EasyChair Preprint

№ 14542

ESMCrystal : Enhancing Protein Crystallization Prediction Through Protein Embeddings

Jayanth Kumar, Kavya Jayakumar and Jayaprakash Sundararaj

August 26, 2024

ESMCrystal : Enhancing Protein Crystallization Prediction through Protein Embeddings

Jayanth Kumar*,[1,2], Kavya Jayakumar[3], Jayaprakash Sundararaj[4]

[1] Computer Science, University of California, Los Angeles, CA, USA. jayanthj@ucla.edu, 0000-0002-4342-9646
[2] Computer Science, Indian Institute of Technology Bombay, Mumbai, India. jayanth8506@iitbombay.org
[3] Tata Institute of Fundamental Research, Hyderabad, India. jkavya@tifrh.res.in, 0000-0002-7849-6259
[4] Computer Science, Indian Institute of Technology Bombay, Mumbai, India. jayaprakash12@cse.iitb.ac.in

*corresponding author

Keywords: Protein crystallization, Structural Biology, Protein Embeddings, ESMCrystal models, ESMFold2, Machine learning, Bioinformatics, Protein structure analysis.

Abstract. Protein crystallization is a critical yet challenging step in determining protein structures, crucial for advancing our understanding of biological mechanisms. This study introduces ESMCrystal, a novel approach leveraging protein embeddings derived from the advanced Meta ESMFold2 architecture to predict protein crystallization. By integrating transfer learning techniques, ESMCrystal models demonstrate enhanced predictive performance across various datasets, highlighting the potential of deep learning in structural biology. This research not only improves the predictability of protein crystallization but also sets the stage for broader applications of machine learning in understanding complex biological systems. The standalone source code and models, along with the inference server are available at
https://huggingface.co/jaykmr/ESMCrystal_t6_8M_v1 and
https://huggingface.co/jaykmr/ESMCrystal_t12_35M_v2.

1 Introduction

The quest to determine protein structures has long been pivotal in the field of biochemistry and molecular biology, fundamentally inspiring advancements in medicine, genetics, and various biotechnologies. Traditionally, the determination of these structures is primarily facilitated through X-ray crystallography. However, this method poses significant challenges, notably its high failure rate and the substantial costs associated with producing diffraction-quality crystals. While the success rates of obtaining such crystals range only between 2 - 10% [1], the costs attributed to unsuccessful attempts account for more than 70% [2] of total expenses. This inefficiency underscores an urgent need for innovative approaches to predict protein crystallization success more accurately and efficiently.

Historically, several machine learning and statistical methods have been developed to predict protein crystallization propensity from sequence data [3, 4]. These models including CrystalP2 [5], PPCpred [6], PredPPCrys [7], XtalPred-RF [8], TargetCrys [9], Crysalis [3], Crysf [10], fDETECT [11], DeepCrystal [12], and BCrystal [13], which span from early statistical analyses to more complex machine learning frameworks, have primarily hinged on feature extraction techniques to identify critical biological and physiochemical features from protein sequences. However, these methods often require intricate feature selection and substantial computational resources, which can be a bottleneck for scalability and practical application in both academic and industrial settings.

Our paper presents ESMCrystal, a model that utilizes deep learning and protein embeddings [14] to predict protein crystallization, aiming to significantly reduce the computational and

financial costs associated with traditional methods. Our study assesses these models across various standard datasets including DeepCrystal Test, Balanced Test, SP Test, and TR Test, employing comprehensive evaluation metrics such as confusion matrices, accuracy, precision, recall, F1-score, PR-AUC and ROC-AUC. The results obtained not only demonstrate high accuracy levels but also highlight the robustness of protein embeddings in enhancing the predictability of successful protein crystallization under diverse experimental conditions.

2 Data and Methods

2.1 *Datasets*

We perform our experiments on publicly available datasets, specifically from BCrystal [13] and DeepCrystal [12] dataset. Furthermore, the training dataset was completely based on the DeepCrystal dataset, resulting in 26,821 training samples, of which 4,420 are crystallizable and 22,401 are non-crystallizable. Out of the 26,821 samples, 5% (1342 samples) were used for the validation dataset, picked randomly before training. Additionally, there are 1,898 test samples, with 898 being crystallizable and 1,000 non-crystallizable.

We treat the crystallization prediction problem as a binary classification problem, distinguishing diffraction-quality crystals from the rest. The positive class or label 1 denotes crystallizable protein sequence, while the negative class or label 0 denotes non-crystallizable protein sequence.

Additionally, we use two independent test sets for further validation and comparison. These datasets, SP final (SwissProt) and TR final (Tremble) from [10], contain sequences with ¡25% sequence similarity with the training set. We also, use a fairly balanced test set from [12], consisting of 891 crystallizable and 896 non-crystallizable proteins for evaluation.

In the SP final dataset, there are 148 proteins belonging to the positive class, while the remaining 89 sequences are non-crystallizable. In the TR final dataset, there are 374 crystallizable proteins and 638 proteins belonging to the negative class.

Dataset	Total Sequences	Crystallizable	Non-crystallizable
DeepCrystal Train	26281	4420	22401
DeepCrystal Test	1898	898	1000
Balanced Test	1787	891	896
SP Test	237	148	89
TR Test	1012	374	638

Table 1: **Dataset Summary**. Details of Datasets Used in the Study.

These datasets contain a diverse range of protein sequences annotated based on their crystallization outcome. This diversity is crucial for testing the robustness and generalizability of our prediction models across different experimental scenarios and protein types.

2.2 *Training Procedure*

The models were trained using a transfer learning approach, starting with pre-trained embeddings from the ESMFold2 architecture [15], which were then adapted to the specific task of protein crystallization prediction. This method leverages the general protein structure understanding of ESMFold2, fine-tuning it further to predict crystallization outcomes from protein sequences.

2.3 *Models*

We trained two state-of-the-art machine learning models based on the Meta ESMFold architecture, fine-tuning it specifically for protein crystallization prediction:

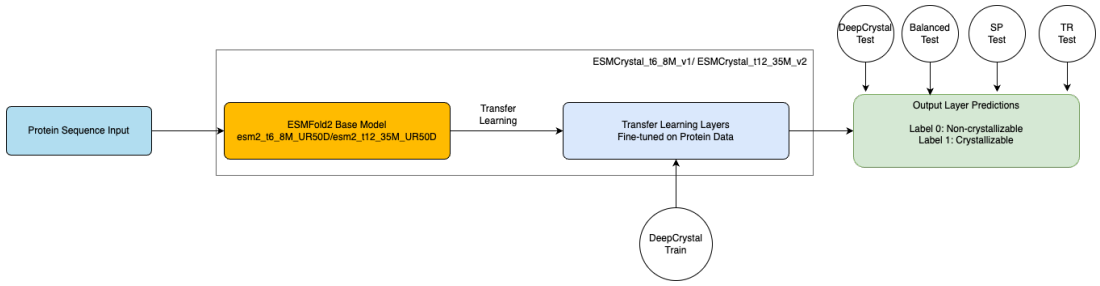

Figure 1: **Transfer Learning Architecture for Protein Crystallization Prediction**. This diagram illustrates the transfer learning process utilizing the ESMFold2 base model to predict protein crystallization. Starting with an input protein sequence, the architecture employs the foundational layers of Meta's ESMFold2, which are fine-tuned with specific protein crystallization datasets. The predictive model outputs two labels: 'Non-crystallizable' (Label 0) and 'Crystallizable' (Label 1), demonstrating an application of advanced machine learning techniques in structural biology.

1. ESMCrystal_t6_8M_v1: This smaller model is finetuned on the esm2_t6_8M_UR50D data, with 6 hidden layers and 8 million parameters. The model size is approximately 31.4MB. It is likely faster to train and requires less computational resources but may capture less complex patterns compared to the larger model.

2. ESMCrystal_t12_35M_v2: This larger and more complex model is fine-tuned on esm2_t12_35M_UR50D, featuring 12 hidden layers and 35 million parameters, with a total size of approximately 136MB. Since it has more hidden layers, it is more capable of capturing complex patterns in the protein sequence.

The dimensionality of the hidden states in smaller model is set to 320 while in the larger model is set to 480, which determines the size of the vector representations learned by the model. Within each transformer block, the dimensionality of the intermediate layer in the feedforward network is set to 1280 in the smaller model and 1920 in the larger model, which processes the output of the attention mechanism. These configurations are used as default from the ESMFold2 architecture.

Both models have 20 attention heads, which means they are equally capable of parallelizing the process of attending to different parts of the input sequence. This feature is particularly useful in tasks like protein sequence analysis where different segments of the sequence might have various functional implications.

We disabled dropout for attention probabilities and hidden layers in both the models as model robustness is not much of an issue due to less variation in protein sequence data. We use "rotary" position embeddings, ideal for maintaining relative positional information, crucial in protein sequences. We also, enable token dropout, which helps improve generalization by randomly dropping tokens during training.

2.4 *Evaluation Metrics*

Confusion matrices were generated for a thorough assessment of true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN) across all test datasets. To assess the performance of our models, several metrics were generated and compared based on confusion metrics, such as precision (PRE), recall (REC), F-score (F), accuracy (ACC), Matthews Correlation Coefficient (MCC), negative predictive value (NPV), receiver operating characteristic - area under curve (ROC), and Precision-Recall Area Under Curve (PR-AUC). A detailed definition of these sets and the importance of each of these evaluation metrics are provided in [16, 12, 17].

Based on the testing across the specified datasets, each model's performance was documented, focusing on their predictive accuracy.

ROC-AUC and PR-AUC curves were plotted for both models across all tests, providing visual insights into model performance and the trade-offs between sensitivity and specificity that each model exhibits.

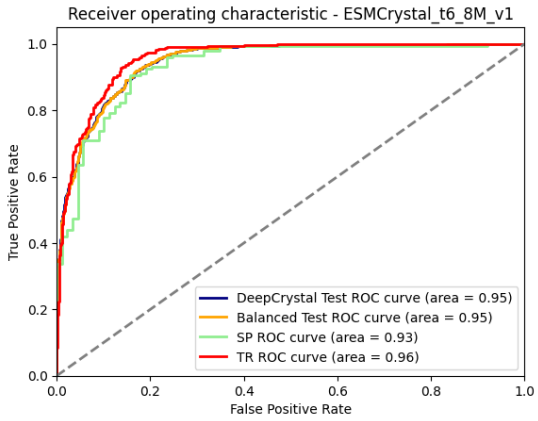

Figure 2: **ROC curve for ESMCrystal_t6_8M_v1**. This diagram plots the Receiver Operating Characteristic curve for ESMCrystal_t6_8M_v1 on the different test datasets.

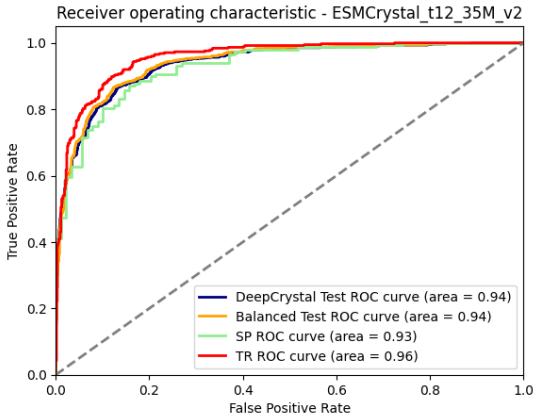

Figure 3: **ROC curve for ESMCrystal_t12_35M_v2**. This diagram plots the Receiver Operating Characteristic curve for ESMCrystal_t12_35M_v2 on the different test datasets.

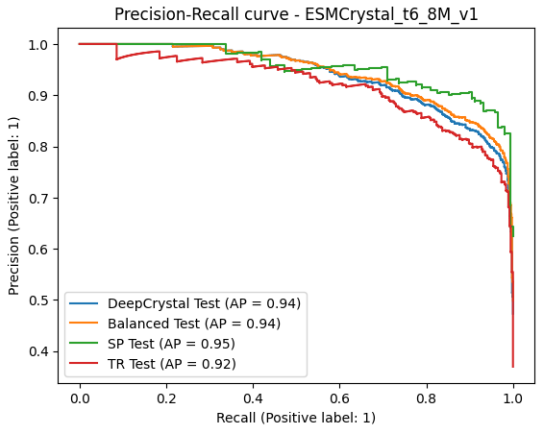

Figure 4: **PR curve for ESMCrystal_t6_8M_v1**. This diagram plots the Precision-Recall curve for ESMCrystal_t6_8M_v1 on the different test datasets.

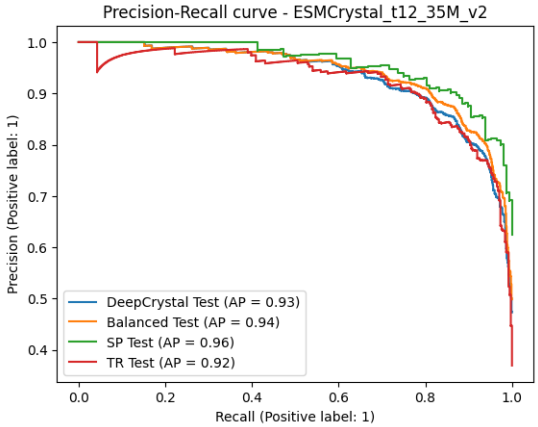

Figure 5: **PR curve for ESMCrystal_t12_35M_v2**. This diagram plots the Precision-Recall curve for ESMCrystal_t12_35M_v2 on the different test datasets.

3 Results

The experimental results obtained from employing the ESMCrystal_t6_8M_v1 and ESMCrystal_t12_35M_v2 models on various test datasets have provided significant insights into the efficacy of using advanced machine learning techniques for predicting protein crystallization. Below, we detail the performance outcomes and findings for each of these models across our testing environments.

Dataset	TP	FP	FN	TN	Precision	Recall	F1	Acc	ROC	MCC	NPV
DeepCrystal	532	362	34	966	0.5951	0.9399	0.7288	0.7909	0.9467	0.6119	0.9660
Balanced	531	360	31	865	0.5960	0.9448	0.7309	0.7812	0.9396	0.6045	0.9654
SP	80	68	4	85	0.5405	0.9524	0.6897	0.6962	0.9328	0.5017	0.9551
TR	207	167	16	622	0.5535	0.9283	0.6935	0.8192	0.9562	0.6153	0.9749

Table 2: **Performance of ESMCrystal_t6_8M_v1**. Performance of the smaller model on different test datasets

3.1 *ESMCrystal_t6_8M_v1 Results*

The model's performance underscored its potential to predict protein crystallization across different datasets effectively, particularly emphasized by its higher performance in more homogeneous test environments (TR Test).

3.2 *ESMCrystal_t12_35M_v2 Results*

Dataset	TP	FP	FN	TN	Precision	Recall	F1	Acc	ROC	MCC	NPV
DeepCrystal	579	319	30	970	0.6448	0.9507	0.7684	0.8161	0.9403	0.6575	0.9700
Balanced	573	318	30	866	0.6431	0.9502	0.7671	0.8053	0.9396	0.6446	0.9665
SP Test	97	51	5	84	0.6554	0.9510	0.7760	0.7637	0.9293	0.5861	0.9438
TR Test	225	149	14	624	0.6016	0.9414	0.7341	0.8389	0.9562	0.6588	0.9781

Table 3: **Performance of ESMCrystal_t12_35M_v2**. Performance of the larger model on different test datasets

3.3 *Observations*

The ESMCrystal_t12_35M_v2 model illustrated a noticeable improvement over the 6-layer model, particularly in handling the dataset complexities more effectively, which is attributed to its more hidden layers and more parameters, i.e. deeper learning structure.

The comparison between ESMCrystal_t6_8M_v1 and ESMCrystal_t12_35M_v2 models clearly shows the advantage of a deeper neural network architecture, as evidenced by the consistently higher performance metrics across all test datasets by the latter.

Both models have shown high predictive reliability, emphasized by high ROC-AUC scores across datasets. The ESMCrystal_t12_35M_v2 model generally reported better precision and recall rates, especially evident in the TR Test dataset.

Despite the varied nature of the test datasets, the robustness of ESMFold-based models under different experimental conditions was commendable. It suggests that transfer learning from a pre-trained model on a broad dataset allows effective generalization across divergent protein sequences.

3.4 *Comparative Analysis*

The performance of ESMCrystal models was compared against several state-of-the-art sequence-based protein crystallization predictors, including BCrystal, DeepCrystal, Crysf, Crysalis I and II, fDETECT, TargetCrys, XtalPred-RF, PPCPred and CrystalP2. The comparison with Crysf was conducted only on the SP and TR datasets as Crysf required Uniprot ids as input, which were available only for these two datasets.

These results underline the significant potential of using deep learning techniques, specifically those harnessing robust pre-trained models like Meta's ESMFold, in advancing protein crystallization predictions. Furthermore, the study advocates for additional explorations into

Models	Accuracy	MCC	AUC	F-score	Recall	Precision	NPV
PPCpred	0.672	0.359	0.754	0.616	0.528	0.740	0.635
fDETECT	0.646	0.355	0.778	0.504	0.360	0.840	0.593
Crysalis I	0.777	0.556	0.865	0.767	0.738	0.799	0.758
Crysalis II	0.804	0.610	0.888	0.796	0.767	0.828	0.784
XtalPred-RF	0.650	0.301	0.710	0.654	0.663	0.645	0.655
TargetCrys	0.627	0.255	0.637	0.637	0.656	0.619	0.593
CrystalP2	0.585	0.177	0.608	0.627	0.700	0.568	0.613
DeepCrystal	0.828	0.658	0.903	0.822	0.795	0.851	0.809
BCrystal	0.954	0.908	0.981	0.954	0.970	0.939	0.969
ESMCrystal_t6_8M_v1	0.7812	0.6045	0.9396	0.7309	0.9448	0.5960	0.9654
ESMCrystal_t12_35M_v2	0.8053	0.6446	0.9396	0.7671	0.9502	0.6431	0.9665

Table 4: **Comparision of ESMCrystal models on balanced dataset**. ESMCrystal performs comparably with other protein crystallization predictors on the balanced test data

Models	Accuracy	MCC	AUC	F-score	Recall	Precision	NPV
Crysf	0.700	0.426	0.811	0.727	0.641	0.840	0.572
PPCpred	0.666	0.403	0.784	0.675	0.554	0.863	0.535
fDETECT	0.616	0.381	0.837	0.580	0.425	0.913	0.494
Crysalis I	0.725	0.448	0.835	0.763	0.709	0.826	0.609
Crysalis II	0.751	0.505	0.851	0.783	0.722	0.856	0.633
XtalPred-RF	0.451	0.149	0.449	0.548	0.553	0.564	0.288
TargetCrys	0.611	0.223	0.641	0.659	0.601	0.729	0.486
CrystalP2	0.658	0.257	0.696	0.734	0.756	0.713	0.550
DeepCrystal	0.759	0.530	0.874	0.788	0.716	0.876	0.637
BCrystal	0.894	0.774	0.951	0.919	0.966	0.877	0.932
ESMCrystal_t6_8M_v1	0.6962	0.5017	0.9328	0.6897	0.9524	0.5405	0.9551
ESMCrystal_t12_35M_v2	0.7637	0.5861	0.9293	0.7760	0.9510	0.6554	0.9438

Table 5: **Comparision of ESMCrystal models on SP dataset**. ESMCrystal performs comparably with other protein crystallization predictors on the SP test data

refining these models, with particular attention on enhancing their ability to manage datasets marked by high variability and complexity.

4 Conclusion

This study highlights the potent capabilities of advanced machine learning, specifically utilizing Meta's ESMFold [18, 15] architecture, to predict protein crystallization from sequence data. The ESMCrystal models, especially the ESMCrystal_t12_35M_v2, have demonstrated exceptional accuracy in forecasting crystallization potential, showcasing the adaptability and robustness of deep learning in tackling complex biological problems. These results validate the effectiveness of these sophisticated models in structural biology and suggest broader applications for these techniques.

The superior performance of the ESMCrystal_t12_35M_v2 model, owing to its deeper architecture and extensive training, underscores the potential for further advancements in this technology. This research opens pathways for refining these models using even Alphafold [19], enhancing their accuracy and reliability. Moreover, the successful application of transfer learning methodologies within this study advocates for their expanded use across the life sciences, reduc-

Models	Accuracy	MCC	AUC	F-score	Recall	Precision	NPV
Crysf	0.841	0.663	0.887	0.747	0.631	0.918	0.817
PPCpred	0.748	0.448	0.819	0.640	0.606	0.677	0.782
fDETECT	0.750	0.447	0.847	0.548	0.411	0.823	0.733
Crysalis I	0.787	0.546	0.870	0.715	0.724	0.707	0.836
Crysalis II	0.816	0.603	0.892	0.748	0.740	0.756	0.849
XtalPred-RF	0.451	0.040	0.525	0.452	0.537	0.390	0.651
TargetCrys	0.634	0.325	0.693	0.614	0.788	0.503	0.733
CrystalP2	0.581	0.241	0.673	0.577	0.775	0.460	0.780
DeepCrystal	0.841	0.657	0.910	0.781	0.762	0.800	0.864
BCrystal	0.963	0.922	0.988	0.951	0.970	0.933	0.982
ESMCrystal_t6_8M_v1	0.8192	0.6153	0.9562	0.6935	0.9283	0.5535	0.9749
ESMCrystal_t12_35M_v2	0.8389	0.6588	0.9562	0.7341	0.9414	0.6016	0.9781

Table 6: **Comparision of ESMCrystal models on TR dataset**. ESMCrystal performs comparably with other protein crystallization predictors on the TR test data

ing the need for extensive manual intervention and facilitating more efficient, high-throughput predictions.

In conclusion, the findings from this study set a solid foundation for the use of deep learning in predicting protein crystallization, emphasizing the transformative impact of AI and machine learning in structural biology. As we continue to innovate and refine these computational strategies, integrating them with broader biological data, we pave the way for significant advancements in understanding protein structures and their complex functions. This ongoing evolution in computational biology promises to accelerate scientific discoveries, pushing the boundaries of our knowledge and capabilities.

Acknowledgments

The authors extend gratitude to Meta ESMFold2 for making their protein embeddings openly available, to Huggingface for hosting the code and providing inference capabilities, and to Paperspace Compute Cloud for supplying the GPU servers essential for computational tasks.

Availability of data and software code

Our software code is available at the following URL:

- ESMCrystal_t6_8M_v1 : https://huggingface.co/jaykmr/ESMCrystal_t6_8M_v1
- ESMCrystal_t12_35M_v2 : https://huggingface.co/jaykmr/ESMCrystal_t12_35M_v2

The dataset and the inference API is also, available for prediction on the above URLs.

References

[1] Thomas C Terwilliger, David Stuart, and Shigeyuki Yokoyama. Lessons from structural genomics. *Annual review of biophysics*, 38:371–383, 2009.

[2] Robert Service. Structural genomics, round 2, 2005.

[3] Huilin Wang, Liubin Feng, Ziding Zhang, Geoffrey I Webb, Donghai Lin, and Jiangning Song. Crysalis: an integrated server for computational analysis and design of protein crystallization. *Scientific reports*, 6(1):21383, 2016.

[4] Jianzhao Gao, Zhonghua Wu, Gang Hu, Kui Wang, Jiangning Song, Andrzej Joachimiak, and Lukasz Kurgan. Survey of predictors of propensity for protein production and crystallization with application to predict resolution of crystal structures. *Current Protein and Peptide Science*, 19(2):200–210, 2018.

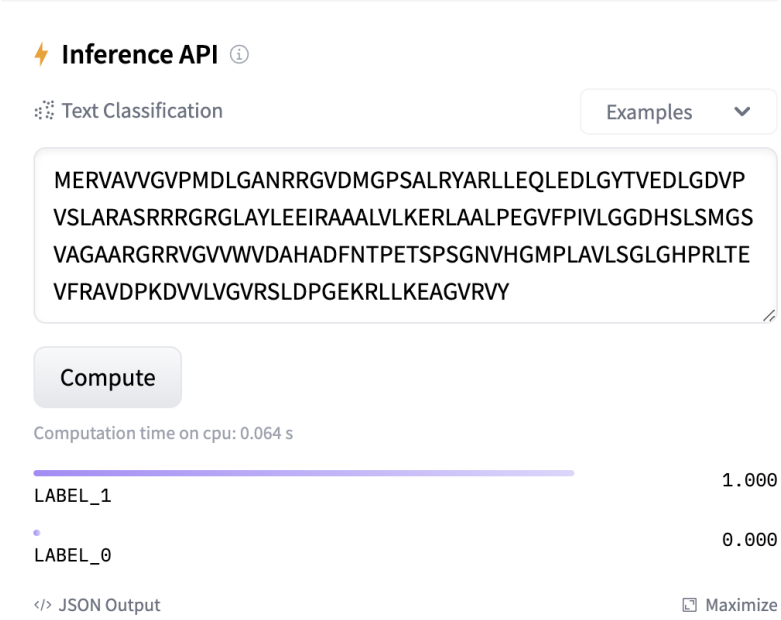

Figure 6: **Inference API** Input the protein sequence.

[5] Lukasz Kurgan, Marcin J Mizianty, et al. Sequence-based protein crystallization propensity prediction for structural genomics: review and comparative analysis. *Natural Science*, 1(02):93, 2009.

[6] Phasit Charoenkwan, Watshara Shoombuatong, Hua-Chin Lee, Jeerayut Chaijaruwanich, Hui-Ling Huang, and Shinn-Ying Ho. Scmcrys: predicting protein crystallization using an ensemble scoring card method with estimating propensity scores of p-collocated amino acid pairs. *PloS one*, 8(9):e72368, 2013.

[7] Huilin Wang, Mingjun Wang, Hao Tan, Yuan Li, Ziding Zhang, and Jiangning Song. 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(8):e105902, 2014.

[8] Samad Jahandideh, Lukasz Jaroszewski, and Adam Godzik. Improving the chances of successful protein structure determination with a random forest classifier. *Acta Crystallographica Section D: Biological Crystallography*, 70(3):627–635, 2014.

[9] Jun Hu, Ke Han, Yang Li, Jing-Yu Yang, Hong-Bin Shen, and Dong-Jun Yu. Targetcrys: protein crystallization prediction by fusing multi-view features with two-layered svm. *Amino acids*, 48:2533–2547, 2016.

[10] Huilin Wang, Liubin Feng, Geoffrey I Webb, Lukasz Kurgan, Jiangning Song, and Donghai Lin. Critical evaluation of bioinformatics tools for the prediction of protein crystallization propensity. *Briefings in bioinformatics*, 19(5):838–852, 2018.

[11] Fanchi Meng, Chen Wang, and Lukasz Kurgan. fdetect webserver: fast predictor of propensity for protein production, purification, and crystallization. *BMC bioinformatics*, 18:1–11, 2017.

[12] Abdurrahman Elbasir, Balasubramanian Moovarkumudalvan, Khalid Kunji, Prasanna R Kolatkar, Raghvendra Mall, and Halima Bensmail. Deepcrystal: a deep learning framework for sequence-based protein crystallization prediction. *Bioinformatics*, 35(13):2216–2225, 2019.

[13] Abdurrahman Elbasir, Raghvendra Mall, Khalid Kunji, Reda Rawi, Zeyaul Islam, Gwo-Yu Chuang, Prasanna R Kolatkar, and Halima Bensmail. Bcrystal: an interpretable sequence-based protein crystallization predictor. *Bioinformatics*, 36(5):1429–1438, 2020.

[14] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *PNAS*, 2019.

[15] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*, 2022.

[16] Sameer Khurana, Reda Rawi, Khalid Kunji, Gwo-Yu Chuang, Halima Bensmail, and Raghvendra Mall. Deepsol: a deep learning framework for sequence-based protein solubility prediction. *Bioinformatics*, 34(15):2605–2613, 2018.

[17] Reda Rawi, Raghvendra Mall, Khalid Kunji, Chen-Hsiang Shen, Peter D Kwong, and Gwo-Yu Chuang. Parsnip: sequence-based protein solubility prediction using gradient boosting machine. *Bioinformatics*, 34(7):1092–1098, 2018.

[18] Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alexander Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. *bioRxiv*, 2021.

[19] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.