores are less than 70. A notable example is shown in Figure S5, which belongs to the lipocalin-like β barrel domain (PDB ID: 8DML, chain B). For this example, among the top 100 hits returned by mTM-align2 and Foldseek-TM, 97 and 20 are true positives, respectively, yielding respective sTM-scores of 60.12 and 13.43. mTM-align2 successfully identifies 77 true positives that are missed by Foldseek-TM. The superimposition of two of these hits against the query structure reveals that the structures in the common core regions (highlighted in red) are highly similar, while other outlier regions differ significantly.

mTM-align2 is Competitive with Other Methods for SCOP Domain Classification. The ground truth for the comparison above relies on structural similarity defined by TM-align, which may introduce bias against structure alignment methods. To address this issue, we further compare mTM-align2 with other methods based on the SCOP[33] domain classification. The data set comprises 379 structure domains randomly selected from the SCOPe40[33] database (version 2.08), which contains 15,172 domains with less than 40% pairwise sequence identity. Hits that match the fold/superfamily/family label of the query are regarded as true positives. A precision−recall curve is plotted by adjusting the scoring thresholds of each method (e.g., predicted TM-score in mTM-align2).

Figure 4 summarizes the precision−recall curves for the fold, superfamily and family classification. Similar to the results in the previous section, mTM-align2, Foldseek-TM, fTM-align, and GTalign exhibit comparable performance and are nearly indistinguishable from one another. However, as previously mentioned, GTalign and fTM-align are not ideal for large-scale retrieval tasks due to their reliance on computationally intensive pairwise structure alignment for identifying similar structures. In contrast, the alignment-based filtering procedure is crucial for enhancing the performance of mTM-align2 and Foldseek-TM. When structure-based filtering is removed, both mTM-align2 and Foldseek-TM experience a noticeable decline in performance. The decrease is more pronounced at the fold level compared to the superfamily and family levels. For instance, at a recall of 62%, the precision for both mTM-align2 and Foldseek-TM drops from over 80% to about 65% after removing the filter. In contrast, the corresponding decrease for superfamily and family classifications is less than 5%. This difference may be attributed to the high correlation between the TM-score (used in filtering) and the definition of the SCOP folds.

The sequence-based PLM methods, PLMSearch and DHR, are less accurate than structure-based methods. Nevertheless, both methods demonstrate a common trend: the performance gap with structure-based methods narrows when the sequence signal becomes stronger from fold level to family level. Specifically, their accuracy improves progressively from the fold level to the superfamily and, subsequently, the family level. We conclude that this behavior is a direct consequence of the inherent sequence-based nature of these methods.

Running Time Analysis of mTM-align2. *Speed for Database Construction.* The running speed of mTM-align2 is compared with Foldseek in Figure 4d,e. Both mTM-align2 and Foldseek were executed on a Linux server equipped with four 24-core CPUs, 2 TB memory, and an NVIDIA A100 GPU. Each method requires a preprocessing step to construct its respective database. For mTM-align2, this involves clustering and generating protein embeddings on the A100, which took approximately 3.5 h for our database of around 730,000 monomers. Constructing the Zernike database for the PDB multimer data set (approximately 310,000 multimers) took about 4 h using 20 CPU cores. Foldseek outperforms mTM-align2 in database preprocessing, requiring about 1 h to generate 3Di sequences with 20 CPU cores.

Speed for Database Search. We compare the speed of structure search against the prebuilt databases using a single CPU core for all programs. The evaluation involved searching

https://doi.org/10.1021/acs.jcim.5c02385
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

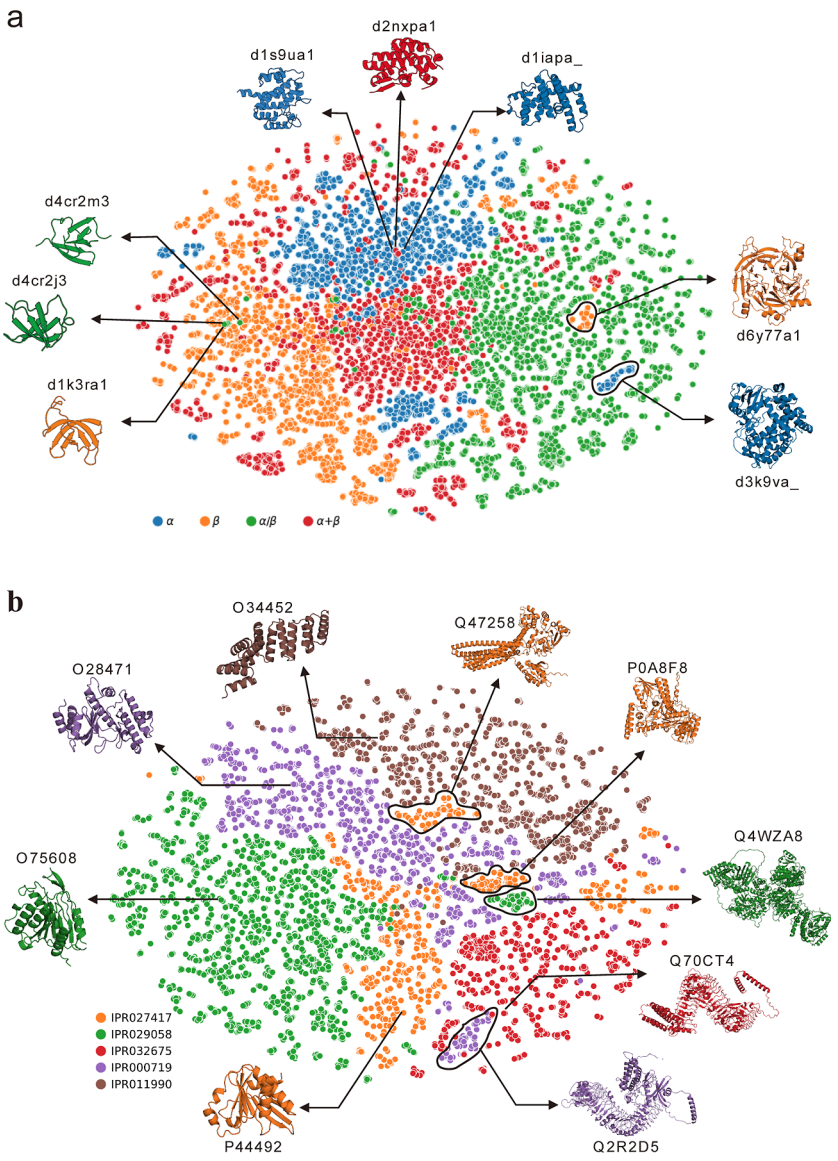


Figure 5. Visualization of mTM-align2 embeddings by t-SNE. (a) Results for the 15,172 SCOPe structures, where each point represents a structure, colored by structural classes defined by SCOPe. (b) Results for 8598 predicted structure models from AFDB. The point colors indicate different InterPro domains.

286 multimeric structures against approximately 310,000 multimers and 500 monomeric structures against around 730,000 monomers. For monomeric queries, mTM-align2 shows slightly faster speed than Foldseek-TM, while for multimeric queries, it achieves a two-fold improvement in computational efficiency compared to Foldseek-MM-TM. This increased speed can be attributed to two key factors: first, mTM-align2 estimates similarity using cosine functions, which is much faster than the *k*-mer based heuristic approach of Foldseek; second, it utilizes a preclustered database, allowing searches against a nonredundant data set and extending results to other cluster members of the returned hits, similar to the strategy used in mTM-align.

Speed for Searching Structures with Different Lengths. We also evaluate mTM-align2's searching time for monomers of varying lengths on a single CPU core. The query processing in mTM-align2 follows a three-stage procedure: generating embeddings for the query structure, computing cosine similarity between the query embedding and those in the database, and structure alignment-based filtering. As shown in Figure 4f, the average running time demonstrates a positive correlation with the length of the structures. Generating the query embedding typically requires a few seconds, while computing the cosine similarity takes approximately 0.5 s. Notably, the structure-based filtering accounts for the majority of the running time, particularly for larger structures, indicating that the structure alignment-based filtering process is the most time-consuming component and serves as the primary determinant of the increased running time for longer structures.

Acceleration with Parallel Computing. The above experiments were conducted using a single CPU core; however, both Foldseek-TM and mTM-align2 can be accelerated through multicore computing. For mTM-align2, this acceleration was achieved by parallelizing the structure alignment-based filtering procedure. As shown in Figure 4g, both methods benefit significantly from parallelization, with Foldseek-TM and mTM-align2 achieving speedups of 76% and 78%, respectively,

https://doi.org/10.1021/acs.jcim.5c02385
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

when the number of CPU cores increases from 1 to 6. Notably, the two methods exhibit different scalability trends: Foldseek-TM showed negligible speed improvement beyond 6 CPU cores, while mTM-align2 continues to improve in speed as more CPU cores are utilized.

t-SNE Visualization of the mTM-align2 Embedding. Given the outstanding performance of mTM-align2, we visualize its embedding using the t-distributed stochastic neighbor embedding (t-SNE)[34] with a perplexity of 5. Figure 5a displays the results for 15,172 structures from SCOPe40. Each point in the figure represents one structure, colored according to the structural classes defined by SCOPe. Here, we focus on the top four largest structural classes: α, β, α/β, and $\alpha+\beta$. The results indicate that the structures are largely clustered correctly according to their classes, with several noteworthy observations outlined below.

mTM-align2 Embeddings Discover Structures that are Atypical for Their Assigned SCOPe Class. In the left panel of Figure 5a, two structures (SCOPe IDs: *d4cr2m3* and *d4cr2j3*) belong to the α/β class (green points) but are clustered alongside structures from the β class (orange points). Visual inspection of the protein structures indicates a resemblance to their neighbors. This is exemplified by the structure shown in the figure (SCOPe ID: *d1k3ra1*), which shares moderate structural similarity with *d4cr2m3* and *d4cr2j3*, yielding TM-scores of 0.46 and 0.51, respectively. We also observe that a few $\alpha+\beta$-class structures (red points) are clustered with α-class structures (blue points). A notable example is the clustering of the $\alpha+\beta$-class structure *d2nxpa1* with two α-class structures, *d1s9ua1* and *d1iapa_* (Figure 5a, top). While *d2nxpa1* is classified as $\alpha+\beta$ due to two minor β-sheets, its secondary structure is dominated by α-helices, making it structurally analogous to its α-class neighbors. These examples indicate that our embeddings could discover structures that are atypical for their assigned SCOPe class.

mTM-align2 Embedding Reveals Structural Fold. Additionally, we note the emergence of distinct subclusters within a single cluster. Two such subclusters from the β class (orange points) and the α class (blue points) are highlighted in the right panel of Figure 5a. They appear within the cluster of α/β class structures (green points). Upon inspecting the structures in these two subclusters, we present two representative examples (*d6y77a1* and *d3k9va*). Both structures contain α helices and β sheets, indicating shared structural similarities with other α/β-class structures in the cluster. Furthermore, these two subclusters correspond to two SCOPe folds, a.104 and b.69, respectively, which have well-defined boundaries with other structures. This suggests that mTM-align2 embedding has the potential to cluster structures from a 'coarse-grained' class level to a 'fine-grained' fold level.

mTM-align2 Embeddings for Predicted Structures in AlphaFold DB. We further analyze the predicted structures from AFDB[6] using the mTM-align2 embeddings. For illustration, we took 8598 high-confidence structures with pLDDT > 80 from five InterPro[35] domains. Figure 5b demonstrates that most structures cluster well, with clear boundaries between different InterPro domains: IPR029058 (*Alpha/Beta hydrolase*, green points), IPR027417 (P-loop containing nucleoside triphosphate hydrolase, orange points), IPR000719 (Protein kinase, purple points), IPR011990 (Leucine-rich repeat domain, red points), and IPR011990 (Tetratricopeptide-like helical domain, brown points). Each InterPro domain is represented by an example structure in the figure, illustrating the structural differences among them.

Notably, the first three InterPro domains (green, orange, and purple points) are closely clustered, likely due to their similar structural features (containing both α helices and β sheets) and functions (enzymes).

In contrast, the other two domains (red and brown points) are well-separated, reflecting their distinct structural characteristics: the Leucine-rich repeat domain (red points) is characterized by a repeated α/β horseshoe fold, while the Tetratricopeptide-like helical domain (brown points) consists of a multihelical fold made up of two curved layers of α-helices. These results further confirm that the mTM-align2 embeddings are structurally informed.

However, we observe the formation of some subclusters, four of which are highlighted in circles. A representative structure is provided for each subcluster. A common characteristic of these structures is that they are significantly larger than other structures within their respective domains. For instance, the structure for Q4WZA8 (shown on the right side of Figure 5b) contains over 2000 amino acids, approximately ten times larger than the example structure O75608 from the same domain. In fact, Q4WZA8 is a multidomain protein. According to the InterPro annotation, this protein contains 21 InterPro domains, including the domain IPR029058 associated with structure O75608. Similarly, the purple structure (Q2R2D5) shown at the bottom of Figure 5b also contains multiple domains, including IPR000719 (purple) and IPR032675 (red). This structure is primarily characterized by the repeated α/β horseshoe fold, leading to its assignment in the cluster of IPR032675 (red) rather than IPR000719 (purple). The clustering is further supported by the structural similarity (TM-score >0.5) between the two representative structures (Q2R2D5 and Q70CT4). These data indicate that the mTM-align2 embeddings effectively capture the global features of protein structures.

Ablation Study. For monomeric structures, mTM-align2 makes use of contrastive learning (Siamese network) and structure alignment-based filter (fTM-align) to enhance the precision. We conducted an ablation study to assess the improvements introduced by these components. On the monomer test data set, the precision of mTM-align2 without fTM-align drops from 95.64% to 70.11% (Figure 3a, top 100 hits), indicating that the structure-based filter is essential. In addition, the precision significantly declines from 70.11% to 43.90% when the Siamese network is removed. These data suggest the critical role of the Siamese network in enhancing the precision of mTM-align2.

For multimeric structures, the monomeric structure search module (IFM) is combined with the ZPM-based module to enhance the search results (Figure 1b). We evaluated the improvement gained from this combination on the multimer test data set. When mapping multimeric structures using the top hits from IFM, an average of 32.86 hits are obtained. In contrast, when using the ZPM-based search for multimeric structures alone, an average of 15.6 hits are retrieved. The Venn diagram in Figure S6 demonstrates that the two modules are complementary to each other. Both modules have overlap for 12.39 hits and their unique hits (20.47 and 3.21 for IFM and ZPM, respectively). Therefore, combining hits from both modules results in the most accurate results in mTM-align2.

https://doi.org/10.1021/acs.jcim.5c02385
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

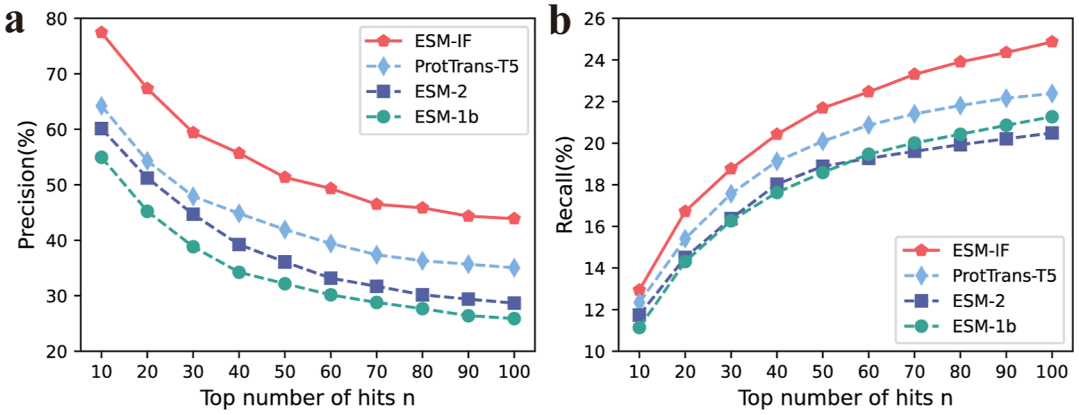


Figure 6. Comparison of zero-shot structure search performance of various Protein Language Models (PLMs) on the monomer data set. (a,b) Precision and recall. Embeddings were generated by average pooling the residue-level features from each pretrained model, and cosine similarity was employed to rank the search results.

To validate the effectiveness of ESM-IF as a structure encoder, we conducted a comparative analysis against several other pretrained sequence-based PLMs: ESM-1b,[36] ESM-2,[37] and ProtT5.[38] We evaluated each model's zero-shot performance on a structure retrieval task using our monomer test data set. For this evaluation, global protein representations were derived by applying average pooling to the raw residue features from each PLM. Structure retrieval was then performed using these global representations based on cosine similarity. The results, illustrated in Figure 6, clearly show that ESM-IF outperforms the other models in both precision and recall. This superior performance suggests that the features generated by ESM-IF effectively capture the structural properties.

CONCLUSION

Protein structure database search has become increasingly important and challenging. Building on advancements in protein engineering and deep learning, we developed mTM-align2, a rapid and accurate approach for protein structure database search and clustering. Unlike other methods, mTM-align2 effectively handles both monomers and multimers within a single framework. Protein structures are converted into embeddings using the inverse folding model and the 3D Zernike polynomials. The embeddings are further optimized through a contrastive learning network trained on ∼7 million structure pairs. The embedding significantly accelerates search speed; while the network enhances the accuracy. Comprehensive benchmarks demonstrate the superior performance of mTM-align2 for both monomers and multimers. mTM-align2 typically completes monomeric structure searches against existing databases within seconds, achieving over 90% precision for the top 10 hits.

The t-SNE visualization of the mTM-align2 embeddings for thousands of protein structures indicates that the mTM-align2 embeddings effectively capture the global features of protein structures, leading to insightful observations, such as identification of similar structures across different classes and recognition of blurred class definitions. These findings highlight the utility of mTM-align2 in advancing our understanding of protein structures and their classifications and functions.

Despite its strengths, we admit certain limitations of mTM-align2. The sum pooling transform of the ESM-IF embedding can result in the loss of structural details, bringing false positive hits. To solve this problem, we apply structure alignment to the top hits to filter out false positives. However, this filtering slows down the search. Potential solution could involve directly comparing the original ESM-IF embeddings with contrastive learning without pooling. We plan to explore this enhancement in our future work.

METHODS

Training and Test Data Sets. *Monomer Test Data Set.* We obtained a data set of ∼ 730,000 monomers from Q-BioLiP[39] (version 2023.01.11) and 15,172 domains from the SCOPe40 database (version 2.08). These structures, comprising both monomers and domains, were clustered using CD-HIT[40] at a 40% sequence identity threshold to eliminate redundancy, resulting in 70,270 distinct clusters. We randomly selected 500 monomers from 500 clusters as the monomer test set. The test set for the SCOP domain data set was constructed by selecting each domain structures in the same cluster, resulting in 379 domains (some of the selected clusters do not contain any domain structures).

Multimer Test Data Set. Given that our multimeric structure search algorithm relies on monomer-based structural searches to retrieve constituent subunits, it is imperative to prevent data leakage between the training and testing data sets. To achieve this, the multimer test data set was constructed based on the 500 monomer structures by mapping them to their corresponding multimers. From this mapping, 286 multimers were obtained, as other monomers do not form multimeric structures.

Training Set. To train the Siamese neural network, we constructed >7 million pairs of protein structures as follows. First, we generated random protein pairs from the non-redundant set of monomeric structures (testing structures were removed), resulting in ∼5 million pairs of structures. Most of them are negative training samples with a TM-score less than 0.5. We then use fTM-align to search for protein pairs with high structure similarity. This step yields a high-similarity set of ∼2 million pairs of structures, each with a TM-score greater than 0.5.

Algorithm for Monomeric Structure Search. The monomeric structure search involves three key steps (see Figure 1):

Step 1. Encode a structure using the inverse folding model. The pretrained protein language model ESM-IF[30] is used to

https://doi.org/10.1021/acs.jcim.5c02385
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

transform a monomeric structure into a raw embedding ($L \times 512$). Note that only the encoder from ESM-IF is used. The embedding of the structure is reduced to a raw embedding (512-D vector) by row-wise sum pooling.

Step 2. Optimize the structure embedding with contrastive learning. The raw embedding from the inverse folding space is then transformed into an updated embedding in the Euclidean space using contrastive learning network (Siamese,[31] Figure 1c, introduced below). As shown in Figure 1c, the Siamese network consists of two branches. For inference, the raw embedding from ESM-IF is fed into the upper branch, producing the updated embedding z.

Step 3. Search and filter similar structures. The similarity between two embeddings z_1 and z_2 is defined as the cosine of the angle (θ) between z_1 and z_2, which is named as IF-score

$$\text{IF-score}(z_1, z_2) = \cos(\theta) = \frac{z_1 \cdot z_2}{\|z_1\| \times \|z_2\|} \tag{1}$$

The IF-score between the query embedding and the precalculated embeddings of all structures in the database can be calculated on the fly. Structures with IF-score greater than 0.4 are returned. To improve the precision and generate pairwise alignment, we apply the fast protein structure alignment program fTM-align to filter the structures, removing those with a TM-score below 0.4. The retained hits are then ranked and mapped to their respective clusters to expand the search to the whole database. Details for the monomer structure search algorithm can be found in the supplementary algorithm A1.

Contrastive Learning Using the Siamese Neural Network. We use contrastive learning, specifically, the Siamese neural network[31] with asymmetric structure to optimize the relationship between vector similarity and structure similarity. The network is shown in Figure 1c. During training, as the network is asymmetric, we input the raw features twice by swapping their order. For a structure pair (x_1,x_2), in the first run, we feed the embeddings of x_1 and x_2 to the upper and lower branch, respectively. The order of x_1 and x_2 is swapped in the second run. As a result, we obtain two different vectors for each structure, that is $z(x_1)$,$p(x_1)$ for x_1, and $z(x_2)$, $p(x_2)$ for x_2. Then the average cosine similarity (i.e., average of $\cos(z(x_1),p(x_2))$ and $\cos(z(x_2),p(x_1))$) is used to estimate the structure similarity (i.e., TM-score) of x_1 and x_2. We show the details of the training process in the supplementary algorithm A2.

Loss Function. To minimize the differences between our predicted score and the TM-score, the MSE loss is used.

$$\text{loss} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2}$$

where y_i is the predicted TM-score, $\hat{y}_i$ is the real TM-score, n is the size of the training batch.

Model Training. The training and test are conducted on a Linux server equipped with 4 Intel Xeon Platinum 8260 CPU 96 cores (2.40 GHz) and 2TB memory. The GPU used for the training was a Nvidia A100 with 40GB of high-bandwidth memory. To achieve better performance, we conduct label-optimization according to the significance of TM-score.[41] For dissimilar protein pairs (TM-score <0.3), we subtract 0.2 from their TM-score, further separating them in the feature vector space; if the TM-score is above 0.7, we add 0.2 to their TM-score, with a maximum value of 1. This adjustment in the label assignment significantly enhances the retrieval accuracy (Figure S7).

The initial learning rate is 0.001, which is scheduled by a cosine function as the training going on. We utilized a batch size of 1024. Stochastic gradient descent (SGD) was applied to optimize the parameter. The network converged after 100 epochs in about 1.5 h.

Algorithm for Multimeric Structure Search. The multimeric structure search consists of two modules. The first module is based on 3D Zernike polynomials (ZPM, introduced below in detail). The second module is based on the monomeric structure search introduced above (denoted by IFM). The hits from both modules are combined to yield the final set of similar multimers. More details are available in the supplementary algorithm A3.

Descriptors from 3D Zernike Polynomials. We use the 3D Zernike moments[15] to describe the shape of multimeric structures, with implementation by the package BioZernike.[12] The first step involves representing the shape in 3D space through a process called voxelization. For a given protein, a 3D grid of size $32 \times 32 \times 32$ is created to convert its structure coordinates into a volumetric representation. For each amino acid, the Cα atom is used as the representative atom. Then, Gaussian density is constructed for each Cα atom, where the weight corresponds to the amino acid's molecular weight, and the size reflects the spherically averaged size of the amino acid. Then the Gaussian densities are placed into the volume, which is subsequently scaled to fit within a unit sphere. Finally, a coordinate system is fixed with the origin as the center of the grid and the axes aligned with the grid axes.

After establishing the coordinate system, any given protein can be represented as a volumetric function $f(\boldsymbol{x})$. The 3D Zernike polynomials $Z_{nl}^{m}(\boldsymbol{x})$ are used as orthonormal basis functions, allowing the volumetric function $f(\boldsymbol{x})$ to be decomposed accordingly.

$$f(\boldsymbol{x}) = \sum_{n}\sum_{l}\sum_{m}\Omega_{nl}^{m}\cdot Z_{nl}^{m}(\boldsymbol{x}) \tag{3}$$

where $\boldsymbol{x}$ is the vector representing the coordinates of grid points. The coefficients of the basis functions Ω_{nl}^{m} are defined as the 3D Zernike moments, which will be transformed into the shape descriptors after normalization. The 3D Zernike polynomials $Z_{nl}^{m}(\boldsymbol{x})$ is defined as[15]

$$\begin{aligned} Z_{nl}^{m}(\boldsymbol{x}) = {} & c_l^m 2^{-m}\sum_{v=0}^{k} q_{kl}^{v}\sum_{\alpha=0}^{v}\binom{v}{\alpha} \\ & \sum_{\beta=0}^{v-\alpha}\binom{v-\alpha}{\beta}\sum_{u=0}^{m}(-1)^{m-u}\binom{m}{u} i^{u} \\ & \sum_{\mu=0}^{\lfloor (l-m)/2\rfloor}(-1)^{\mu}2^{-2\mu}\binom{l}{\mu}\binom{l-\mu}{m+\mu}\sum_{v=0}^{\mu}\binom{\mu}{v} \\ & \cdot x^{2(v+\alpha)+u}\cdot y^{2(\mu-v+\beta)+m-u}\cdot z^{2(v-\alpha-\beta-\mu)+l-m} \end{aligned} \tag{4}$$

where i is the imaginary unit, n is the predefined maximum polynomial order, $l \in [0,n]$, $m \in [-l,l]$, $n - l$ is even number, and $k = \frac{n-l}{2}$. The term c_l^m and q_{kl}^{v} are defined as follows

$$c_l^m = \frac{\sqrt{(2l+1)(l+m)!(l-m)!}}{l!} \tag{5}$$

https://doi.org/10.1021/acs.jcim.5c02385
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

$$q_{kl}^{v} = \frac{(-1)^k}{2^{2k}}\sqrt{\frac{2l+4k+3}{3}}\binom{2k}{k}(-1)^v \frac{\binom{k}{v}\binom{2(k+l+v)+1}{2k}}{\binom{k+l+v}{k}} \quad (6)$$

thus, the 3D Zernike moments are defined based on the volume function $f(\boldsymbol{x})$ as

$$\Omega_{nl}^{m} = \frac{3}{4\pi}\int_{|\boldsymbol{x}|<1} f(\boldsymbol{x})\cdot\overline{Z_{nl}^{m}(\boldsymbol{x})}\,\mathrm{d}\boldsymbol{x} \quad (7)$$

where $\overline{Z_{nl}^{m}(x)}$ is the conjugate of $Z_{nl}^{m}(x)$.

Note that the moment Ω_{nl}^{m} is not invariant under rotation. To obtain rotation invariant shape descriptors, the 3D Zernike descriptors D_{nl} are defined as the norms of Ω_{nl}^{m}, that is

$$D_{nl} = \|\Omega_{nl}^{-l}, \Omega_{nl}^{-l+1}, \Omega_{nl}^{-l+2}, \cdots, \Omega_{nl}^{l}\| \quad (8)$$

Moments up to order 20 ($n \leq 20$) are utilized in our experiments and the resulting moment is transformed into a 121-D descriptor. This vector describes the overall shape of the structure. The similarity (denoted by ZP-score) between two descriptor vectors is calculated using the cosine function, similar to eq 1. The structures in the database are ranked by ZP-score and a maximum of 1000 multimeric structures are returned by ZPM.

Identify Multimeric Structures Based on Monomeric Structure Search. We make use of the monomeric structure search module to identify multimeric structure (see Figure S8 for more details). The top 6 longest subunit structures are first extracted from the query structure. Each subunit structure is then fed into the monomeric structure search pipeline. The returned monomers are then mapped to their respective multimers, resulting in a maximum of 1000 multimeric hits. The IF-score for each subunit in the mapped multimers is taken from the previous set of similar subunits, which is set to 0 if not existed in the set. The IF-score for each mapped multimer is then calculated as the mean IF-scores over all subunits.

Strategy for Combining Multimeric Structure Hits from ZPM and IFM. The candidate multimeric structures from ZPM and IFM are combined to generate the final set of multimeric structures based on a consensus score called Q-score (see Figure S8 for more details).

$$\text{Q-score} = \frac{1}{\alpha+\beta}(\alpha \times \text{IF-score} + \beta \times \text{ZP-score}) \quad (9)$$

where the weights α and β are empirically set to 1 and 0.3, respectively. Multimeric structures are ranked based on the Q-score and up to 1000 top hits are returned.

Controlled Baselines. We compare mTM-align2 with several existing methods, including fTM-align,[2] a fast implementation of TM-align; GTalign,[19] a spatial index-driven protein structure alignment method; PLMSearch,[17] a PLM-based method to perform homology detection. DHR,[22] a PLM-based method based on dual-encoder architecture, DALI,[7] a popular structure alignment-based method using distance matrix; BioZernike,[12] an alignment-free method that is officially used by PDB; Foldseek,[4] the state-of-the-art method for efficient protein structure database search; MMseqs2,[27] an efficient method for fast sequence database search. As the model weight for BioZernike is not publicly available, we manually submitted the query structures to its web server (http://shape.rcsb.org/) to conduct the structure search. For all other methods, we downloaded the packages and ran them locally with default settings.

EVALUATION METRICS

Similar to previous studies,[9] the average precision and recall are used to evaluate the performance

$$\text{precision}\ (n) = \frac{1}{q}\sum_{i=1}^{q}\frac{\text{TP}(m_i)}{m_i} \quad (10)$$

$$\text{recall}\ (n) = \frac{1}{q}\sum_{i=1}^{q}\frac{\text{TP}(m_i)}{P_i} \quad (11)$$

where q is the total number of structures in a test set, n is the number of top hits to be assessed, $m_i = \min(N_i, n)$ and N_i is the number of returned hits, P_i is the number of hits that have TM-score >0.5 with the query. A structure is defined as a true positive (TP) if it appears in the top n hits and its TM-score with the query exceeds the specific thresholds. For monomeric/multimeric structures, the threshold is defined as 0.5/0.65. In this study, the TM-score between two structures is defined as the average TM-score, that is, the average of the two TM-scores normalized by the lengths of two structures. Note that TM-align/US-align is used to calculate the TM-score between two monomeric/multimeric structures.

ASSOCIATED CONTENT

Data Availability Statement

The data sets supporting this research are all open-source. Monomer and multimer data are available from the Protein Data Bank (https://www.rcsb.org), with domain data sourced from SCOP (https://scop.berkeley.edu/astral/pdbstyle/ver=2.08). AlphaFold 2 predicted structures were downloaded from the AlphaFold Protein Structure Database (https://alphafold.ebi.ac.uk/download). All test data sets generated and analyzed for this study are available at Zenodo: https://zenodo.org/records/15818043. The web server is available at: https://yanglab.qd.sdu.edu.cn/mTM-align/.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.5c02385.

The Supporting Information includes algorithms for monomeric and multimeric structure search, details about the Siamese network, performance and case study on multimeric structure search, the statistics on the number of similar structures for the monomeric test data set, an example monomeric structure for which mTM-align2 detects more similar structures than Foldseek, a Venn diagram for the average number of results returned by ZPM and IFM on the multimer test data set, an ablation study for label optimization, and a flowchart for oligomeric structure search (PDF)

AUTHOR INFORMATION

Corresponding Authors

Zhenling Peng − *MOE Frontiers Science Center for Nonlinear Expectations, Research Center for Mathematics and*

https://doi.org/10.1021/acs.jcim.5c02385
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Interdisciplinary Sciences, Shandong University, Qingdao 266237, China; Email: zhenling@email.sdu.edu.cn

Jianyi Yang − *MOE Frontiers Science Center for Nonlinear Expectations, Research Center for Mathematics and Interdisciplinary Sciences, Shandong University, Qingdao 266237, China*; orcid.org/0000-0003-2912-7737; Email: yangjy@sdu.edu.cn

Authors

Qiuyi Lyu − *MOE Frontiers Science Center for Nonlinear Expectations, Research Center for Mathematics and Interdisciplinary Sciences, Shandong University, Qingdao 266237, China*; orcid.org/0000-0002-0104-9549

Hong Wei − *Department of Bioinformatics, School of Basic Medical Sciences, Tianjin Medical University, Tianjin 300070, China*

Shuaishuai Chen − *School of Information Science and Engineering, Shandong University, Qingdao 266237, China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.5c02385

Author Contributions

J.Y. conceived and supervised the project. Q.L. performed the experiments. Z.P. cosupervised the project. H.W. and S.C. analyzed the data. J.Y. and Q.L. wrote the manuscript. All authors read and approved the final version of the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work is supported by the National Key Research and Development Program of China (2023YFF1204003), the National Natural Science Foundation of China (NSFC T2225007, T2222012, 32430063), the Shandong Provincial Natural Science Foundation Youth Found (ZR2023QF156), and the Fundamental Research Funds for the Central Universities.

■ REFERENCES

(1) Berman, H. M.; et al. The protein data bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.

(2) Zhang, Y.; Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **2005**, *33*, 2302−2309.

(3) Zhang, C.; Shine, M.; Pyle, A. M.; Zhang, Y. US-align: universal structure alignments of proteins, nucleic acids, and macromolecular complexes. *Nat. Methods* **2022**, *19*, 1109−1115.

(4) Van Kempen, M.; et al. Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.* **2024**, *42*, 243−246.

(5) Jumper, J.; et al. Highly accurate protein structure prediction with AlphaFold. *nature* **2021**, *596*, 583−589.

(6) Varadi, M.; et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **2022**, *50*, D439−D444.

(7) Holm, L.; Rosenström, P. Dali server: conservation mapping in 3D. *Nucleic Acids Res.* **2010**, *38*, W545−W549.

(8) Dong, R.; Pan, S.; Peng, Z.; Zhang, Y.; Yang, J. mTM-align: a server for fast protein structure database search and multiple protein structure alignment. *Nucleic Acids Res.* **2018**, *46*, W380−W386.

(9) Dong, R.; Peng, Z.; Zhang, Y.; Yang, J. mTM-align: an algorithm for fast and accurate multiple protein structure alignment. *Bioinformatics* **2018**, *34*, 1719−1725.

(10) Liu, Z.; Zhang, C.; Zhang, Q.; Zhang, Y.; Yu, D.-J. TM-search: an efficient and effective tool for protein structure database search. *J. Chem. Inf. Model.* **2024**, *64*, 1043−1049.

(11) Aderinwale, T.; Bharadwaj, V.; Christoffer, C.; Terashi, G.; Zhang, Z.; Jahandideh, R.; Kagaya, Y.; Kihara, D. Real-time structure search and structure classification for AlphaFold protein models. *Commun. Biol.* **2022**, *5*, 316.

(12) Guzenko, D.; Burley, S. K.; Duarte, J. M. Real time structural search of the Protein Data Bank. *PLoS Comput. Biol.* **2020**, *16*, No. e1007970.

(13) Suzuki, H.; Kawabata, T.; Nakamura, H. Omokage search: shape similarity search service for biomolecular structures in both the PDB and EMDB. *Bioinformatics* **2016**, *32*, 619−620.

(14) La, D.; et al. 3D-SURFER: software for high-throughput protein surface comparison and analysis. *Bioinformatics* **2009**, *25*, 2843−2844.

(15) Novotni, M.; Klein, R. 3D Zernike descriptors for content based shape retrieval. In *Proceedings of the eighth ACM symposium on Solid modeling and applications*, 2003, pp 216−225,

(16) Hamamsy, T.; et al. TM-Vec: template modeling vectors for fast homology detection and alignment. *bioRxiv* **2022**, 501437.

(17) Liu, W.; Wang, Z.; You, R.; Xie, C.; Wei, H.; Xiong, Y.; Yang, J.; Zhu, S. PLMSearch: Protein language model powers accurate and fast sequence search for remote homology. *Nat. Commun.* **2024**, *15*, 2775.

(18) Liu, Y.; Ye, Q.; Wang, L.; Peng, J. Learning structural motif representations for efficient protein structure search. *Bioinformatics* **2018**, *34*, i773−i780.

(19) Margelevičius, M. GTalign: Spatial index-driven protein structure alignment, superposition, and search. *Nat. Commun.* **2024**, *15*, 7305.

(20) Liu, Y.; Shen, H.-B. FoldExplorer: Fast and Accurate Protein Structure Search with Sequence-Enhanced Graph Embedding. *arXiv* **2023**, arXiv:2311.18219.

(21) Kandathil, S. M.; Lau, A.; Buchan, D.; Jones, D. Foldclass and Merizo-search: embedding-based deep learning tools for protein domain segmentation, fold recognition and comparison. *bioRxiv* **2024**, 586696.

(22) Hong, L.; et al. Fast, sensitive detection of protein homologs using deep dense retrieval. *Nat. Biotechnol.* **2025**, *43*, 983−995.

(23) Greener, J. G.; Jamali, K. Fast protein structure searching using structure graph embeddings. *bioRxiv* **2022**, 518224.

(24) Xia, C.; Feng, S.-H.; Xia, Y.; Pan, X.; Shen, H.-B. Fast protein structure comparison through effective representation learning with contrastive graph neural networks. *PLoS Comput. Biol.* **2022**, *18*, No. e1009986.

(25) Procházka, D.; et al. AlphaFind: discover structure similarity across the proteome in AlphaFold DB. *Nucleic Acids Res.* **2024**, *52*, W182−W186.

(26) Olha, J.; Slanináková, T.; Gendiar, M.; Antol, M.; Dohnal, V. Learned indexing in proteins: substituting complex distance calculations with embedding and clustering techniques. In *International Conference on Similarity Search and Applications*; Springer, 2022, pp 274−282.

(27) Steinegger, M.; Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **2017**, *35*, 1026−1028.

(28) Kim, W.; et al. Rapid and sensitive protein complex alignment with foldseek-multimer. *Nat. Methods* **2025**, *22*, 469−472.

(29) Edgar, R. C. Protein structure alignment by Reseek improves sensitivity to remote homologs. *Bioinformatics* **2024**, *40*, btae687.

(30) Hsu, C.; et al. Learning inverse folding from millions of predicted structures. In *International conference on machine learning*; PMLR, 2022, pp 8946−8970.

(31) Chen, X.; He, K. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp 15745−15753.

(32) Levy, E. D.; Pereira-Leal, J. B.; Chothia, C.; Teichmann, S. A. 3D complex: a structural classification of protein complexes. *PLoS Comput. Biol.* **2006**, *2*, No. e155.

(33) Chandonia, J.-M.; et al. SCOPe: improvements to the structural classification of proteins−extended database to facilitate variant

interpretation and machine learning. *Nucleic Acids Res.* **2022**, *50*, D553−D559.
(34) Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579−2605.
(35) Weisberg, A. J.; Kim, G.; Westwood, J. H.; Jelesko, J. G. Sequencing and de novo assembly of the Toxicodendron radicans (poison ivy) transcriptome. *Genes* **2017**, *8*, 317.
(36) Rives, A.; Meier, J.; Sercu, T.; Goyal, S.; Lin, Z.; Liu, J.; Guo, D.; Ott, M.; Zitnick, C. L.; Ma, J.; et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. U.S.A.* **2021**, *118*, No. e2016239118.
(37) Lin, Z.; et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **2023**, *379*, 1123−1130.
(38) Elnaggar, A.; et al. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 7112−7127.
(39) Wei, H.; Wang, W.; Peng, Z.; Yang, J. Q-biolip.: A comprehensive resource for quaternary structure-based protein−ligand interactions. *Genom. Proteom. Bioinform.* **2024**, *22*, qzae001.
(40) Fu, L.; Niu, B.; Zhu, Z.; Wu, S.; Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **2012**, *28*, 3150−3152.
(41) Xu, J.; Zhang, Y. How significant is a protein structure similarity with TM-score= 0.5? *Bioinformatics* **2010**, *26*, 889−895.

https://doi.org/10.1021/acs.jcim.5c02385
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX