Variant Effect Predictions Leveraging Structure Information

Yuxuan Liu, Atharva Agashe, Omnia Sarhan
Department of Electrical & Computer Engineering
Texas A&M University
188 Bizzell St, 77843 Tx, USA

May 6, 2024

Abstract

Proteins play vital roles in life, guided by their sequences and structures, and mutations offer insights into their functions. Deep Mutation Scanning (DMS) experiments have expanded our understanding of protein landscapes, yet data remains limited. Protein Language Models (PLMs) show promise in predicting variant effects, but integrating structural information poses challenges. Moreover, due to variations in baseline models and datasets used in previous studies, accurately comparing their efficacy is difficult, with each claiming superior performance over others. Our paper suggests the establishment of a benchmark model and dataset to facilitate comparative assessments of previous work in terms of performance. Then, we propose combining previous models into an ensemble model by stacking these individual predictors, aiming to achieve superior performance through supervised ensemble model training compared to individual zero-shot methods.

Code is available in GitHub, and dataset can be downloaded from zenodo

1 Introduction

Proteins, integral to life, perform a myriad of functions. These functions are encoded by protein sequences that fold into various structures. Mutations in protein sequences can provide insights into the relationship between sequence and function. However, clinical data is somewhat limited when compared to the possible types of variants. In recent years, the advent of Deep Mutation Scanning (DMS) experiments has expanded the dataset of sequence variants. Through measurements of protein function across a large number of variants, DMS can provide a comprehensive view of the protein landscape to illustrate the effects of various mutations. Nevertheless, DMS data remains limited due to the time-consuming and challenging nature of such experiments. Therefore, predicting the effects of variants based on current clinical and experimental data is an area of research that urgently needs attention.

Drawing upon methodologies from Natural Language Processing (NLP) (1), Protein Language Models (PLMs) have demonstrated significant advantages in the realm of protein representation. Through the application of self-supervised training on extensive datasets of masked protein sequences, transformer-based PLMs exhibit proficiency in discerning long-range residue correlations and evolutionary information. Notable PLMs such as ESM (2; 3; 4; 5) and ProtT5 (6) have exhibited exceptional performance across a broad spectrum of tasks related to protein structure and function.

Despite the progress in Protein Language Models (PLMs), the vast potential of incorporating protein structure for variant effect prediction remains largely unexplored. The repository of available protein structures has significantly expanded in recent years, largely due to the advent of deep learning-based structure prediction algorithms such as AlphaFold2(7) and ESMFold(4). Concurrently, various protein language models have been developed, facilitating the interpretation of relationships between residues. Capitalizing on these advancements, several models have been proposed to predict mutation effects by integrating structural information into protein language models (SI-PLMs). PST(8) employed a structure extractor trained using a Graph Neural Network (GNN) to adjust the self-attention matrices in the ESM-2 model. Conversely, ProSSN(9) utilized the ESM-2 embedding as node features of the protein graph, with a 6-layer Equivariant GNN updating the node features based on structural information. SI-PLMs(10) addressed the issue of previous models' tendencies to over-finetune to specific family sequences by modifying previous PLMs using structural decoders and adjusting their training losses through an auxiliary sequence-to-structure denoising task. Both SaProt(11) and ProstT5(12) employed Foldseek's(13) 3Di-alphabet to encode the input token sequences derived from the protein sequences. SaProt retrained the ESM-2 based PLM using the 3Di-alphabet in conjunction with the amino-acid token. ProstT5, on the other hand, fine-tuned an existing PLM (ProtT5) to integrate the sequence and structure during the modelling process.

While all these methodologies assert superior performance compared to others, it remains challenging to evaluate the efficiency with which structural information is incorporated into the language model. This is due to the fact that they are predicated on different baseline models and employ diverse evaluation methods. There is a pressing need for a benchmark that compares all these methods using the same baseline model and training dataset. In addition to differences in overall performance, as suggested in the SI-PLMs paper, it is necessary to evaluate performance across different protein families and functions. This would provide a more comprehensive understanding of the strengths and weaknesses of each method, thereby guiding future improvements in the field.

In this report, we first evaluate three SI-PLMs for the task of mutation effect prediction by utilizing the ProteinGym DMS dataset(14). These three models are based on the same PLM (ESM2-650M), which enables fair comparison between different structure presentation methods by zero-shot experiment. We propose a new method by feature-enamelling based on the insight in the benchmark experiment result. Additionally, we propose an ensemble model by stacking individual predictors, aiming to achieve superior performance through supervised ensemble model training compared to individual zero-shot methods.

2 Methods

This section discusses the different methodologies that have been conducted throughout the project. This includes exploring each of the chosen PLM models individually and their working mechanism. Additionally, there are details on the training of concept of these models using Zero-shot variant effect prediction method. Moreover, there is a demonstration of our proposed model for this project. Finally, utilized experimental dataset and techniques are scrutinized.

2.1 Individual Models

This subsection is about each of the chosen PLM models and their working mechanism. The models are Saprot, PST and ProteinSSN as well as the main baseline model used by these three PLMs, called the ESM2 model. The illustration of the architecture of each model is shown in Appendix A.

2.1.1 Baseline: ESM2

In the traditional approach to understanding proteins, multiple sequence alignment (MSA) is a critical step for deducing structure and function. However, this method is time-consuming due to its high computational complexity, requiring both the identification of related sequences and their subsequent alignment. The innovation presented in the ESM-2, shown in Figure 2, methodology sidesteps the need for this intermediate representation entirely. By training on a vast dataset of around 65 million unique protein sequences sourced from the UniRef50 database using Masked Language Modeling, ESM-2 directly learns the interdependencies of amino acids, potentially rendering the MSA process obsolete. This represents a significant advancement in bioinformatics, as it suggests that the conclusions traditionally drawn from the laborious MSA process can be reached more efficiently without the bottleneck of sequence alignment (4; 15).

2.1.2 SaProt

Structure-aware Protein language model (SaProt) is a Large-scale PLM that uses a novel concept of "structure-aware (SA) vocabulary," which integrates both residue and structural data of proteins. Vector quantization techniques is utilized to discretize protein structures into 3D tokens, following similar principles to residue tokens. Leveraging the Foldseek tool, we merge these 3D tokens with residue tokens, creating the SA alphabet (13). This facilitates the transformation of the original residue sequence into an SA-token sequence, which serves as input for existing residue-based PLMs that will be trained in an unsupervised manner to become a Structure-aware PLM. SaProt utilizes an identical network architecture and parameter size to the 650M version of ESM-2. However, the key difference lies in the enlarged embedding layer, which incorporates 441 SA tokens instead of the original 20 residue tokens. This model is trained using the BERT-style Masked Language Modeling (MLM) objective, akin to ESM-1b and ESM-2, enabling it to support both protein-level and residue-level tasks (11) This PLM is shown in Figure 4.

2.1.3 PST

The Protein Structure Transformer (PST) methodology, in Figure 3, augments the ESM-2 PLM with a structure-aware self-attention mechanism, utilizing advances in graph representation learning. Proteins are represented as graphs by treating each amino acid as a node, with edges drawn between nodes that are spatially close to each other, typically within a distance threshold of 8 angstroms. This graph is utilized by the structure extractor function $\phi\_\theta(X, G)$, which modifies the ESM-2 self-attention layers, infusing each with a structural context. To optimize computation, a Graph Isomorphism Network (GIN) is used at each self-attention layer in the ESM-2 model. The output of this function is a set of structural embeddings that, once linearly transformed, recalibrate the self-attention mechanism's query (Q), key (K), and value (V) matrices with spatial structural information. This structure extractor modifies each self-attention calculation within the ESM-2 model. It is fine-tuned on a protein structure database using the same masked language modelling objective as ESM-2. This model is pre-trained on AlphaFold's SwissProt subset of 542,378 predicted structures (8).

2.1.4 ProteinSSN

In a manner akin to the PST, ProteinSSN represents proteins at the amino acid level through graph structures, as in Figure 5. Unlike the modulation of weights in ESM2, ProteinSSN employs ESM2 embeddings as node features. The graph's geometry is a residue graph based on the k-nearest neighbor principle. Edge attributes encapsulate the feature relationships of connected nodes, taking into account inter-atomic distances, local N-C positions, and sequential position encoding. In addition to these invariant features, the graph also incorporates the 3D coordinates of amino acids (AAs) directly in Euclidean space. The constructed graphs are subsequently input into an Equivariant Graph Neural Network to update the hidden node representation and node coordinates. During the pretraining phase, a multi-layer perceptron transforms this hidden representation into the probability distribution of AA types. The model's learnable parameters are optimized by minimizing the cross-entropy between the recovered AAs and the ground-truth AAs in wild-type proteins (9).

This method utilizes the CATH 4.3 dataset for model training, which comprises 30,948 experimental protein structures with less than 40% sequence identity. For efficiency, proteins exceeding 2,000 AAs in length are excluded from the dataset.

2.2 Zero-shot variant effect prediction

Building upon the concepts introduced by ProteinGym(14) and ESM-variance(16), we employ the masked-margin score for calculating the Deep Mutational Scanning (DMS) score. For each mutant sequence, masks are applied to the mutation sites. The models then predict the probability of each amino acid (AA) at these mutation sites. The DMS score is computed as the difference in probabilities between the wild-type AA and the mutant-type AA.

To assess the accuracy of mutation effect prediction, we utilize the Spearman ranking score. Initially, both the ground truth and predicted DMS scores are converted into rankings. Subsequently, the covariance between these two rank variables is calculated and normalized by the product of their standard deviations. The Spearman rank correlation coefficient ranges from -1 to 1, with a coefficient of 1 indicating a perfect positive correlation between the two variables.

At present, given the constraints of time, our focus is limited to single-site mutations. Our dataset comprises 140 Deep Mutational Scanning (DMS) arrays, encompassing a total of 567,983 mutant sequences. Future work may extend this analysis to multi-site mutations to provide a more comprehensive understanding of mutation effects.

2.3 Proposed Model

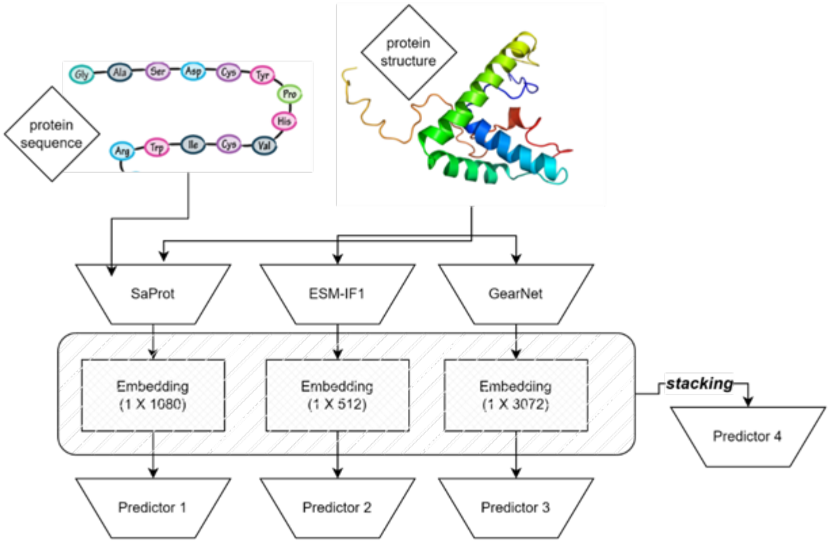

Figure 1: Proposed model: individual predictors and stacking embedding predictor

Based on the benchmark results, it is hypothesized that the performance of Deep Mutational Scanning (DMS) prediction, particularly in the stability function, can be enhanced by employing alternative protein structure encoders such as ESM-IF(17) or GearNet(18) for feature extraction. This hypothesis led to the proposition of our feature-ensemble models, as depicted in Figure 1.

Given a pair of mutant protein sequence and its corresponding wild-type protein structure, SaProt encodes them into a sequence embedding matrix of dimensions equal to the sequence length multiplied by the embedding dimension (1080). The features at the mutant site are utilized to train Predictor1, which is designed to predict the DMS score of the specific mutant.

ESM-IF is employed as a protein structure encoder with a focus on the protein backbone. The protein backbone structure information is initially converted into structure embeddings of dimensions equal to the sequence length times the embedding dimension (512). Predictor2 is trained using the mutant site presentation vector to predict all the DMS scores for different amino acid types (1x20).

GearNet and Predictor3 operate similarly to ESM-1F and Predictor2, with the added expectation that GearNet encodes the information of the entire protein, including the side chain. This is crucial for certain DMS score classes such as binding and activity.

For each mutant sequence, the sequence embeddings from SaProt, along with the wild-type structure embeddings from ESM-IF and GearNet, are stacked together to further train Predictor4. It is anticipated that by stacking these features, information from different encoders can be utilized, thereby enhancing the prediction accuracy for the DMS score.

2.4 Experimental Dataset and Techniques

In this Experiment, we have discussed that we are depending on the ProteinGym (14) dataset that is depicted in Appendix B. Basically, it is a collection of benchmarks aiming at comparing the ability of models to predict the effects of protein mutations. The benchmarks in ProteinGym are divided according to mutation type (substitutions vs. indels), ground truth source (DMS assay vs. clinical annotation), and training regime (zero-shot vs. supervised).

These Deep Mutational Scanning (DMS) assays cover a broad spectrum of functional properties 156 (e.g., ligand binding, aggregation, viral replication, and drug resistance) and span various protein 157 families (e.g., kinases, ion channel proteins, transcription factors, and tumor suppressors) across 158 different taxa (e.g., humans, other eukaryotes, prokaryotes, and viruses)

For the five-fold cross-validation experiment, we adhere to the Random data split in the ProteinGym dataset to facilitate easier comparison. The Deep Mutational Scanning (DMS) scores are initially normalized by subtracting the mean and dividing by the standard deviation within each DMS array.

All four predictors are based on three-layer perceptrons with the Rectified Linear Unit (ReLU) function serving as the activation function. The Mean Square Error (MSE) loss and the Adam optimizer are employed during the training phase. Each model undergoes training for 500 epochs, with the best checkpoint selected based on the MSE of the validation set. This rigorous methodology ensures the robustness and reliability of our predictive models.

For this experiment, one of the used techniques is also the Spearman's Rank correlation coefficient technique which can be used to summarise the strength and direction (negative or positive) of a relationship between two variables. The result will always be between 1 and minus 1. The equation is shown in equation 1

$$\rho = 1 - \frac{6\sum(d_i^2)}{n(n^2-1)} \tag{1}$$

Where n is the number of observations and ρ is Spearman's rank correlation coefficient between a model's predicted scores and an experimentally acquired outcomes that is then used to benchmark models.

Also, $di = R(X) - R(Y)$ is the difference between two ranks of each observation. $\rho S \approx 1$ or $\rho S \approx -1$ indicates that the ranks of the two observations have a strong statistical dependence, $\rho S \approx 0$ indicates no statistical dependence between the ranks of two observations.

3 Experimental Results

This section contains the experimental results of the Zero-shot benchmarking results as well as the ensemble supervised learning methods. It also includes a subsection for the analysis and insight of these results.

3.1 Benchmark result

3.1.1 General performance and spearman score by function

		Spearman by function				
Method	Avg. Spearman	Activity	Binding	Expression	Organismal Fitness	Stability
ESM2	0.414	0.425	0.337	0.415	0.369	0.523
SaProt	0.454	0.471	0.384	0.486	0.417	0.562
PST	0.420	0.468	0.304	0.447	0.387	0.491
ProtSSN	0.430	0.471	0.343	0.427	0.401	0.476

Table 1: Average and Function Specified Spearman Score for different models

We initially compare the general performance of the models using the average Spearman's rank. As indicated in Table 1, all three methods outperform ESM2. Among these, SaProt significantly enhances model performance, which aligns with our expectations but also surprises us. SaProt employs the largest structure dataset and utilizes more training resources than the other models. However, it only uses FoldSeek to convert structure into 20 alphabets, which should result in a loss of structural information compared to other complex Graph Neural Network (GNN) based structure extractors. This could be either because FoldSeek itself is highly efficient, or because a larger dataset yields better results. Further evaluation is required to answer this question.

In the ProteinGym dataset, DMS scores can be categorized into five functions: Activity, Binding, Expression, Organismal Fitness, and Stability. We subsequently evaluated the models' performances across these different functions. While the results for other functions generally follow the same trend as the average performance, the stability results reveal an intriguing phenomenon. Two methods, PST and ProtSSN, actually significantly decrease performance compared to ESM2, which contradicts the results of structure-only methods. For instance, two inverse-folding models, ESM-IF1 and ProteinMPNN, have very high stability scores (0.624 and 0.565, respectively) even when their general performance is not good (average Spearman 0.422 and 0.258, respectively). (These results are not reported in the report, but can be found in the ProteinGym Paper). This might suggest that model stability is highly related to the protein backbone structure, but these three models failed to capture the relevant information.

3.1.2 MSA depth and Sequence Length

	Spearman by MSA Depth			Spearman by sequence length		
Method	Low depth	Medium depth	High depth	<500	500-1000	>1000
ESM2	0.335	0.406	0.515	0.433	0.399	0.375
SaProt	0.427	0.435	0.513	0.462	0.433	0.455
PST	0.394	0.401	0.477	0.434	0.389	0.405
ProtSSN	0.412	0.415	0.473	0.442	0.400	0.425

Table 2: Spearman Score by different MSA Depth and sequence length

In the context of protein structure prediction, such as with AlphaFold, Multiple Sequence Alignment (MSA) depth is utilized. MSA depth is calculated by tallying the number of non-gap residues for each position in the MSA, thereby quantifying the diversity of sequences in a Multiple Sequence Alignment. The ProteinGym dataset categorizes the MSA depth of its dataset into three classes: high, medium, and low. In table 2Structure-based models have shown an increase in performance at low depths, but no improvement at high depths.

Language models generally struggle with extremely long sequences. Therefore, we also evaluated performance across different sequence lengths. As indicated in Table 2, all four models exhibit optimal performance with sequences less than 500 amino acids (AAs) in length. Interestingly, all structure models significantly enhance the performance of long sequences (more than 1000 AAs), but sequences of medium length do not benefit substantially from the protein structure. This observation warrants further investigation to understand the underlying reasons and potential implications.

3.2 Supervised Learning

		Function				
Method	Avg. Spearman	Activity	Binding	Expression	Fitness	Stability
ProteinNPT (SOTA)	0.730 (0.192)	0.697(0.132)	0.656(0.218)	**0.757(0.214)**	0.665(0.223)	**0.892(0.134)**
ESM-IF	0.717(0.172)	0.645(0.168)	0.674(0.229)	0.667(0.122)	0.648(0.167)	0.857(0.046)
GearNet	0.673(0.189)	0.606(0.160)	0.644(0.248)	0.616(0.131)	0.595(0.182)	0.833(0.045)
SaProt	0.724(0.190)	0.742(0.154)	0.713(0.257)	0.739(0.135)	0.702(0.175)	0.785(0.103)
Stacking Embedding	**0.737(0.187)**	**0.758(0.143)**	**0.721(0.274)**	0.740(0.154)	**0.692(0.182)**	0.854(0.104)

Table 3: Average and Function Specified Spearman for different trained predictors. (standard deviation in the parentheses

In alignment with the benchmark results, we present the Spearman score results for various models across the entire dataset and different functional classes. The first row represents the state-of-the-art (SOTA) method, which is pre-trained with a protein language model and Multiple Sequence Alignment (MSA) information, and subsequently fine-tuned with the ProteinGym dataset(19).

The second to fourth rows represent three basic predictors. The average Spearman score and the Spearman score across all five functions exhibit an increase compared to the zero-shot result for the ESM-IF and SaProt models. This suggests that our supervised training effectively aids the features in predicting the DMS score via a Multilayer Perceptron (MLP).

The fifth model represents our stacking embedding models. It is evident that by stacking the features from different encoders, we significantly improve the performance for the average Spearman score and four functions, with the exception of stability, which is only slightly worse than the ESM-IF.

When compared to the SOTA method, our model exhibits a slight improvement in average performance and outperforms in three functions. This demonstrates the efficacy of our approach in enhancing the predictive accuracy of DMS scores.

3.3 Analysis and Insights

In the benchmark section of our study, we conducted a comparative analysis of the performance of various Sequence-based Protein Language Models (Si-PLMs) within the context of the ProteinGym dataset. The results indicated that all the models outperformed the original ESM2-650M model. This suggests that the incorporation of protein structure information can enhance the protein language model's ability to predict the effects of mutations.

However, it's important to note that different methodologies employed to represent the model yielded varying results. Among the tested models, SaProt emerged as the most effective. This was achieved by converting the structure into a structure alphabet using the Foldseek method. This transformation allowed for a more efficient representation of the protein structure, thereby improving the model's predictive capabilities.

In the supervised-learning section of our study, we explored the potential benefits of stacking the embeddings from different models. Our model, which utilized GearNet and ESM-IF as additional structure encoders, outperformed the state-of-the-art (SOTA) methods. This indicates that our stacked features contain a richer set of information about the protein structure, which in turn enhances the prediction of mutation effects.

By integrating the features from different encoders, we were able to capture a broader spectrum of structural information. This comprehensive representation of the protein structure proved to be instrumental in improving the accuracy of our mutation effect predictions. It underscores the potential of our approach in advancing the field of protein structure prediction and its applications in various domains such as drug discovery and genetic engineering.

4 Conclusion

4.1 Conclusion

The project of leveraging structure information in the prediction of the variant and mutation effects has been completed successfully. The project involved the establishment of a benchmark model and dataset to compare previously utilized protein language models' performance. Firstly, three different SI-PLMs models that usePLM (ESM2-650M) model as a baseline were chosen, which are SaProt, PST and ProteinSSN. Each was trained individually on the ProteinGym DMS dataset and then compared using the Zero-shot variant effect prediction method. After benchmarking these models, an ensemble model is created by stacking the three previously trained models.Finally, five-fold cross-validation experiment is conducted to compare our proposed methods with baseline and state-of-art models.

From the results, we conclude that by individually training predictors for each base model, we observe an enhancement in performance across all models, surpassing the zero-shot models. The training performance showed that SaProt model is the best at predicting DMS scores without prior training. But for stability, it's not as good as ESM-IF and ProteinMPNN. Finally, by combining three models to make the ensemble model, we get the best results, even better than the start-of-the-art methods. Hence, through supervised ensemble model training, we achieve superior performance compared to individual zero-shot methods.

4.2 Limitations and Future Work

Due to time constraints, our work has the following limitations:

- **Single-Site Substitution Mutations:** Our models currently only consider single-site substitution mutations. We exclude multiple-site substitutions and indels mutations. This is a significant limitation as multiple-site substitutions and indels mutations can have profound effects on protein function.
- **Mutant-Type Structures:** Our models only use the wild-type structures predicted from AlphaFold2. This is due to the insufficient prediction results from AlphaFold2 for single site mutations. Further work can be done to use other models to predict mutant structure.
- **Embedding at the Mutant Site:** Our models currently only use the embedding at the mutant site. Considering the embedding at the mutation site and its surroundings might enhance model performance. This is because even a single site mutation can influence its surroundings.
- **Predictor Architecture:** Our models only use MLP as the predictor. We do not test with convolution layers or attention layers. These layers could potentially capture more complex patterns and improve the model's performance.
- **Ablation Study:** We need an ablation study to further compare the difference between our model and the SOTA methods, beyond the Spearman performance. This would provide a more comprehensive understanding of the strengths and weaknesses of our model.

These limitations provide avenues for future research to improve the performance and applicability of our models in predicting the effects of protein mutations.

Hence, for future work, it would be valuable to aggregate the results obtained from the Spearman coefficient analysis to derive an average performance metric. By calculating the average performance across multiple experiments, we can gain a more comprehensive understanding of the model's predictive capabilities and assess its consistency across different datasets or conditions.

Moreover, our focus lies on expanding the scope of our models to encompass multiple-site substitutions and indels mutations, thus providing a more holistic understanding of their impact on protein function. Additionally, extending the analysis to include predictions generated by other ensemble methods, embedding strategies and predictor architectures presents an opportunity to explore alternative modeling approaches. This holistic evaluation framework can contribute to refining predictive models and optimizing their performance for various applications in the field.

5 Computing resources

All the models were trained and tested on NVIDIA A100 GPUs provided by the Texas A&M High Performance Research Computing (HPRC) facility. In the benchmarking phase, PST, SaProt, and ProteinSSN took approximately 20 to 24 hours for result prediction, while ESM2 necessitated approximately 48 hours. For the ensemble model training phase, individual predictors like ESM-IF, GearNet, and SaProt demanded approximately 8, 8, and 18 hours, respectively, on the ProteinGym dataset. The stacked model necessitated around 18 hours for training.

6 Authors contributions

All authors, A.A., Y.L., and O.S., played integral roles in conceptualizing, executing, and analyzing the research project. Collectively, they conducted an extensive literature survey, reviewing relevant papers and studying state-of-the-art models in the field. They meticulously analyzed how existing models extract structural information and incorporate it into their methodologies by benchmarking on the ProteinGym dataset. Together, they performed benchmarking, training, and evaluation of individual predictors and the ensemble model, ensuring thorough analysis. Additionally, all authors collectively prepared the project's presentation, progress report, and final report. Additionally, the timeline for these tasks have been summarized using a gantt chart in Appendix C.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[2] R. M. Rao, J. Liu, R. Verkuil, J. Meier, J. Canny, P. Abbeel, T. Sercu, and A. Rives, "Msa transformer," in *International Conference on Machine Learning*, pp. 8844–8856, PMLR, 2021.

[3] J. Meier, R. Rao, R. Verkuil, J. Liu, T. Sercu, and A. Rives, "Language models enable zero-shot prediction of the effects of mutations on protein function," *Advances in neural information processing systems*, vol. 34, pp. 29287–29303, 2021.

[4] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, A. dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, *et al.*, "Language models of protein sequences at the scale of evolution enable accurate structure prediction," *BioRxiv*, vol. 2022, p. 500902, 2022.

[5] R. Rao, J. Meier, T. Sercu, S. Ovchinnikov, and A. Rives, "Transformer protein language models are unsupervised structure learners," *Biorxiv*, pp. 2020–12, 2020.

[6] K. Weissenow, M. Heinzinger, and B. Rost, "Protein language-model embeddings for fast, accurate, and alignment-free protein structure prediction," *Structure*, vol. 30, no. 8, pp. 1169–1177, 2022.

[7] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, *et al.*, "Highly accurate protein structure prediction with alphafold," *Nature*, vol. 596, no. 7873, pp. 583–589, 2021.

[8] D. Chen, P. Hartout, P. Pellizzoni, C. Oliver, and K. Borgwardt, "Endowing protein language models with structural knowledge," *arXiv preprint arXiv:2401.14819*, 2024.

[9] Y. Tan, B. Zhou, L. Zheng, G. Fan, and L. Hong, "Semantical and topological protein encoding toward enhanced bioactivity and thermostability," *bioRxiv*, pp. 2023–12, 2023.

[10] Y. Sun and Y. Shen, "Structure-informed protein language models are robust predictors for variant effects," *Research Square*, 2023.

[11] J. Su, C. Han, Y. Zhou, J. Shan, X. Zhou, and F. Yuan, "Saprot: protein language modeling with structure-aware vocabulary," *bioRxiv*, pp. 2023–10, 2023.

[12] M. Heinzinger, K. Weissenow, J. G. Sanchez, A. Henkel, M. Steinegger, and B. Rost, "Prostt5: Bilingual language model for protein sequence and structure," *bioRxiv*, pp. 2023–07, 2023.

[13] M. van Kempen, S. S. Kim, C. Tumescheit, M. Mirdita, C. L. Gilchrist, J. Söding, and M. Steinegger, "Foldseek: fast and accurate protein structure search," *Biorxiv*, pp. 2022–02, 2022.

[14] P. Notin, A. Kollasch, D. Ritter, L. Van Niekerk, S. Paul, H. Spinner, N. Rollins, A. Shaw, R. Orenbuch, R. Weitzman, *et al.*, "Proteingym: large-scale benchmarks for protein fitness prediction and design," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[15] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, *et al.*, "Evolutionary-scale prediction of atomic-level protein structure with a language model," *Science*, vol. 379, no. 6637, pp. 1123–1130, 2023.

[16] W. Lin, J. Wells, Z. Wang, C. Orengo, and A. C. Martin, "Varipred: Enhancing pathogenicity prediction of missense variants using protein language models," *bioRxiv*, pp. 2023–03, 2023.

[17] C. Hsu, R. Verkuil, J. Liu, Z. Lin, B. Hie, T. Sercu, A. Lerer, and A. Rives, "Learning inverse folding from millions of predicted structures," in *International conference on machine learning*, pp. 8946–8970, PMLR, 2022.

[18] Z. Zhang, M. Xu, A. Lozano, V. Chenthamarakshan, P. Das, and J. Tang, "Enhancing protein language model with structure-based encoder and pre-training," in *ICLR 2023-Machine Learning for Drug Discovery workshop*, 2023.

[19] P. Notin, R. Weitzman, D. Marks, and Y. Gal, “Proteinnpt: improving protein property prediction and design with non-parametric transformers,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[20] N. Brandes, G. Goldman, C. H. Wang, C. J. Ye, and V. Ntranos, “Genome-wide prediction of disease variant effects with a deep protein language model,” *Nature Genetics*, vol. 55, no. 9, pp. 1512–1522, 2023.

Appendix

Appendix A: Protein Language Models

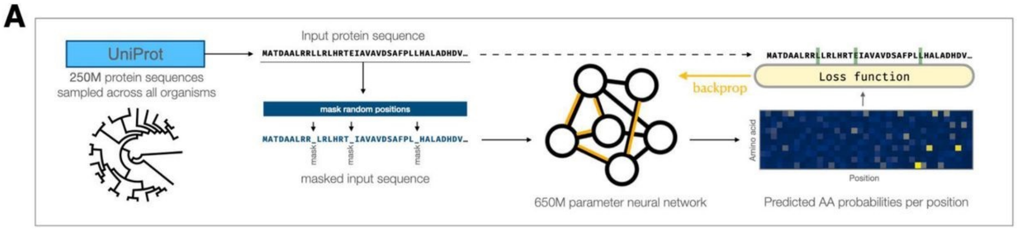
A
UniProt
250M protein sequences sampled across all organisms
Input protein sequence
MATDAALRRLLRLHRTEIAVAVDSAFPLLHALADHDV…
mask random positions
MATDAALRR_LRLHRT_IAVAVDSAFPL_HALADHDV…
mask
masked input sequence
650M parameter neural network
backprop
Loss function
Amino acid
Position
Predicted AA probabilities per position

Figure 2: ESM-variance: Benchmarking Mutation Impact with LLR. (20).

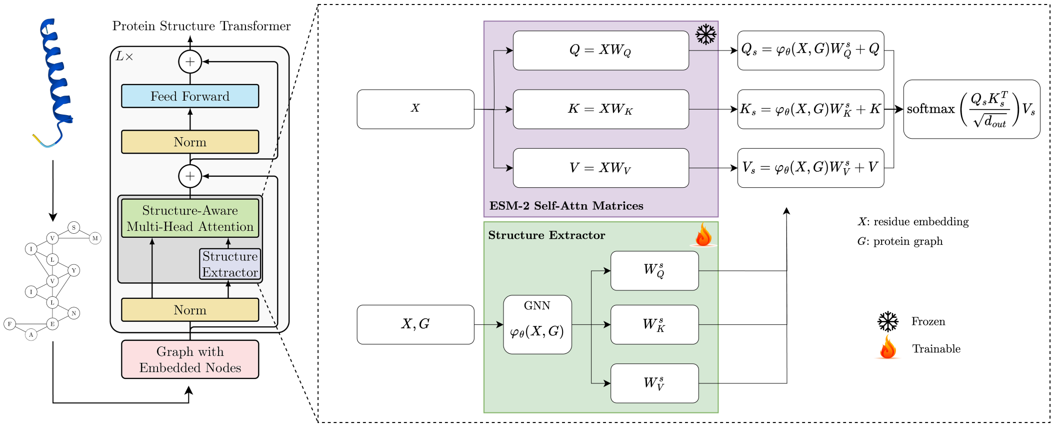
Protein Structure Transformer
L×
Feed Forward
Norm
Structure-Aware Multi-Head Attention
Structure Extractor
Norm
Graph with Embedded Nodes
X
$Q = XW_Q$
$K = XW_K$
$V = XW_V$
ESM-2 Self-Attn Matrices
$Q_s = \varphi_\theta(X,G)W_Q^s + Q$
$K_s = \varphi_\theta(X,G)W_K^s + K$
$V_s = \varphi_\theta(X,G)W_V^s + V$
softmax $\left(\frac{Q_s K_s^T}{\sqrt{d_{out}}}\right)V_s$
Structure Extractor
X, G
GNN
$\varphi_\theta(X,G)$
W_Q^s
W_K^s
W_V^s
X: residue embedding
G: protein graph
Frozen
Trainable

Figure 3: PST: Enhancing ESM-2 with GNN-Based Structure Extraction (8).

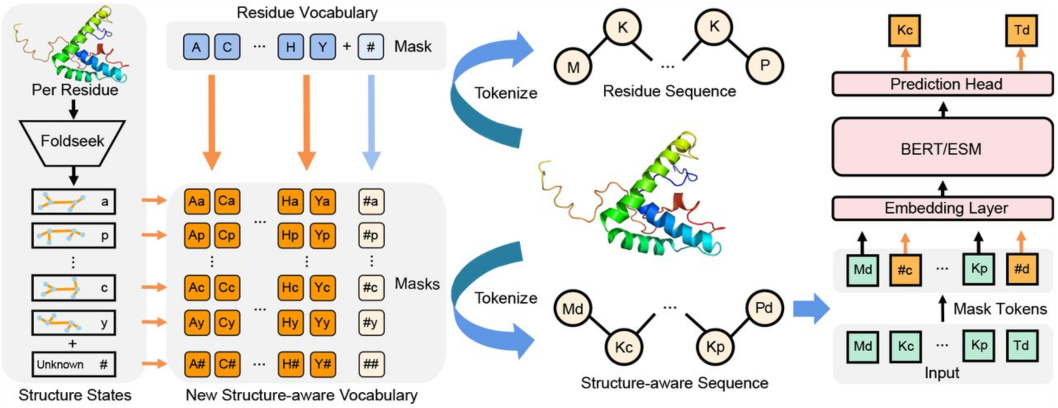

Figure 4: SaProt: Protein Language Modeling with Structure-aware Vocabulary (11).

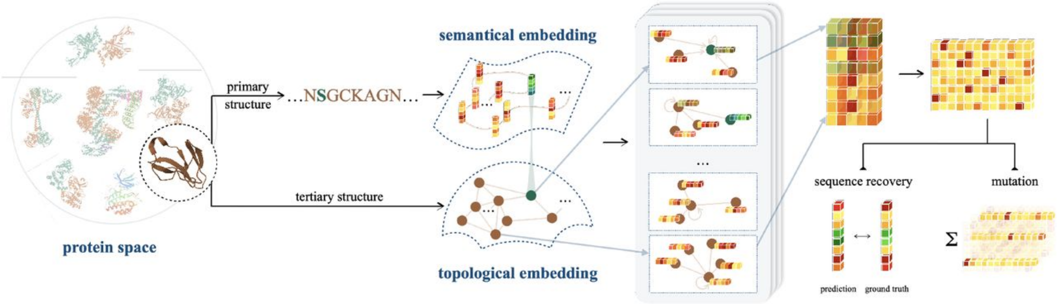

Figure 5: ProtSSN: Learning Protein Structures and Semantics (9).

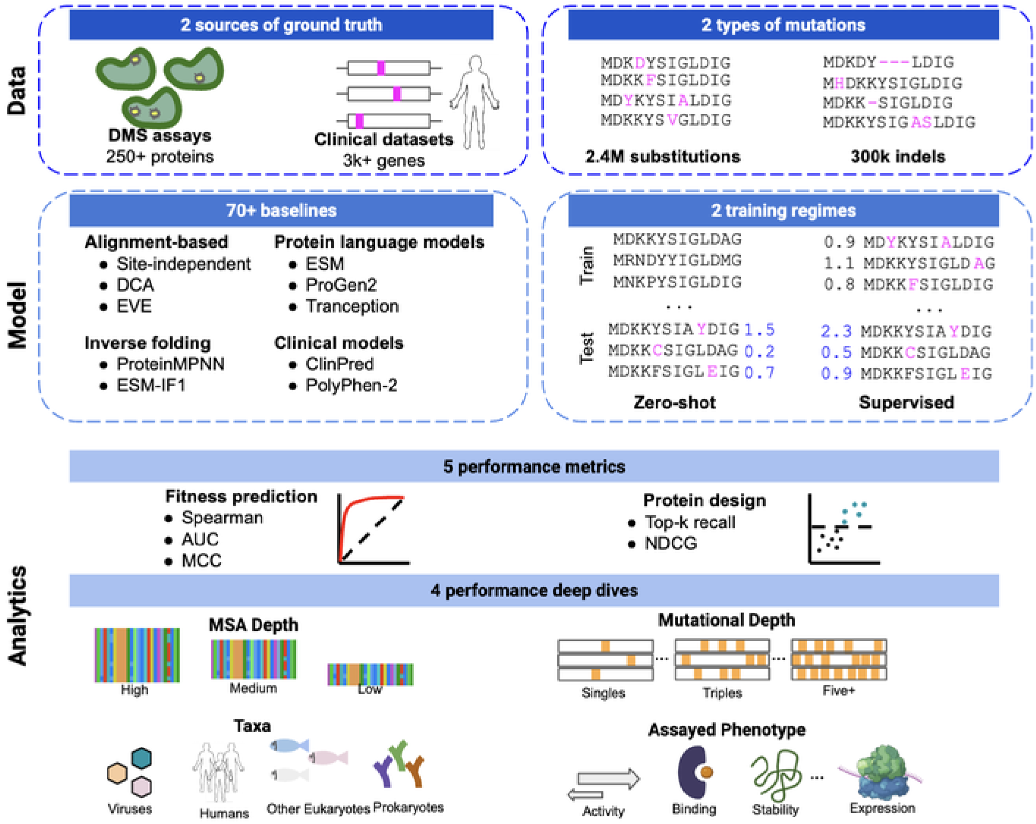

Figure 6: ProteinGym Dataset Details.

Appendix B: ProteinGym Dataset

Appendix C: Milestones and Timeline

In this section, we have summarized our milestones and timelines as seen in the Gantt Chart in Figure 7, and the details are discussed below.

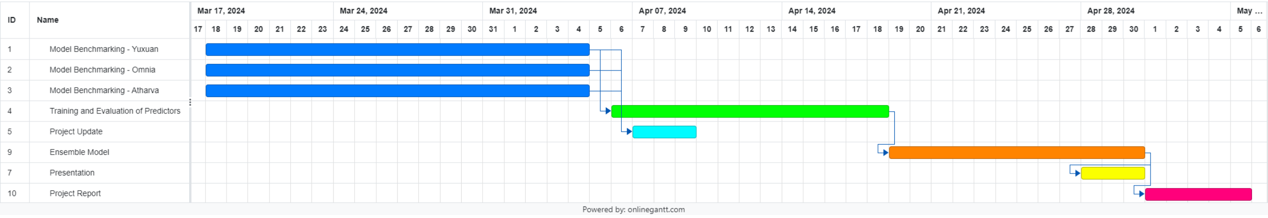

Figure 7: A Gantt Chart depicting our milestones and timelines.

Individual Model Benchmarking

- **Timeline**: March 18, 2024, to April 04, 2024
- **Responsibilities**: Each team member independently benchmarked a separate existing model that uses protein structure information for variant effect prediction. This parallel approach provided us with diverse insights and a comprehensive comparative framework for our subsequent evaluation phase.

Training and Evaluation Assessment

- **Timeline**: April 06, 2024, to April 18, 2024
- **Responsibilities**: We discussed and assessed the benchmarking results of the individual models. The team compared and contrasted the performance of each model, determining the best aspects of each. Our collaborative efforts in this phase were critical to identifying a model with the most potential for improvement and impact.

Project Update

- **Timeline**: April 07, 2024, to April 09, 2024
- **Responsibilities**: We collectively prepared a concise report detailing the progress of our benchmarking efforts, the evaluation methods we're employing, and any initial observations or challenges we have encountered.

Ensemble Model

- **Timeline**: April 19, 2024, to April 30, 2024
- **Responsibilities**: The development of the ensemble model was a collaborative effort, with all team members contributing to the selection, training, and integration of individual predictors to enhance model performance.

Presentation

- **Timeline**: April 27, 2024, to April 30, 2024
- **Responsibilities**: The presentation preparation and delivery required teamwork, with each team member contributing to the creation of materials and actively participating in delivering the content.

Project Report

- **Timeline**: May 1, 2024, to April 5, 2024
- **Responsibilities**: The completion of the final project report was a combined effort, with all team members contributing to the documentation of research objectives, methodologies, results, and conclusions.