Semantical and Topological Protein Encoding Toward Enhanced Bioactivity and Thermostability

Yang Tan[1,2] *, Bingxin Zhou[1] *†, Lirong Zheng[1], Guisheng Fan[2] and Liang Hong[1]
[1]Shanghai Jiao Tong University
{bingxin.zhou, lirong_zheng, hongl3liang}@sjtu.edu.cn
[2]East China University of Science and Technology
tyang@mail.ecust.edu.cn, gsfan@ecust.edu.cn

***Abstract*—Protein engineering is a pivotal aspect of synthetic biology, involving the modification of amino acids within existing protein sequences to achieve novel or enhanced functionalities and physical properties. Accurate prediction of protein variant effects requires a thorough understanding of protein sequence, structure, and function. Deep learning methods have demonstrated remarkable performance in guiding protein modification for improved functionality. However, existing approaches predominantly rely on protein sequences, which face challenges in efficiently encoding the geometric aspects of amino acids' local environment and often fall short in capturing crucial details related to protein folding stability, internal molecular interactions, and bio-functions. Furthermore, there lacks a fundamental evaluation for developed methods in predicting protein stability, although it is a key physical property that is frequently investigated in practice. To address these challenges, this paper introduces a novel pre-training framework that integrates sequential and geometric encoders for protein primary and tertiary structures. This framework guides mutation directions toward desired traits by simulating natural selection on wild-type proteins and evaluates variant effects based on their fitness to perform specific functions. We assess the proposed approach using four benchmarks comprising approximately 200 deep mutational scanning assays. These include two refined benchmarks for protein activities and two novel benchmarks for thermostability. The prediction results showcase exceptional performance across extensive experiments when compared to other zero-shot learning methods, all while maintaining a minimal cost in terms of trainable parameters. This study not only proposes an effective framework for more accurate and comprehensive predictions to facilitate efficient protein engineering, but also enhances the *in silico* assessment system for future deep learning models to better align with empirical requirements.**

***Index Terms*—protein engineering, directed evolution, deep mutation scanning dataset, self-supervised learning, zero-shot prediction, graph neural networks, protein language model**

I. Introduction

The diverse roles proteins play in nature necessitate their involvement in varied bio-functions, such as catalysis, binding, scaffolding, and transport [9]. Advances in science and technology have positioned proteins, particularly those with desired catalytic and binding properties, to pivotal positions in medicine, biology, and scientific research. For example, endonucleases (*e.g.*, CRISPR/Cas and Argonaute proteins) and antibodies are extensively employed in gene editing [45], nucleic acid diagnostics [14], and targeted cancer therapy [56]. While wild-type proteins exhibit optimal bio-functionality in their native environments, industrial applications often demand adaptation to conditions such as high temperature, high pressure, strong acidity, and strong alkalinity. The constrained efficacy of proteins in meeting the stringent requirements of industrial functioning environments hinders their widespread applications [2], [18], [19], [68], [78]. Consequently, there arises a need to engineer these natural proteins to enhance their functionalities, aligning them with the demands of both industrial and scientific applications.

Analyzing the relationship between protein sequence and function yields valuable insights for engineering proteins with new or enhanced functions in synthetic biology. The intrinsic goal of protein engineering is to unveil highly functional sequences by navigating the intricate, high-dimensional surface of the fitness landscape [72], which delineates the relationship between amino acid (AA) sequences and the desired functional state. Predicting the effects of AA alternation in proteins, *i.e.*, mutation, thus allows researchers to dissect how changes in the AA sequence can impact the protein's catalytic efficiency, stability, and interactions with other molecules [58]. However, the extensive, non-contiguous search space of AA combinations poses challenges for conventional methods, such as rational design [1] or directed evolution [67], in efficiently and effectively identifying protein variants with desired properties. In response, deep learning has emerged as a modern solution that proposes favorable protein variants with high fitness.

Deep learning approaches have been instrumental in advancing scientific insights into proteins, predominatly categorieed into two primary domains: sequence-based [30], [34], [60] and structure-based [33], [80], [81] methods. Analogous to the analysis of semantics and syntax in natural language, protein language models (PLM) interpret protein sequences as raw text and employ autoregressive inference techniques [34], [43] with self-attention mechanisms [3], [41], [51]. These methods reveal hidden long-range sequential dependencies that imply sophisticated projections to protein functionalities. However, both sequence-to-structure and structure-to-function projections present a many-to-one relationship. While data-intensive methods with computationally demanding propagation modules demonstrate promising performance across diverse inference tasks, concerns arise regarding the efficiency

* Equal contribution first authors
† Corresponding author

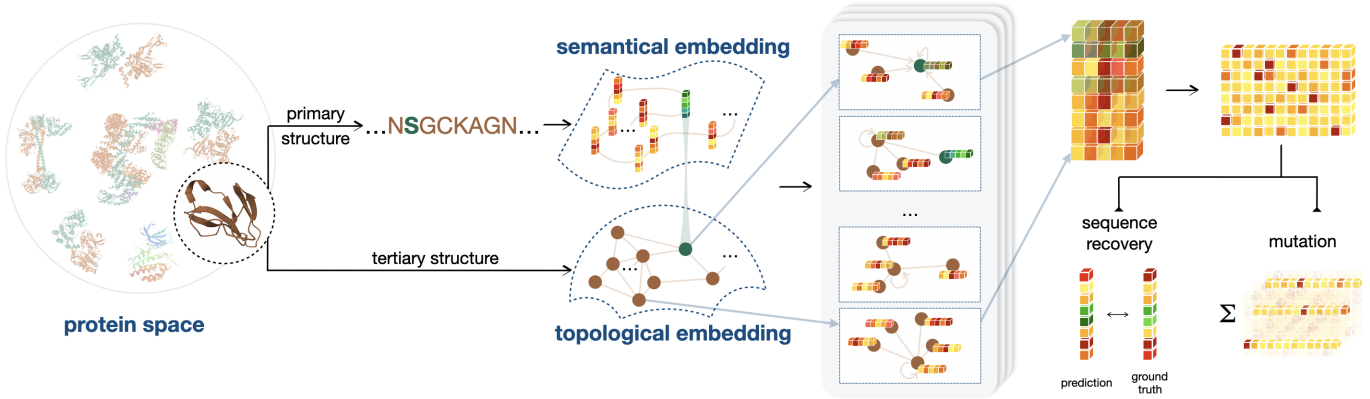

Fig. 1. An illustration of PROTSSN that extracts the semantics and topology of a protein by learning its primary and tertiary structures. The hidden representation can be decoded for variants effect prediction that recognizes the impact of mutating a few sites of a protein on its functionality.

of these pathways when navigating the complexities of the vast sequence space associated with a target function. Alternatively, structure-based modeling methods serve as an effective enhancement to complement sequence-oriented inferences [22], [62], [74], [79]. The structure of a protein, determined by its AA sequence, not only guides molecular interactions but also dictates its functionality. An accurate description of a protein's microenvironment is particularly crucial for determining its properties or functionality. Given that core mutations often induce functional defects through subtle disruptions to structure or dynamics [53], incorporating protein geometry into the learning process can offer valuable insights into stabilizing protein functioning. Nevertheless, structure encoders fall short in capturing long-range sequential connections for an AA beyond its contact region and overlook correlations that do not conform to the 'structure-function determination' heuristic. Consequently, there is a pressing need to develop a novel framework that overcomes limitations inherent in individual implementations of sequence or structure-based investigations.

To evaluate the capability of an established deep learning framework in identifying favorable protein variants, a common *in silico* assessment involves predicting mutation effects. **ProteinGym** [43] comprehends fitness scoring of mutants to catalytic activity and binding affinity across over 80 well-studied proteins of varying deep mutational scanning (DMS) assays and taxa. Each protein includes tens of thousands of mutants documented from high-throughput screening techniques. Other evaluation benchmarks predominantly center around supervising tasks with fixed data partitioning strategy, such as **FLIP** [5] and **TAPE** [46]. Although **ProteinGym** provides comprehensive evaluation benchmarks for mutation effects predictions in both supervised and self-supervised learning models, its primary focus lies on protein activities, with limited assessment of protein stability or thermostability. In practice, protein engineering extends beyond enhancing catalytic activities to maintaining or increasing thermostability for prolonged functional lifetimes and applications under elevated temperatures [32]. A trade-off relationship emerges between activity and stability, where enhanced activity often correlates with diminished protein thermostability, and vice versa. This phenomenon arises due to the structural flexibility induced by increased activity, which reduces the energy barrier towards the functional form but concurrently diminishes intramolecular interactions such as hydrogen bonds and salt bridges, ultimately leading to reduced thermostability [78]. Consequently, there is a need to enrich mutation effect prediction benchmarks that evaluate the model's efficacy in capturing variant effects concerning thermostability.

In this study, we address the mutation effect prediction tasks on thermostability with **DTm** and **DDG**, two new single-site mutation benchmarks that measure thermodynamic stability using ΔTm and $\Delta\Delta$G values, respectively. Differing from activity-related evaluations, the thermostability of a particular mutant depends on both the protein and the experimental environment. Consequently, both benchmarks group experimental assays based on the protein-condition combinations. In total, we establish $3,349$ mutants across 102 protein-condition combinations, with 66 in **DTm** and 36 in **DDG**. These two datasets supplement the existing publicly available DMS assay benchmarks and facilitate ready-to-use assessment for future deep learning methods on improving protein stability. On the other hand, the activity of proteins can be roughly classified into interactions and catalysis, where the functionality of scaffold proteins mainly depends on the binding affinity and enzymes are more focused on the catalytic efficiency. To provide a finer and more insightful analysis of the prediction performance from deep learning models, we separate **ProteinGym** into two subsets by protein functions, *i.e.*, interaction (**ProteinGym**-int) and catalysis (**ProteinGym**-cat).

The prediction of protein functionality benefits from integrating information from both sequence and structure, as they both play pivotal roles in impacting protein's thermostability, interaction with substrates, and catalytic process [28], [77]. To address the limitations exhibited in previous sequence or structure-based deep learning models, we introduce a new framework named PROTSSN, which assimilates the **S**emantics and **T**opology of proteins using language and graph models. PROTSSN incorporates the intermediate state of protein

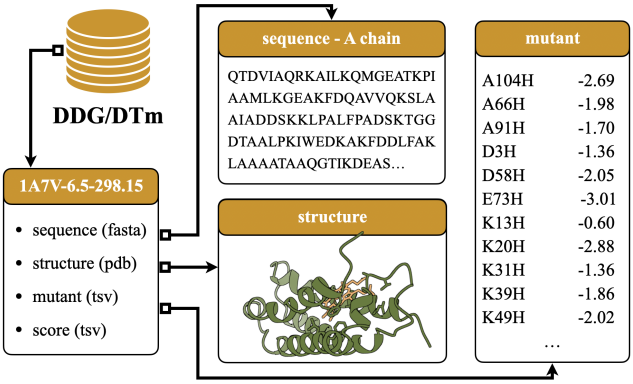

Fig. 2. An example source record of the mutation assay.

structures and facilitates the discovery of an efficient and effective trajectory for mapping protein sequences to functionalities. The developed model extends the generalization and robustness of self-supervised protein language models while maintaining low computational costs, thereby facilitating self-supervised training and task-specific customization. A funnel-shaped learning pipeline, as depicted in Fig. 1, is designed due to the limited availability of crystallized protein structures compared to observed protein sequences. Initially, the linguistic embedding establishes the semantic and grammatical rules in AA chains by inspecting millions of protein sequences [31]. Then, the topological embedding encodes the microenvironment of AAs, supplementing sequential relationships with spatial connections. Since a geometry can be placed or observed by different angles and positions in space, we represent proteins' topology by graphs and enhance the model's robustness and efficiency with a rotation and translation equivariant graph representation learning scheme. Consequently, the trained model can interpret the characterization of AAs closely related to their functionality. The pre-trained PROTSSN demonstrates its feasibility across a broad range of variant effect prediction benchmarks covering catalysis, interaction, and thermostability.

This work fulfills practical necessities in the development of deep learning methods for protein engineering from three distinct perspectives. Firstly, the developed PROTSSN employs self-supervised learning during training to obviate the necessity for additional supervision in downstream tasks. This zero-shot scenario is desirable due to the scarcity of experimental results, as well as the 'cold-start' situation common in many wet lab experiments. Secondly, the trained model furnishes robust and meaningful approximations to the joint distribution of the entire amino acid chain, thereby augmenting the *epistatic effect* [25], [54] in deep mutations. This augmentation stems from the model's consideration of the nonlinear combinatorial effects of amino acid sites. Additionally, we complement the existing dataset of DMS assay with additional benchmark assays related to protein-environment interactions, specifically focusing on thermostability–an integral objective in numerous protein engineering projects.

TABLE I
STATISTICAL DETAILS OF **DTM** AND **DDG**.

pH Range	DTm			DDG		
	# Assays	avg. mut	avg. AA	# Assays	avg. mut	avg.AA
Acid (pH $<$ 7)	29	29.1	272.8	21	22.6	125.3
Neutral (pH $=$ 7)	14	37.4	221.2	10	23.1	78.3
Alkaline (pH $>$ 7)	23	50.1	233.3	5	24.6	101.4
sum	66	38.2	221.2	36	23.0	108.8

II. BENCHMARKS FOR VARIANT EFFECT PREDICTION

We enrich the current mutation effect prediction benchmarks by protein biochemical functionality and physical properties. In particular, we propose two novel benchmark datasets, **DTm** and **DDG**, to assess protein thermostability, where mutations are evaluated through a rigorous screening process under consistent control environments. We also introduce **ProteinGym**-int and **ProteinGym**-cat with DMS assays for interaction and catalysis–the two main categories of protein functions.

A. New Benchmarks for Thermostability: DTm and DDG

To evaluate model accuracy in predicting protein thermostability, we introduce two novel benchmarks: **DTm** and **DDG** (refer to Fig. 2). The assays are sourced from single-site mutations in **ProThermDB** [42] and **DDMut** [82], respectively. Details of the new datasets are provided in Table I. Considering the significant influence of experimental conditions on protein temperature, we explicitly note the pH environment and assign each protein-environment pairing a name in the format '*UniProt_ID-pH level*' (*e.g.*, '*O00095-8.0*' in **DTm** signifies mutations conducted and evaluated under pH 8.0 for protein O00095). For **DDG**, in addition to experimental pH, we also consider the measurement at the same temperature, naming the protein-environment pairing in the format '*PDB_ID-pH-TM level*'.

*a) **DTm**:* Fig. 2 depicts an example of 'protein-environment' assay in **DTm**. We define the benchmark with attributes 'mutant', 'score', and 'UniProt_ID' by curating assays in **ProThermDB** with at least 10 mutants to guarantee meaningful evaluation. In order to concentrate on single-site mutations compared to wild-type proteins, we exclude records with continuous mutations. To avoid dealing with partial or misaligned protein structures from incomplete wet experimental procedures, we employ UniProt ID to pair protein sequences with folded structures predicted by ALPHAFOLD2 [23] from https://alphafold.ebi.ac.uk. In total, **DTm** consists of 66 such protein-environment pairs and $2,520$ mutations.

*b) **DDG**:* Different from other datasets, mutants in **DDMut** [82] are paired with their PDB documents. Consequently, we employ crystal structures for proteins in **DDG** and conduct additional preprocessing steps for environmental consistency. In particular, we group the mutants of the same protein under identical experimental conditions and examine the alignment between mutation sites and sequences extracted from crystal structures. We discard mutant records that cannot

be aligned with the sequence loaded from the associated PDB document. After removing data points with fewer than 10 mutations, **DDG** comprises 36 data points and 829 mutations.

B. Refined Benchmarks for Functionalities: ProteinGym-cat and ProteinGym-int

We assess our model's capacity to predict deep mutation effects using **ProteinGym** [43]. It is currently the most extensive protein substitution benchmark, comprising roughly $1.5M$ missense variants across 87 DMS assays. These DMS assays cover a broad spectrum of functional properties (*e.g.*, ligand binding, aggregation, viral replication, drug resistance) and spanning various protein families (*e.g.*, kinases, ion channel proteins, G-protein coupled receptors, polymerases, transcription factors, tumor suppressors) across different taxa (*e.g.*, humans, other eukaryotes, prokaryotes, viruses). The evaluation in this study is conducted on 86 variants from the full set, excluding `A0A140D2T1_ZIKV_Sourisseau_growth_2019` that failed to be folded by both AlphaFold2 and ESMFold.

Given the diverse aspects influencing different protein functionalities, we divide **ProteinGym**-int and **ProteinGym**-cat from the full **ProteinGym** dataset. The former encompasses 45 DMS assays related to protein binding affinity, and the latter comprises 41 assays focused on catalytic activity. These assays are categorized based on literature associated with the proteins. Separating assays by the two main categories of functions help investigate the effectiveness of a deep learning model (see also Section V-B for additional discussion). Additionally, as zero-shot protein engineering in practice typically targets a specific protein and function, it becomes essential to select an optimal model based on empirical requirements, *e.g.*, enhancing the catalytic activity of a hydrolase.

III. Multi-level Protein Representation Learning

Labeled data are usually scarce in biomolecule research, which demands designing a general self-supervised model for predicting variant effects on unknown proteins and protein functions. This study utilizes a pre-training scheme that recovers AA types from a given protein backbone structure with noisy local environment observations. It simulates nature's iterative selection of well-functioning mutated proteins from harmful random mutations.

A. Multi-level Protein Representation

a) Protein Primary Structure (Noisy): For a given observation with an AA type $\tilde{\boldsymbol{v}}$, we assume it is in the process of random mutation toward a stable state. The model is then trained to learn a revised state $\boldsymbol{v}$ that is less likely to be eliminated by natural selection due to unfavorable properties such as instability or inability to fold. Formally, we define the perturbed observation by a multinomial distribution as follows:

$$\boldsymbol{\pi}(\tilde{\boldsymbol{v}} \mid \boldsymbol{v}) = p\boldsymbol{\Theta}(\pi_1, \pi_2, \ldots, \pi_{20}) + (1-p)\delta(\tilde{\boldsymbol{v}} - \boldsymbol{v}), \quad (1)$$

where an AA in a protein chain has a chance of p to mutate to one of 20 AAs following the *replacement distribution* $\boldsymbol{\Theta}(\cdot)$ and $(1-p)$ of remaining unchanged. We consider p as a tunable parameter and define $\boldsymbol{\Theta}(\cdot)$ based on the frequency of AA types observed in wild-type proteins in the training dataset.

b) Protein Tertiary Structure: The geometry of a protein is described by $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \boldsymbol{W}_V, \boldsymbol{W}_E, \boldsymbol{X}_V)$, a residue graph based on the k-nearest neighbor (kNN) [80]. Each node $v_i \in \mathcal{V}$ represents an AA in the protein connected to up to k other nodes in the graph that are the closest in Euclidean distance within a contact region of 30Å. Node attributes $\boldsymbol{W}_V$ are hidden semantic embeddings of AA types, and edge attributes $\boldsymbol{W}_E \in \mathbb{R}^{93}$ feature relationships of connected nodes based on inter-atomic distances, local N-C positions, and sequential position encoding. Additionally, $\boldsymbol{X}_V$ records 3D coordinates of AAs in the Euclidean space, which plays a crucial role in the subsequent topological embedding stage to preserve roto-translation equivariance.

B. Semantic Encoding of Protein Sequence

Although it is generally believed by the community that a protein's sequence determines its biological function via the folded structure, following strictly to this singular pathway risks overlooking other unobserved yet potentially influential inter-atomic communications impacting protein fitness. In line with this reasoning, our proposed model, PROTSSN, begins by extracting pairwise relationships for residues through an analysis of proteins' primary structure from $\tilde{\boldsymbol{V}}$ and embed them to hidden representations $\boldsymbol{W}_V$ for residues. At each update, the information and representations of the noisy AA sequence are encoded from the noisy input via an *Evolutionary Scale Modeling* ESM-2 [31][1]. This approach employs a BERT-style masked language modeling (MLM) objective that predicts the identity of randomly selected AAs in a protein sequence by observing their context within the remainder of the sequence. Note that during training, the sequence embedding operates in every epoch as AA types are subject to independent random perturbations. See discussion in Section V-D for alternative encoding strategies.

C. Topological Encoding of Protein Structure

Proteins are structured in 3D space, which requires the geometric encoder to possess roto-translation equivariance to node positions as well as permutation invariant to node attributes. This design is vital to avoid the implementation of costly data augmentation strategies. We practice *Equivariant Graph Neural Networks* (EGNN) [55] to acquire the hidden node representation $\boldsymbol{W}_V^{l+1} = \left\{\boldsymbol{w}_{v_1}^{l+1}, \ldots, \boldsymbol{w}_{v_n}^{l+1}\right\}$ and node coordinates $\boldsymbol{X}_{\text{pos}}^{l+1} = \left\{\boldsymbol{x}_{v_1}^{l+1}, \ldots, \boldsymbol{x}_{v_n}^{l+1}\right\}$ at the $l+1$th layer

$$\begin{aligned}
\boldsymbol{m}_{ij} &= \phi_e\left(\boldsymbol{w}_{v_i}^l, \boldsymbol{w}_{v_j}^l, \left\|\boldsymbol{x}_{v_i}^l - \boldsymbol{x}_{v_j}^l\right\|^2, \boldsymbol{w}_{e_{ij}}\right), \\
\boldsymbol{x}_{v_i}^{l+1} &= \mathbf{x}_{v_i}^l + \frac{1}{n}\sum_{j\neq i}\left(\mathbf{x}_{v_i}^l - \mathbf{x}_{v_j}^l\right)\phi_x\left(\mathbf{m}_{ij}\right), \\
\boldsymbol{w}_{v_i}^{l+1} &= \phi_v\big(\boldsymbol{w}_i^l, \sum_{j\neq i}\boldsymbol{m}_{ij}\big).
\end{aligned} \quad (2)$$

[1]Official implementation at https://github.com/facebookresearch/esm.

In these equations, $\boldsymbol{w}_{e_{ij}}$ represents the input edge attribute on $\mathcal{V}_{ij}$, which is not updated by the network. The propagation rules ϕ_e, ϕ_x and ϕ_v are defined by differentiable functions, *e.g.*multi-layer perceptrons (MLPs). The final hidden representation on nodes $\boldsymbol{W}_V^L$ embeds the microenvironment and local topology of AAs, and it will be carried on by readout layers for label predictions.

IV. Zero-Shot Variant Effect Prediction

ProtSSN is designed for protein engineering with a self-supervised learning scheme. The model is capable of conducting zero-shot variant effect prediction on an unknown protein, and it can generate the joint distribution for the type of all AAs along the protein sequence of interest. This process accounts for the epistatic effect and concurrently returns all AA sites in a sequence. Below we detail the workflow for training the zero-shot model and scoring the effect of a specific mutant.

A. Model Pipeline

a) Training: The fundamental model architecture cascades a frozen sequence encoding module and a trainable tertiary structure encoder. Initially, a protein language model encodes pairwise hidden relationships of AAs by analyzing the input protein sequence and produces a vector representation $\boldsymbol{w}_{v_i} \in \boldsymbol{W}_V$ for an arbitrary AA, where

$$\boldsymbol{W}_V = \mathrm{LM}_{\text{frozen}}(\tilde{\boldsymbol{V}})$$

with $\tilde{\boldsymbol{V}}$ be the perturbed initial AA-type encoding. The language model $\mathrm{LM}_{\text{frozen}}(\cdot)$, ESM-2 [31] for instance, has been pre-trained on a massive protein sequence database (*e.g.*, **UniRef50** [61]) to understand the semantic and grammatical rules of wild-type proteins. It conceals high-dimensional AA-level long-short-range interactions that may or may not have been investigated and explained by scientists. Next, we represent proteins by kNN graphs with model-encoded node attributes, handcrafted edge attributes, and 3D positions of the corresponding AAs. This representation is embedded using a stack of L EGNN [55] layers to yield

$$\boldsymbol{W}_V^L = \mathrm{EGNN}(\mathcal{G}).$$

This process extracts the geometric and topological embedding for protein graphs with AAs represented by $\boldsymbol{w}_{v_i}$. During the pre-training phase for protein sequence recovery, the output layer $\phi(\cdot)$ provides the probability of AA types, *i.e.*,

$$\boldsymbol{Y} = \phi(\boldsymbol{W}_V^L) \in \mathbb{R}^{n \times 20}$$

for a protein comprising n AAs. The model's learnable parameters are refined by minimizing the cross-entropy of the recovered AAs with respect to the ground-truth AAs in wild-type proteins.

b) Inferencing: For a given mutant, its fitness score is derived from the joint distribution of the altered AA types on associated nodes that provides a preliminary assessment based on natural observations. We consider the AA type in the wild-type protein as a reference state and compare it with the predicted probability of AAs at the mutated site. Formally, for a mutant with mutated sites $\mathcal{T}$ ($|\mathcal{T}| \geq 1$), we define its fitness score by the corresponding *log-odds-ratio*, *i.e.*,

$$\sum_{t \in \mathcal{T}} \log p(\boldsymbol{y}_t) - \log p(\boldsymbol{v}_t),$$

where $\boldsymbol{y}_t$ and $\boldsymbol{v}_t$ denote the mutated and the wild-type AA at site t, respectively.

B. Evaluation Metrics

It is critical to evaluate the developed model's effectiveness using quantitative, computable measurements before proceeding to biochemical experiment validations. Within the scope of mutation effect prediction, each DMS assay in the dataset includes dozens to tens of thousands of mutants with varying depths of mutation sites. Considering that protein functions are sensitive to the external environment and experimental methods, the absolute values measured by individual labs are typically not directly comparable. Consequently, we evaluate the performance of pre-trained models on a diverse set of proteins and protein functions using Spearman's correlation. It is commonly applied in mutation effect prediction tasks [10], [30], [37], [43], [80] that measures the strength and direction of the monotonic relationship between two ranked sequences, *i.e.*, experimentally evaluated mutants and model-inferred mutants. This non-parametric rank measure is robust to outliers and asymmetry in mutation scores. It does not assume any specific distribution of mutation fitness scores, and it captures non-linear correlations between the two sequences. The scale of ρ ranges from -1 to 1, which indicates to what extent the predicted sequence is negatively or positively related to the ground truth. Ideally, the prediction is preferred if its ρ to the experimentally examined ground truth is close to 1.

V. Numerical Experiments

We validate the efficacy of ProtSSN on zero-shot mutation effect prediction on 188 diverse proteins scored on bioactivity and thermostability. The performance is compared with other SOTA models of varying numbers of trainable parameters. The implementations (https://github.com/tyang816/ProtSSN) are programmed with `PyTorch-Geometric` (ver 2.2.0) and `PyTorch` (ver 1.12.1) and executed on an NVIDIA® Tesla A100 GPU with $6,912$ CUDA cores and 80GB HBM2 installed on an HPC cluster.

A. Experimental Protocol

a) Training Setup: ProtSSN is trained on a non-redundant subset of **CATH v4.3.0** [44] domains, which contains $30,948$ experimental protein structures with less than 40% sequence identity. We also remove proteins that exceed $2,000$ AAs in length for efficiency. Each protein domain is transformed into a kNN graph following Section III, with node features extracted by a frozen ESM2-t33 [31] prefix model. Protein topology is inferred by a 6-layer EGNN [55] with the hidden dimension tuned from $\{512, 768, 1280\}$. Adam [26] is used for backpropagation with the learning rate set to 0.0001. To avoid training instability or CUDA out-of-memory, we limit

TABLE II
DETAILS OF BASELINE MODELS.

Model	Type		Description	Source Code
	seq	struct		
RITA [15]	✓		a generative protein language model with billion-level parameters	github.com/lightonai/RITA
PROGEN2 [41]	✓		a generative protein language model with billion-level parameters	github.com/salesforce/progen
PROTTRANS [8]	✓		TRANSFORMER-based models trained on large protein sequence corpus	github.com/agemagician/ProtTrans
TRANCEPTION [43]	✓		an autoregressive model for variant effect prediction task with retrieve machine	github.com/OATML-Markslab/Tranception
CARP [71]	✓		pretrained CNN protein sequence masked language models of various sizes	github.com/microsoft/protein-sequence-models
ESM-1B [51]	✓		masked language model-based pre-train method with various pre-training dataset and positional embedding strategies	github.com/facebookresearch/esm
ESM-1V [37]	✓			
ESM2 [31]	✓			
ESM-IF1 [17]		✓	an inverse folding method with mask language modeling and geometric vector perception [21].	
MIF-ST [73]	✓	✓	pretrained masked inverse folding models with sequence pretraining transfer	github.com/microsoft/protein-sequence-models

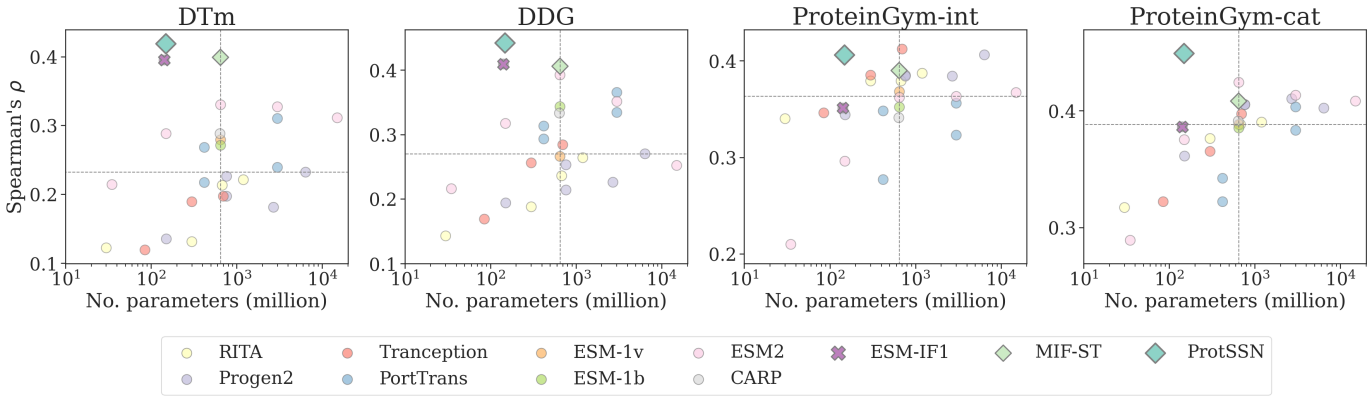

Fig. 3. Number of trainable parameters versus Spearman's ρ correlation on **DTm**, **DDG**, **ProteinGym**-int, and **ProteinGym**-cat. Dashed lines indicate the medium value on the trainable parameters and prediction performance. Dot, cross, and diamond markers represent sequence-based, structure-based, and sequence-structure models, respectively.

the maximum input to $8,192$ AA tokens per batch, constituting approximately 32 residue graphs.

b) Baseline Methods: We undertake an extensive comparison with baseline methods of self-supervised SOTA models on the fitness of mutation effects prediction. These methods, as listed in Table II, utilize protein sequences and/or structures for learning. Sequence models employ position embedding strategies such as autoregression (RITA [15], TRANCEPTION [43], and PROGEN2 [41]), masked language modeling (ESM-1B [51], ESM-1V [37], and ESM2 [31]), a combination of the both (PROTTRANS [8]), and convolutional modeling (CARP [71]). As our model acquires structural encoding, we also compare with ESM-IF1 [17] which incorporates mask language modeling objectives with GVP [21], as well as MIF-ST [73] that encodes both sequence and structure information.

c) Benchmark Datasets: We conduct a comprehensive comparison of deep learning predictions on hundreds of deep mutational scanning (DMS) assays concerning different protein types and biochemical assays. We prioritize experimentally measured properties that possess a monotonic relationship with protein fitness, including catalytic activity, binding affinity, and thermostability [43], [50]. In particular, the fitness prediction associated with interaction with substrates and catalytic process are respectively assessed on **ProteinGym**-int and **ProteinGym**-cat, and thermostability scoring are examined by **DTm** and **DDG**. In total, 188 assays with ~ 1.5 million mutants are scored.

B. Variant Effect Prediction

We commence our investigation by assessing the predictive performance of PROTSSN on four benchmark datasets using state-of-the-art (SOTA) protein learning models. In this analysis, PROTSSN learns residue graphs with and deploys $\{512, 768, 1,280\}$ hidden neurons in each of the 6 EGNN layer. As visualized in Fig. 3 and reported in Table III, PROTSSN demonstrated exceptional predictive performance with a significantly smaller number of trainable parameters

TABLE III
VARIANT EFFECT PREDICTION ON **DTM**, **DDG**, **PROTEINGYM**-INT, AND **PROTEINGYM**-CAT.

	Model	Version	# Params (million)	Spearman's ρ ↑			
				DTm	**DDG**	**ProteinGym**-int	**ProteinGym**-cat.
sequence	RITA [15]	small	30	0.122	0.143	0.340	0.317
		medium	300	0.131	0.188	0.379	0.376
		large	680	0.213	0.236	0.379	0.388
		xlarge	1,200	0.221	0.264	0.387	0.390
	PROGEN2 [41]	small	151	0.135	0.194	0.344	0.361
		medium	764	0.226	0.214	0.385	0.405
		base	764	0.197	0.253	0.384	0.405
		large	2,700	0.181	0.226	0.384	0.410
		xlarge	6,400	0.232	0.270	**0.406**	0.402
	PORTTRANS [8]	bert	420	0.268	0.313	0.348	0.322
		bert_bfd	420	0.217	0.293	0.277	0.342
		t5_xl_uniref50	3,000	0.310	0.365	0.356	0.403
		t5_xl_bfd	3,000	0.239	0.334	0.326	0.383
	TRANCEPTION [43]	small	85	0.119	0.169	0.346	0.322
		medium	300	0.189	0.256	0.385	0.365
		large	700	0.197	0.284	**0.412**	0.397
	ESM-1V [37]	-	650	0.279	0.266	0.368	0.388
	ESM-1B [51]	-	650	0.271	0.343	0.352	0.385
	ESM-2 [31]	t12	35	0.214	0.216	0.210	0.289
		t30	150	0.288	0.317	0.296	0.375
		t33	650	0.330	0.392	0.362	**0.424**
		t36	3,000	0.327	0.351	0.363	**0.413**
		t48	15,000	0.311	0.252	0.367	0.408
	CARP [71]	-	640	0.288	0.333	0.341	0.391
structure	ESM-IF1 [17]	-	142	**0.395**	**0.409**	0.351	0.386
structure+ sequence	MIF-ST [73]	-	643	**0.400**	**0.406**	0.390	0.408
	PROTSSN	k20_h512	148	**0.419**	**0.442**	**0.406**	**0.449**

† The top three are highlighted by **First**, **Second**, **Third**.

when predicting the activity and thermostability of protein mutation sequences.

Compared to protein language models (marked as circles), PROTSSN benefits from abundant structure information to more accurately capture the overall protein characteristics, resulting in enhanced thermostability, thus achieving higher correlation scores on **DTm** and **DDG**. This is attributed to the presence of stable secondary structures, such as alpha helices and beta sheets, and the compactness of the overall architectur, both of which are crucial to protein thermostability. These structural elements and packing strategies provide a framework that resists unfolding at elevated temperatures [52]. Consequently, the other two structures-involved models, *i.e.*, ESM-IF1 and MIF-ST, also exhibits higher performance on the two thermostability benchmarks.

In comparison, the analysis of protein activity enhancement are more related to AA sequences. While sequence-based analysis cannot entirely replace the role of structure-based analysis, compared to a fully-structure-based deep learning method, a protein language model is more likely to capture sufficient information from sequences by increasing the model scale, *i.e.*, the number of trainable parameters. This observation aligns with the mechanism of protein functionality. In the interaction and catalysis between proteins and substrates, not only both local structures (such as binding pockets and catalytic pockets) and the overall structure (such as conformational changes and allosteric effect) are crucial for binding and catalysis [57]. Unlike the thermostability benchmarks, the discrepancies between structure-involved models and protein language models are mitigated in **ProteinGym** by increasing the model scale. Specifically, TRANCEPTION-large and ESM2-650M outperform the two structure-based benchmarks, ESM-IF1 and MIF-ST. However, the outperformance is less pronounced on **ProteinGym**-cat, and these methods are still surpassed by PROTSSN.

Furthermore, Fig.4 provides a detailed breakdown of the rank of model predictions on individual protein assays. Across all proteins and protein-environment combinations in the four datasets, PROTSSN consistently achieves the top rank among all baseline methods. These comprehensive evaluations col-

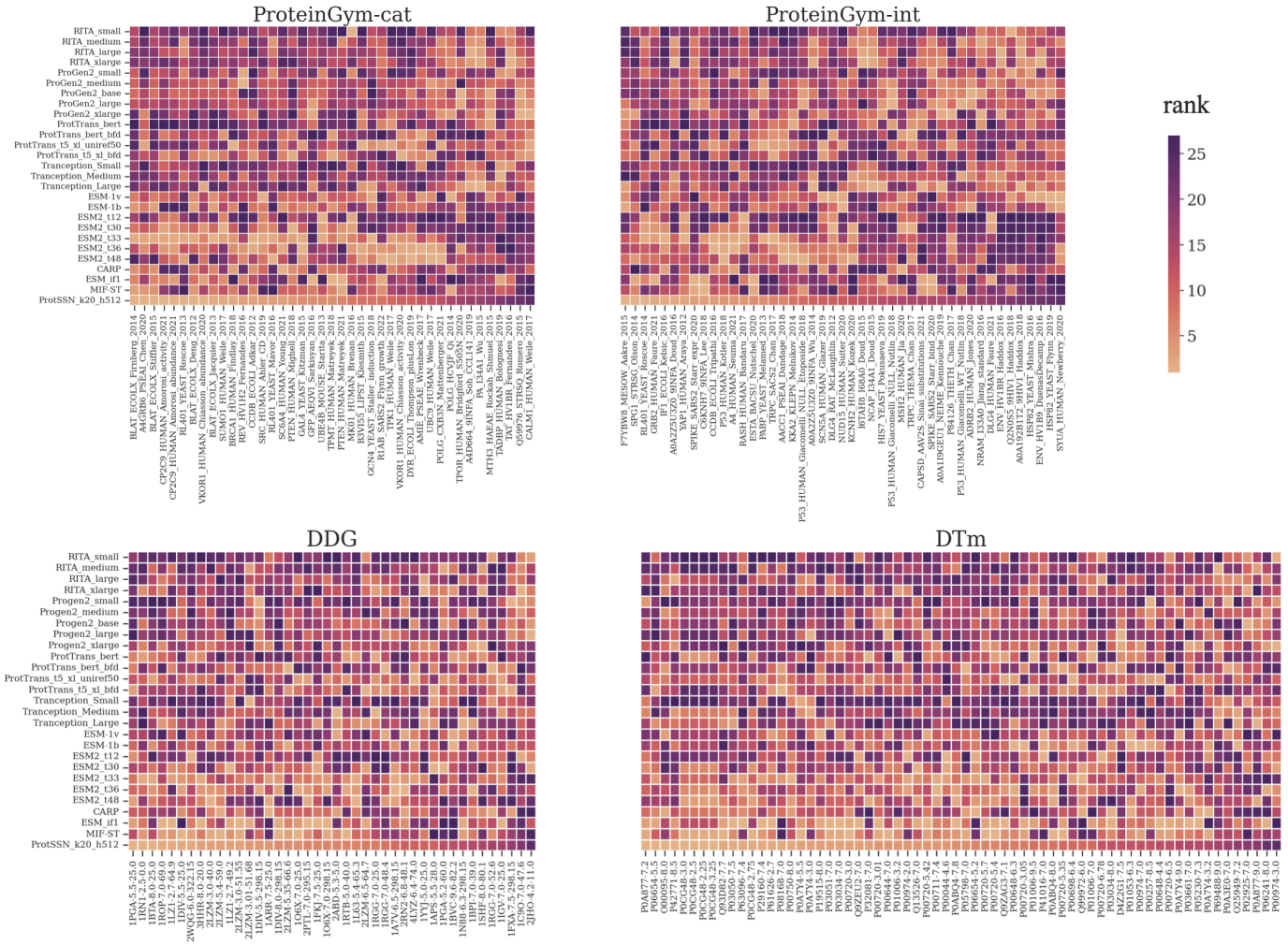

Fig. 4. Ranks of baseline methods based on the Spearman's ρ correlation on individual DMS assays in the four datasets. The best-performed model is ranked at 1 with the lightest color.

lectively underscore the significance of effectively integrating both sequence and structure features in proteins for more accurate mutation effect predictions.

C. Ablation Study

We next evaluate the effectiveness of PROTSSNwith different modular designs by its performance on **ProteinGym**, the public dataset of all 86 DMS assays. The results are summarized in Fig. 5(a)-(d), where each record is subject to a combination of key arguments under investigation. For instance, in the top orange box of Fig. 5(a), we report all ablation results that utilize 6 EGNN layers for graph convolution, regardless of the different scales of ESM-2 or the definitions of node attributes. For all modules investigated in this section, we separately discuss their influence on predicting mutation effects when modifying a single site or an arbitrary number of sites. We mark them as 'single' and 'all' on the y-axis, respectively.

a) Inclusion of Roto-Translation Equivariance: We assess the effect of incorporating rotation and translation equivariance during protein geometric and topological encoding. Three types of graph convolutions are compared, including GCN [27], GAT [65], and EGNN [55]. The first two are classic non-equivariant graph convolutional methods, while the last one, which we apply in the main algorithm, preserves roto-translation equivariance. We fix the number of EGNN layers to 6 and examine the performance of the other two methods with either 4 or 6 layers. We find that integrating equivariance when embedding protein geometry significantly improves prediction performance.

b) Sequence Encoding: We next investigate the benefits of defining data-driven node attributes for protein representation learning. We compare the performance of models trained on two sets of graph inputs: the first set defines its AA node attributes through trained ESM2 [31], while the second set uses one-hot encoded AA types for each node. A clear advantage of using hidden representations by prefix models over hardcoded attributes is evident from the results presented in Fig. 5(b).

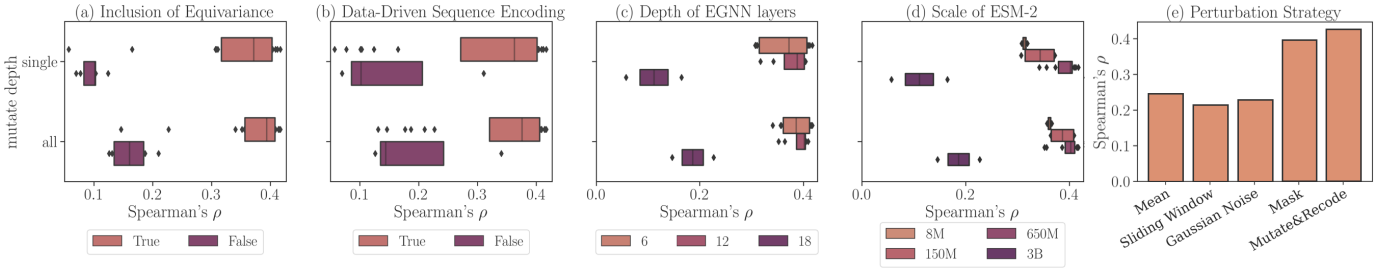

Fig. 5. Ablation Study on **ProteinGym** for different modular settings of PROTSSN. Each record represents the average Spearman's correlation on single-site and deep mutations of proteins in ProteinGym.

c) Depth of EGNN: Although graph neural networks can extract topological information from geometric inputs, it is vital to select an appropriate number of layers for the module to deliver the most expressive node representation without encountering the oversmoothing problem. We investigate a wide range of choices for the number of EGNN layers among $\{6, 12, 18\}$. As reported in Fig. 5(c), embedding graph topology with deeper networks does not lead to performance improvements. A moderate choice of 6 EGNN layers is sufficient for our learning task.

d) Scale of ESM: We also evaluate our models on different choices of language embedding dimensions to study the trade-off between the computational cost and input richness. Various scales of prefix models, including $\{8, 150, 650, 3000\}$ millions of parameters, have been applied to produce different sequential embeddings with $\{320, 640, 1280, 2560\}$ dimensions, respectively. Fig. 5(d) reveals a clear preference for ESM-2-t33, which employs 650 million parameters to achieve optimal model performance with the best stability. Notably, a higher dimension and richer semantic expression do not always yield better performance. In fact, performance degradation is observed when using the t36 version of the prefix model with 3 billion parameters.

D. AA Perturbation Strategies

PROTSSN utilizes noise sampling at each epoch on the node attributes to emulate random mutations in nature. This introduction of noise directly affects the node attribute input to the graphs. Alongside the 'mutate-then-recode' method we implemented in the main algorithm, we examined four additional strategies to perturb the input data during training. The construction of these additional strategies is detailed below, and their corresponding Spearman's ρ on **ProteinGym** with all 86 assay is reported in Fig. 5(e).

a) Mean: Suppose the encoded sequential representation of a node is predominantly determined by its AA type. In essence, the protein sequence encoder will return similar embeddings for nodes of the same AA type, albeit at different positions within a protein. With this premise, the first method replaces the node embedding for perturbed nodes with the average representations from the same AA types. For example, a random node $\boldsymbol{v}_i$ in a protein is of type L. If it is altered in the current epoch, $\tilde{\boldsymbol{v}}_i$ is designated as the average sequential embedding of all other nodes of type L.

b) Sliding Window: The presumption in the previous method neither aligns with the algorithmic design nor biological heuristics. Self-attention discerns the interaction of the central token with its neighbors across the entire document (protein) and AAs, inclusive of their types and properties, are thought to be closely aligned with its local environment. Thus, averaging embeddings of AAs from varying positions is likely to forfeit positional information of the modified AA. Consequently, the second method designs a sliding window along the protein sequence for perturbing node noise. By setting the window size to k, a corrupted node will update its representation by averaging the node representation of the nearest k neighbor AAs along the AA chain.

c) Gaussian Noise: The subsequent method regards node embeddings of AA as hidden vectors and imposes white noise on the vector values. We define the mean and variance of the noise as 0 and 0.5, respectively, making the revised node representation $\tilde{\boldsymbol{v}} = \boldsymbol{v} + \mathcal{N}(0, 0.5)$.

d) Mask: Finally, we employ the masking technique prevalent in masked language modeling and substitute the perturbed AA token with a special `<mask>` token. The prefix language model will interpret it as a novel token and employ self-attention mechanisms to assign a new representation to it. We utilize the same hyperparameter settings as that of BERT [6] and choose 15% of tokens per iteration as potentially influenced subjects. Specifically, 80% of these subjects are replaced with `<mask>`, 10% of them are replaced randomly (to other AA types), and the rest 10% stay unchanged.

E. Folding Strategy

The analysis of protein structure through topological embeddings poses challenges due to the inherent difficulty in obtaining accurate structural observations through experimental techniques such as NMR [69] and cryo-EM [7]. As a substitution, we use *in silico* folding models such as ALPHAFOLD2 [23] and ESMFOLD [31]. Although these folding methods may introduce additional errors due to the inherent uncertainty or inaccuracy in structural predictions, we investigate the robustness of our model in the face of potential inaccuracies introduced by these folding strategies.

Table IV examines the impact of different folding strategies on the performance of DMS predictions for ESM-IF1, MIF-

TABLE IV
INFLUENCE OF FOLDING STRATEGIES (ALPHAFOLD2 AND ESMFOLD) ON PREDICTION PERFORMANCE FOR STRUCTURE-INVOLVED MODELS.

	DTm			**DDG**			**ProteinGym**-int			**ProteinGym**-int		
	ALPHAFOLD2	ESMFOLD	**diff** ↓	ALPHAFOLD2	ESMFOLD	**diff** ↓	ALPHAFOLD2	ESMFOLD	**diff** ↓	ALPHAFOLD2	ESMFOLD	**diff** ↓
avg. pLDDT	90.86	83.22	-	95.19	86.03	-	82.86	65.69	-	85.80	73.45	-
ESM-IF1 [17]	0.395	0.371	0.024	0.457	0.488	-0.031	0.351	0.259	0.092	0.386	0.368	0.018
MIF-ST [73]	0.400	0.378	0.022	0.438	0.423	-0.015	0.390	0.327	0.063	0.408	0.388	0.020
k30_h1280	0.384	0.370	0.014	0.396	0.390	**0.006**	0.398	0.373	**0.025**	0.443	0.439	**0.004**
k30_h768	0.359	0.356	**0.003**	0.378	0.366	0.012	0.400	0.374	**0.026**	0.442	0.436	**0.006**
k30_h512	0.413	0.399	0.014	0.408	0.394	0.014	0.400	0.372	**0.028**	0.447	0.442	**0.005**
k20_h1280	0.415	0.391	0.024	0.429	0.410	0.019	0.399	0.365	0.034	0.446	0.441	**0.005**
k20_h768	0.415	0.403	**0.012**	0.419	0.397	0.022	0.401	0.370	0.031	0.449	0.442	0.007
k20_h512	0.419	0.395	0.024	0.441	0.432	**0.009**	0.406	0.371	0.035	0.449	0.439	0.010
k10_h1280	0.406	0.391	0.015	0.426	0.411	0.015	0.396	0.365	0.031	0.440	0.434	**0.006**
k10_h768	0.400	0.391	**0.009**	0.414	0.400	0.014	0.379	0.349	0.030	0.431	0.421	0.010
k10_h512	0.383	0.364	0.019	0.424	0.414	**0.010**	0.389	0.364	**0.025**	0.440	0.432	0.008

TABLE V
VARIANT EFFECT PREDICTION ON **PROTEINGYM** WITH BOTH MSA AND NON-MSA METHODS. RESULTS ARE RETRIEVED FROM [43].

	Model	**Type**	**# Params** (million)	ρ **(by mutation depth)** ↑			ρ **(by MSA depth)** ↑			ρ **(by taxon)** ↑			
				Single	Double	All	Low	Medium	High	Prokaryote	Human	Eukaryote	Virus
MSA	SITEINDEP [16]	single	-	0.378	0.322	0.378	**0.431**	0.375	0.342	0.343	0.375	0.401	0.406
MSA	EVMUTATION [16]	single	-	0.423	0.401	0.423	0.406	0.403	0.484	0.499	0.396	0.429	0.381
MSA	WAVENET [59]	single	-	0.399	0.344	0.400	0.286	0.404	0.489	0.492	0.373	0.442	0.321
MSA	DEEPSEQUENCE [50]	single	-	0.411	0.358	0.416	0.386	0.391	0.505	0.497	0.396	0.461	0.332
MSA	DEEPSEQUENCE [50]	ensemble	-	0.433	0.394	0.435	0.386	0.411	**0.534**	0.522	0.405	0.480	0.361
MSA	EVE [10]	single	-	**0.451**	**0.406**	**0.452**	0.417	**0.434**	0.525	0.518	0.411	0.469	**0.436**
MSA	EVE [10]	ensemble	-	**0.459**	**0.409**	**0.459**	**0.424**	**0.441**	**0.532**	**0.526**	0.419	**0.481**	**0.437**
MSA	MSA-TRANSFOMER [48]	single	100	0.405	0.358	0.426	0.372	0.415	0.500	0.506	0.387	0.468	0.379
MSA	MSA-TRANSFOMER [48]	ensemble	500	0.440	0.374	0.440	0.387	0.428	0.513	0.517	0.398	0.467	0.406
MSA	TRANCEPTION-R [43]	single	700	**0.450**	**0.427**	**0.453**	**0.442**	**0.438**	0.500	0.495	0.424	**0.485**	**0.434**
Non-MSA	RITA [15]	ensemble	2,210	0.393	0.236	0.399	0.350	0.414	0.407	0.391	0.391	0.405	0.417
Non-MSA	ESM-1V [37]	ensemble	3,250	0.416	0.309	0.417	0.390	0.378	**0.536**	**0.521**	**0.439**	0.423	0.268
Non-MSA	PROGEN2 [41]	ensemble	10,779	0.421	0.312	0.423	0.384	0.421	0.467	0.497	0.412	0.459	0.373
Non-MSA	PROTTRANS [8]	ensemble	6,840	0.417	0.360	0.413	0.372	0.395	0.492	0.498	0.419	0.400	0.322
Non-MSA	ESM2 [31]	ensemble	18,843	0.415	0.316	0.413	0.391	0.381	0.509	0.508	**0.456**	0.461	0.213
Non-MSA	PROTSSN	ensemble	1,476	0.433	0.381	0.433	0.406	0.402	**0.532**	**0.530**	**0.436**	**0.491**	0.290

ST, and PROTSSN. Although the performance based on ESMFOLD-derived protein structures generally lags behind structures folded by ALPHAFOLD2, the minimal difference between the two strategies across all four benchmarks validates the robustness of our PROTSSN in response to variations in input protein structures. Notably, the reported performance difference for both ESM-IF1 and MIF-ST is $3-5$ times higher than that of PROTSSN. Note that the optimal results for **DDG** is inconsistent with the scores reported in Table III due to the utilization of crystal structures for the proteins in the main results. Another noteworthy observation pertains to the performance of ESM-IF1 and MIF-ST on **DDG**. In this case, predictions based on ESMFOLD surpass those based on ALPHAFOLD2, even outperforming predictions derived from crystal structures. However, the superior performance of these methods with ESMFOLD hinges on inaccurate protein structures, rendering them less reliable.

In summary, PROTSSN emerges as the most robust method, maintaining consistent performance across different folding strategies, and underscores the importance of our model's resilience to variations in input protein structures.

TABLE VI
VARIANT EFFECT PREDICTION ON **DTM** AND **DDG** FOR NON-MSA METHODS WITH MODEL ENSEMBLE.

Model	**# Params** (million)	**Spearman's** ρ ↑	
		DTm	**DDG**
RITA [15]	2,210	0.195	0.255
ESM-1V [37]	3,250	0.300	0.310
TRANCEPTION [43]	1,085	0.202	0.277
PROGEN2 [41]	10,779	0.293	0.282
PROTRANS [8]	6,840	0.323	0.389
ESM2 [31]	18,843	0.346	0.383
PROTSSN	1,476	**0.425**	**0.440**

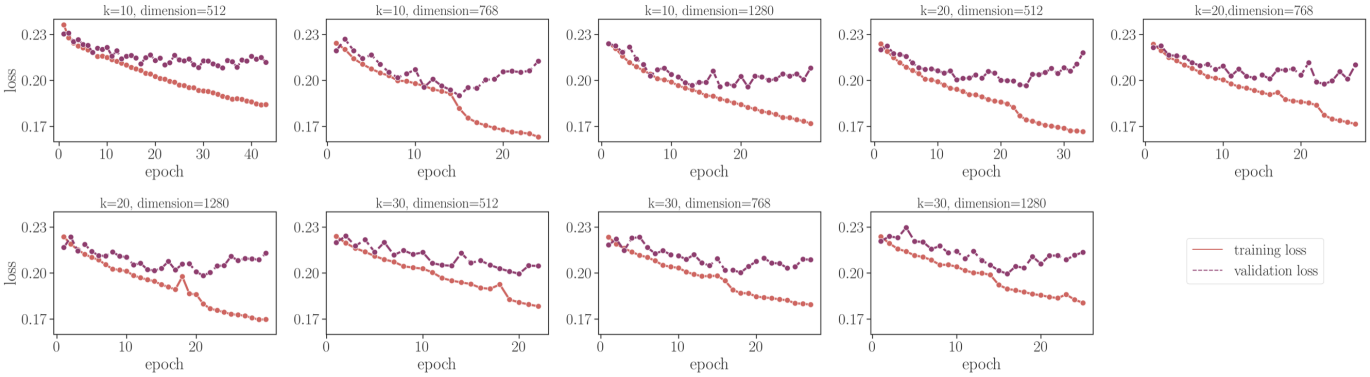

Fig. 6. Training and validation losses in 9 combinations of $k = 10, 20, 30$ nearest neighbor AAs and the hidden dimension of $512, 768, 1280$.

F. MSA-based Few-Shot Prediction

In addition to the previously compared zero-shot protein encoders, we extend our evaluation to include few-shot learning methods that leverage MSA information of proteins [37]. The comprehensive results are detailed in Table V, which are retrieved from **ProteinGym**'s repository[2]. For PROTSSN, we employ an ensemble of nine models with varying k values ($k \in 10, 20, 30$) and hidden neurons in $512, 768, 1280$, as previously discussed. The MSA depth reflects the number of MSA sequences that can be found for the protein, which influences the quality of MSA methods.

PROTSSN demonstrates superior performance among non-MSA zero-shot methods in predicting both single-site and multi-site mutants, underscoring its pivotal role in guiding protein engineering towards deep mutation. Furthermore, PROTSSN outperforms MSA-based methods across taxa, except for viruses, where precise identification of conserved sites is crucial for positive mutations, especially in spike proteins [24], [38]. In essence, PROTSSN provides an efficient and effective solution for variant effects prediction when compared to both non-MSA and MSA-based methods. Note that although MSA-based methods generally consume fewer trainable parameters than non-MSA methods, they incur significantly higher costs in terms of search time and memory usage. Moreover, the inference performance of MSA-based methods relies heavily on the quality of the input MSA, where the additionally involved variance makes impacts the stability of the model performance.

G. Learning Curve

We compared the learning curve of PROTSSN with 9 combinations of the number of k nearest neighbors used in constructing protein graphs from $\{10, 20, 30\}$, and hidden dimension from $\{512, 768, 1280\}$, with respect to both training and validation. In all cases, we have used early stopping to ensure the convergence of training loss and sufficient learning. However, the validation loss varies depending on the scenario. Notably, for each value of k, there exists a corresponding hidden dimension scenario that converges consistently with training. There is typically a gap between validation and training loss, indicating that there is room for improving the generalizability of the model. One possible approach to address this issue is to increase the size of both the encoder and training datasets.

[2]https://github.com/OATML-Markslab/ProteinGym

VI. Related Work

a) Deep Mutational Scanning Benchmarks: With the continuous evolution of deep mutation scanning techniques and other high-throughput technologies, benchmark datasets have been curated to document fitness changes in mutants across diverse proteins, encompassing a spectrum of AA sequence alterations, including substitutions, insertions, and deletions [13], [43], [50]. These datasets span mutations with varying degrees of combinatorial and random modifications. Complementing these extensive datasets, additional resources such as **SoluProtMutDB** [64], **SKEMPI** [39], and **ProThermDB** [42] focus on evaluating changes in functional properties associated with single AA variants (SAVs). Beyond datasets proposed in bioinformatic, there are also benchmark dataset partitioning methods introduced in the realm of computer science. For instance, **FLIP** [5] provides different data partitioning methods for three types of protein fitness data, aimed at evaluating supervised model performance across various scenarios. The **TAPE** [46] benchmark comprises five tasks, three of which are focused on structure prediction while the remaining two address fitness prediction. **PEER** [70] encompasses seventeen biology-related tasks, covering multiple aspects of protein understanding such as function prediction, localization prediction, and solubility prediction. **ProteinGLUE** [4] includes seven downstream tasks related to protein domains, primarily designed for self-supervised protein representation learning.

b) Protein Primary Structure Embedding: Self-supervised protein language models play the predominant role in the training of large quantities of protein sequences for protein representation learning. These methodologies draw parallels between protein sequences and natural language, encoding amino acid tokens using the TRANSFORMER model [63] to extract pairwise relationships among tokensemble These methods typically pre-train on extensive protein

sequence databases to autoregressively recover protein sequences [34], [43]. Alternatively, masked language modeling objectives develop attention patterns that correspond to the residue-residue contact map of the protein [31], [37], [47], [51], [66]. Other methods start from a multiple sequence alignment, summarizing the evolutionary patterns in target proteins [11], [48], [50]. Both aligned and non-aligned methods result in a strong capacity for discovering the hidden protein space, but this often comes at the expense of excessive training input or the use of substantial learning resources. This trade-off underlines the need for efficient and cost-effective approaches in self-supervised protein modeling.

c) Protein Tertiary Structure Embedding: Protein structures directly dictate protein functions and are essential to *de novo* protein design, which is a critical challenge in bioengineering. The remarkable success of accurate protein folding by ALPHAFOLD2 [23] and the subsequent enrichment of the structure-aware protein repository have motivated a series of research initiatives focused on learning protein geometry. Recent efforts have been made to encode geometric information of proteins [12], [21], [76] for topology-sensitive tasks such as molecule binding [20], [29], [40], protein interface analysis [35], [49], and protein properties prediction [75].

d) Variant Effect Prediction: Variant effect predictions quantify the fitness of a mutant protein in comparison to its wild-type counterpart. For well-studied proteins, it is feasible to fit a supervised model on the hidden representation of amino acid local environments to infer fitness scores [10], [33], [36], [79]. However, in many cases, labeled data is either scarce or inaccessible. To overcome this, zero-shot methods have been developed to infer the fitness of a mutation from the evolutionary landscape of the original protein using sequence alignments [16], [50], [59]. Alternatively, hybrid models [43], [48] utilize retrieval or attention mechanisms to extract patterns from Multiple Sequence Alignments (MSAs) in conjunction with protein language or structure models.

VII. CONCLUSION AND DISCUSSION

The development of dependable computational methodologies for protein engineering is a crucial facet of *in silico* protein design. Accurately assessing the fitness of protein mutants not only supports cost-effective experimental validations but also guides the modification of proteins to enhance existing or introduce new functions. Most recent deep learning solutions employ a common strategy that involves establishing a hidden protein representation and masking potential mutation sites for amino acid generation. Previous research has primarily focused on extracting protein representations from either their sequential or structural modalities, with many treating the prediction of mutational effects merely as a secondary task following inverse folding or *de novo* protein design. These approaches often overlook the importance of comprehending various levels of protein structures that are critical for determining protein function. Furthermore, they seldom implement model designs tailored specifically for mutation tasks. In this work, we introduce PROTSSN, a denoising framework that effectively cascades protein primary and tertiary structure embedding for the specific task of predicting mutational effects. This framework first employs a prefix protein language model to decode sequence representation and identify residue-wise intercommunications. This is subsequently enhanced by a roto-translation equivariant graph neural network, which encodes geometric representations for amino acid microenvironments. We have extensively validated the efficacy of PROTSSNacross various protein function assays and taxa, including two thermostability datasets that were prepared by ourselves. Our approach consistently demonstrates substantial promise for protein engineering applications, particularly in facilitating the design of mutation sequences with improved thermal stability.

a) Broader Impact: The intersection of deep learning and structural biology, as showcased in this study, has the potential to transform our approach to protein engineering challenges, paving the way for sustainable and efficient solutions. Algorithms, such as PROTSSN, are primarily designed to enhance enzymes to support initiatives in drug discovery, biofuel production, and other relevant industries. However, given the pervasive presence of proteins across numerous scenarios and organisms, it is feasible that these methods could be employed to modify dangerous species, such as harmful viruses. Therefore, it is crucial to regulate the use of these deep learning methods, akin to the oversight required for any other powerful tools. Interestingly, our model demonstrates suboptimal performance when applied to such categories of proteins (refer to Table V), suggesting an inherent limitation in its potential misuse.

b) Limitation: The consumption of training resources for AI-driven protein engineering techniques has surged considerably nowadays. For instance, ESM-IF1, which is another geometric model that utilizes structural information of proteins, necessitates months of processing time and hundreds of machines to integrate sequence and topological data. Owing to these computational cost constraints, our approach does not train on such an extensive corpus from scratch. Instead, we harness publicly-available language models to extract hidden representations for amino acids. Nevertheless, training and inference in such an integrated model require geometric information from proteins in addition to sequential data. The current data repositories are rich with protein structures experimentally solved by biologists and supplemented by high-quality protein geometries from contemporary techniques such as ALPHAFOLD2, and they are adequate for training our model. However, it's plausible that a revised protein could have an excessively long sequence that lacks a crystallized structure and cannot be folded by computational tools. An example of such a limitation is evident in our experiment, where a protein with an extended length was removed from the **ProteinGym** benchmark dataset.

REFERENCES

[1] Francesco A Aprile, Pietro Sormanni, Marina Podpolny, Shianne Chhangur, Lisa-Maria Needham, Francesco S Ruggeri, Michele Perni, Ryan Limbocker, Gabriella T Heller, Tomas Sneideris, et al. Rational design of a conformation-specific antibody for the quantification of

aβ oligomers. *Proceedings of the National Academy of Sciences*, 117(24):13509–13518, 2020.

[2] Uwe T Bornscheuer, GW Huisman, RJ Kazlauskas, S Lutz, JC Moore, and K Robins. Engineering the third wave of biocatalysis. *Nature*, 485(7397):185–194, 2012.

[3] Nadav Brandes, Dan Ofer, Yam Peleg, Nadav Rappoport, and Michal Linial. ProteinBERT: A universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8):2102–2110, 2022.

[4] Henriette Capel, Robin Weiler, Maurits Dijkstra, Reinier Vleugels, Peter Bloem, and K Anton Feenstra. Proteinglue multi-task benchmark suite for self-supervised protein modeling. *Scientific Reports*, 12(1):16047, 2022.

[5] Christian Dallago, Jody Mou, Kadina E Johnston, Bruce J Wittmann, Nicholas Bhattacharya, Samuel Goldman, Ali Madani, and Kevin K Yang. Flip: Benchmark tasks in fitness landscape inference for proteins. *bioRxiv*, pages 2021–11, 2021.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805*, 2018.

[7] Dominika Elmlund, Sarah N Le, and Hans Elmlund. High-resolution cryo-em: the nuts and bolts. *Current opinion in structural biology*, 46:1–6, 2017.

[8] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Wang Yu, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. ProtTrans: Towards cracking the language of lifes code through self-supervised deep learning and high performance computing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[9] Hans Frauenfelder and Ben McMahon. Dynamics and function of proteins: the search for general concepts. *Proceedings of the National Academy of Sciences*, 95(9):4795–4797, 1998.

[10] Jonathan Frazer, Pascal Notin, Mafalda Dias, Aidan Gomez, Joseph K Min, Kelly Brock, Yarin Gal, and Debora S Marks. Disease variant prediction with deep generative models of evolutionary data. *Nature*, 599(7883):91–95, 2021.

[11] Jonathan Frazer, Pascal Notin, Mafalda Dias, Aidan Gomez, Joseph K Min, Kelly Brock, Yarin Gal, and Debora S Marks. Disease variant prediction with deep generative models of evolutionary data. *Nature*, 599(7883):91–95, 2021.

[12] Vladimir Gligorijević, P Douglas Renfrew, Tomasz Kosciolek, Julia Koehler Leman, Daniel Berenberg, Tommi Vatanen, Chris Chandler, Bryn C Taylor, Ian M Fisk, Hera Vlamakis, et al. Structure-based protein function prediction using graph convolutional networks. *Nature Communications*, 12(1):3168, 2021.

[13] Vanessa E Gray, Ronald J Hause, Jens Luebeck, Jay Shendure, and Douglas M Fowler. Quantitative missense variant effect prediction using large-scale mutagenesis data. *Cell systems*, 6(1):116–124, 2018.

[14] Jorrit W Hegge, Daan C Swarts, and John van der Oost. Prokaryotic argonaute proteins: novel genome-editing tools? *Nature Reviews Microbiology*, 16(1):5–11, 2018.

[15] Daniel Hesslow, Niccoló Zanichelli, Pascal Notin, Iacopo Poli, and Debora Marks. Rita: a study on scaling up generative protein sequence models. *arXiv:2205.05789*, 2022.

[16] Thomas A Hopf, John B Ingraham, Frank J Poelwijk, Charlotta PI Schärfe, Michael Springer, Chris Sander, and Debora S Marks. Mutation effects predicted from sequence co-variation. *Nature Biotechnology*, 35(2):128–135, 2017.

[17] Chloe Hsu, Robert Verkuil, Jason Liu, Zeming Lin, Brian Hie, Tom Sercu, Adam Lerer, and Alexander Rives. Learning inverse folding from millions of predicted structures. In *International Conference on Machine Learning*, pages 8946–8970. PMLR, 2022.

[18] Abdallah R Ismail, Hamdy Kashtoh, and Kwang-Hyun Baek. Temperature-resistant and solvent-tolerant lipases as industrial biocatalysts: Biotechnological approaches and applications. *International Journal of Biological Macromolecules*, 187:127–142, 2021.

[19] Fan Jiang, Jiahao Bian, Hao Liu, Song Li, Xue Bai, Lirong Zheng, Sha Jin, Zhuo Liu, Guang-Yu Yang, and Liang Hong. Creatinase: Using increased entropy to improve the activity and thermostability. *The Journal of Physical Chemistry B*, 127(12):2671–2682, 2023.

[20] Wengong Jin, Jeremy Wohlwend, Regina Barzilay, and Tommi S Jaakkola. Iterative refinement graph neural network for antibody sequence-structure co-design. In *International Conference on Learning Representations*, 2021.

[21] Bowen Jing, Stephan Eismann, Patricia Suriana, Raphael JL Townshend, and Ron Dror. Learning from protein structure with geometric vector perceptrons. *arXiv:2009.01411*, 2020.

[22] Derek Jones, Hyojin Kim, Xiaohua Zhang, Adam Zemla, Garrett Stevenson, WF Drew Bennett, Daniel Kirshner, Sergio E Wong, Felice C Lightstone, and Jonathan E Allen. Improved protein–ligand binding affinity prediction with structure-based deep fusion inference. *Journal of Chemical Information and Modeling*, 61(4):1583–1592, 2021.

[23] 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.

[24] Kazutaka Katoh and Daron M Standley. Emerging sars-cov-2 variants follow a historical pattern recorded in outgroups infecting non-human hosts. *Communications Biology*, 4(1):1134, 2021.

[25] Olga Khersonsky, Rosalie Lipsh, Ziv Avizemer, Yacov Ashani, Moshe Goldsmith, Haim Leader, Orly Dym, Shelly Rogotner, Devin L Trudeau, Jaime Prilusky, et al. Automated design of efficient and functionally diverse enzyme repertoires. *Molecular Cell*, 72(1):178–186, 2018.

[26] Diederik P Kingma and Jimmy Ba. ADAM: A method for stochastic optimization. In *Proceedings of International Conference on Learning Representation (International Conference on Learning Representations)*, 2015.

[27] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

[28] Julia Koehler Leman, Pawel Szczerbiak, P Douglas Renfrew, Vladimir Gligorijevic, Daniel Berenberg, Tommi Vatanen, Bryn C Taylor, Chris Chandler, Stefan Janssen, Andras Pataki, et al. Sequence-structure-function relationships in the microbial protein universe. *Nature communications*, 14(1):2351, 2023.

[29] Xiangzhe Kong, Wenbing Huang, and Yang Liu. Conditional antibody design as 3D equivariant graph translation. In *The Eleventh International Conference on Learning Representations*, 2023.

[30] Mingchen Li, Liqi Kang, Yi Xiong, Yu Guang Wang, Guisheng Fan, Pan Tan, and Liang Hong. Sesnet: sequence-structure feature-integrated deep learning method for data-efficient protein engineering. *Journal of Cheminformatics*, 15(1):1–13, 2023.

[31] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

[32] Qian Liu, Guanhua Xun, and Yan Feng. The state-of-the-art strategies of protein engineering for enzyme stabilization. *Biotechnology advances*, 37(4):530–537, 2019.

[33] Hongyuan Lu, Daniel J Diaz, Natalie J Czarnecki, Congzhi Zhu, Wantae Kim, Raghav Shroff, Daniel J Acosta, Bradley R Alexander, Hannah O Cole, Yan Zhang, et al. Machine learning-aided engineering of hydrolases for PET depolymerization. *Nature*, 604(7907):662–667, 2022.

[34] Ali Madani, Ben Krause, Eric R Greene, Subu Subramanian, Benjamin P Mohr, James M Holton, Jose Luis Olmos Jr, Caiming Xiong, Zachary Z Sun, Richard Socher, et al. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, pages 1–8, 2023.

[35] Sazan Mahbub and Md Shamsuzzoha Bayzid. EGRET: edge aggregated graph attention networks and transfer learning improve protein–protein interaction site prediction. *Briefings in Bioinformatics*, 23(2):bbab578, 2022.

[36] Céline Marquet, Michael Heinzinger, Tobias Olenyi, Christian Dallago, Kyra Erckert, Michael Bernhofer, Dmitrii Nechaev, and Burkhard Rost. Embeddings from protein language models predict conservation and variant effects. *Human Genetics*, 141(10):1629–1647, 2022.

[37] Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alex Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. In *Advances in Neural Information Processing Systems*, volume 34, pages 29287–29303, 2021.

[38] Ivan Mercurio, Vincenzo Tragni, Francesco Busto, Anna De Grassi, and Ciro Leonardo Pierri. Protein structure analysis of the interactions between sars-cov-2 spike protein and the human ace2 receptor: from conformational changes to novel neutralizing antibodies. *Cellular and Molecular Life Sciences*, 78:1501–1522, 2021.

[39] Iain H Moal and Juan Fernández-Recio. Skempi: A structural kinetic and energetic database of mutant protein interactions and its use in empirical models. *Bioinformatics*, 28(20):2600–2607, 2012.

[40] Yoochan Myung, Douglas EV Pires, and David B Ascher. CSM-AB: graph-based antibody–antigen binding affinity prediction and docking scoring function. *Bioinformatics*, 38(4):1141–1143, 2022.

[41] Erik Nijkamp, Jeffrey Ruffolo, Eli N Weinstein, Nikhil Naik, and Ali Madani. ProGen2: exploring the boundaries of protein language models. *arXiv:2206.13517*, 2022.

[42] Rahul Nikam, A Kulandaisamy, K Harini, Divya Sharma, and M Michael Gromiha. ProThermDB: thermodynamic database for proteins and mutants revisited after 15 years. *Nucleic Acids Research*, 49(D1):D420–D424, 2021.

[43] Pascal Notin, Mafalda Dias, Jonathan Frazer, Javier Marchena Hurtado, Aidan N Gomez, Debora Marks, and Yarin Gal. Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval. In *International Conference on Machine Learning*, pages 16990–17017. PMLR, 2022.

[44] CA Orengo, AD Michie, S Jones, DT Jones, MB Swindells, and JM Thornton. CATH – a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–1109, 1997.

[45] Adrian Pickar-Oliver and Charles A Gersbach. The next generation of crispr–cas technologies and applications. *Nature reviews Molecular cell biology*, 20(8):490–507, 2019.

[46] Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Peter Chen, John Canny, Pieter Abbeel, and Yun Song. Evaluating protein transfer learning with tape. *Advances in neural information processing systems*, 32, 2019.

[47] Roshan Rao, Joshua Meier, Tom Sercu, Sergey Ovchinnikov, and Alexander Rives. Transformer protein language models are unsupervised structure learners. In *International Conference on Learning Representations*, 2021.

[48] Roshan M Rao, Jason Liu, Robert Verkuil, Joshua Meier, John Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. MSA transformer. In *International Conference on Machine Learning*, pages 8844–8856. PMLR, 2021.

[49] Manon Réau, Nicolas Renaud, Li C Xue, and Alexandre MJJ Bonvin. DeepRank-GNN: a graph neural network framework to learn patterns in protein–protein interfaces. *Bioinformatics*, 39(1):btac759, 2023.

[50] Adam J Riesselman, John B Ingraham, and Debora S Marks. Deep generative models of genetic variation capture the effects of mutations. *Nature Methods*, 15(10):816–822, 2018.

[51] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021.

[52] Marc Robinson-Rechavi, Andreu Alibés, and Adam Godzik. Contribution of electrostatic interactions, compactness and quaternary structure to protein thermostability: lessons from structural genomics of thermotoga maritima. *Journal of molecular biology*, 356(2):547–557, 2006.

[53] Benjamin P Roscoe, Kelly M Thayer, Konstantin B Zeldovich, David Fushman, and Daniel NA Bolon. Analyses of the effects of all ubiquitin point mutants on yeast growth rate. *Journal of Molecular Biology*, 425(8):1363–1377, 2013.

[54] Karen S Sarkisyan, Dmitry A Bolotin, Margarita V Meer, Dinara R Usmanova, Alexander S Mishin, George V Sharonov, Dmitry N Ivankov, Nina G Bozhanova, Mikhail S Baranov, Onuralp Soylemez, et al. Local fitness landscape of the green fluorescent protein. *Nature*, 533(7603):397–401, 2016.

[55] Vıctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks. In *International Conference on Machine Learning*, pages 9323–9332, 2021.

[56] Andrew M Scott, Jedd D Wolchok, and Lloyd J Old. Antibody therapy of cancer. *Nature reviews cancer*, 12(4):278–287, 2012.

[57] Gang Sheng, Hongtu Zhao, Jiuyu Wang, Yu Rao, Wenwen Tian, Daan C Swarts, John van der Oost, Dinshaw J Patel, and Yanli Wang. Structure-based cleavage mechanism of thermus thermophilus argonaute dna guide strand-mediated dna target cleavage. *Proceedings of the National Academy of Sciences*, 111(2):652–657, 2014.

[58] Jung-Eun Shin, Adam J Riesselman, Aaron W Kollasch, Conor McMahon, Elana Simon, Chris Sander, Aashish Manglik, Andrew C Kruse, and Debora S Marks. Protein design and variant prediction using autoregressive generative models. *Nature Communications*, 12(1):2403, 2021.

[59] Jung-Eun Shin, Adam J Riesselman, Aaron W Kollasch, Conor McMahon, Elana Simon, Chris Sander, Aashish Manglik, Andrew C Kruse, and Debora S Marks. Protein design and variant prediction using autoregressive generative models. *Nature Communications*, 12(1):2403, 2021.

[60] Raghav Shroff, Austin W Cole, Daniel J Diaz, Barrett R Morrow, Isaac Donnell, Ankur Annapareddy, Jimmy Gollihar, Andrew D Ellington, and Ross Thyer. Discovery of novel gain-of-function mutations guided by structure-based deep learning. *ACS Synthetic Biology*, 9(11):2927–2935, 2020.

[61] Baris E Suzek, Yuqi Wang, Hongzhan Huang, Peter B McGarvey, Cathy H Wu, and UniProt Consortium. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6):926–932, 2015.

[62] Wen Torng and Russ B Altman. High precision protein functional site detection using 3D convolutional neural networks. *Bioinformatics*, 35(9):1503–1512, 2019.

[63] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

[64] Jan Veleckỳ, Marie Hamsikova, Jan Stourac, Milos Musil, Jiri Damborsky, David Bednar, and Stanislav Mazurenko. Soluprotmutdb: A manually curated database of protein solubility changes upon mutations. *Computational and Structural Biotechnology Journal*, 20:6339–6347, 2022.

[65] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

[66] Jesse Vig, Ali Madani, Lav R Varshney, Caiming Xiong, Nazneen Rajani, et al. BERTology meets biology: Interpreting attention in protein language models. In *International Conference on Learning Representations*, 2021.

[67] Yajie Wang, Pu Xue, Mingfeng Cao, Tianhao Yu, Stephan T Lane, and Huimin Zhao. Directed evolution: methodologies and applications. *Chemical reviews*, 121(20):12384–12444, 2021.

[68] John M Woodley. Protein engineering of enzymes for process applications. *Current opinion in chemical biology*, 17(2):310–316, 2013.

[69] Kurt Wüthrich. Protein structure determination in solution by nmr spectroscopy. *Journal of Biological Chemistry*, 265(36):22059–22062, 1990.

[70] Minghao Xu, Zuobai Zhang, Jiarui Lu, Zhaocheng Zhu, Yangtian Zhang, Ma Chang, Runcheng Liu, and Jian Tang. Peer: A comprehensive and multi-task benchmark for protein sequence understanding. *Advances in Neural Information Processing Systems*, 35:35156–35173, 2022.

[71] Kevin K Yang, Nicolo Fusi, and Alex X Lu. Convolutions are competitive with transformers for protein sequence pretraining. *bioRxiv*, pages 2022–05, 2022.

[72] Kevin K Yang, Zachary Wu, and Frances H Arnold. Machine-learning-guided directed evolution for protein engineering. *Nature methods*, 16(8):687–694, 2019.

[73] Kevin K Yang, Niccolò Zanichelli, and Hugh Yeh. Masked inverse folding with sequence transfer for protein representation learning. *bioRxiv*, pages 2022–05, 2022.

[74] Kai Yi, Bingxin Zhou, Yiqing Shen, Pietro Liò, and Yu Guang Wang. Graph denoising diffusion for inverse protein folding. In *NeurIPS*, 2023.

[75] Ningyu Zhang, Zhen Bi, Xiaozhuan Liang, Siyuan Cheng, Haosen Hong, Shumin Deng, Jiazhang Lian, Qiang Zhang, and Huajun Chen. Ontoprotein: Protein pretraining with gene ontology embedding. *arXiv:2201.11147*, 2022.

[76] Zuobai Zhang, Minghao Xu, Arian Jamasb, Vijil Chenthamarakshan, Aurelie Lozano, Payel Das, and Jian Tang. Protein representation learning by geometric structure pretraining. *arXiv:2203.06125*, 2022.

[77] Weishu Zhao, Bozitao Zhong, Lirong Zheng, Pan Tan, Yinzhao Wang, Hao Leng, Nicolas de Souza, Zhuo Liu, Liang Hong, and Xiang Xiao. Proteome-wide 3d structure prediction provides insights into the ancestral metabolism of ancient archaea and bacteria. *Nature Communications*, 13(1):7861, 2022.

[78] Lirong Zheng, Hui Lu, Bing Zan, Song Li, Hao Liu, Zhuo Liu, Juan Huang, Yongjia Liu, Fan Jiang, Qian Liu, et al. Loosely-packed dynamical structures with partially-melted surface being the key for

thermophilic argonaute proteins achieving high dna-cleavage activity. *Nucleic Acids Research*, 50(13):7529–7544, 2022.

[79] Bingxin Zhou, Outongyi Lv, Kai Yi, Xinye Xiong, Pan Tan, Liang Hong, and Yu Guang Wang. Accurate and definite mutational effect prediction with lightweight equivariant graph neural networks. *arXiv:2304.08299*, 2023.

[80] Bingxin Zhou, Lirong Zheng, Banghao Wu, Yang Tan, Outongyi Lv, Kai Yi, Guisheng Fan, and Liang Hong. Protein engineering with lightweight graph denoising neural networks. *bioRxiv*, 2023.

[81] Bingxin Zhou, Lirong Zheng, Banghao Wu, Kai Yi, Bozitao Zhong, Pietro Lio, and Liang Hong. Conditional protein denoising diffusion generates programmable endonucleases. *bioRxiv*, pages 2023–08, 2023.

[82] Yunzhuo Zhou, Qisheng Pan, Douglas EV Pires, Carlos HM Rodrigues, and David B Ascher. DDMut: predicting effects of mutations on protein stability using deep learning. *Nucleic Acids Research*, page gkad472, 2023.