Benchmarking Protein Language Models for Protein Crystallization

Raghvendra Mall[1,*], Rahul Kaushik[1], Zachary A. Martinez[2], Matt W. Thomson[2], and Filippo Castiglione[1,3,*]

[1]Biotechnology Research Center, Technology Innovation Institute, P.O. Box 9639, Abu Dhabi, United Arab Emirates
[2]Division of Biology and Bioengineering, California Institute of Technology, Pasadena, 91125, California, United States of America
[3]Institute for Applied Computing, National Research Council of Italy, Rome, 00185, Italy
[*]Corresponding authors: raghvendra.mall@tii.ae, filippo.castiglione@tii.ae

ABSTRACT

The problem of protein structure determination is usually solved by X-ray crystallography. Several *in silico* deep learning methods have been developed to overcome the high attrition rate, cost of experiments and extensive trial-and-error settings, for the predicting the crystallization propensities of proteins based on their sequences. In this work, we benchmark the power of open protein language models (PLMs) through the TRILL platform, a be-spoke framework democratizing the usage of PLMs for the task of predicting crystallization propensities of proteins.
By comparing LightGBM / XGBoost classifiers built on the embedding representations learned by different PLMs, such as ESM2, Ankh, ProtT5-XL, ProstT5, with the performance of state-of-the-art sequence-based methods like DeepCrystal, ATTCrys and CLPred, we identify the most effective methods for predicting crystallization outcomes. The LightGBM classifiers utilizing embeddings from ESM2 model with 30 and 36 transformer layers and 150 and $3,000$ million parameters respectively have performance gains by 3-5% then all compared models for various evaluation metrics, including AUPR (Area Under Precision-Recall Curve), AUC (Area Under the Receiver Operating Characteristic Curve), and F1 on independent test sets.
Furthermore, we fine-tune the ProtGPT2 model available via TRILL to generate crystallizable proteins. Starting with $3,000$ generated proteins and through a step of filtration processes including consensus of all open PLM-based classifiers, sequence identity through CD-HIT, secondary structure compatibility, aggregation screening, homology search and foldability evaluation, we identified a set of 5 novel proteins as potentially crystallizable.

Introduction

Protein structure at atomic resolution is usually determined by X-ray crystallography[1] or nuclear magnetic resonance (NMR)[2]. However, this is an expensive process where $> 70\%$ of the total cost is spent on attempts that do not produce crystals of diffraction quality[3]. Crystallization of proteins is a prerequisite for structural determination. Yet, it has been a daunting challenge, with the overall rate of success ranging between 2 and 10%[4]. The determination of important biological features that help increase the propensity for protein crystallization remains a great challenge. Several machine learning methods and statistical techniques have been developed for sequence-based protein crystallization prediction[5–11]. These approaches utilize feature-based protein representations including physicochemical and k-mer frequency features from amino acid sequences and corresponding structures. Most of these techniques undergo a feature selection procedure(s), followed by traditional machine learning techniques such as support vector machines[12,13], random forests[14] and gradient-boosting machines[15].

The availability of large-scale protein datasets through public databases such as PepcDB[16], enables the use of deep learning techniques for the problem of protein crystallization prediction. DeepCrystal, a deep neural network (DNN) based model was proposed by Elbasir et al.[17] to predict protein crystallization propensity **using just the protein AA sequence as input** without the need to extract additional physio-chemical and k-mer features by implementing convolutional neural networks (CNNs)[18] as backbone. DeepCrystal captures frequently occurring amino acid (AA) k-mers of different lengths driving the crystallization prediction and outperforms state-of-the-art (*sota*) feature-based methods. Furthermore, techniques such as ATTCry[19] design a CNN framework based on multi-scale and multi-head self-attention for crystallization prediction. CLPred[20] uses a bidirectional recurrent neural network with long- and short-term memory (BLSTM) to capture long-range interaction patterns between the k-mers of AA sequence to predict protein crystallizability using the protein AA sequence as input.

A new deep learning pipeline, GCmapCrys[21], was proposed for multi-stage crystallization propensity prediction by integrating the graph attention networks with the predicted protein contact map. Moreover, it uses BLAST[22] to generate a position-specific scoring matrix, SCRATCH-1D (https://scratch.proteomics.ics.uci.edu/) to use predicted

solvent accessibility and secondary structure, and HHblits[23] for multiple sequence alignment (MSA). A similar technique, namely BCrystal[24], utilizes homology, secondary structure, solvent accessibility, torsion angle features in combination with an XGBoost model. However, these techniques, in particular those using MSA are extremely slow ($\approx$ 30 minutes for one protein sequence) and cannot be used for high-throughput screening of proteins.

Since, the goal of our work was to compare the crystallization propensity of a protein using just their AA sequence and the ability of the model to perform high-throughput screening, hence we focus on methods such as DeepCrystal, ATTCrys and CLPred during our experimental comparisons.

In recent years, application of natural language processing (NLP) methods to protein sequences has led to remarkable breakthroughs for *sota* protein structure and property prediction. The driving force for these breakthroughs is the transformer, a deep learning architecture[25], which uses the concept of self-attention to efficiently capture long-range dependencies and intricate patterns in protein sequences that were previously difficult to discern using traditional deep learning methods[25].

Analogous to using words and sentences to train typical large language models (LLMs), transformer-based models such as ESM2 use individual AAs, peptides, and protein sequences[26] to learn the "language" of life. These protein language models (PLMs) follow a self-supervised learning framework, where the model attempts to predict the identity of randomly masked AAs (usually 15% of the AAs per protein sequence) using the unmasked potions of the protein sequence. For example, ESM2 was pre-trained on the masked language training task with $\approx$65 million unique protein sequences from UniRef[26]. After this extensive training, scientists are able to use these pre-trained models to extract high-dimensional representations for their proteins of interest. These vectors can be used for downstream tasks such as protein property prediction, protein clustering, and functional comparisons[24,27–32].

In the present work, we perform efficacy assessments of several open source PLMs for the task of predicting protein crystallization using the TRILL platform[33]. TRILL is a comprehensive resource designed to democratize access to *sota* open PLMs, eliminating the requirement for advanced computational skills. Using robust deep learning frameworks such as Pytorch Lightning[34] and HuggingFace Accelerate[35], TRILL provides access to several PLMs such as ESM2[26], Ankh[36] and ProstT5[37], specifically for tasks such as protein design and property analysis. Moreover, TRILL facilitates the usage of these PLMs with different model configurations and parameter space. These PLMs in TRILL are complemented by a suite of utilities that enhance user experience and functionality.

For protein sequence classification, the platform provides functionalities to embed protein sequences into vector representations, visualize the embedded protein sequence representation, train custom classifiers, and predict class labels for unseen protein sequences. These diverse tools and functionalities are encapsulated within a command-line interface, organized through ten commands as detailed in the original TRILL paper[33]. In the present work, we utilize the TRILL platform to determine the vector representation of proteins for each PLM using just the AA sequence as input. These vector representations are then passed as training data to classifiers which are optimized through hyper-parameter tuning. This results in optimal crystallization propensity predictor for individual PLM. We then perform a comprehensive comparison of these PLM-based predictors on several independent test sets. Finally, we generate 3000 proteins through a fine-tuned ProtGPT2 model (on the crystallizable class) and through a series of computational filtration steps identify a reduced set of 5 novel proteins as potentially crystallizable.

The key contributions of the manuscript are:

- Benchmarking different ESM2 models for the task of protein crystallization prediction using raw protein sequences on external balanced, SwissProt and TrEMBL test sets;
- Benchmark PLMs such as Ankh, Ankh-Large, ProstT5 and ProtT5-XL for the task of protein crystallization prediction on external balanced, SwissProt and TrEMBL test sets;
- Comprehensive comparison of open-source PLMs to predict diffraction-quality crystals with superior performance on aforementioned test sets;
- Provide all the code used for benchmarking open-PLMs for crytallization prediction task via github (`https://github.com/raghvendra5688/crystallization_benchmark`) for reproducibility and enabling community to utilize TRILL for their protein property prediction task.
- Fine-tune a protein generator namely ProtGPT2[38] to generate *de novo* protein sequences from the crystallizable class;
- Evaluate, screen and validate the generated proteins to identify a unique set of stable and well-folded proteins.

Figure 1 provides a flow diagram of the proposed framework for predicting protein crystallization propensity.

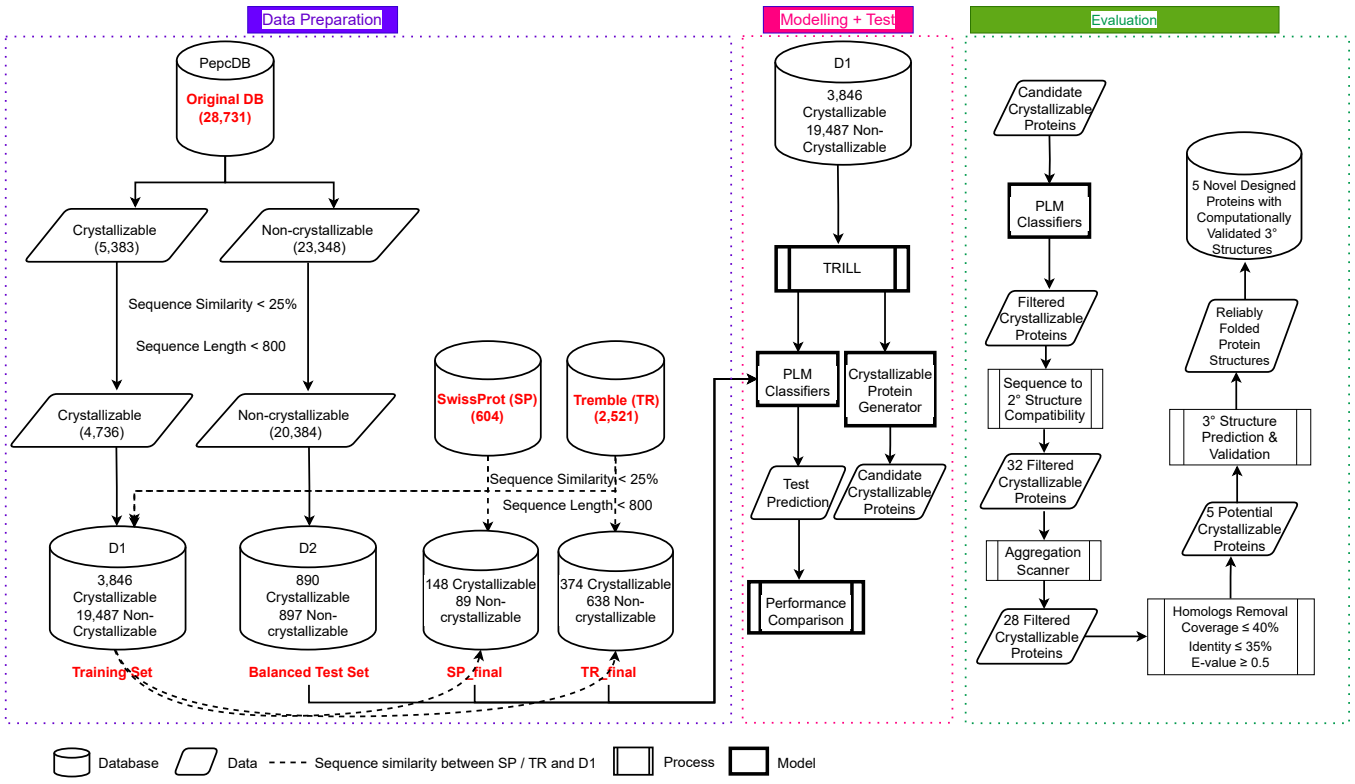


Figure 1. Flowchart of the proposed PLM benchmarking framework for protein crystallization propensity prediction.

Materials and Methods

Overview

The problem of predicting the crystallization propensity of a protein is a binary classification task. A protein sequence, is given by a sequence of AAs $x = (x_1, x_2, \ldots, x_L)$, where x_i, is the i^{th} amino acid in the sequence and is part of a vocabulary comprising 20 amino acids, while L is the length of the protein sequence. A given PLM uses its encoder referred as "tokenizer" ($t(\cdot)$) that encodes the AA sequence x to an encoded representation ($t(x) \in \mathbb{R}^L$) that is then ingestible for deep learning technique. This is a widely used encoding scheme in natural language processing (NLP) to have a vector representation for words in a sentence[39,40].

The encoded representation $t(x)$ is then given as input to the PLM and the final transformer layer of the PLM generates an embedding representation of the protein, preserving meaningful inter-residue relationships and contextual information within the original protein sequence. In mathematical terms $e(t(x))$ is the embedding of the protein x, with $e : \mathbb{R}^L \rightarrow \mathbb{R}^d$, where d represents the embedding dimension of the transformer layer of the PLM (note: for comparison reasons, we use different PLMs, thus d changes). Our aim is to learn a function $c(\cdot)$ that takes as input the embedded protein sequence $e(t(x))$ and outputs a probability, i.e., $c : \mathbb{R}^L \rightarrow [0, 1]$, where $c(\cdot)$ is the function computed by the nonlinear classifier. In this work, $c(\cdot)$ is an XGBoost[41] or a LightGBM[42] classifier.

While fine-tuning individual PLM (either all layers or few layers) with a classification head is an option, some of the PLMs tested in this work are extremely large i.e. ESM2 with 36 transformer layers and $\approx$ 3 billion parameters. Thus, it is impossible to fine-tune such PLM even with a batch size of 2, given the configuration of the available GPU - NVIDIA RTX A6000 with 48 Gb RAM. Hence, to have a fair evaluation given our GPU capacity, and to understand the learning representation capacity of these PLMs, we considered all these PLMs in a zero-shot learning framework to generate the embedded vector representations for proteins using their AA sequence.

Data Partitioning

We perform our experiment on the processed PepcDB dataset (`http://pepcdb.rcsb.org`) following the protocols set by Wang et al.[11]. The data set comprises proteins which have been classified into five groups, namely i) diffraction-quality crystals, ii) protein cloning failure, iii) protein material production failure, iv) purification failure, and v) crystallization failure. We

consider the proteins labeled as diffraction-quality crystals to be *the* crystallizable class, while other proteins are assigned to the non-crystallizable class. The final dataset comprises 28,731 sequences of which 5,383 proteins belong to the crystallizable class, and the remaining 23,348 are non-crystallizable. As in[11,17], all sequences in each class are passed through a filter of sequence identity $> 25\%$ with other proteins in that class to remove redundant and similar protein sequences within each class.

To divide our dataset into training and test sets, we follow a simple protocol. The maximum length of a protein sequence considered for our model is $L_{\max} = 800$. This is done to be compliant with methods like DeepCrystal[17] and CLPred[20], which use the same L as the maximum length of the protein sequence. Proteins with $L < L_{\max}$ are padded with the symbolic representation of gaps. By performing this protein filtering step, the total number of proteins in the dataset is reduced to 25,120.

We follow the procedure used in DeepCrystal[17], ATTCrys[19] and CLPred[20] to divide this dataset into two parts: $\mathbb{D}_1$ and $\mathbb{D}_2$ such that $\mathbb{D}_2$ consists of $\mathbb{D}_2^1 = 891$ crystallizable and $\mathbb{D}_2^0 = 896$ non-crystallizable proteins. Here 1 corresponds to crystallizable and 0 corresponds to non-crystallizable class. Thus, $\mathbb{D}_2$ represents the fairly balanced test set for performance evaluation as used in DeepCrystal, ATTCrys and CLPred methods. $\mathbb{D}_1$ has a total of 23,333 protein sequences, where $\mathbb{D}_1^1 = 3,846$ proteins belong to crystallizable class while remaining $\mathbb{D}_1^0 = 19,487$ proteins fall are non-crystallizable.

We also use two independent test sets generated in[1] as external validation sets. The two external datasets, referred as SP_final and TR_final were obtained from SwissProt and TrEMBL databases respectively, following the protocol detailed in Elbasir et al.[17]. In the SP_final dataset, we have 148 proteins belonging to the positive class while remaining 89 sequences are non-crystallizable, whereas in the TR_final dataset there are 374 crystallizable proteins and 638 proteins belonging to the negative class. We compare our methods with sota web-servers such as fDETECT[8], DeepCrystal[17], ATTCrys[19] and CLPred[20] on these datasets. For all performance comparison, we provide our test protein sequences to these web-servers to obtain corresponding prediction scores.

Benchmarking Models

The TRILL platform[33] provides access to several PLMs, such as ESM2[26], Ankh[36], ProstT5[37] and ProtT5-XL[43], which can generate protein embedding representations via a zero-shot learning framework. Moreover, there are several pretrained PLMs, such as ESM2[26], ProtGPT2[38] and ZymCTRL[44], which can either directly generate proteins in a zero-shot fashion or first by fine-tuning these models and then proceed with protein generation. Here we provide a summary of several PLMs used in the present work. For further details of these PLMs, reader's indulgence is sought.

Evolutionary Scale Modeling (ESM2)

ESM2 is a sota transformer-based protein language model trained on ≈65 million unique protein sequences[26]. ESM2 has been shown to outperform all tested single-sequence PLMs across a range of structure prediction tasks, enabling atomic resolution structure prediction. While the ESM2 model has been benchmarked for structure prediction, it has not been gauged for protein property prediction and has been shown to not scale for protein function prediction[45]. Moreover, the ESM2 models are available with different architectural configurations, that is, with an increase in number of transformer layers leading to an increase in number of model parameters. The ESM2 models are available with 6, 12, 30, 33 and 36 transformer layers having ≈ 8, 12, 150, 650 and 3,000 million parameters respectively.

Ankh

The Ankh is an optimized general-purpose PLM, as a first version for future specialized high-impact protein modeling tasks. Ankh is pre-trained on the UniRef50 dataset[46], that provides more variability and representation compared to UniRef100[46] and BFD[47]. The model is tested on a comprehensive set of downstream tasks spanning protein function prediction, structure prediction, and localization prediction. Ankh demonstrated superior performances on the tasks such as fluorescence prediction, solubility prediction, contact prediction, fold prediction, and secondary structure prediction. Additionally, Ankh used Google's latest TPU v4 hardware and JAX/Flax software for efficient training. Thus, Ankh is presented as a powerful general-purpose PLM that can serve as a foundation for specialized protein modeling tasks, with outstanding performances demonstrated on a wide range of benchmarks. Ankh-Large has ≈2 billion parameters and is trained using the encoder-decoder architecture, while Ankh base has $< 10\%$ parameters when compared to the sota models.

ProstT5

ProstT5 is a bilingual language model for protein sequences and structures that leverages the AlphaFold Protein Structure Database (AFDB)[48]. ProstT5 was pre-trained using 34.6 million proteins. It can translate between 1-D amino acid sequences and 1-D structure sequences (3Di tokens). ProstT5 demonstrated the improved performance in various protein function prediction tasks compared to sota sequence-based models such as ProtT5, ESM2 and Ankh. It can perform inverse folding, generate novel AA sequences that adopt a desired structural template, and assess the quality of its own predictions. ProstT5 exemplifies how language modeling techniques and transformers can be used to leverage the wealth of information from protein structure databases such as AFDB. Finally, ProstT5 is a proof-of-concept bilingual PLM that showcases the potential of integrating sequence and structure information for various protein modeling tasks.

ProtT5-XL

ProtT5-XL uses an encoder-decoder framework for training[25]. ProtT5-XL has 3 billion parameters and is trained using an 8-way model parallelism. ProtT5-XL is trained on BFD for 1.2 million steps, followed by a fine-tuning on UniRef50 for 991k steps. Contrary to the original T5 model[48] that masks spans of multiple tokens, ProtT5-XL adopts BERT's denoising objective to corrupt and reconstruct single tokens using a masking probability of 15%. ProtT5-XL uses the AdaFactor optimizer with inverse square root learning rate schedule for pretraining. Using embeddings from ProtT5-XL as the input to supervised models to predict secondary structure and subcellular localization, it outperformed previous methods on these tasks.

ProtGPT2

ProtGPT2 is a PLM that can generate novel protein sequences which are structurally and functionally similar to natural proteins[38]. ProtGPT2 effectively generates sequences that are distantly related to natural ones but are not a consequence of memorization and repetition. Majority of ProtGPT2 sequences (93%) have significant sequence similarity to natural proteins[38]. AlphaFold predictions show 37% of ProtGPT2 sequences have high confidence (pLDDT > 70) for being ordered structures, comparable to 66% for natural sequences. Molecular dynamics simulations indicate ProtGPT2 sequences have similar dynamic properties as natural proteins[38] .

Integrating ProtGPT2 sequences into a structural network representation of the protein universe reveals they bridge separate "islands" of known protein structures. ProtGPT2 generates sequences across different structural classes like all-α, all-β, α/β, etc. The model can be conditioned to design proteins for specific families, functions or structural classes. Thus, the unsupervised ProtGPT2 model effectively learns the "protein language" and generates novel sequences that populate unexplored regions of protein structure space while maintaining key structural and functional properties. This highlights the potential of PLMs for *de novo* protein design.

Model Building & Test

We follow a simple protocol to use the TRILL platform for our task of benchmarking PLMs for protein crystallization propensity prediction. Starting with the training sequences $x \in \mathbb{D}_1$, we obtain embedding representations $e(t(x))$ for each of the following 9 protein language models: ESM2 T6-8M, ESM2 T12-35M, ESM2 T30-150M, ESM2 T33-650M, ESM2 T36-3B, Ankh, Ankh Large, ProstT5, ProtT5-XL PLMs using the `embed` function.

The embedding representations $e_k(t(x)), k = 1 \ldots 9$ are generated in a zero-shot learning setting. These embedding representations of the training set $\mathbb{D}_1$ are then passed to the XGBoost classifier using the `classify` utility, where a 10-fold cross-validation technique is used for hyper-parameter optimization. The details of the hyperparameters are available via xgboost classifier script.

The XGBoost classifiers optimizes a weighted average F1-metric during the classification step to address the problem of class-imbalance. We also pass the embedding representations $e_k(t(x))$ from each PLM to custom LightGBM models[42] in 10-fold cross-validation setting to generate LightGBM classifiers. We performed a randomized search over a grid of parameters including number of estimators, maximum depth of a tree, number of leaves, minimum child samples, learning rate, subsampling rate, L1 and L2 regularizers during hyper-parameter optimization. The details of the parameter space for LightGBM classifiers are available at hyperparameter tuning script.

Thus, in total we have 9 XGBoost classifiers and 9 LightGBM classifiers, where each classifier is built on top of embedding representations $(e_k(t(x)))$ obtained from a PLM. After obtaining the XGBoost / LightGBM classifier for each of the 9 PLMs, we pass the test sets to each PLM to obtain embedding representations for the respective set of proteins. Finally, the class label and probability $c\left(e_k(t(x))\right)$ for each protein sequence x in a given test set and the k^{th} PLM is obtained by passing its embedding representation $e_k(x)$ to the classifier $c(\cdot)$. We utilize the `classify` function with '–preComputed_Embs' and '–preTrained' utilties in TRILL to obtain the class probability as shown in Figure 2. A consensus of the predictions from the 18 classifiers is obtained by taking average of the probabilities estimated by these classifiers.

A detailed workflow of building the classifiers and obtaining predictions on test sets is highlighted in Figure 2.

Protein Generation

We fine-tune the ProtGPT2 PLM on the crystallizable class ($\mathbb{D}_1^1$) using the `fine-tune` function available in TRILL for 10 epochs. In[33] it was shown that 10 epochs are sufficient to generate synthetic cell penetrating peptides and anti-crispr proteins using ProtGPT2. Thus, the fine-tuned ProtGPT2 model learns the underlying distribution of crystallizable proteins. We then generate a total of 3,000 proteins using the fine-tuned ProtGPT2 model via the `lang_gen` utility. Once we have generated the synthetic proteins, we obtain the embedding representation for the same using the PLMs and visualize these embeddings in a low-dimensional space (2 dimensions) using the `visualize` function. This function utilizes the Unified Manifold and Approximation (UMAP) algorithm[49] to project the embeddings into a two-dimensional space. Then, the embedding representation for a generated protein is obtained and classified by each of the 18 classifiers. This protein generation and classification process is illustrated in Figure 3.

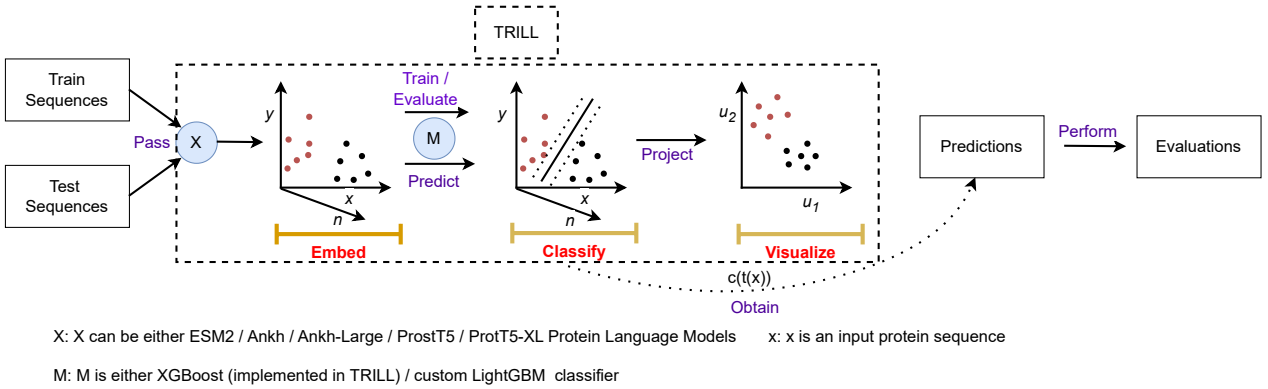


Figure 2. Workflow of building the crystallization propensity prediction classifiers for each PLM and obtaining test set predictions using the TRILL platform. Here the 'red' colored dots represent crystallizable proteins and 'black' colored dots correspond to non- crystallizable proteins.

We then follow a series of filtration steps to determine the most promising candidates:

Step 1: A consensus of all PLM-based classifiers consistently identified 706 out of the 3,000 generated proteins as crystallizable proteins.

Step 2: To remove generated sequences with high sequence identity with training set, we perform CD-HIT-2D[50] with a identity cut-off of $\leq 40\%$, resulting in 700 protein sequences.

Step 3: CD-HIT is then performed to cluster proteins with $> 25\%$ sequence identity into groups, leading to a total of 347 proteins with low sequence identity within the group and with the training set.

Step 4: Filtered protein sequences are screened by sequence to secondary structure compatibility scores[51,52]. The secondary structural characterization of the designed protein sequences is performed by utilizing PSIPRED (standalone ver. 4.02)[53]. This reduces the generated protein set from 347 to 32 candidate sequences.

Step 5: The screened proteins are further evaluated on the basis of presence of aggregation prone regions[54] and 4 sequences are filtered out.

Step 6: The screened proteins are subjected for the availability of any homolog(s) in known protein sequence database, UniRef100[46], resulting in a reduced set of 5 proteins.

Step 7: The 5 filtered proteins are modeled using a consensus approach by implementing RoseTTAFold2[55], and AlphaFold2[56], resulting in 6 model structures (5 from AlphaFold2 and 1 from RoseTTAFold end-2-end prediction) for each protein.

Step 8: Each model structure is refined by implementing GalaxyRefine[57] to generate 5 refined model structures, resulting in 30 candidate model structure for each protein.

Step 9: The modeled structure for each protein are thoroughly analyzed to identify the best model structure (1 out of 30) among the candidate structures using ModFold (ver. 9.0)[23] and ProFitFun[51,52].

Step 10: Finally, the stereo-chemical quality (all atoms contact and geometry) of the best model structure for each protein is assessed by passing it through ProCheck[58], Errat[59], and MolProbity[60].

By following the aforementioned steps, we filter an initial set of 3,000 proteins generated from crystallizable class to the set of 5 most likely and high confidence crystallizable proteins.

Evaluation Metrics

The performance of benchmark classifiers is compared with various other sota techniques using quality metrics such as accuracy, Matthew's correlation coefficient (MCC) as in[17,31]. We assessed other evaluation metrics, based on TP, TN, false positives (FP) and false negative (FN). We highlight that TP represents the set of proteins which are crystallizable (the true label is 1) and are correctly identified by a given method as crystallizable, i.e., $c\left(e(t(x))\right) \geq 0.5$. Similarly, TN represents the set of proteins which are non-crystallizable (true label is 0) and are correctly identified by a given method as non-crystallizable $c\left(e(t(x))\right)) < 0.5$.

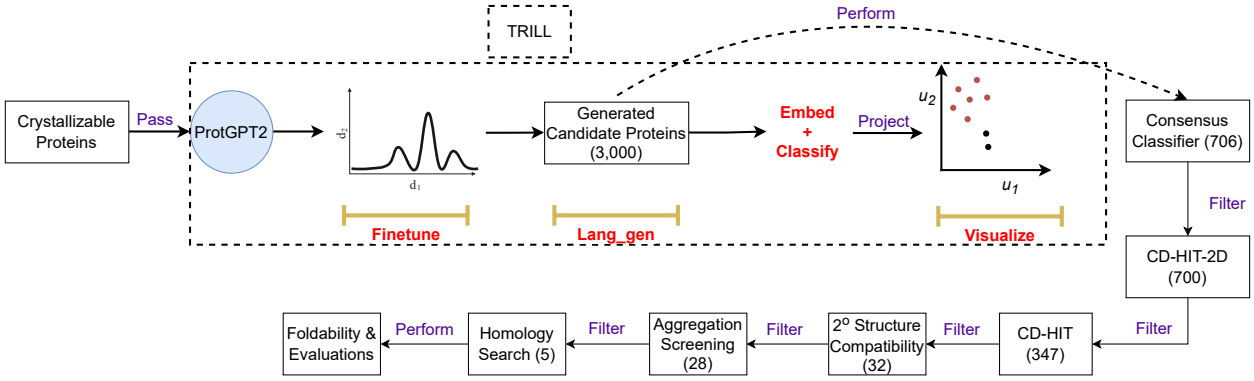


Figure 3. The protocol followed to generate crystallizable proteins using fine-tuned ProtGPT2 PLM and further downstream filtering and evaluation.

The metrics for evaluation include:

$$\begin{aligned}
\text{Accuracy (ACC)} &= \frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{TN}+\text{FN}} \\
\text{MCC} &= \frac{\text{TP}\cdot\text{TN}-\text{FP}\cdot\text{FN}}{\sqrt{(\text{TP}+\text{FP})\cdot(\text{TP}+\text{FN})\cdot(\text{TN}+\text{FP})\cdot(\text{TN}+\text{FN})}} \\
\text{Recall (Rec)} &= \frac{\text{TP}}{\text{TP}+\text{FN}} \\
\text{Precision (Prec)} &= \frac{\text{TP}}{\text{TP}+\text{FP}} \\
\text{F1-score (F1)} &= \frac{2\cdot\text{Prec}\cdot\text{Rec}}{\text{Prec}+\text{Rec}}
\end{aligned} \tag{1}$$

Experimental Results

We benchmark the predictive performance of the PLMs on the $\mathbb{D}_2$ test set extracted from the publicly available dataset[11] as described earlier (see Section II *B*). Moreover, we evaluate the quality of predictions from these models on two independent datasets obtained from SwissProt and TrEMBL, the SP_final and TR_final datasets, respectively. A comprehensive comparison of the PLMs of varying size and configurations including ESM2 T6-8M, ESM2 T12-35M, ESM2 T30-150M, ESM2 T33-650M, ESM2 T36-3B, Ankh, Ankh Large, ProstT5, ProtT5-XL was done against methods like fDETECT, DeepCrystal, ATTCrys and CLPred across these test sets. The evaluation metric values for fDETECT and CLPred were obtained from[17] and[20] respectively. Finally, the cross-validation performance of the XGBoost and LightGBM classifiers built on embedding representations learnt via each PLM on various evaluation metrics is highlighted in Supp. Figures 1 and 2. From Supp. Figures 1 and 2 and Tables 1, 2 and 3, we observe that the XGBoost models are over-fitting on the training set and have poor generalization performance. On the other hand, the LightGBM classifiers have better generalization performance as their cross-validation performance aligns with the performance attained on multiple independent test sets (see Supp. Figure 2 and Tables 1, 2 and 3).

Balanced Test Set Results

On the balanced test set consisting of 1787 proteins (891 crystallizable and 896 non-crystallizable), the ESM2 T30-150M PLM (with LightGBM classifier) achieves a prediction accuracy of 85.7%. This is better than the current sota method, CLPred (85.1%). The ESM2 T30-150M (LightGBM) also reaches the best performance of 0.854 and 0.715 for quality metrics such as F1 score and MCC, respectively, as observed from Table 1. These quality metrics take into account the class imbalance in the data set. The performance of ESM2 T30-150M (LightGBM) is 0.4% and 1.5% better in absolute terms than the current sota sequence-based crystallization predictor i.e., CLPred. Moreover, ESM2 T30-150M is 3.2%, 2.9%, and 5.7% better than DeepCrystal for F1 score, accuracy, and MCC metrics, respectively.

However, with respect to quality metrics such as AUPR and AUC, the ESM2 T30-150M (with XGBoost classifier) model leads when compared to all other benchmark models as observed from Table 1 and Figures 4a, 4b, 5a, and 5b. The ESM2 T30-150M (XGBoost) model reaches AUPR = 0.929 and AUC = 0.936. This is 4.3% and 3.3% better than DeepCrystal for

AUPR and AUC metrics, respectively, as observed in Table 1. Furthermore, from Table 1, we observe that PLMs with XGBoost classifier available via TRILL tend to handle the class-imbalance worse than PLMs with custom LightGBM classifier. This is highlighted from the superior performance of PLMs with LightGBM classifier on F1-score and MCC metrics when compared to their equivalent XGBoost classifiers available via TRILL as depicted in Table 1. Overall, PLMs trained with either LightGBM classifier outperform CLPred, ATTCrys and DeepCrystal across all metrics on balanced test set.

Table 1. Benchmarking of PLMs in TRILL on the balanced test set against sota methods.

Model	Method	F1	ACC	MCC	Prec	Rec	AUPR	AUC
fDETECT	RF	0.504	0.646	0.355	0.840	0.360	0.777	0.778
DeepCrystal	CNN	0.822	0.828	0.658	0.851	0.795	0.886	0.903
ATTCrys	Multi-Stage CNN	0.811	0.810	0.621	0.805	0.817	0.850	0.876
CLPred	CNN + Bi-LSTM	0.850	0.851	0.700	0.849	0.852	0.900	0.928
ESM2 T6-8M	XGBoost	0.674	0.746	0.546	0.934	0.527	0.9	0.916
ESM2 T12-35M	XGBoost	0.643	0.726	0.51	0.921	0.494	0.905	0.916
ESM2 T30-150M	XGBoost	0.803	0.826	0.669	**0.92**	0.713	**0.929**	**0.936**
ESM2 T33-650M	XGBoost	0.754	0.794	0.618	0.928	0.635	0.91	0.928
ESM2 T36-3B	XGBoost	0.716	0.767	0.571	0.914	0.588	0.908	0.92
Ankh	XGBoost	0.764	0.792	0.602	0.883	0.672	0.893	0.913
Ankh Large	XGBoost	0.783	0.804	0.619	0.874	0.709	0.906	0.917
ProstT5	XGBoost	0.761	0.791	0.6	0.885	0.667	0.907	0.924
ProtT5-XL	XGBoost	0.757	0.791	0.606	0.903	0.651	0.913	0.924
ESM2 T6-8M	LightGBM	0.828	0.837	0.676	0.869	0.791	0.9	0.914
ESM2 T12-35M	LightGBM	0.803	0.821	0.652	0.891	0.731	0.916	0.92
ESM2 T30-150M	LightGBM	**0.854**	**0.857**	**0.715**	0.871	0.838	0.916	0.932
ESM2 T33-650M	LightGBM	0.845	0.845	0.69	0.843	0.846	0.9	0.917
ESM2 T36-3B	LightGBM	0.829	0.833	0.666	0.843	0.816	0.904	0.916
Ankh	LightGBM	0.848	0.843	0.687	0.82	**0.877**	0.896	0.91
Ankh Large	LightGBM	0.831	0.832	0.663	0.83	0.833	0.907	0.918
ProstT5	LightGBM	0.85	0.851	0.702	0.855	0.845	0.916	0.929
ProtT5-XL	LightGBM	0.838	0.842	0.685	0.86	0.817	0.919	0.928

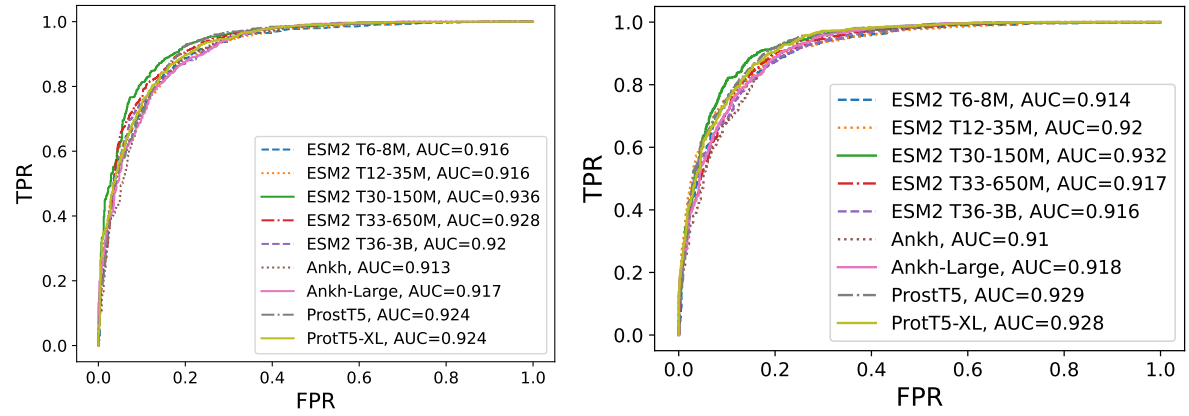
SP_final Test Set Results

A second experiment is performed on the reduced SP_final dataset obtained from SP_Pre dataset[1]. The ESM2 T36-3B model (with LightGBM classifier) outperforms sota sequence-based crystallization predictors like CLPred and DeepCrystal for the majority of the metrics, including F1, accuracy, MCC and precision as depicted in Table 2. The ESM2 T36-3B (LightGBM) model also outperforms other PLMs available via TRILL for these quality metrics as shown in Table 2. ESM2 T36-3B model (LightGBM) achieves a prediction accuracy of 89%, which is 9% and 14% better than CLPred and DeepCrystal respectively (see Table 2). From Table 1, we observe ESM2 T36-3B model (LightGBM) attains an MCC of 0.769 and F1-score of 0.911, whereas CLPred obtains an MCC of 0.599 and F1-score of 0.832 indicating 17% and 8% improvement in performance. The ProstT5 model (with LightGBM classifier) achieves the best AUC (0.940) and AUPR (0.964) compared to other PLM-based classifiers as depicted in Figures 4c, 4d, 5c and 5d.

We observe from Table 2 that small sized ESM2 models such as ESM2 T6-12M and ESM2 T12-35M cannot outperform CLPred for several quality metrics but bigger sized ESM2 models easily surpass sota models like fDETECT, DeepCrystal, ATTCrys and CLPred. Finally, the SP_final test set comprises 237 proteins with very little sequence similarity with training set and still ESM2 T36-3B classifiers (desgined with XGBoost / LightGBM) outperforms majority of sequence-based predictors on several evaluation metrics highlighting their effectiveness for crystallization propensity prediction.

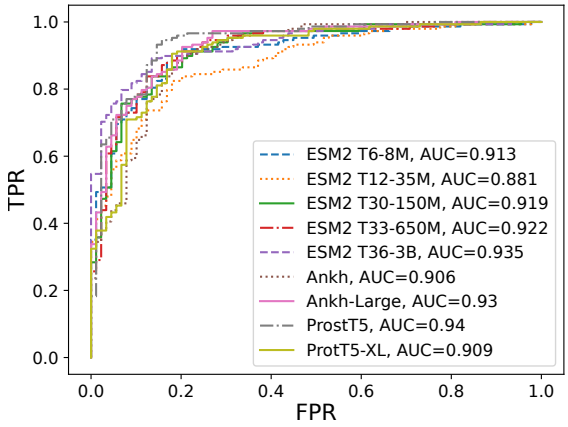
TR_final Test Set Results

We perform a final experiment to test for crystallization propensities of proteins using sota crystallization tools and benchmark PLM-based classifiers available via TRILL platform on the TR_final dataset[1]. ESM2 T30-150M model (with LightGBM classifier) achieves a prediction accuracy of 89.4%, which is 4% better than CLPred (85.4%), 5.3% better than DeepCrystal (84.1%) and fDETECT (84.1%). It is also 0.9% better than the next-best ESM2 T6-8M (LightGBM) model that attains an accuracy of 88.5% as depicted in Table 3. The ESM2 T30-150M model (LightGBM) achieves the best F1 (0.862) and MCC (0.778) as highlighted in Table 3 and second best performance for AUC (0.929) and AUPR (0.959) when compared to ESM2

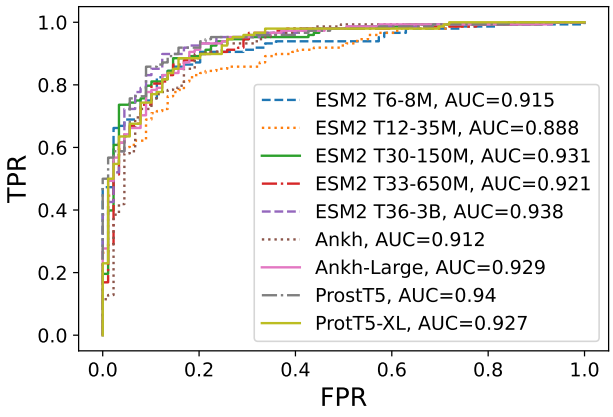


(a) Balanced Test Set using XGBoost

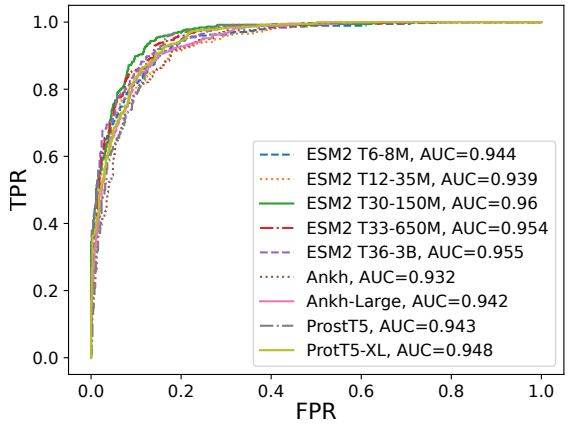
(b) Balanced Test Set using LightGBM



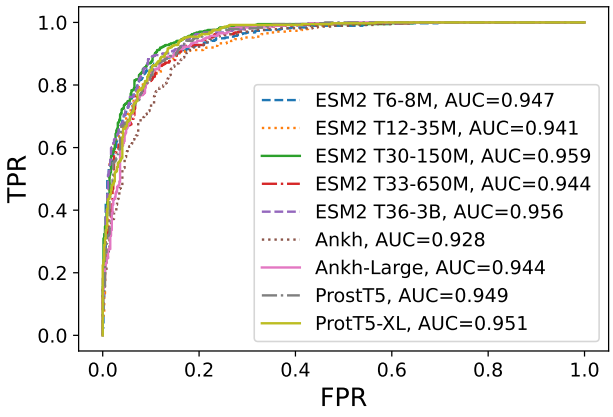
(c) SP_final Test Set using XGBoost



(d) SP_final Test Set using LightGBM



(e) TR_final Test Set using XGBoost

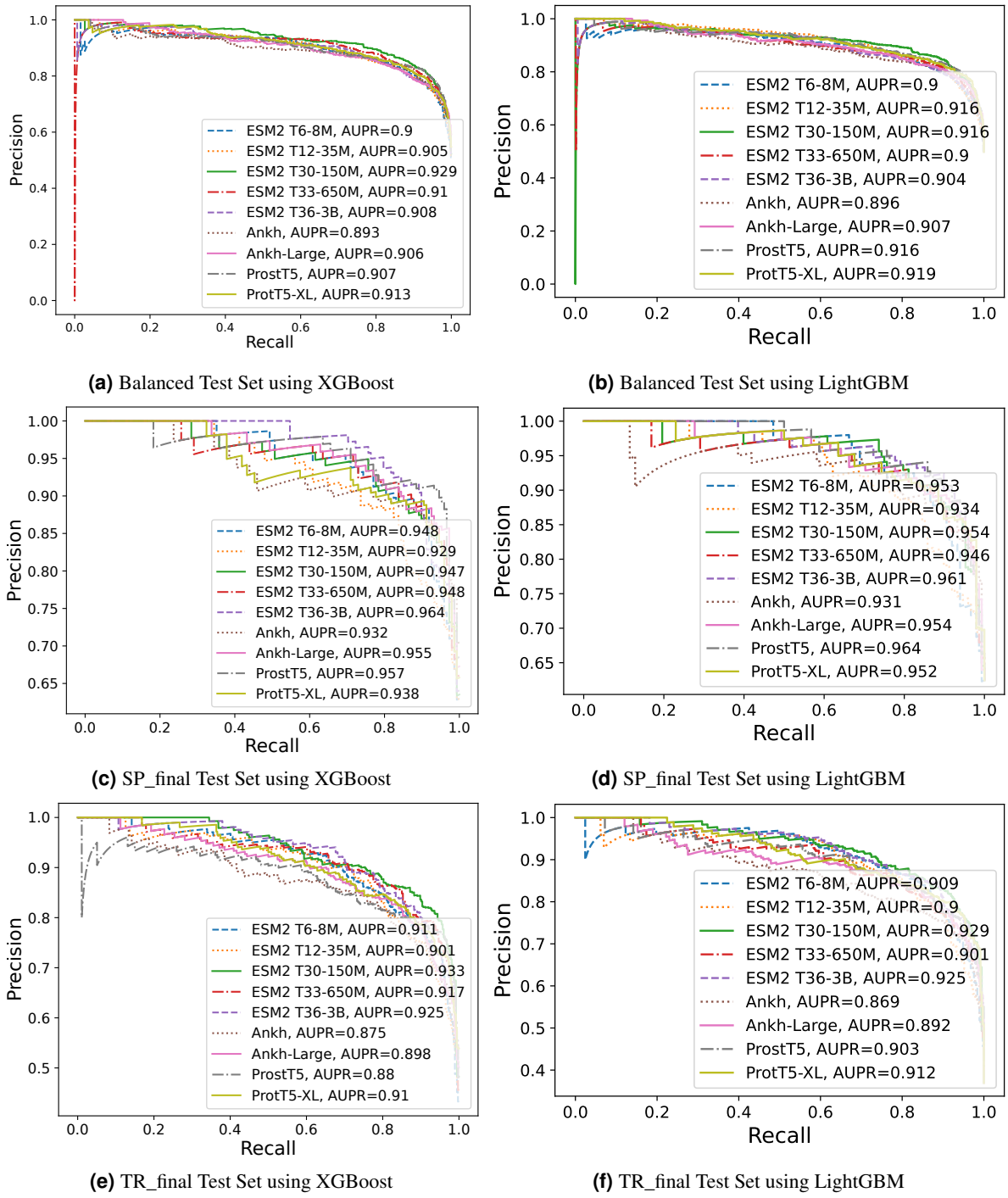


(f) TR_final Test Set using LightGBM

Figure 4. Comparison of area under receiver operating curve (AUC) of benchmark PLMs for the crystallization prediction task across the three different test sets. (a) AUC for fairly balanced test set using XGBoost, (b) AUC for SP_final dataset using XGBoost, (c) AUC for TR_final dataset using XGBoost, (d) AUC for fairly balanced test set using LightGBM, (e) AUC for SP_final dataset using LightGBM, and (f) AUC for TR_final dataset using LightGBM.

T30-150M (XGBoost), which achieves an AUC of 0.933 and AUPR of 0.960 as indicated in Table 3 and Figures 4e, 4f, 5e and 5f.

Interestingly, we observe from Table 3 that LightGBM classifiers are superior than their counterpart XGBoost classifiers for the same PLM models and configurations highlighting their generalization capability (see Supp. Figure 2). Finally, on



(a) Balanced Test Set using XGBoost

(b) Balanced Test Set using LightGBM

(c) SP_final Test Set using XGBoost

(d) SP_final Test Set using LightGBM

(e) TR_final Test Set using XGBoost

(f) TR_final Test Set using LightGBM

Figure 5. Comparison of area under precision-recall curve (AUPR) of benchmark PLMs for the crystallization prediction task across the three different test sets. (a) AUPR for fairly balanced test set using XGBoost, (b) AUPR for SP_final dataset using XGBoost, (c) AUPR for TR_final dataset using XGBoost, (d) AUPR for fairly balanced test set using LightGBM, (e) AUPR for SP_final dataset using LightGBM, and (f) AUPR for TR_final dataset using LightGBM.

the TR_final dataset comprising 1012 proteins (far more than SP_final test set), the PLM-based classifiers are superior than DeepCrystal, ATTCrys and CLPred w.r.t. several evaluation metrics.

Table 2. Benchmarking of PLMs in TRILL on the SP_final test set against sota methods.

Model	Method	F1	ACC	MCC	Prec	Rec	AUPR	AUC
fDETECT	RF	0.580	0.616	0.381	0.913	0.425	0.882	0.837
DeepCrystal	CNN	0.788	0.759	0.53	0.876	0.716	0.877	0.874
ATTCrys	Multi-Stage CNN	0.814	0.772	0.521	0.831	0.797	0.856	0.827
CLPred	CNN + Bi-LSTM	0.832	0.801	0.599	0.885	0.783	0.880	0.887
ESM2 T6-8M	XGBoost	0.712	0.713	0.524	0.955	0.568	0.948	0.913
ESM2 T12-35M	XGBoost	0.615	0.646	0.445	0.957	0.453	0.929	0.881
ESM2 T30-150M	XGBoost	0.836	0.814	0.646	0.933	0.757	0.947	0.919
ESM2 T33-650M	XGBoost	0.795	0.781	0.61	0.953	0.682	0.948	0.922
ESM2 T36-3B	XGBoost	0.814	0.802	0.657	**0.981**	0.696	**0.964**	0.935
Ankh	XGBoost	0.761	0.743	0.528	0.907	0.655	0.932	0.906
Ankh Large	XGBoost	0.84	0.819	0.653	0.934	0.764	0.955	0.93
ProstT5	XGBoost	0.829	0.81	0.648	0.948	0.736	0.957	**0.94**
ProtT5-XL	XGBoost	0.794	0.776	0.593	0.936	0.689	0.938	0.909
ESM2 T6-8M	LightGBM	0.871	0.848	0.694	0.924	0.824	0.953	0.915
ESM2 T12-35M	LightGBM	0.803	0.781	0.585	0.914	0.716	0.934	0.888
ESM2 T30-150M	LightGBM	0.873	0.848	0.688	0.912	0.838	0.954	0.931
ESM2 T33-650M	LightGBM	0.883	0.857	0.699	0.901	0.865	0.946	0.921
ESM2 T36-3B	LightGBM	**0.911**	**0.89**	**0.769**	0.924	**0.899**	0.961	0.938
Ankh	LightGBM	0.885	0.857	0.694	0.885	0.885	0.931	0.912
Ankh Large	LightGBM	0.876	0.848	0.681	0.894	0.858	0.954	0.929
ProstT5	LightGBM	0.898	0.878	0.751	0.941	0.858	**0.964**	**0.94**
ProtT5-XL	LightGBM	0.873	0.848	0.688	0.912	0.838	0.952	0.927

Table 3. Benchmarking of PLMs in TRILL on the TR_final test set against sota methods.

Model	Method	F1	ACC	MCC	Prec	Rec	AUPR	AUC
fDETECT	RF	0.747	0.841	0.663	0.918	0.631	0.768	0.887
DeepCrystal	CNN	0.781	0.841	0.657	0.800	0.762	0.815	0.910
ATTCrys	Multi-Stage CNN	0.758	0.810	0.605	0.718	0.802	0.793	0.880
CLPred	CNN + Bi-LSTM	0.807	0.854	0.690	0.787	0.829	0.865	0.930
ESM2 T6-8M	XGBoost	0.729	0.835	0.648	0.926	0.602	0.911	0.944
ESM2 T12-35M	XGBoost	0.692	0.819	0.616	0.932	0.551	0.901	0.939
ESM2 T30-150M	XGBoost	0.816	0.875	0.73	0.9	0.746	**0.933**	**0.96**
ESM2 T33-650M	XGBoost	0.772	0.854	0.685	0.912	0.668	0.917	0.954
ESM2 T36-3B	XGBoost	0.783	0.863	0.708	**0.94**	0.671	0.925	0.955
Ankh	XGBoost	0.756	0.839	0.649	0.858	0.676	0.875	0.932
Ankh Large	XGBoost	0.797	0.858	0.69	0.844	0.754	0.898	0.942
ProstT5	XGBoost	0.762	0.84	0.65	0.846	0.693	0.88	0.943
ProtT5-XL	XGBoost	0.776	0.852	0.678	0.878	0.695	0.91	0.948
ESM2 T6-8M	LightGBM	0.846	0.885	0.755	0.841	0.85	0.909	0.947
ESM2 T12-35M	LightGBM	0.807	0.868	0.712	0.873	0.751	0.9	0.941
ESM2 T30-150M	LightGBM	**0.862**	**0.894**	**0.778**	0.833	0.893	0.929	0.959
ESM2 T33-650M	LightGBM	0.829	0.867	0.723	0.787	0.877	0.901	0.944
ESM2 T36-3B	LightGBM	**0.862**	**0.894**	0.777	0.835	0.89	0.925	0.956
Ankh	LightGBM	0.82	0.853	0.706	0.748	**0.906**	0.869	0.928
Ankh-Large	LightGBM	0.835	0.87	0.732	0.785	0.89	0.892	0.944
ProstT5	LightGBM	0.839	0.875	0.739	0.799	0.882	0.903	0.949
ProtT5-XL	LightGBM	0.844	0.88	0.749	0.814	0.877	0.912	0.951

Protein Generation Results

The selected crystallizable candidates ($n = 347$) were trimmed on the basis of sequence to secondary structural compatibility (CS-Score ≥ 40 and CSS-Scores ≥ 20)[51,52], resulting in a dataset of 32 proteins. The cut-off values for CS- and CSS-Scores

were adopted from their benchmarking of successfully designed proteins[51]. These proteins were further tapered to 28 proteins, based on presence of aggregation protein region screening[61], and to 5 proteins based on screening against UniRef100[46].

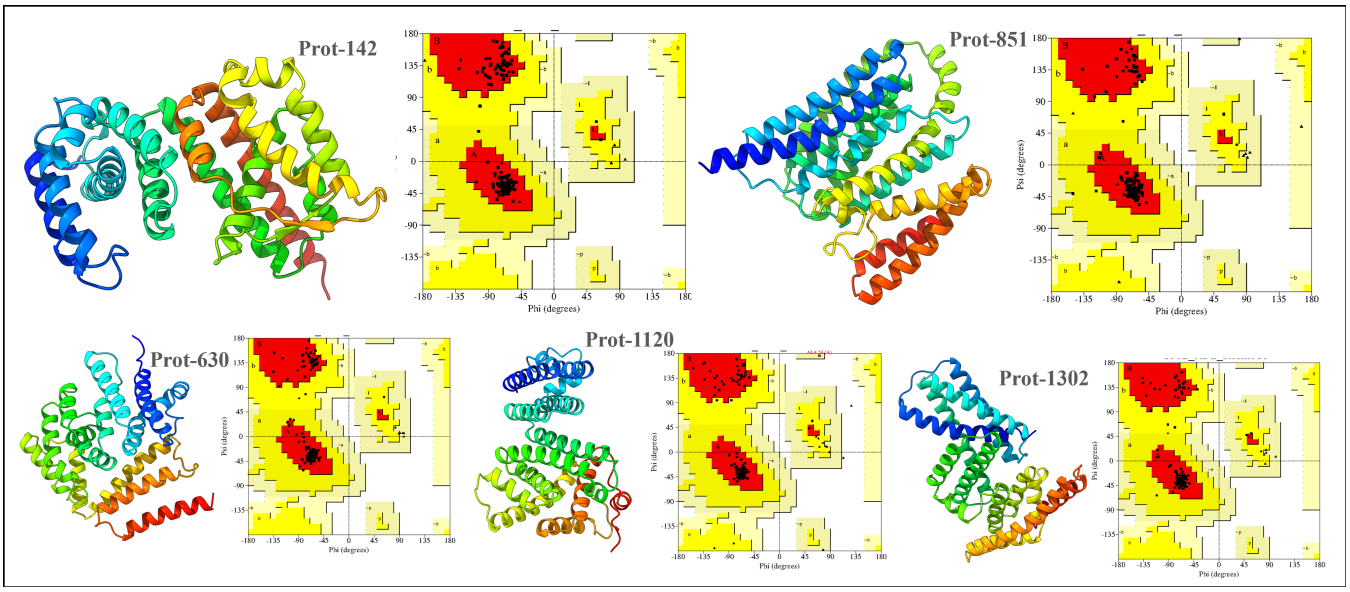


Figure 6. Best model structures for the 5 candidate proteins identified through our crystallizable protein generator workflow.

The proteins with pairwise sequence coverage $\geq 40\%$, sequence identity $\geq 35\%$ and e-value ≤ 0.5 were discarded while screening for available homolog(s) in known protein sequence database (UniRef100), resulting in the set of 5 proteins. These protein were modeled by implementing RoseTTAFold (end-2-end prediction; 1 candidate structure for each protein)[62] and AlphaFold2 ($n = 5$ candidate structures for each protein)[56], followed by structure refinement by using GalaxyRefine ($n = 30$; 5 refined candidate structures for each candidate structure)[63]. The best model structure for each protein, selected on the basis of consensus score from ModFold[23] and ProFitFun[52]. The best model structure for each protein along with the distribution of backbone di-hedrals (Ramachandran Map) are depicted in Figure 6. A summary of different quality assessment statistics of the best model structures is provided in Table 4. Additionally, the predicted Global Distance Test - Template Score (GDT-TS), Template Modeling Score (TMS), Global Quality Score (GQS), and Average Quality Score (OAQS) for all the candidate model structures are provided in Table 4.

The quality metrics for the best model structure of selected proteins (Prot-142, Prot-630, Prot-851, Prot-1120, and Prot-1302) ensured the accuracy of the tertiary structure prediction (Table 4). For all the model structures, the Ramachandran distribution of backbone di-hedral angles (ϕ and ψ) is found to be distributed in the allowed regions, mainly the core region (colored 'red'), as shown in Table 4 and Figure 6. The predicted model structure for Prot-630 and Prot-1302 had the highest quality score (=0.69), followed by Prot-142 (=0.67), Prot-1120 (=0.65), and Prot-851 (=0.58). Notably, the predicted GDT-TS (0.84 for Prot-630 and 0.88 for Prot-1302) and predicted TM Score (0.83 for Prot-630 and 0.82 for Prot-1302) for these protein structure fall in the highly reliable range for predicted model structure (0.8 - 1.0). The GDT-TS and TM Score varies from 0-1, where 1 shows the highest level of structural prediction. The relative predicted quality of the model structure for Prot-851 was observed to be lower as compared to the model structures of other proteins. The secondary and tertiary structures of the selected protein revealed them to be mainly α-proteins, except for Prot-142 which has fraction of residues (about 4%) part of β-strands.

The functional annotations including biological processes (BP), molecular functions (MF) and cellular components (CC) associated with the generated proteins are provided in Supp. Table 1. Additionally, the two proteins with the maximum functional annotations were Prot-1120 and Prot-1302. The functional annotations associated with these proteins is depicted in Figure 7. We observed that Prot-142 and Prot-630 are localized in cytoplasm, associated to different membranes such as cellular anatomical entity and mainly involved in different metabolic processes and bio-synthetic processes as depicted in Supp. Table 1. The designed protein, Prot-851, while being associated with plasma membrane and cell peripheries such as cellular anatomical entity, was predicted to perform diverse transporter activities by its involvement in different metabolic and transport processes. In contrast to the functional characterization of Prot-142, Prot-630, and Prot-851, the designed proteins Prot-1120 and Prot-1302 were predicted to be involved in the highly diverse set of molecular functions and biological processes as illustrated in Figure 7. For instance, Prot-1120, with the similar cellular localization of other designed proteins, was predicted to be involved in a wider range of metabolic processes, viz. phosphorous, phosphate-containing, and organo-nitrogen

Table 4. Summary of different quality evaluation parameters for the best model structure for each of the selected protein.

Quality Parameters	Prot-142	Prot-630	Prot-851	Prot-1120	Prot-1302
Ramachandran Distribution					
Core Region	98.9%	98.1%	98.8%	98.5%	97.4%
Allowed Region	1.1%	1.9%	1.2%	1.1%	2.6%
Generously Allowed Region	0.0%	0.0%	0.0%	0.4%	0.0%
Disallowed Region	0.0%	0.0%	0.0%	0.0%	0.0%
Bond Lengths within Limits	96.6%	97.3%	96.6%	97.9%	96.5%
Bond Angles within Limits	93.8%	94.3%	92.6%	94.0%	92.9%
Planner Groups within Limits	98.8%	100.0%	98.1%	100.0%	99.0%
Favored Rotamers	98.7%	98.6%	97.8%	97.1%	99.5%
Errat Score	97.3%	99.0%	98.1%	98.0%	96.4%
MolProbity Score	1.72	1.21	1.57	1.31	1.40
Predicted GDT-TS Score	0.75	0.84	0.63	0.72	0.88
Predicted TM-Score	0.83	0.83	0.65	0.80	0.82
Global Quality Score	0.44	0.41	0.45	0.44	0.35
Average Quality Score	0.67	0.69	0.58	0.65	0.69

compound metabolic processes, primary and cellular metabolic processes, and overall regulation of cellular processes. The Prot-1120 was predicted to be involved catalytic activity, calcium-dependent phospholipid binding, transferase activity, purine ribonucleoside triphosphate binding, small molecule binding, phosphoric ester hydrolase activity, ion binding, organic cyclic compound binding, carbohydrate derivative binding, and heterocyclic compound binding. Further, Prot-1302 is computationally characterized to perform metabolic and biosynthetic process along with trans-membrane transport of various compounds. With the involvement in a diverse set of biological processes, the Prot-1302 was predicted to perform ion channel activity, ATP binding, trans-membrane transporter activity, transferase activity, phosphotransferase activity, purine ribonucleoside triphosphate binding, small molecules and ions binding, organic cyclic compound binding, carbohydrate derivative binding, and heterocyclic compound binding.

With a comprehensive computational functional characterization, we believe that experimental validation of Prot-1120 and Prot-1302 can lead to the novel functional proteins that can be fine-tuned to have desired functions.

Discussion & Conclusion

One of the main challenges for protein structure determination is that only about 2-10% of pursued protein targets yield high-resolution protein structures[64]. Upon investigating these estimates in the TargetDB database[6], it was observed that among the 150,727 cloned targets that were deposited into TargetDB, only 37,398 (24.8%) were successfully purified, 12,923 (8.6%) further successfully crystallized, and 6,942 (4.6%) resulted in diffraction quality crystals[65]. Additionally, majority of the cost of structure determination is consumed by the failed attempts[7] as crystallization is a process that is characterized by a significant rate of attrition. The reasons for this attrition include the need for the crystals to be sufficiently large ($>$ 50 micrometers), pure in composition, regular in structure, and without significant internal imperfections. Furthermore, to produce diffraction-quality crystals, an empirical or trial-and-error approach is commonly used, in which a large number of experiments are brute-forced to find a suitable setup[66], often resulting in failure. Thus, the above provides strong motivation to develop accurate and efficient *in silico* sequence-based protein crystallization predictors that allow high-throughput screening of candidate protein sequences for favorable crystallization propensity.

In this paper, we benchmark open-PLMs accessed via the TRILL platform, a framework enabling democratization of protein language models, for sequence-based protein crystallization propensity prediction. The main objective is to determine whether PLMs trained on hundreds of millions of protein sequences can discriminate crystallizable proteins from non-crystallizable ones without fine-tuing using just the raw protein sequences as input. These PLMs encode the raw protein sequences and generate embedding (vector) representations. We then built optimized tree-based classifiers (XGBoost / LightGBM) on top of these embedding representations to estimate their discriminative capacity without the need to manually engineered biological and physiochemical features. By implementing a thorough benchmarking on a set of independent test sets, we observe that these open-PLM based classifiers consistently outperform state-of-the-art deep learning techniques, such as DeepCrystal, ATTCrys and CLPred, on several evaluation metrics.

DeepCrystal[17] captures frequent amino acid k-mers in the input sequence using a set of parallel convolution filters of varying sizes with the CNN design providing the freedom of calculating local dependencies with different filter sizes. Conversely,

(a) BP associated with Prot-1120 **(b)** CC associated with Prot-1120

(c) MF associated with Prot-1120 **(d)** BP associated with Prot-1302

(e) CC associated with Prot-1302 **(f)** MF associated with Prot-1302

Figure 7. Functional annotations associated with Prot-1120 and Prot-1302.

CLPred[20] uses a BiLSTM deep learning architecture to capture high-order, long-range interaction patterns between *k*-mers making it better than the CNN-based DeepCrystal as indicated in Tables 1, 2 and 3. However, open source protein language models trained on several million protein sequences are much better than smaller and crystallization specific deep learning models like DeepCrystal, ATTCrys and CLPred (see Tables 1, 2 and 3), even with no additional fine-tuning and a simple linear probing approach i.e. building classifiers on top of embedding representations. In particular, the **ESM2 T30-150M** and **ESM2 T36-3B** based models (with LightGBM classifier) outperform every other benchmark model on the three independent test sets

for quality metrics such as *F1-score*, *accuracy*, *MCC*, and *precision*.

This success can be attributed to the huge amount of data on which these PLMs are trained, the underlying transformer architecture which can capture local and long-range contextual dependencies in protein sequences through attention mechanism[25] and generate meaningful and discriminative embedding representations for the downstream crystallization task.

The proposed methodology illustrates its ability to generate and filter unique crystallizable proteins as well as engineer proteins to achieve desired properties and functions. These proteins may aid in the better understanding of biological processes, as well as the rapid development of new medicines and materials. For example, a designed protein with certain mutations could aid in understanding the roles of specific amino acid residue(s) in the natural protein. Similarly, protein-based therapeutic regimes that involve improvements in the efficacy, stability, solubility, or specificity of certain enzymes, antibodies, and hormones may be accelerated with computational engineering with the help of proposed workflow. Furthermore, computational design may help in the development of more efficient, stable, and selective enzymes that can considerably boost industrial output in the fields of bio-catalysis, food industry, and bio-fuels.

Data availability

All the code used for the analysis in this study is available at https://github.com/raghvendra5688/crystallization_benchmark/

Conflicts of Interest

None Declared

Acknowledgements

The authors would like to acknowledge Dr. Thomas Launey for his valuable feedback which helped to better position the paper.

Author Contributions

R.M., M.T. and F.C. conceived the study. R.M. and R.K. performed the data curation. R.M., Z.M. and R.K. designed the methodology. R.M. and R.K. performed the experiments and visualizations. All authors contributed in writing, reviewing and editing the manuscript.

References

1. Wang, H. & Wang, J. How cryo-electron microscopy and x-ray crystallography complement each other. *Protein Sci.* **26** (2017).
2. Wüthrich, K. Protein structure determination in solution by nmr spectroscopy. *J. Biol. Chem.* **265**, 22059–22062 (1990).
3. Service, R. Structural genomics, round 2. *Science* **307**, 1554–1558, DOI: 10.1126/science.307.5715.1554 (2005). https://www.science.org/doi/pdf/10.1126/science.307.5715.1554.
4. Terwilliger, T. C., Stuart, D. I. & Yokoyama, S. Lessons from structural genomics. *Annu. review biophysics* **38**, 371–83 (2009).
5. Gao, J. *et al.* Survey of predictors of propensity for protein production and crystallization with application to predict resolution of crystal structures. *Curr. protein & peptide science* **19 2**, 200–210 (2017).
6. Hu, J. *et al.* Targetcrys: protein crystallization prediction by fusing multi-view features with two-layered svm. *Amino Acids* **48**, 2533–2547 (2016).
7. Kurgan, L. *et al.* Crystalp2: sequence-based protein crystallization propensity prediction. *BMC Struct. Biol.* **9**, 50 – 50 (2009).
8. Meng, F., Wang, C. & Kurgan, L. fdetect webserver: fast predictor of propensity for protein production, purification, and crystallization. *BMC Bioinforma.* **18** (2017).
9. Mizianty, M. J. & Kurgan, L. Cryspred: accurate sequence-based protein crystallization propensity prediction using sequence-derived structural characteristics. *Protein peptide letters* **19 1**, 40–9 (2012).
10. Jahandideh, S. & Mahdavi, A. Rfcrys: sequence-based protein crystallization propensity prediction by means of random forest. *J. theoretical biology* **306**, 115–9 (2012).
11. Wang, H. *et al.* Predppcrys: Accurate prediction of sequence cloning, protein production, purification and crystallization propensity from protein sequences using multi-step heterogeneous feature fusion and selection. *PLoS ONE* **9** (2014).
12. Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A. & Vapnik, V. N. Support vector regression machines. In *Neural Information Processing Systems* (1996).
13. Suykens, J. A. K., Gestel, T. V., Brabanter, J. D., Moor, B. D. & Vandewalle, J. Least squares support vector machines (2002).

14. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).

15. Friedman, J. H. Greedy function approximation: A gradient boosting machine. *Annals Stat.* **29**, 1189–1232 (2001).

16. Kouranov, A. *et al.* The rcsb pdb information portal for structural genomics. *Nucleic acids research* **34**, D302–D305 (2006).

17. Elbasir, A. *et al.* Deepcrystal: a deep learning framework for sequence-based protein crystallization prediction. *Bioinformatics* **35**, 2216–2225 (2019).

18. LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **86**, 2278–2324 (1998).

19. Jin, C., Gao, J., Shi, Z. & Zhang, H. Attcry: Attention-based neural network model for protein crystallization prediction. *Neurocomputing* **463**, 265–274 (2021).

20. Xuan, W., Liu, N., Huang, N., Li, Y. & Wang, J. Clpred: a sequence-based protein crystallization predictor using blstm neural network. *Bioinformatics* **36**, i709–i717 (2020).

21. Wang, P.-H., Zhu, Y.-H., Yang, X. & Yu, D.-J. Gcmapcrys: integrating graph attention network with predicted contact map for multi-stage protein crystallization propensity prediction. *Anal. Biochem.* **663**, 115020 (2023).

22. Altschul, S. F. *et al.* Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research* **25**, 3389–3402 (1997).

23. McGuffin, L. J. & Alharbi, S. M. A. Modfold9: a web server for independent estimates of 3d protein model quality. *J. Mol. Biol.* (2024).

24. Elbasir, A. *et al.* Bcrystal: an interpretable sequence-based protein crystallization predictor. *Bioinformatics* **36**, 1429–1438 (2020).

25. Vaswani, A. *et al.* Attention is all you need. *Adv. neural information processing systems* **30** (2017).

26. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).

27. Bernhofer, M. & Rost, B. Tmbed: transmembrane proteins predicted through language model embeddings. *BMC bioinformatics* **23**, 326 (2022).

28. Goffinet, E., Mall, R., Singh, A., Kaushik, R. & Castiglione, F. Mate-pred: Multimodal attention-based tcr-epitope interaction predictor. *bioRxiv* 2024–01 (2024).

29. Mall, R. *et al.* A modeling framework for embedding-based predictions for compound–viral protein activity. *Bioinformatics* **37**, 2544–2555 (2021).

30. Mall, R. Solxplain: An explainable sequence-based protein solubility predictor. *BioRxiv* 651067 (2019).

31. Rawi, R. *et al.* Parsnip: sequence-based protein solubility prediction using gradient boosting machine. *Bioinformatics* **34**, 1092–1098 (2018).

32. Khurana, S. *et al.* Deepsol: a deep learning framework for sequence-based protein solubility prediction. *Bioinformatics* **34**, 2605–2613 (2018).

33. Martinez, Z. A., Murray, R. M. & Thomson, M. W. Trill: Orchestrating modular deep-learning workflows for democratized, scalable protein analysis and engineering. *bioRxiv* (2023).

34. Falcon, W. *et al.* Pytorchlightning/pytorch-lightning: 0.7. 6 release. *Zenodo* (2020).

35. Gugger, S. *et al.* Accelerate: Training and inference at scale made simple, efficient and adaptable. https://github.com/huggingface/accelerate (2022).

36. Elnaggar, A. *et al.* Ankh: Optimized protein language model unlocks general-purpose modelling. *arXiv preprint arXiv:2301.06568* (2023).

37. Heinzinger, M. *et al.* Prostt5: Bilingual language model for protein sequence and structure. *bioRxiv* 2023–07 (2023).

38. Ferruz, N., Schmidt, S. & Höcker, B. Protgpt2 is a deep unsupervised language model for protein design. *Nat. communications* **13**, 4348 (2022).

39. Zhang, Q.-s. & Zhu, S.-C. Visual interpretability for deep learning: a survey. *Front. Inf. Technol. & Electron. Eng.* **19**, 27–39 (2018).

40. Zhang, X., Zhao, J. & LeCun, Y. Character-level convolutional networks for text classification. *Adv. neural information processing systems* **28** (2015).

41. Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794 (2016).

42. Ke, G. *et al.* Lightgbm: A highly efficient gradient boosting decision tree. *Adv. neural information processing systems* **30** (2017).

43. Elnaggar, A. *et al.* Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE transactions on pattern analysis machine intelligence* **44**, 7112–7127 (2021).

44. Munsamy, G., Lindner, S., Lorenz, P. & Ferruz, N. Zymctrl: a conditional language model for the controllable generation of artificial enzymes. In *NeurIPS Machine Learning in Structural Biology Workshop* (2022).

45. Li, F.-Z., Amini, A. P., Yue, Y., Yang, K. K. & Lu, A. X. Feature reuse and scaling: Understanding transfer learning with protein language models. *bioRxiv* 2024–02 (2024).

46. Suzek, B. E. *et al.* Uniref clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**, 926–932 (2015).

47. Elnaggar, A. *et al.* Prottrans: Towards cracking the language of life's code through self-supervised deep learning and high performance computing. arxiv 2020. *arXiv preprint arXiv:2007.06225* (2007).

48. Tunyasuvunakool, K. *et al.* Highly accurate protein structure prediction for the human proteome. *Nature* **596**, 590–596 (2021).

49. McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).

50. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).

51. Kaushik, R. & Zhang, K. Y. J. A protein sequence fitness function for identifying natural and nonnatural proteins. *Proteins: Struct.* **88**, 1271 – 1284 (2020).

52. Kaushik, R. & Zhang, K. Y. J. Profitfun: a protein tertiary structure fitness function for quantifying the accuracies of model structures. *Bioinformatics* (2021).

53. Buchan, D. W. A. & Jones, D. T. The psipred protein analysis workbench: 20 years on. *Nucleic Acids Res.* **47**, W402 – W407 (2019).

54. Kaushik, R. & Launey, T. Decoding protein aggregation through computational approach: Identification and scoring of aggregation-prone regions in protein sequences. *bioRxiv* DOI: 10.1101/2024.06.11.598423 (2024). https://www.biorxiv.org/content/early/2024/06/12/2024.06.11.598423.full.pdf.

55. M, B. *et al.* Accurate prediction of protein structures and interactions using a 3-track neural network. *Sci. (New York, N.Y.)* **373**, 871 – 876 (2021).

56. Jumper, J. M. *et al.* Highly accurate protein structure prediction with alphafold. *Nature* **596**, 583 – 589 (2021).

57. Lee, G. R., Won, J., Heo, L. & Seok, C. Galaxyrefine2: simultaneous refinement of inaccurate local regions and overall protein structure. *Nucleic acids research* **47**, W451–W455 (2019).

58. Laskowski, R., MacArthur, M. & Thornton, J. Procheck: validation of protein-structure coordinates. international tables for crystallography. *Vol F Chapter* **21**, 684–687 (2012).

59. Colovos, C. & Yeates, T. O. Verification of protein structures: patterns of nonbonded atomic interactions. *Protein science* **2**, 1511–1519 (1993).

60. Kaushik, R. & Zhang, K. Y. An integrated protein structure fitness scoring approach for identifying native-like model structures. *Comput. Struct. Biotechnol. J.* **20**, 6467–6472 (2022).

61. Cima, V. *et al.* Prediction of aggregation prone regions in proteins using deep neural networks and their suppression by computational design. *bioRxiv* DOI: 10.1101/2024.03.06.583680 (2024). https://www.biorxiv.org/content/early/2024/03/11/2024.03.06.583680.full.pdf.

62. Krishna, R. *et al.* Generalized biomolecular modeling and design with rosettafold all-atom. *bioRxiv* (2023).

63. Heo, L., Park, H. & Seok, C. Galaxyrefine: protein structure refinement driven by side-chain repacking. *Nucleic Acids Res.* **41**, W384 – W388 (2013).

64. Service, R. F. Structural genomics, round 2. *Science* **307**, 1554 – 1558 (2005).

65. Kurgan, L. & Mizianty, M. J. Sequence-based protein crystallization propensity prediction for structural genomics: Review and comparative analysis. *Nat. Sci.* **1**, 93–106 (2009).

66. Chayen, N. E. Turning protein crystallisation from an art into a science. *Curr. opinion structural biology* **14 5**, 577–83 (2004).