Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Rapid and accurate prediction of protein homo-oligomer symmetry with Seq2Symm

Meghana Kshirsagar

Meghana.Kshirsagar@microsoft.com

Microsoft AI for Good Research Lab https://orcid.org/0000-0002-4673-614X

Artur Meller

Washington University in St. Louis https://orcid.org/0000-0002-5504-2684

Ian Humphreys

University of Washington

Samuel Sledzieski

Massachusetts Institute of Technology

Yixi Xu

Microsoft AI for Good research lab

Rahul Dodhia

Microsoft AI for Good research lab

Eric Horvitz

Microsoft

Bonnie Berger

Massachusetts Institute of Technology

Gregory Bowman

University of Pennsylvania https://orcid.org/0000-0002-2083-4892

Juan Lavista Ferres

Microsoft AI for Good Research Lab

David Baker

University of Washington https://orcid.org/0000-0001-7896-6217

Minkyung Baek

Seoul National University https://orcid.org/0000-0003-3414-9404

Article

Keywords:

Posted Date: April 26th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-4215086/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

Abstract

The majority of proteins must form higher-order assemblies to perform their biological functions. Despite the importance of protein quaternary structure, there are few machine learning models that can accurately and rapidly predict the symmetry of assemblies involving multiple copies of the same protein chain. Here, we address this gap by training several classes of protein foundation models, including ESM-MSA, ESM2, and RoseTTAFold2, to predict homo-oligomer symmetry. Our best model named Seq2Symm, which utilizes ESM2, outperforms existing template-based and deep learning methods. It achieves an average PR-AUC of 0.48 and 0.44 across homo-oligomer symmetries on two different held-out test sets compared to 0.32 and 0.23 for the template-based method. Because Seq2Symm can rapidly predict homo-oligomer symmetries using a single sequence as input (~ 80,000 proteins/hour), we have applied it to 5 entire proteomes and ~ 3.5 million unlabeled protein sequences to identify patterns in protein assembly complexity across biological kingdoms and species.

Introduction

Across nature, proteins often form assemblies involving multiple subunits to perform their biological functions. When multiple identical protein subunits are held together by non-covalent interactions, the resulting protein complex is called a homo-oligomer. Homo-oligomers can range in size from dimers, which have two identical subunits, to large oligomeric complexes with hundreds of subunits. Homo-oligomerization can be essential for the protein' stability, folding, and function. For instance, some enzymes require the formation of a homo-oligomer to recognize their substrates [1].

The global arrangement of the identical subunits (> = 95% sequence identity over 90% of the length of the subunits) in a homo-oligomer defines their symmetry. This can be either *point group symmetry*, involving the placements of subunits along one or more axes of rotation, or a *helical symmetry*, which involves both rotation and translation of the subunits along the axis of rotation [2]. The most common type of point group symmetry is cyclic (C_n symmetry) where the complex consists of n subunits rotated around a central axis. For example, this type of symmetry is often found in membrane proteins [3] which require a central pore, such as the β-Barrel pore-forming toxins (β-PFT), a large family of bacterial toxins [4]. Another common point group symmetry is dihedral symmetry (D_n symmetry), in which homo-oligomers contain both a rotational axis of symmetry and perpendicular axes of two-fold symmetry. Dihedral symmetry is common among cytoplasmic enzymes because it facilitates a variety of protein-protein interfaces, enabling allosteric control [2]. In addition, homo-oligomers may adopt a cubic symmetry that combines 3-fold rotational axes with non-perpendicular rotational axes such as icosahedral symmetry seen in viral capsules.

Despite the importance of homo-oligomerization for protein function, predicting the quaternary state and symmetry group of a protein given a single chain remains challenging. Currently, annotations of oligomeric states in the Protein Data Bank (PDB) are based on predictions from the PISA algorithm [5, 6], supplemented by the assignments made by the researchers who deposit the structure. Although PISA is

recognized for its high accuracy [6], this method relies upon an experimentally determined structure to extract assembly information and inform the most likely oligomeric state.

Methods that predict oligomeric state without experimental data often rely on homology template searches (such as HHSearch [7]) against known assemblies or employ docking-based symmetric transformations of monomers to model complexes [8]. One such method, GalaxyHomomer [9], combines template-based and docking-based approaches, and incorporates loop refinement to improve structure prediction. Recently, as a result of methods for highly accurate protein structure prediction [10, 11], AlphaFold has been shown to predict homo-dimers at a proteome-scale, and in select cases higher order oligomeric assemblies [12]. However, using AlphaFold [11] or RoseTTAFold [10] for *ab initio* oligomeric state prediction poses significant computational challenges, as it requires running inference for each potential number of chains to score various copy number models, and is generally limited to proteins with high-quality MSAs. More computationally efficient methods to fold large protein oligomers, such as Uni-Fold Symmetry [13] still require the pre-specified symmetry group as input to make predictions. Protein embeddings from ESM2 [14] have been used to predict the most likely quaternary state of a protein chain (QUEEN [15]); however, in this approach, the model only predicts the multiplicity of the oligomer thereby giving no clue as to global symmetry of the protein.

We set out to fine tune protein foundation models (pFMs) to predict homo-oligomer symmetry. We define as "*pre-trained*" any approach that involves a protein model being used as a feature extractor feeding a classifier model (for instance, a logistic regression or neural network classifier) which is then trained on homo-oligomer symmetry prediction. We use *"fine-tuning"* to refer to any approach that involves modifying any parameters from the protein model by training them explicitly for oligomer symmetry prediction. Our approach, outlined in Fig. 1, can be applied to diverse protein families and its rapid runtime (see Fig. 2f) enables proteome-scale annotations.

Results

We evaluate the various methods on our PDB-derived benchmarking dataset consisting of 129,014 structures which are split into training, validation and test splits in a sequence-aware manner (30% sequence identity over 80% coverage is used to define sequence-similar proteins). We use two additional datasets for evaluation: a UniFold test set [13] and a denovo set of proteins (see Methods for details).

An ESM2 fine-tuned model better predicts protein homo-oligomer symmetry than a template-based method and other protein language models

We evaluated pre-trained and fine-tuned variants of ESM-MSA [16], ESM2 [14] and RoseTTAFold2 [17] (RF2) against a template-based method, HHSearch [7], using metrics suited for class-imbalanced datasets and multi-label classification: Area Under Precision-Recall Curve (AUC-PR), confusion matrices, F1-scores, and Precision-Recall curve plots (detailed results in Supplementary Tables S1-S3 and Supplementary Figs. S1-S4). We experiment with fine-tuning a varying number of layers in the pFM and

try different feed-forward neural network architectures for the “classifier head” block shown in Fig. 1 (see Methods).

We find that an ESM2 fine-tuned model (which we call Seq2Symm) with a modified version of the language modeling head used in RoBERTa, trained with margin loss (see Methods) performs the best on all datasets (Fig. 2a,b), with further improvements seen with an expanded training dataset (i.e. with distillation). The next best macro averaged AUC-PR is obtained by the ESM2 pre-trained model with 0.47, 0.40 and 0.38 on the validation, test, and UniFold [13] test sets respectively (see Methods for dataset details) while ESM-MSA fine-tuned and RF2 fine-tuned models perform worse (test AUC-PR of 0.36 and 0.34 respectively). In comparison, the template-based HHSearch method achieved a far lower performance than most of the pFMs with a AUC-PR of 0.38, 0.32 and 0.23 on the validation, test, and UniFold sets respectively, with RF2 pre-trained performing the worst.

Next, we interrogated the strengths and weaknesses of Seq2Symm, by looking at class-wise AUC-PR (Fig. 2d and class distribution of the test set shown in Fig. 2g). We find that it accurately identifies proteins across most cyclic symmetries (except C7-C9), across D symmetries such as D2, D3, D5, as well as helical and icosahedral symmetries. Looking at the confusion matrices, Seq2Symm has a lower tendency to overpredict the majority C1 and C2 class (Fig. 2c; bottom), unlike many other models (Supplementary Fig. S4) and compared to the template-based method HHSearch (Fig. 2c; top), it is more likely to correctly predict dihedral, higher order cyclical, helical, and icosahedral symmetry groups. The only symmetries it performs worse on are ‘O’ (octahedral) and ‘T’ (tetrahedral). Some confusion categories are more easily rationalized: C4 confused as D2, C10-C17 are confused as helical, C7-C9 are confused as D6-D12 (likely due to the co-occurrence of some higher-order C and D symmetry groups) (Supplementary Table S6), while some others are unclear: O confused as C6. Finally, visualizing Seq2Symm’s embedding space in 2D using t-SNE (Supplementary Fig. S5) reveals several clusters of points with the same class label (e.g., C5), suggesting that the model learned to segregate protein sequences by their homo-oligomer symmetry.

We observed that ESM2-based models outperform RF2 and ESM-MSA (Fig. 2a) and analyze this difference further in Fig. 2e, where we average the performance of the various sequence-based models and compare it with the average of the various MSA-based models. We find that the sequence-based models (blue bars) outperform the MSA-based models (gray bars) on all oligomer symmetries except ‘O’ and ‘I’. To understand the reason behind the strength of sequence-based models, we construct another training regime that features no homology between examples in the train and validation/test splits (e-value < 0.1; see Supplementary material Table S8), that we call the no-homology split. Interestingly, we find that while ESM2 still marginally outperforms other methods on this split, the difference between methods and the overall performance significantly decreases (Fig. S10a,b). This, along with model performance on *de novo* designed proteins (Fig. S8), indicates that pLM approaches struggle on proteins without homology to the training dataset (see Fig. S10c,d for homology-related statistics in the two splits). Orthologous proteins can adopt different oligomeric states in different organisms (Fig. S11b,e and Fig. S12), and our PDB-derived dataset may underestimate the diversity of homo-oligomer symmetries

within a protein family. These diverse oligomeric symmetries may contribute significant noise in MSA-based methods where homology to a query sequence is sufficient to predict the oligomeric symmetry of the query protein.

Fine-tuning protein language models improves performance on minority classes

Given that fine-tuned models outperform pre-trained models on our test dataset (Fig. 2a), we investigated which homo-oligomer classes explain the difference in performance (Fig. 3a). We find that fine-tuning improves performance on higher-order oligomer symmetries, which are also rarer in the dataset, except in the cases of class C5, where the pre-trained models substantially outperform the fine-tuned model. For the most frequent symmetries in our dataset, such as C1, C2, C3, D2, we find that pre-trained models are comparable, suggesting that the effects of fine-tuning are most significant for minority data classes. We also note that the benchmarking test split which contains a greater percentage of higher-order oligomer symmetries demonstrates the highest gains from fine-tuning as compared to the Unifold test split, where many of these rarer symmetries are missing (see Table S7). These trends are representation-agnostic, as we see a similar behavior with both the MSA-based (ESM-MSA) and the sequence-based (ESM2) models.

To further investigate the benefits of model fine-tuning, we compare the quaternary state predictions of Seq2Symm (our best ESM2 fine-tuned model) with those of a prior approach, QUEEN, which uses a pre-trained ESM2 model. Specifically, QUEEN uses ESM2 feature embeddings and a supervised head with a single layer (equivalent to logistic regression) to predict one of several quaternary state classes in a multi-class classification setting (see supplementary Table S4-S5, and Fig S7 for detailed results). We created a filtered version of our test dataset for comparison purposes by removing proteins that are homologous to any proteins in the QUEEN training set (30% identity, 80% coverage, 1e-3 e-value). This results in a 66% reduction in test structures (from 64,723 to 21,441). Since QUEEN only predicts quaternary state, we convert our test data labels from homo-oligomer symmetries to quaternary states (ex: ‘C1’ is mapped to 1) producing a many-to-many mapping (‘D6’ and ‘C12’ mapped to 12; ‘C14’ and ‘D7’ mapped to 14; ‘O’ and ‘I’ mapped to 24). This is difficult for some symmetries (‘H’ can be any of 5,6,7…10,12,13–18,24 and ‘T’ can be either 12 or 24) and as a result, some structures (848), were discarded due to the lack of a unique label match. Mapping Seq2Symm’s output to a unique quaternary state, for comparison to QUEEN, is complicated by the fact that we coalesce some higher-order symmetries into a single class (Fig. 3c). Nonetheless, Seq2Symm shows superior performance to QUEEN’s pre-trained ESM2 model on all quaternary states except 24. This suggests that fine-tuning protein language models for a specific task, rather than simply using their embeddings as inputs to a trainable classifier, can improve performance.

Rapid predictions of homo-oligomer symmetry across proteomes reveal patterns across biological kingdoms

Given the rapid inference time of Seq2Symm, we apply it to five proteomes (*Pyrococcus furiosus*, *Escherichia coli*, *Saccharomyces cerevisiae*, *Homo sapiens*, and *Exaiptasia pallida*) and to a large set of

~ 3.5 million unlabeled sequences from UniRef50 and metagenomic sources, spanning diverse life forms (see Table S9 for details). We show the distribution of various homo-oligomer symmetries, shown as a percentage of the proteome, among the five proteomes in Fig. 4a. We find that the distribution of homo-dimers in our predictions for the four proteomes, 45%, 42%, 35%, 35% in *P. furiosus, E. coli, S. cerevisiae, H. sapiens* respectively, aligns with the findings from [12] (which reported 43%, 44%, 21%, and 21% of the four proteomes respectively). Across the five proteomes, the prevalence of higher order symmetries is similar among simpler organisms (*P. furiosus* and *E. coli*) and among the complex organisms (*S. cerevisiae*, *H. sapiens*, *E. pallida*), except in the case of proteins with Helical ('H'), Octahedral ('O') and Icosahedral ('I') symmetries. In Fig S14, we see the prevalence of multiple homo-oligomeric symmetries per protein in the five proteomes. ~20% of the proteins from *P. furiosus* and *E. coli* have more than one symmetry, while this statistic is ~ 13% for *S. cerevisiae* and *H. sapiens.*

To analyze Seq2Symm's homo-oligomer symmetry predictions over the ~ 3.5 million unlabeled proteins, we assign each protein to a superkingdom / kingdom using annotations from UniprotKB and the Taxonomy database. Bacterial proteins constitute 53% of the proteins in this set and the rest come from other organisms (see the "Overall" bar). In Fig. 4(b), we show the percentage of proteins in each symmetry class from the various life-forms. We find a significantly higher representation of simpler organisms, mainly bacteria, in the lower-order symmetries C1, C2, C3, D2, D3, with exceptions seen for 'C6', 'H', 'O', 'T'. Reassuringly, viral proteins are overrepresented among 'I' (icosahedral) homo-oligomers. Higher-order C symmetries ('C4','C5', 'C7-C9') see significantly higher representation, and D symmetries ('D4', 'D5', 'D6-D12') are, to some extent more prominent, in higher-order organisms.

To demonstrate the utility of Seq2Symm in generating structures for higher order oligomeric symmetries, we use Seq2Symm's highest confidence predictions as chain copy number inputs to AlphaFold2 Multimer [18, 19] and generate structures which are depicted in Fig. 4(c). By using Seq2Symm, it is possible to bypass an exhaustive search of different homo-oligomer quaternary states as is historically done [12] and instead predict a single homo-oligomer structure based on the output from Seq2Symm (see Supplementary information for details).

Discussion

We describe a rapid deep learning method for accurate prediction of homo-oligomer symmetry. Our approach is computationally efficient and, unlike template-based approaches, does not rely on the availability of symmetry annotations for homo-oligomers on homologous structures. We explore various configurations involving different pFMs and find that the sequence-based ESM2 model upon fine-tuning (Seq2Symm) outperforms template-based homology searches, as well as approaches that finetune or use ESM-MSA or RF2 as pre-trained feature extractors.

Seq2Symm outperforms previous methods and pre-trained approaches by an average of approximately 19% on higher-order oligomer symmetries, which are less common in our dataset. Its performance on rare classes, such as C6, C10-C17, and D5, is notable, with AUC-PRs of 0.71, 0.78, and 0.49 on the test dataset,

respectively. This suggests that Seq2Symm has likely learned unique characteristics of these more complex oligomeric symmetry proteins. Our training setup, which involves 1) grouping similar and very rare higher order symmetries into broader classes (e.g., C10-C17 and D6-D12), 2) oversampling the minority classes, and 3) undersampling the majority classes, leads to a reasonable performance on C3, C4, C5, C6, D2, D3, D5, H, and I. We also include a hierarchical loss term for 'coarse classes', grouping all higher-order C and D symmetries into single CX and DX classes, respectively. These strategies have improved the Seq2Symm's performance from 0.50 to 0.52 on the validation set and the ESM-MSA fine-tuned model's performance from 0.27 to 0.45 on the same set. Further improvement is achieved through distillation (see Methods), boosting performance from 0.52 to 0.58 on the validation set.

We find that proteins with identical sequences, MSAs, and very similar structures often have different labels. For example, the same bovine seminal ribonuclease is assigned different symmetry labels across similar PDB entries (e.g., C2 for 11ba, 'C2, D2' for 11bg). This diversity in labels acts as "noise" for machine learning models and presents a significant challenge. Our analysis of the errors made by Seq2Symm reveals that incorrect predictions are often made on proteins that belong to clusters with heterogeneous oligomer symmetries. For instance, in Fig. S9a which shows the model's predictions for one such heterogeneous cluster, several 'H' symmetry class examples are predicted to be 'C1' or 'CX' (some higher order C symmetry) possibly because the cluster contains many structures with 'C1' and 'C6' labels. Prior work has estimated that such discrepancies in oligomer-symmetry annotations indicate that ~ 10% of biological assembly labels in the PDB are potentially incorrect [20]. Although we do not speculate on the accuracy of biological assembly labels here, we aim to estimate the extent to which such discrepancies might affect a machine learning model's prediction performance. Towards this end, we cluster our dataset at 90% sequence identity with 90% coverage and analyze label similarity within each cluster. We find that approximately 8.8% of validation proteins and 11% of test proteins have different symmetries from other homologous proteins with which they share very high sequence identity (see Fig S15 and the accompanying text).

Several avenues exist for improving the performance of models in this study. Currently, all errors are assigned the same penalty during training, but adjusting loss based on class relationships could offer a more nuanced approach, for instance, a misprediction from C3 to C4 being penalized less than one to C17 or D10. This approach, facilitated by a matrix of size = (# labels) x (# labels), could allow for expressing and optimizing misprediction penalties in a context-aware manner. Our models consider sequence and MSAs as input representations; however, we can also incorporate the structure of the single chain as input (the true structure from PDB where available or the predicted structure otherwise). This is straightforward for models such as RoseTTAFold2, which are designed to input 3D representations. One possibility to extend the applicability of sequence-based models like ESM2, is to use embeddings of 3D structures from structure prediction models, as additional inputs.

Lastly, our work predicts homo-oligomer symmetry for a protein, which does not always explicitly encode the quaternary state of the protein (number of subunits), especially in the case of helical ('H') and icosahedral ('I') symmetries. Predicting the symmetry type and quaternary state simultaneously with a

single model (e.g., ‘H’ with 6 chains, ‘I’ with 180 chains) could improve its utility, possibly without compromising performance.

Nonetheless, Seq2Symm, in its current form, accelerates the modeling of homo-oligomer structural models and the annotation of symmetry groups at the proteome scale. By integrating the output from Seq2Symm with protein structure prediction algorithms, it becomes possible to generate physically realistic 3D structural models of complicated homo-oligomers (Fig. 2b, 4c). Furthermore, Seq2Symm’s rapid runtime facilitates the comparison of symmetry group distributions across different species and kingdoms. Thus, Seq2Symm has the potential to become a valuable tool for both proteomic-scale protein structure prediction and comparative analysis.

Online Methods

Datasets

We derive a dataset of 298,771 homo oligomeric labels over 129,014 structures from the PDB. For each structure, the global symmetry annotations assigned to all the deposited biological assemblies are considered while defining the homo-oligomer symmetry of the structure. We use all annotations in a multi-label prediction setting and try different approaches to incorporate the multiple labels of a single structure such as using “soft labels” for all symmetry annotations other than the one from “biological assembly 1”, lower misclassification penalty for annotations from later biological assemblies. We find that treating all labels equally results in a model with the best performance on validation data.

The symmetry annotations in the PDB are assigned by the depositing authors and/or computed by the PISA algorithm [6], which computes various statistics using the deposited atomic structures obtained from X-ray crystallography experiments. PISA uses a scoring function that combines several criteria such as interface contact area, number of interfacial buried residues, salt bridges, disulfide bonds etc. [21] to distinguish the biologically relevant interfaces that define an oligomeric complex from the irrelevant lattice contacts in protein crystals. While there are several other newer tools in the field [22], PISA is still considered the gold-standard for estimating the quaternary state [23].

There are 45 different homo-oligomer symmetry labels in our dataset, the most frequent being the ‘monomeric’ (C1) and the ‘cyclic dimeric’ (C2) symmetries while higher order symmetry labels are less well represented (Supplementary Table S6). Since certain structures have multiple assemblies, these can have multiple homomer symmetries in our dataset. For example, 6nal [24] has ‘C1’ and ‘C2’ labels. There are 17,758 such structures (~ 6% of the dataset), with some structures having as many as 4 labels and a total of 131 different label combinations (Supplementary Table S6).

Data splits

To ensure that the test data does not contain homologs of proteins seen during training, we create the train/validation/test splits based on sequence similarity. We use MMSeqs2 [25] with a threshold of > 30%

sequence identity and > 80% sequence coverage to cluster the structures, which results in a total of 19,200 clusters. Each cluster is then assigned to one of the train, validation or test splits. This is a more relaxed criterion for clustering proteins as compared to the > 80% coverage and > 50% sequence identity cut-offs used in the ESM models [14, 26], thereby resulting in data splits that are better separated (less “leakage” between train and evaluation splits). The multi-domain structure of proteins is most likely preserved when using a coverage of > 80%.

We select 70% of the clusters to be the training data (13,433 clusters), 10% for the validation split (1860 clusters) and 20% for the test split (3907 clusters). In terms of protein structures, this is equivalent to 205,548 training, 28,509 validation and 64,723 test structures. In MSAs this equates to 49,584 training, 7,304 validation and 15,710 test MSAs (one MSA made per unique protein sequence, which results in fewer MSAs due to homologous proteins). All machine learning methods were trained using the same data splits. We discuss the “no homology” split in the Supplementary material.

Evaluation

We use the training split for training all models and the validation split for hyper-parameter and model selection. The test split and other evaluation sets were unseen until the final models were selected and were then used to evaluate final performance.

UniFold test

In addition to our curated dataset, we use another completely unseen set of protein structures for the final evaluation of the models, curated from the test set of the UniFold structure prediction model [13], which has 163 structures. After filtering for hetero-oligomer labels we get 96 structures, of which, we were able to construct MSAs for 94 structures using the HHblits algorithm [27] with an e-value cut-off 1e-3 (searching over the following databases: protein sequences from the UniRef30_2023_02 [28] version and BFD [29]) and filtered for quality using hhfilter with 90% identity and 75% coverage. We remove proteins that were sequence-similar to our homo-oligomer dataset (30% identity, 80% coverage, 1e-3 e-value); this gives us 83 structures with 85 labels (Supplementary Table S7). All methods are evaluated on these 83 structures as we have both sequence and MSA for these.

De novo test

To expand the scope of inference and test the transferability of methods to examples which have highly divergent amino acid sequences from those in the training data, we curated a small test-set of *de novo* designed proteins. Additionally, in protein design, an *in silico* method to screen for oligomeric symmetry prediction would assist in oligomeric design. We collect experimentally resolved symmetric oligomers generated using *hallucination* [30] and *RFdiffusion* [31] and sequences fit with *ProteinMPNN* [32]. Symmetry groups of designs were validated using one or more of the following methods: size-exclusion

chromatography (SEC-MALS), negative stain electron microscopy (nsEM), cryo-EM, or X-ray crystallography (Supplementary Table S10).

Class imbalance

Our dataset of protein homomer symmetries is heavily class imbalanced due to the high prevalence of certain symmetries such as C1 (monomers) and C2 (cyclic dimers) (Supplementary Table S6). Given the dearth of labels on several higher order symmetry categories, we either prune very rare classes or merge the rarer categories into larger groups. The following classes are merged: C7-C9, C10-C17, and D6-D12.

While the merging of the rarer classes addresses this to an extent, further techniques are needed to adjust for class imbalance during training. We use under-sampling of the majority class, where we under-sample the “majority” classes C1 and C2 to 70% of their original sizes, followed by oversampling of the minority classes. We over-sample the “extreme minority” classes with fewer than 10,000 protein structures to 5 times their original size (for instance C5 with ~ 4,000 training examples is upsampled to be ~ 20,000 examples) and the “moderate minority” classes which have more than 10,000 examples to twice their original size. For example, C3 with ~ 7,000 training examples is upsampled to ~ 14,000 examples. This sampling is done as a pre-processing step prior to training, and is only done on the training split of the data and all models were trained using the same sampled dataset.

Pre-trained models as feature encoders:

The pre-trained models: ESM-MSA, ESM2, and RoseTTAFold2 are used as feature encoders, whereby input proteins are embedded using the hidden layer representations from the neural network models.

The ESM-MSA Transformer uses only the multiple sequence alignment (MSA) of the given protein as input and produces 768-dimensional embeddings for each input residue, which we aggregate by averaging into a single 768-dimensional embedding. To prevent out-of-memory errors during inference, we crop input MSAs by truncating the N-terminal portion of the sequence at 1024 residues. Further, we select no more than 128 protein sequences per MSA using a greedy selection algorithm based on pairwise Hamming-distances between the protein sequences from the input MSA, as prescribed in the original work.

We obtain 256-dimensional embeddings from RoseTTAFold-2 (RF2), by excluding the 3D track of the model and using as inputs: the MSA of the given protein (1D track), a default structure template (2D track), and averaging the embeddings over the residues of the input protein. The input MSA is cropped to a length of 1024 residues for computational efficiency, by taking a random region of the MSA of length 1024.

The ESM2 model operates on single protein amino acid sequences as input and the embeddings produced by this model have a dimensionality of 1,280. We specifically use the esm2_t33_650M_UR50D version of the model consisting of 33 layers and 650M parameters that was trained on the UniRef50

database. Analogous to ESM-MSA, we truncate protein sequences longer than 1024 amino acids, by deleting the N-terminal.

Given the embeddings obtained from these pre-trained models as the “features” for an input protein, we train supervised models for predicting the protein’s homomer symmetry using both linear (logistic regression) and non-linear model architectures.

Fine Tuning protein language models

In addition to using the protein language models as pre-trained feature extractors, we also fine-tune the weights from the original models to adapt to the task of homo-oligomer symmetry prediction. While fine-tuning these models, we do not use the loss functions (such as masked amino-acid prediction or pLDDT, etc.) that were used to train the original models and instead optimize the model for predicting the homomer symmetry. Towards this, we experiment with the following supervised neural network architectures and loss functions.

We fine-tune 1, 2, 4, 8 layers from all protein language models, and the following additional options for the number of layers from ESM-MSA and ESM2: 12 and ‘all’ layers, with gpu memory and compute-time setting the limit on how many layers were possible to fine-tune from each model.

Architecture of the supervised head

Multilayer perceptron:

This is a simple one- or two-layer feedforward neural network with linear or ReLU activation. We do not incorporate layer normalization or drop-out here as we did not see any changes to the performance on the validation set.

RobertaLMHead:

The architecture of this module is an extension of the masked language modeling prediction head from ESM. This module begins with a linear transformation, followed by the application of a GELU (Gaussian Error Linear Unit) [33] activation function, introducing non-linearity. A dropout layer, with a configurable rate, is applied post-activation to enhance model robustness. Subsequently, a custom layer normalization is applied (ESM1bLayerNorm) [34]. Next, we average over the protein residues, creating a summary representation. Finally, the summary representation is linearly mapped to the output dimension (number of classes) using a dense layer.

Multitask RobertaLMHead:

We train one supervised head per class, where each head has a RobertaLMHead architecture. There are thus separate parameters for each class, like in a multitask learning setting, with only the protein language model parameters being shared between them.

Loss functions

Since our goal is multi-label multi-class classification, we use the binary cross entropy with logits loss function (BCEWithLogits). BCE with logits treats each class label independently, where for each label, the loss is computed based on the predicted probability and the true label and the total loss is a summation of the independent class-level loss terms, thereby making it possible for an example to have multiple labels.

Margin loss function:

For each example, we compute a pairwise loss inspired by contrastive learning that constructs pairs of positive and negative labels (all oligomer-symmetries that are not the correct symmetry are considered "negative") and calculate the hinge-loss for each pair as defined below. Given a protein sample x with oligomer symmetry class vector $y = [0,1]^C$ where C is the number of homo-oligomer symmetries, our model F predicts $y\prime \in \Re^C = F(x)$. We then compute the sample loss as

$$L(y', y) = \sum_{\forall(i,c),\, y_i \neq 1,\, y_c = 1} max(0, -(y'_c - y'_i) + m)$$

where m is the margin, here 1.0.

The loss encourages the model to rank the positive labels of an example higher than negative labels by at least the margin. This loss is unlike typical contrastive learning losses, where the positives and negatives are examples, rather than different labels of the same example.

Template based prediction of homomer symmetry

We also implement a template matching procedure using HHSearch [7]. The homomer symmetry label is assigned based on the homomer labels of the matched structure in the PDB. A matched structure is considered only if the sequence identity between the query and the target satisfies the specified threshold '*t*'. We assign all labels from the top-k matches which results in a multi-label output. To be consistent with the comparison to the machine learning models, we only consider "hits" with sequence identity of at most 30%. Note that this allows HHSearch access to all proteins from the validation/test splits that have a sequence identity < 30%. We vary the number of top matches '*K*' that we consider from the returned hits, to get a trade-off between the precision and recall for this approach. The other parameters used to run HHSearch are: maximum number of hits of 1000, e-value threshold of 0.001 (see supplementary material and Table S13 for details). No labels are predicted for proteins where no hits were found matching the specified thresholds. No matches are found for 736 and 1063 proteins from the validation and test sets respectively.

Distillation

Distillation, also called pseudo-labeling in semi-supervised learning, involves training an initial model on labeled data and using it to assign new "pseudo" labels on a very large unlabeled dataset. These "newly labeled" examples are subsequently used to expand the training dataset, on which a secondary model is

trained, which can often generalize better. We use Seq2Symm to generate new labels on our distillation dataset which contains ~ 7.6 million proteins from UniRef50. We first select proteins satisfying the cut-off of pLDDT > 0.8 and pick one random protein per cluster (where clusters were protein sequence based at 30% identity, 80% coverage, 1e-3 e-value) giving us ~ 2.8 million input proteins. We run inference on this set to get predicted probabilities per oligomer symmetry class and select all structures that satisfy our class-specific classifier thresholds and exclude C1 or C2 predictions, on account of their over-representation in our gold standard dataset.

Declarations

Acknowledgements

We acknowledge funding from Bill and Melinda Gates Foundation #OPP1156262 (I.R.H.).

M.B. was supported by IITP/MSIT (RS-2023-00220628), NRF/MSIT (RS-2023-00210147), and the New Faculty Startup Fund from Seoul National University.

D.B. is a Howard Hughes Medical Institute investigator.

S.S. was supported by the NSF Graduate Research Fellowship under Grant No. 2141064.

AM was supported by the National Institutes of Health F30 Fellowship (1F30HL162431-01A1).

Data and source availability

All code, models, datasets, predictions from this work will be available at https://github.com/microsoft/seq2symm

References

1. Luo, M., & Tanner, J. J. (2015). Structural basis of substrate recognition by aldehyde dehydrogenase 7A1. *Biochemistry*, *54*(35), 5513-5522.
2. Goodsell, D. S., & Olson, A. J. (2000). Structural symmetry and protein function. *Annual review of biophysics and biomolecular structure*, *29*(1), 105-153.
3. Forrest, L. R. (2015). Structural symmetry in membrane proteins. *Annual review of biophysics*, *44*, 311-337.
4. Leone, P., Bebeacua, C., Opota, O., Kellenberger, C., Klaholz, B., Orlov, I., ... & Roussel, A. (2015). X-ray and cryo-electron microscopy structures of monalysin pore-forming toxin reveal multimerization of the pro-form. *Journal of Biological Chemistry*, *290*(21), 13191-13201.
5. Krissinel, E., & Henrick, K. (2007). Inference of macromolecular assemblies from crystalline state. *Journal of molecular biology*, *372*(3), 774-797.

6. Krissinel, E. (2015). Stock-based detection of protein oligomeric states in jsPISA. *Nucleic acids research*, *43*(W1), W314-W319.
7. Steinegger, M., Meier, M., Mirdita, M., Vöhringer, H., Haunsberger, S. J., & Söding, J. (2019). HH-suite3 for fast remote homology detection and deep protein annotation. *BMC bioinformatics*, *20*(1), 1-15.
8. Yan, Y., Tao, H., & Huang, S. Y. (2018). HSYMDOCK: a docking web server for predicting the structure of protein homo-oligomers with Cn or Dn symmetry. *Nucleic acids research*, *46*(W1), W423-W431.
9. Baek, M., Park, T., Heo, L., Park, C., & Seok, C. (2017). GalaxyHomomer: a web server for protein homo-oligomer structure prediction from a monomer sequence or structure. *Nucleic acids research*, *45*(W1), W320-W324.
10. Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., ... & Baker, D. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, *373*(6557), 871-876.
11. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583-589.
12. Schweke, H., Pacesa, M., Levin, T., Goverde, C. A., Kumar, P., Duhoo, Y., Dornfeld, L. J., Dubreuil, B., Georgeon, S., Ovchinnikov, S., Woolfson, D. N., Correia, B. E., Dey, S., & Levy, E. D. (2024). An atlas of protein homo-oligomerization across domains of life. In *Cell*. https://doi.org/10.1016/j.cell.2024.01.022
13. Li, Z., Yang, S., Liu, X., Chen, W., Wen, H., Shen, F., ... & Zhang, L. (2022). Uni-Fold Symmetry: harnessing symmetry in folding large protein complexes. *bioRxiv*, 2022-08.
14. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., ... & Rives, A. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, *379*(6637), 1123-1130.
15. Avraham, O., Tsaban, T., Ben-Aharon, Z., Tsaban, L., & Schueler-Furman, O. (2023). Protein language models can capture protein quaternary state. *BMC Bioinformatics 24,* 433, 2023.
16. Rao, R. M., Liu, J., Verkuil, R., Meier, J., Canny, J., Abbeel, P., ... & Rives, A. (2021, July). MSA transformer. In *International Conference on Machine Learning* (pp. 8844-8856). PMLR.
17. Baek, M., Anishchenko, I., Humphreys, I., Cong, Q., Baker, D., & DiMaio, F. (2023). Efficient and accurate prediction of protein structure using RoseTTAFold2. *bioRxiv*, 2023-05.
18. Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., ... & Hassabis, D. (2021). Protein complex prediction with AlphaFold-Multimer. *biorxiv*, 2021-10.
19. Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., & Steinegger, M. (2022). ColabFold: making protein folding accessible to all. *Nature methods*, *19*(6), 679-682.
20. Dey, S., Ritchie, D. W., & Levy, E. D. (2018). PDB-wide identification of biological assemblies from conserved quaternary structure geometry. *Nature methods*, *15*(1), 67-72.
21. Henrick, K., & Thornton, J. M. (1998). PQS: a protein quaternary structure file server. *Trends in biochemical sciences*, *23*(9), 358-361.

22. Luo, J., Guo, Y., Fu, Y., Wang, Y., Li, W., & Li, M. (2014). Effective discrimination between biologically relevant contacts and crystal packing contacts using new determinants. *Proteins: Structure, Function, and Bioinformatics*, *82*(11), 3090-3100.
23. Yueh, C., Hall, D. R., Xia, B., Padhorny, D., Kozakov, D., & Vajda, S. (2017). ClusPro-DC: Dimer classification by the CLUSPRO server for protein–protein docking. *Journal of molecular biology*, *429*(3), 372-381.
24. Wade, K. R., Lawrence, S. L., Farrand, A. J., Hotze, E. M., Kuiper, M. J., Gorman, M. A., ... & Tweten, R. K. (2019). The structural basis for a transition state that regulates pore formation in a bacterial toxin. *MBio*, *10*(2), 10-1128.
25. Steinegger, M., & Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology*, *35*(11), 1026-1028.
26. Meier, J., Rao, R., Verkuil, R., Liu, J., Sercu, T., & Rives, A. (2021). Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in neural information processing systems*, *34*, 29287-29303.
27. Remmert, M., Biegert, A., Hauser, A., & Söding, J. (2012). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature methods*, *9*(2), 173-175.
28. Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., Wu, C. H., & UniProt Consortium. (2015). UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, *31*(6), 926-932.
29. Steinegger, M., & Söding, J. (2018). Clustering huge protein sequence sets in linear time. *Nature communications*, *9*(1), 2542.
30. Wicky, B. I. M., Milles, L. F., Courbet, A., Ragotte, R. J., Dauparas, J., Kinfu, E., ... & Baker, D. (2022). Hallucinating symmetric protein assemblies. *Science*, *378*(6615), 56-61.
31. Watson, J. L., Juergens, D., Bennett, N. R., Trippe, B. L., Yim, J., Eisenach, H. E., ... & Baker, D. (2023). De novo design of protein structure and function with RFdiffusion. *Nature*, *620*(7976), 1089-1100.
32. Dauparas, J., Anishchenko, I., Bennett, N., Bai, H., Ragotte, R. J., Milles, L. F., ... & Baker, D. (2022). Robust deep learning–based protein sequence design using ProteinMPNN. *Science*, *378*(6615), 49-56.
33. Hendrycks, D., & Gimpel, K. (2016). Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*.
34. esm/esm/modules.py at main · facebookresearch/esm · GitHub
35. Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., ... & Velankar, S. (2022). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic acids research*, *50*(D1), D439-D444.

Figures

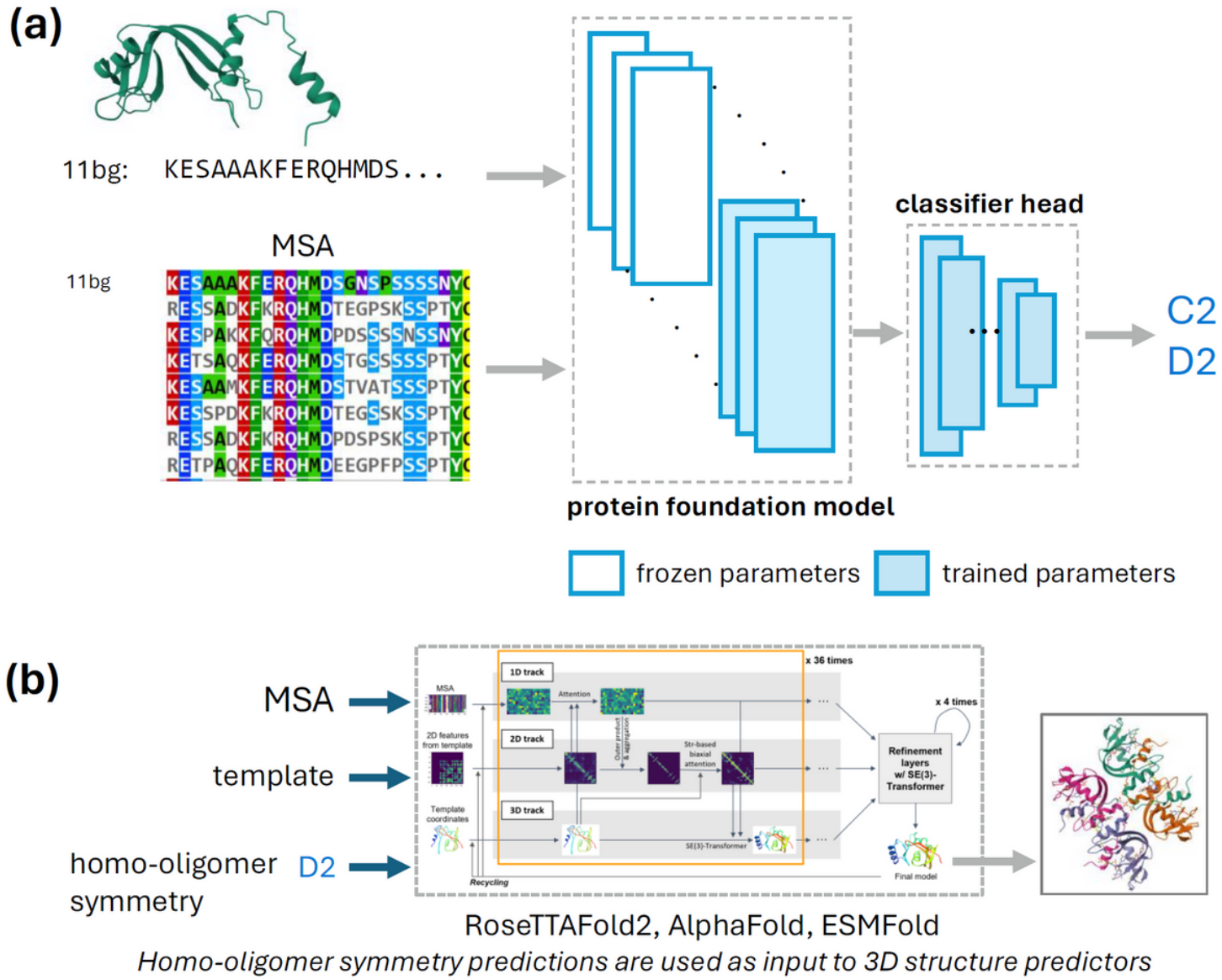

Figure 1

Protein foundation models can be fine-tuned to predict a protein's homo-oligomer symmetry. (a) Schematic showing our modeling setup for multi-label prediction of homo-oligomer symmetry, illustrated for the protein **11bg**. The input can be either the protein amino-acid sequence and/or the multiple sequence alignment (MSA). The 'protein foundation model' (pFM) can be ESM-MSA, ESM2, or RoseTTAFold2 (RF2). We experiment with various architectures for the 'classifier head' (see Methods). We vary the number of layers we fine-tune in the pFM, from a fully frozen model with a single trainable prediction head (i.e., "pre-trained only") to a model with all weights freely tunable (i.e., "fine-tuned"). (b) The homo-oligomer symmetry prediction can then be supplied to a structure prediction algorithm (e.g., AlphaFold, RoseTTAFold2, or ESMFold) to guide the generation of an atomic-resolution homo-oligomer structure.

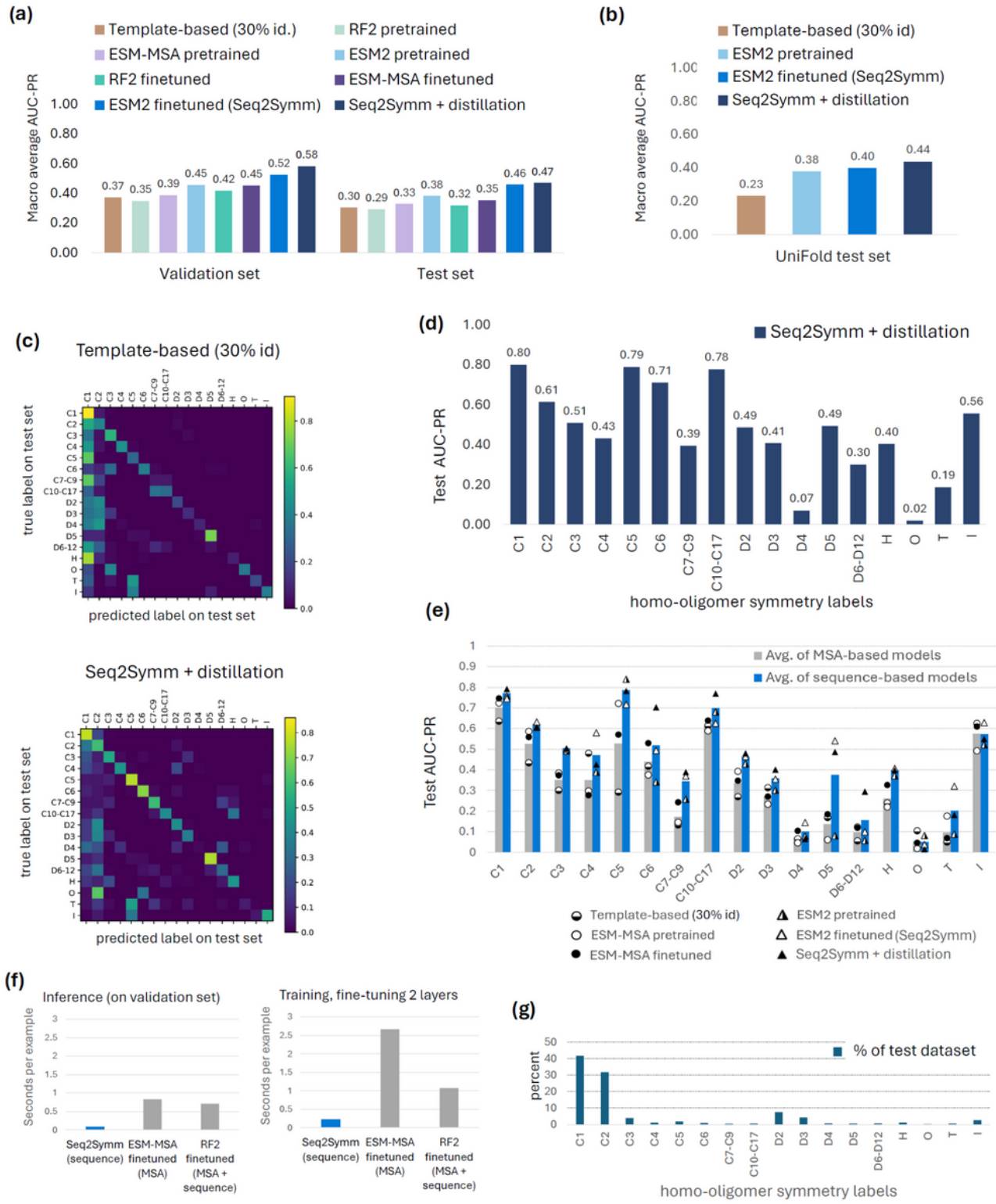

Figure 2

Protein foundation models predict homo-oligomer symmetry more accurately than current template-based methods. (a) Performance, measured using area under the precision-recall curve (AUC-PR), for the various methods on the validation split and the held-out test split of our dataset. The AUC-PR shown is the macro-average over class-wise AUC-PR, with class-weighted AUC-PR results in Supplementary Fig. S1. **(b)** Performance of representative models on another completely unseen dataset from prior work, the

"UniFold test set" (see Methods for dataset details). The AUC-PR is a macro-average over class-wise AUC-PR for the classes in this dataset. **(c)** Confusion matrix of one of the baselines: HHSearch and Seq2Symm (a fine tuned ESM2-based model), showing the symmetries where there is confusion. This matrix is shown for only proteins with a single label (i.e. multi-label examples are excluded). **(d)** Test AUC-PR for each homo-oligomer symmetry shown for the best model, an ESM2-based fine tuned model.**(e)** Class-wise AUC-PR on the test set, averaged over sequence-based models (orange bars) and MSA-based models (blue bars) in the bar chart, with individual model performances in each category shown by the points. We find that the models using a sequence-only representation (triangle points, n=3) achieve a higher AUC-PR for nearly every symmetry class, as compared to the MSA representation based models (circular points, n=3). The biggest gains are seen on higher-order symmetries such as C4, C5, C7-C9, D5. **(f)** Inference and training time taken by each protein foundation model, shown per input example. **(g)** Distribution of homo-oligomer symmetries in our test set.

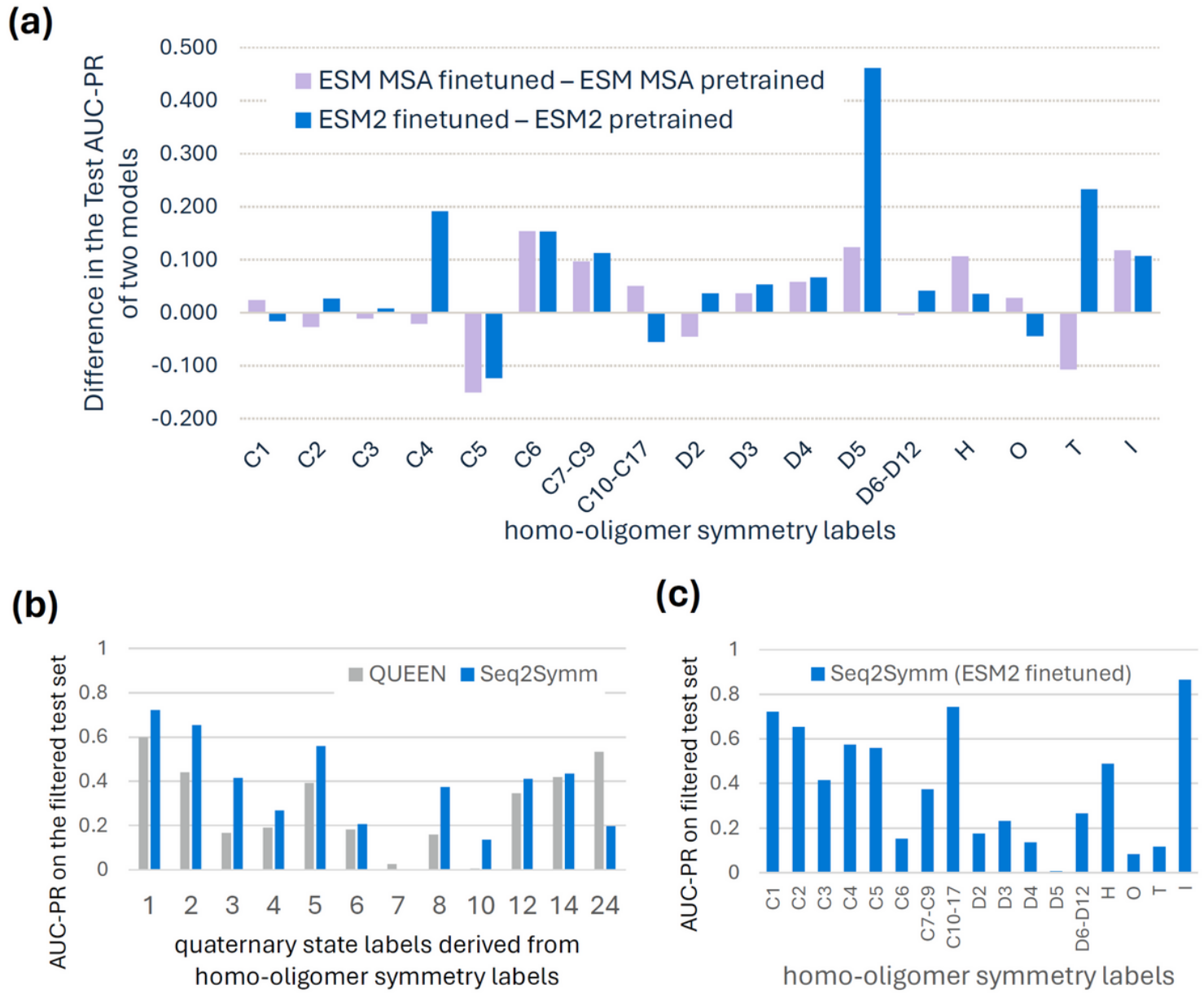

Figure 3

Fine-tuning protein foundation models improves homo-oligomer symmetry prediction and quaternary state prediction. (a) Fine-tuning improves model performance across nearly every symmetry group, with the most improvements over pre-trained performance seen in rarer classes. **(b)** Seq2Symm (ESM2 fine-tuned) outperforms QUEEN, a pre-trained model from prior work on quaternary state prediction **(c)** The class-wise AUC-PR of Seq2Symm on the filtered test set.

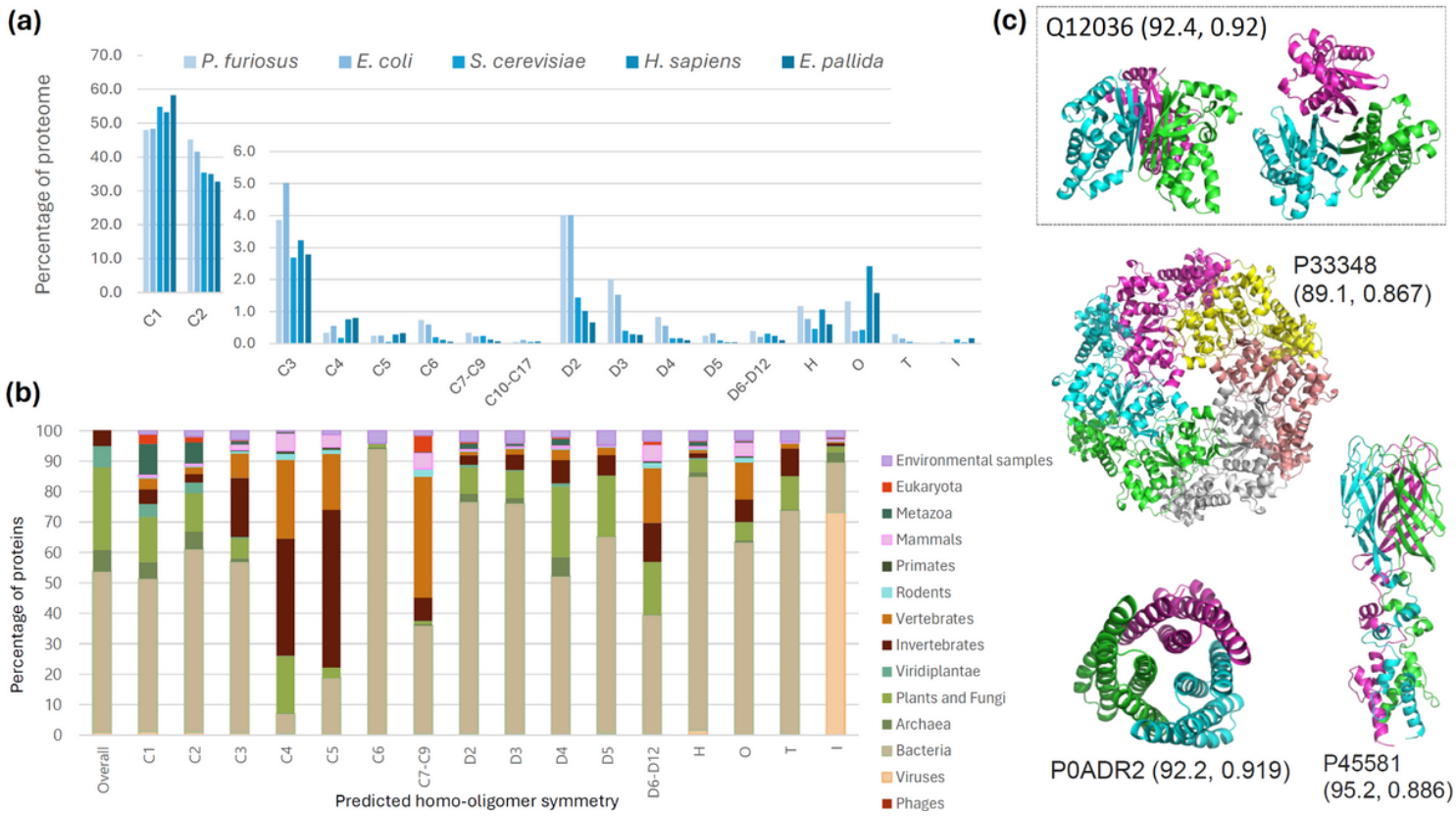

Figure 4

Seq2Symm's rapid predictions enable proteome-wide annotation of homo-oligomer symmetry. (a) Proteome-wide distribution of predicted homo-oligomer symmetries in five different organisms depicted as the percentage of all proteins in the proteome reveals a higher percentage of D2 complexes in *P. furiosus*(an archaea) and *E.coli* and a higher percentage of complexes with octahedral symmetry in *H. sapiens* and *E. pallida* (a sea anemone species). **(b)** Homo-oligomer symmetry predictions for ~3.5 million unlabeled protein sequences across several biologic kingdoms reveal differences in symmetry propensities (e.g., icosahedral symmetry is overrepresented in viruses). For each predicted symmetry, we show the proportion of proteins with that predicted label from each animal kingdom. The leftmost column shows the prevalence of the different kingdoms in the dataset. **(c)** Homo-oligomer structures generated using AlphaFold2 based on some of Seq2Symm's homo-oligomer symmetry predictions with (pLDDT, iPTM) scores shown in brackets: Q12036 (*S. cerevisiae*) with 'C3', P33348 (*E. coli*) with 'C6',

P0ADR2 (*E. coli*) with ‘C3’, P45581 (*E. coli*) with ‘C3’. High structure quality metrics suggest that Seq2Symm’s predictions can aid in generating accurate structures for homo-oligomers.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementaryinformation.docx