Just Add Structure: Protein Language Models Combined with Structural Equivariance Excel at Protein Tasks

Qurat-ul-ain [1] **Carlos Outeiral** [1] **Matteo Cagiada** [1 2] **Yee Whye Teh** [1] **Charlotte M. Deane** [1]

Abstract

Accurate *in silico* prediction of protein properties, functional fitness, and mutational effects remains a central challenge in protein engineering and therapeutic design. While Protein Language Models (PLMs) successfully capture rich evolutionary and functional constraints from sequence data, they only indirectly encode the spatial and geometric information that fundamentally governs protein function. Consequently, state-of-the-art approaches typically rely on extensive fine-tuning, ensembling, or the incorporation of handcrafted structural features to achieve competitive accuracy, making them computationally expensive and difficult to scale. In this work, we demonstrate that explicit geometric modeling can substitute for, and in most cases outperform, large-scale PLM fine-tuning, with much higher parameter efficiency. Our approach, ProtEGNN, pairs PLM residue representations with a lightweight $E(3)$-Equivariant Graph Neural Network, competing with or achieving state-of-the-art performance across seven different benchmarks in protein property, mutational effect and function prediction, while needing 100–1000$\times$ fewer parameters than competing approaches. Notably, even when paired with the smallest readily available PLM, ESM2-T6 (8M parameters), ProtEGNN matches fine-tuned, sequence-only methods on mutational effect prediction, despite training orders of magnitude fewer parameters. Together, these results highlight geometric inductive bias as a powerful and scalable alternative to task-specific fine-tuning of large PLMs for protein modeling.

[1]Department of Statistics, University of Oxford [2]Department of Biology, University of Copenhagen. Correspondence to: Charlotte M. Deane <deane@stats.ox.ac.uk>.

Accepted at the 2026 Workshop on Generative and Agentic AI for Biology (ICML 2026)

1. Introduction

Understanding how amino acid sequence gives rise to protein function and biophysical properties lies at the heart of protein engineering: the science of tailoring proteins for improved stability (Tsuboyama et al., 2023), solubility (Thumuluri et al., 2021), catalytic performance (Yu et al., 2023) and other therapeutic traits. Recent advances in protein design (Watson et al., 2023; Butcher et al., 2025; Stark et al., 2025) have expanded our ability to generate novel protein backbones and binding scaffolds, shifting the challenge towards optimising early designs into viable therapeutic and industrial products. This optimisation remains constrained by the cost and throughput of experimental protein property measurements, which are typically labour-intensive and slow. Computational approaches capable of protein analysis *in silico* are therefore essential to accelerating protein engineering and to reduce major bottlenecks in drug discovery and synthetic biology.

The most common tools used for protein analysis are Protein Language Models (PLMs) trained on large-scale sequence databases, which achieve strong performance in tasks such as property prediction, mutational-effect estimation and protein engineering (Weissenow et al., 2022; Stärk et al., 2021; Marquet et al., 2022; Lin et al., 2022). Models such as the ESM suite (ESM2, ESM3, ESMc) (Rives et al., 2019; ESM Team, 2024), ProtTrans (Elnaggar et al., 2020), Ankh3 (Alsamkary et al., 2025) and others (Olsen et al., 2022; Outeiral & Deane, 2024) capture evolutionary and functional constraints from protein sequences and have even been shown to implicitly reflect aspects of the tertiary structure, including fold classes, residue–residue contact patterns and molecular motions (Rao et al., 2020; Vig et al., 2020; Lombard et al., 2024). However, since PLMs are trained only on amino acid sequences, this structural knowledge is indirect: it emerges from statistical regularities in sequence co-variation rather than symmetry-aware geometric reasoning. This distinction is important because many protein properties are fundamentally governed by spatial interactions (Meng et al., 2025). For example, solubility depends on surface exposure and charge distribution, and thermodynamic stability depends on core packing and hydrogen-bond networks. Numerous studies suggest that structural data (Abramson et al., 2024;

Lin et al., 2022) can inform protein properties in ways that sequences alone cannot, as the three-dimensional configuration of a protein is a major determinant of its biophysical and functional characteristics (Ferreira & Castro, 2023; Vendruscolo et al., 2011).

In contrast, the predominant strategy used by state-of-the-art models for protein tasks has been to apply increasingly sophisticated fine-tuning pipelines (Jiang et al., 2024) and parameter-efficient adaptation methods on ever-larger PLMs (Schmirler et al., 2024). Furthermore, many of these approaches rely on ensembling (Thumuluri et al., 2021) or mix-and-match style frameworks that combine embeddings from multiple PLMs with diverse downstream architectures (Zhang et al., 2024; Yuan et al., 2026), often requiring thousands of experimental configurations to achieve competitive performance (Bikias et al., 2025). Moreover, extensive fine-tuning can induce catastrophic forgetting (McCloskey & Cohen, 1989), where specialization for a particular downstream task degrades the general representations learned during pretraining and leads to worse performance on other tasks compared to the original frozen model (Heinzinger et al., 2024). As the scale of PLMs has increased rapidly in recent years, growing from on the order of 10^6 parameters to over 10^9 parameters, these methods make task-specific adaptation of billion-parameter models neither straightforward nor computationally efficient, substantially increasing engineering complexity, training cost, and reproducibility challenges.

This emphasis on elaborate adaptation of PLMs treats sequence representations as the dominant, and often sufficient, lever for improvement, while principled geometric modeling remains underutilized. Even when structure is incorporated into PLMs, it is often done indirectly through handcrafted features, coarse distance bins, or quantazied 3D token representations that are appended to sequence embeddings rather than modeled as continuous 3D geometry (Su et al., 2023; Li et al., 2024). As a result, many methods are costly to train and adapt yet still lack explicit structural inductive biases. We challenge this fine-tune-first approach and contend that joint modeling of static PLMs representations with explicit protein structure using Equivariant Graph Neural Networks (EGNNs) (Satorras et al., 2021) is a direct and efficient path to improved performance in protein tasks.

In this paper, we present ProtEGNN, a simple method to pair explicit structural information with PLMs and demonstrate that it is competitive with the leading baselines across diverse benchmarks, including many with vastly superior complexity and number of parameters. Our approach constructs graphs for each protein, where nodes correspond to 3D coordinates of carbon-alpha (C_α) atoms of the predicted structure backbone. Each node is initialized with a feature vector given by embeddings from the final layer of a pretrained PLM; these embeddings are extracted once and treated as fixed inputs during training. Edges connect spatially proximal residues, and an EGNN updates the node features in an equivariant manner, combing multimodal signals from both sequence-derived embeddings and explicit geometry. Our contributions are as follows:

- Across 7 datasets testing protein property prediction (solubility, thermostability), mutational fitness prediction (GB1, GFP), and protein function annotation (subcellular localization), we provide a direct comparison between ProtEGNN and prior work. ProtEGNN outperforms or rivals leading methods on all evaluated tasks while using 100–1000$\times$ fewer parameters and establishes new state-of-the-art results on solubility and thermostability by substantial margins.
- We evaluate whether incorporating explicit protein structure can compensate for, or reduce the need for, adaptation of PLMs by pairing EGNNs with static residue embeddings from two pretrained PLMs of vastly different scales: the smallest readily available PLM we could find, ESM2-T6 (8M parameters) (Rives et al., 2019), and a large 6B parameter ESMc model (ESM Team, 2024). At each scale, we compare ProtEGNN to the pretrained and fine-tuned PLM, isolating the effects of model scale, fine-tuning, and explicit geometry.
- Through experiments in mutational effect prediction (GB1 and GFP), we show that modeling protein structure with an $E(3)$-equivariant graph, *even when paired with a small PLM* (ESM2-T6; 8M parameters) and a *single wild-type structure*, delivers gains that rival state-of-the-art methods relying on large-scale PLMs and task-specific fine-tuning. These results demonstrate that principled geometric inductive biases can be a more effective lever for performance than additional model scale or fine-tuning.

2. Related Work

2.1. Protein Large Language Models

Many predictive methods build on pretrained PLMs by combining their embeddings with task-specific architectures. For example, PLMSol (Zhang et al., 2024) integrated embeddings from multiple PLMs using attention and recurrent modules, while LMProtein (Yuan et al., 2026) fed ESM-2 representations into a hybrid CNN, LSTM and MLP architecture. Other approaches, such as PLM-Fit (Bikias et al., 2025) used combinations of ESM2 (Rives et al., 2019), ProGen2 (Nijkamp et al., 2023), ProteinBERT (Brandes et al., 2022)) with transfer-learning methods such as feature extraction and bottleneck adapters, requiring more

than 3,000 experiments. Similarly, NetSolP (Thumuluri et al., 2021) relied on extensive fine-tuning and ensembling of ESM models (Rao et al., 2020) to improve performance. While effective, these approaches often require complex pipelines and substantial computational effort. As PLMs scale, adapting them to downstream tasks has become increasingly resource-intensive, motivating parameter-efficient fine-tuning approaches. Schmirler et al. (Schmirler et al., 2024) provide detailed, "recipe"-like guidelines for parameter efficient model adaptation. Structural information has also been incorporated into protein models. Earlier work such as DeepSol (Khurana et al., 2018) used handcrafted physicochemical and structural features, whereas more recent methods embed structure directly into PLMs through discretization. For example, SaProt (Su et al., 2023) augmented the tokenizer with quantized 3D structural tokens, and ProSST (Li et al., 2024) learnt a structure-quantized latent space with sequence–structure attention. These methods inject structural cues but do so indirectly via engineered or discrete representations.

2.2. Equivariant Graph Neural Networks

Graph Neural Networks (GNNs) incorporate structure explicitly by modeling proteins as residue- or atom-level graphs with edges defined by spatial proximity (Duval et al., 2023; Zhang et al., 2022). These models capture local and global geometric patterns and have been used for diverse tasks including protein binding site, protein-protein interaction and enzyme class prediction (Zhang et al., 2025; Hu & Ohue, 2025; Zhang et al., 2022; van der Weg et al., 2025). More recently, $E(3)$-equivariant architectures, such as Geometric Vector Perceptrons (GVPs) (Jing et al., 2020) and EGNNs (Satorras et al., 2021) have been designed to respect the inherent symmetries of data. The term 'equivariant' refers to how these networks maintain consistent behavior under transformations such as rotations, translations, and reflections. When combined with PLMs, these geometric models form multimodal pipelines in which PLM representations and structural features are fused, often via joint training of the PLM and the GNN or by augmenting graphs with handcrafted physicochemical descriptors (Wang et al., 2022; Roche et al., 2024; Tan et al., 2024).

3. Methodology

3.1. Sequence and Structure Representations

For sequence information, we extract last-layer, per-residue embeddings from two pretrained PLMs of different scales: ESM2-T6 (8M parameters) and ESMc (6B parameters). For a protein of length n, these embeddings form a matrix in $\mathbb{R}^{n\times d}$, where the embedding dimension d is 320 for ESM2-T6 and 2560 for ESMc. During training, gradients from the downstream loss are not propagated back to the PLM; the sequence representations are extracted once and used as fixed node features.

Structural information is obtained by predicting protein 3D coordinates using ESMFold (Lin et al., 2022). Each residue is represented by its Cα atom, which provides a consistent and compact representation of the protein backbone. Using the Cα coordinates, we compute an $n \times n$ pairwise distance matrix, where each entry d_{ij} denotes the Euclidean distance between residues i and j.

3.2. Equivariant Graph Construction

Following (Satorras et al., 2021), we represent each protein as a graph

$$G = (V, E, \mathbf{x}),$$

where $V = \{v_0, \ldots, v_n\}$ denotes the set of nodes corresponding to residues, $E \subseteq V \times V$ denotes the set of edges, and $\mathbf{x} = \{x_i \in \mathbb{R}^3\}_{i=0}^n$ represents the 3D coordinates of the Cα atom associated with node v_i. Each layer of ProtEGNN takes as input the set of node embeddings $h^t = \{h_0^t, \ldots, h_n^t\}$, the coordinate representations $x^t = \{x_0^t, \ldots, x_{n-1}^t\}$, and the edge information $\mathcal{E} = (e_{ij})$, and outputs updated node and coordinate representations h^{t+1} and x^{t+1}, respectively. $t \in \{0, \ldots, T-1\}$ refers to the layer number, where T is a hyper-parameter. Each node v_i is associated with an initial coordinate $x_i^{(0)} \in \mathbb{R}^3$, corresponding to the three-dimensional position of the C_α atom representing the residue in protein structure. Feature vectors for each node v_i are derived from PLM embeddings and initialized as $h_i^{(0)} \in \mathbb{R}^d$. Edges are defined based on spatial proximity in 3D space using the distance matrix described above. Specifically, an edge $(v_i, v_j) \in E$ exists if the Euclidean distance between the C_α atoms of residues i and j satisfies

$$\|x_i^{(0)} - x_j^{(0)}\| < \tau.$$

where $\tau \in \{5, 10, 20\}$ Angstroms (Å) and is selected as a hyper-parameter. Equivalently, we define the neighbor set $\mathcal{N}(i) = \{j \neq i : \|x_i^{(0)} - x_j^{(0)}\| < \tau\}$ and let $E = \{(i,j) : j \in \mathcal{N}(i)\}$. We optionally allow edge attributes e_{ij}; in our experiments we set $e_{ij} = \varnothing$.

The layerwise equivariant message-passing is defined by the following equations:

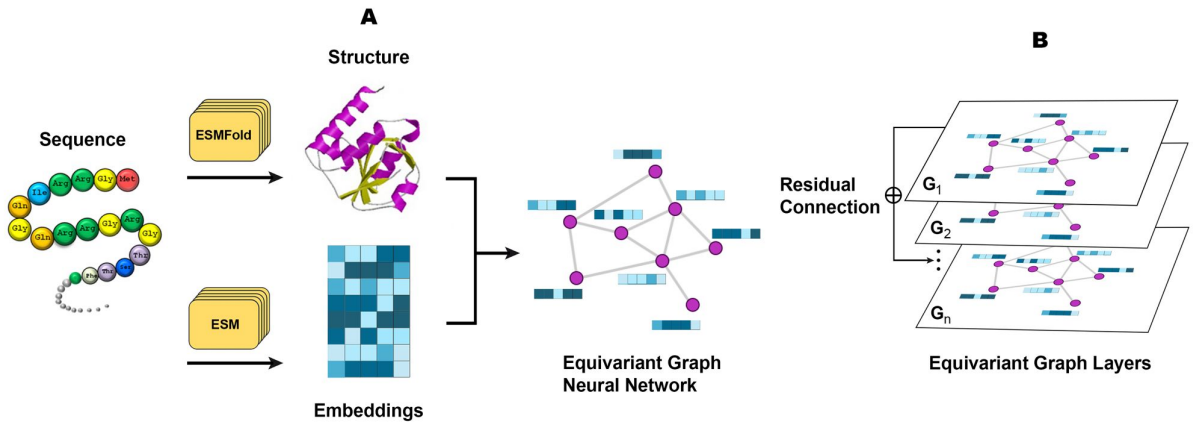

Figure 1. Integration of protein sequence and structure data using an EGNN architecture in ProtEGNN. In **Panel A**, a protein sequence is processed through two distinct pathways: ESMFold predicts the 3D structure of the protein, while ESM (ESMc or ESM2-T6) generates sequence embeddings. The predicted structure and sequence embeddings are then combined to construct a graph representation, where nodes correspond to amino acids and edges represent spatial relationships. In **Panel B**, the graph is passed through multiple EGNN layers. Residual connections are used between layers to maintain information flow and improve learning stability.

$$m_{ij}^{(t)} = \phi_e\Big(h_i^{(t)},\, h_j^{(t)},\, \|x_i^{(t)} - x_j^{(t)}\|^2,\, e_{ij}\Big), \quad j \in \mathcal{N}(i), \tag{1}$$

$$m_i^{(t)} = \sum_{j \in \mathcal{N}(i)} m_{ij}^{(t)}, \tag{2}$$

$$x_i^{(t+1)} = x_i^{(t)} + C \sum_{j \in \mathcal{N}(i)} (x_i^{(t)} - x_j^{(t)})\, \phi_x\big(m_{ij}^{(t)}\big), \tag{3}$$

$$h_i^{(t+1)} = \phi_h\Big(h_i^{(t)},\, m_i^{(t)}\Big). \tag{4}$$

Here ϕ_e, ϕ_x, ϕ_h are edge, coordinate and node operations respectively, implemented as learnable functions using MLPs. Using relative squared distance $\|x_i^t - x_j^t\|^2$ in the edge function ϕ_e provides rotation and translation-invariant geometric inputs and the aggregation in (2) preserves permutation invariance. (3) is the crucial coordinate update step: coordinates are moved by a weighted radial vector field where each neighbor contributes the relative displacement $(x_i - x_j)$ weighted by a learned scaler $\phi_x(m_{ij})$. The constant C normalizes the update to stabilize magnitudes. (4) performs a permutation-equivariant update of node features using the aggregated messages $m_i^{(t)}$, allowing feature representations to be refined in each layer.

4. Datasets

We evaluate ProtEGNN on seven datasets spanning three complementary protein learning regimes:

Protein Property Prediction Solubility (Bhandari et al., 2020) is a protein-level binary classification task, measured by ROC-AUC. To assess generalization beyond the training distribution, we additionally evaluate zero-shot solubility prediction on the Price dataset (Price et al., 2011), which shares no close homologs with PSI-Biology.

Thermostability is evaluated using two protein-level regression datasets: Meltome, from the FLIP benchmark (Dallago et al., 2021), which measures melting temperatures, T_m, and Stability dataset Rocklin et al. (2017), which measures protease resistance of de novo–designed mini-proteins in local mutational neighborhoods. Both tasks are evaluated using Spearman's rho (SPR).

Mutational Effect Prediction We evaluate on GFP from the TAPE benchmark (Rao et al., 2019) and GB1 (binding fitness) from the FLIP benchmark (Dallago et al., 2021), where each variant differs by a small number of mutations from a wild-type sequence. Both are formulated as regression tasks over variants and evaluated using SPR. In both settings, we provide only a single wild-type structure and vary only residue embeddings to reflect realistic experimental conditions.

Protein Function Annotation Subcellular localization is evaluated on the DeepLoc's (Almagro Armenteros et al., 2017) SetHard version (Stärk et al., 2021) as a 10-class protein-level classification task, measured by accuracy (ACC).

These benchmarks jointly probe global and local sequence–structure relationships under both sequence-diverse and mutation-centric settings. Full dataset descriptions, preprocessing steps, splits, and evaluation protocols are provided in the Appendix A

5. Results

Our goal with ProtEGNN is to show that incorporating explicit 3D protein structure into pretrained PLMs is the most effective and efficient strategy for protein tasks, and should be prioritized before resorting to costly fine-tuning of large PLMs. Furthermore, to study how PLM capacity interacts with geometric modeling, we instantiated ProtEGNN with representations from two pretrained PLMs of vastly different scales: a large 6B parameter ESMc (ProtEGNN(ESMc)) and a small 8M-parameter ESM2-T6 (ProtEGNN(T6)). We trained separate models for each variant on each task. All ProtEGNN models are evaluated using 3-fold cross-validation over random seeds 97, 98, 99; except solubility which used 5-fold cross-validation following (Thumuluri et al., 2021).

5.1. Baseline Experiments

To contextualize the performance of ProtEGNN and isolate the contribution of explicit geometry, we evaluate three settings for each PLM backbone size: a pretrained PLM used as a static feature extractor, a task-adapted PLM, and ProtEGNN, which augments fixed PLM residue embeddings with a protein structure. For the ESM2-T6 backbone, pretrained and fine-tuned results (except for solubility) are reported from (Schmirler et al., 2024). For ESMc, fine-tuning is performed using LoRA-based adaptation (Hu et al., 2022), while the pretrained baseline corresponds to frozen embeddings without task-specific updates. The second solubility dataset, Price (Price et al., 2011), is used as a holdout set to test generalization and hence, not included in this experiment.

Table 1 demonstrates that adding explicit geometric modeling consistently improves performance over sequence-only baselines, with ProtEGNN achieving the best results on five of six tasks. While increasing PLM scale benefits global property prediction, we find that on mutational tasks (GFP, GB1), the small ESM2-T6 backbone augmented with structure matches the larger ESMc. This suggests that geometric inductive biases can effectively compensate for reduced PLM capacity, particularly in local mutational contexts.

5.2. Protein Property Prediction

5.2.1. Solubility

Table 2a reports solubility prediction performance on the PSI-Biology dataset, together with model sizes for direct comparison. We compare ProtEGNN against Prot-T5 (Elnaggar et al., 2020), a large 1.2B parameter protein language model, ESM-MSA (Rao et al., 2021), which leverages Multiple Sequence Alignments and NetSolP (Thumuluri et al., 2021), a previous state-of-the-art solubility predictor specifically trained and tuned on the PSI-Biology dataset. NetSolP relied on both fine-tuning and ensembling of multiple ESM models (Rao et al., 2020). PSI-Biology results for Prot-T5, ESM-MSA, and NetSolP are as reported in (Thumuluri et al., 2021). Despite being orders of magnitude smaller, both ProtEGNN variants perform competitively; with ProtEGNN(ESMc) setting the new state-of-the-art with a large margin while using 100x less number of parameters.

To assess out-of-distribution performance, we tested leading solubility predictors on the Price solubility dataset (Price et al., 2011), which was not used in training ProtEGNN. As shown in 2b, ProtEGNN(ESMc) achieved the best zero-shot performance with ROC_AUC of 0.77 using only 1.2M parameters, outperforming substantially larger models. Even our smaller variant, ProtEGNN(t6), remains competitive (0.74, 500K), surpassing ProtSolM (0.55, 3.2M) and PLMSoL (0.60, 7.3M). Notably, ProtSolM (Tan et al., 2024) integrates an EGNN directly with a PLM, together with handcrafted physicochemical structural descriptors, and performs end-to-end training with gradients propagated through the entire combined architecture. Similarly, PLMSol (Zhang et al., 2024) aggregated embeddings from multiple large PLMs and coupled them with an ensemble of classifiers (e.g. MLPs, Light Attention, CNN–BiLSTM), resulting in a resource-intensive approach.

5.2.2. Stability

On the Meltome thermostability benchmark (Table 3a), we observe a strong dependence on PLM capacity: while ProtEGNN(t6) lags (0.61 SPR), scaling to ESMc achieves state-of-the-art performance (0.79 SPR). ProtEGNN outperforms resource-intensive pipelines such as PLM-Fit (Bikias et al., 2025), which required over 3,000 experiments to optimize adapters across multiple PLMs and PRIME (Jiang et al., 2024), which necessitated specialized pre-training on 96M bacterial sequences. ProtEGNN also surpasses Prot-T5-FT (Schmirler et al., 2024), which fine-tuned 3.5M parameters of the 1.2B Prot-T5 (Elnaggar et al., 2020) using LoRA (Hu et al., 2022). Crucially, ProtEGNN(ESMc) also surpasses structure-aware baselines such as ProSST (Li et al., 2024), which learnt disentangled sequence-structure attention, and SaProt (Su et al., 2023), which fused discrete structural tokens into PLM training. In contrast, ProtEGNN achieves top performance by training only a lightweight EGNN (≈600K parameters) on a frozen PLM backbone, demonstrating that explicit geometric modeling is a far more efficient alternative to extensive architectural search or specialized pre-training.

On the second stability benchmark (Table 3b), ProtEGNN(t6) achieves an SPR of 0.77, matching the performance of the 1.2B-parameter finetuned Ankh-FT despite using only 400K trainable parameters. Scaling to

Table 1. Effect of PLM scale and explicit geometric modeling across tasks. Bold values indicate the best method.

Sequence Model	Method	Meltome(↑)	Stability(↑)	Solubility(↑)	GFP(↑)	GB1(↑)	Sub-loc(↑)
ESM2-T6 (8M)	Pre-trained	56.40 ± 0.46	75.00 ± 2.02	54.42 ± 0.02	63.90 ± 0.20	81.80 ± 0.41	52.00 ± 0.61
	Fine-tuned	58.40 ± 0.40	76.50 ± 1.96	70.31 ± 0.04	68.80 ± 0.30	88.30 ± 0.95	55.90 ± 1.50
	ProtEGNN	$\mathbf{60.85 \pm 0.48}$	$\mathbf{77.13 \pm 0.03}$	$\mathbf{72.62 \pm 2.05}$	$\mathbf{69.48 \pm 0.12}$	$\mathbf{89.95 \pm 0.58}$	$\mathbf{57.21 \pm 0.62}$
ESMc (6B)	Pre-trained	58.03 ± 0.16	78.10 ± 0.09	61.00 ± 0.03	64.30 ± 0.35	83.01 ± 0.21	60.12 ± 0.01
	Fine-tuned	74.30 ± 0.30	$\mathbf{83.80 \pm 0.36}$	75.90 ± 0.05	68.01 ± 0.32	89.44 ± 0.07	65.98 ± 0.80
	ProtEGNN	$\mathbf{78.65 \pm 0.08}$	82.40 ± 0.02	$\mathbf{78.31 \pm 2.03}$	$\mathbf{69.13 \pm 0.02}$	$\mathbf{89.51 \pm 0.30}$	$\mathbf{66.81 \pm 0.09}$

Table 2. Solubility prediction performance and model size comparison. Results are shown in ascending order of performance (ROC_AUC). Higher is better. Bold values indicate our method.

(a) PSI-Biology dataset (5-fold cross-validation)

Method	# Parameters	ROC-AUC(↑)
Prot-T5	1.2B	0.73
NetSolP	650M	0.73
ProtEGNN(t6)	**500K**	**0.73**
ESM-MSA	100M	0.75
ProtEGNN(ESMc)	**1.2M**	**0.78**

(b) Zero-shot solubility prediction on Price dataset

Method	# Parameters	ROC-AUC(↑)
ProtSolM	3.2M	0.55
PLMSoL	7.3M	0.60
Prot-T5	1.2B	0.73
ProtEGNN(t6)	**500K**	**0.74**
ESM-MSA	100M	0.75
NetSolP	650M	0.76
ProtEGNN(ESMc)	**1.2M**	**0.77**

Table 3. Thermostability prediction performance and model size comparison. Results are shown in ascending order of performance (SPR). Higher is better. Bold values indicate our method.

(a) Meltome Dataset

Model	# Parameters	SPR(↑)
ProtEGNN(t6)	**400K**	**0.61**
PLM-Fit	6.4B	0.72
Prime	653M	0.72
SaProt	650M	0.72
Prot-T5-FT	3.5M	0.72
ProSST	110M	0.72
ProtEGNN(ESMc)	**600K**	**0.79**

(b) Stability Dataset

Method	# Parameters	SPR (↑)
CARP	640M	0.72
Ankh-FT	1.2B	0.77
ProtEGNN(t6)	**400K**	**0.77**
LMProtein	Unavailable	0.79
ProtEGNN(ESMc)	**500K**	**0.82**
ESM2-T36-FT	7.7M	0.84

ProtEGNN(ESMc) further improves performance to 0.82 SPR, surpassing CARP (0.72) (Yang et al., 2024) which combined PLMs with Convolutional architecture, and LM-Protein (0.79) (Yuan et al., 2026), a hybrid architecture that stacks CNNs, LSTMs, and MLPs on top of ESM-2 embeddings. The LoRA finetuned ESM2-T36-FT (Schmirler et al., 2024) achieves the highest score (0.84), albeit using orders-of-magnitude more parameters.

5.3. Mutational Effect Prediction

For each dataset, we use a *single wild-type structure* to represent all variants, fixing the underlying geometry across the mutational landscape. Even without modeling the explicit mutation-induced structural rearrangements, our equivariant geometric model captures local physical constraints that remain informative under sequence perturbations, enabling state-of-the-art mutational effect prediction.

5.3.1. GFP

Table 4a reports results on the GFP fluorescence landscape. ProtEGNN(T6) achieves state-of-the-art performance (SPR = 0.70), tying with Ankh-FT model (Schmirler et al., 2024), while training only 500K task-specific parameters on top of a small 8M-parameter PLM backbone. Notably, Ankh-FT applied LoRA on top of the 1.9B parameter Ankh backbone, resulting in millions of trainable parameters. ProtEGNN achieves the same performance using fewer parameters (4.9M vs 500K) and much smaller PLM backbone (1.9B vs 8M).

Both ProtEGNN variants outperform substantially larger models such as Prot-T5 (1.2B parameters) (Elnaggar et al., 2020) and Ankh3 (5.7B parameters) (Alsamkary et al., 2025). ProtEGNN also performs better than LM-GVP (Wang et al., 2022) (0.68), which stacked an EGNN in front of a PLM (Elnaggar et al., 2020) and jointly trained the models, modifying and fine-tuning the PLM embeddings.

Table 4. GFP and GB1 mutational effect prediction performance and model size comparison. Results are shown in ascending order of performance (SPR). Higher is better. Bold values indicate our method.

(a) GFP

Model	# Parameters	SPR (↑)
Prot-T5	1.2B	0.61
Ankh3	5.7B	0.65
LM-GVP	Unavailable	0.68
ProtEGNN(ESMc)	**650K**	**0.69**
Ankh-FT	4.9M	0.70
ProtEGNN(t6)	**500K**	**0.70**

(b) GB1

Model	# Parameters	SPR(↑)
SaProt	650M	0.81
PRIME	653M	0.82
ESM2-T48	15B	0.85
PLM-Fit	6.4B	0.88
ESM2-T36-FT	7.7M	0.89
ProtEGNN(ESMc)	**1M**	**0.90**
ProtEGNN(t6)	**400K**	**0.90**
Ankh3	5.7B	0.90

5.3.2. GB1

We observe a similar pattern in the GB1 binding landscape (Table 4b). Both ProtEGNN(t6) and ProtEGNN(ESMc) reach an SPR of 0.90, matching the best-performing 5.7B parameter Ankh3 model (Alsamkary et al., 2025). This parity is achieved with as few as 400K trainable parameters in ProtEGNN(t6). Notably, ESM2-T36-FT, as reported in (Schmirler et al., 2024) applies LoRA to a 3B parameter PLM backbone, reducing the number of trainable parameters to only 7.7M; however, despite this adaptation, it still underperforms ProtEGNN, highlighting the limitations of parameter-efficient fine-tuning when structural inductive biases are absent. Crucially, our results show that adding a single wildtype structure to even the smallest 8M parameter backbone delivers predictive power equivalent to multi-billion parameter sequence models without the need for extensive fine-tuning or structure predictions for every variant.

5.4. Protein Function Annotation

5.4.1. Subcellular Localization

On the subcellular localization benchmark, Table 5, ProtEGNN achieves competitive accuracy with orders-of-magnitude fewer parameters than large PLM baselines. In particular, ProtEGNN matches the performance of Prosst-T5 (Heinzinger et al., 2024) (ACC = 0.57) while using only 400K parameters. Notably, even though Prosst, a fine-tuned derivative of ProtT5 (Elnaggar et al., 2020), incorporates structural information, its reduced performance relative to ProtT5 and LA-Prot-T5 (Stärk et al., 2021) (0.57 vs. 0.65) highlights the effect of task specialization through fine-tuning, where adapting a PLM for one objective can degrade its performance on others.

Table 5. Subcellular Localization performance and model size comparison. Results are shown in ascending order of performance (ACC). Higher is better. Bold values indicate our method.

Subcellular Localization

Model	# Parameters	ACC(↑)
Prost-T5	2.8B	0.57
ProtEGNN(t6)	**400K**	**0.57**
Prot-T5	3B	0.65
LA-Prot-T5	3B	0.65
ProtEGNN(ESMc)	**500K**	**0.67**
ESM2-T36	7.7M	0.68

In contrast, ProtEGNN(ESMc) attains an accuracy of 0.67 with just 500K parameters, closely approaching the performance of the fine-tuned ESM2-T36 (0.68) reported in (Schmirler et al., 2024). It is important to note that (Schmirler et al., 2024) employed parameter efficient fine tuning techniques to reduce the trainable parameters of ESM2-T36 from 3B to 7.7M. However, even parameter-efficient fine-tuning techniques approaches incur significantly higher parameters compared to our method, which simply adds a lightweight geometric head to frozen PLM representations. These results reinforce the idea that adding explicit structural inductive biases is a more robust and efficient alternative to fine-tuning large PLMs, even when parameter-efficient techniques are used.

6. Discussion

Our experiments convey a clear and actionable message: explicit, equivariant geometric modeling can substitute for, and in most cases outperform, large-scale PLM fine-tuning, while being 100-1000× more parameter-efficient. Our findings expose a broader limitation of current practice: fine-tuning large PLMs for each downstream task is not a viable long-term strategy. Beyond the computational burden, extensive fine-tuning sometimes leads to task specialization at the cost of generality, as seen in the case of ProSST where subcellular localization performance degrades after fine-tuning. Our results on mutational landscapes demonstrate that geometric deep learning offers distinct advantages for modeling proteins, enabling lightweight models to capture local physical constraints and deliver much of the predictive benefit usually attributed to massive, fine-tuned PLMs

We emphasize that our efficiency claims refer specifically

to the number of trainable parameters and task-specific optimization cost. Structure prediction is performed once per protein as an offline preprocessing step and can be amortized across tasks, in contrast to repeated fine-tuning of large PLMs for each downstream objective. Moreover, as our pipeline relies on predicted structures, future work should explore uncertainty-aware inference to mitigate sensitivity to structural quality; especially for disordered or membrane proteins which are poorly predicted by current structure prediction methods. We also encourage exploration into multi-task EGNN heads and interpretability methods that connect EGNN message passing to concrete biophysical mechanisms. Broadly, our findings underscore the critical importance of appropriate inductive biases in model design. We contend that prioritizing higher-fidelity representations and biology-aware architectures will yield significantly greater dividends in protein modeling than indiscriminate parameter scaling or elaborate adaptation of PLMs.

Impact Statement

This work aims to advance machine learning methods for protein modeling by showing that explicit geometric inductive biases can improve protein property, mutational-effect, and function prediction while requiring substantially fewer task-specific trainable parameters than large-scale protein language model fine-tuning. The main positive impacts are scientific and practical: more efficient protein models may reduce the computational cost of adapting large protein language models, lower barriers for groups with limited compute, and support faster in silico screening in protein engineering, therapeutic design, and synthetic biology.

At the same time, improved protein modeling tools can contribute to dual-use capabilities if incorporated into workflows for designing or optimizing biologically active molecules. Our work is predictive rather than generative: it does not introduce a new protein design system, propose wet-lab protocols, or release optimized biological sequences. Nevertheless, we recommend that future applications of this approach in protein engineering be paired with appropriate biosafety review, sequence screening, experimental validation, and domain-expert oversight, particularly when applied to pathogens, toxins, immune modulation, or other sensitive biological functions.

There are also limitations relevant to responsible deployment. Because the method relies on predicted or available protein structures, its reliability may decrease for proteins whose structures are uncertain, disordered, membrane-bound, or poorly represented in current structural datasets. Predictions should therefore not be interpreted as substitutes for experimental characterization. More broadly, benchmark improvements in protein machine learning may not translate directly into safe or effective biological products without careful validation in the intended experimental context.

References

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., Bodenstein, S. W., Evans, D. A., Hung, C.-C., O'Neill, M., Reiman, D., Tunyasuvunakool, K., Wu, Z., Žemgulytė, A., Arvaniti, E., Beattie, C., Bertolli, O., Bridgland, A., Cherepanov, A., Congreve, M., Cowen-Rivers, A. I., Cowie, A., Figurnov, M., Fuchs, F. B., Gladman, H., Jain, R., Khan, Y. A., Low, C. M. R., Perlin, K., Potapenko, A., Savy, P., Singh, S., Stecula, A., Thillaisundaram, A., Tong, C., Yakneen, S., Zhong, E. D., Zielinski, M., Žídek, A., Bapst, V., Kohli, P., Jaderberg, M., Hassabis, D., and Jumper, J. M. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, May 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07487-w. URL http://dx.doi.org/10.1038/s41586-024-07487-w.

Almagro Armenteros, J. J., Sønderby, C. K., Sønderby, S. K., Nielsen, H., and Winther, O. Deeploc: prediction of protein subcellular localization using deep learning. *Bioinformatics*, 33(21):3387–3395, 2017.

Alsamkary, H., Elshaffei, M., Elkerdawy, M., and Elnaggar, A. Ankh3: Multi-task pretraining with sequence denoising and completion enhances protein representations. *arXiv preprint arXiv:2505.20052*, 2025.

Bhandari, B. K., Gardner, P. P., and Lim, C. S. Solubility-weighted index: fast and accurate prediction of protein solubility. *Bioinformatics*, 36(18):4691–4698, June 2020. ISSN 1367-4811. doi: 10.1093/bioinformatics/btaa578. URL http://dx.doi.org/10.1093/bioinformatics/btaa578.

Bikias, T., Stamkopoulos, E., and Reddy, S. T. Plmfit: benchmarking transfer learning with protein language models for protein engineering. *Briefings in Bioinformatics*, 26(4), July 2025. ISSN 1477-4054. doi: 10.1093/bib/bbaf381. URL http://dx.doi.org/10.1093/bib/bbaf381.

Brandes, N., Ofer, D., Peleg, Y., Rappoport, N., and Linial, M. Proteinbert: a universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8): 2102–2110, February 2022. ISSN 1367-4811. doi: 10.1093/bioinformatics/btac020. URL http://dx.doi.org/10.1093/bioinformatics/btac020.

Butcher, J., Krishna, R., Mitra, R., Brent, R. I., Li, Y., Corley, N., Kim, P. T., Funk, J., Mathis, S., Salike, S., et al. De novo design of all-atom biomolecular interactions with rfdiffusion3. *bioRxiv*, 2025.

Dallago, C., Mou, J., Johnston, K. E., Wittmann, B. J., Bhattacharya, N., Goldman, S., Madani, A., and Yang, K. K. Flip: Benchmark tasks in fitness landscape inference for proteins. November 2021. doi: 10.1101/2021.11.09.467890. URL http://dx.doi.org/10.1101/2021.11.09.467890.

Duval, A., Mathis, S. V., Joshi, C. K., Schmidt, V., Miret, S., Malliaros, F. D., Cohen, T., Liò, P., Bengio, Y., and Bronstein, M. A hitchhiker's guide to geometric gnns for 3d atomic systems, 2023. URL https://arxiv.org/abs/2312.07511.

Edgar, R. C. Search and clustering orders of magnitude faster than blast. *Bioinformatics*, 26(19):2460–2461, August 2010. ISSN 1367-4803. doi: 10.1093/bioinformatics/btq461. URL http://dx.doi.org/10.1093/bioinformatics/btq461.

Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., Bhowmik, D., and Rost, B. Prottrans: Towards cracking the language of life's code through self-supervised learning. *Cold Spring Harbor Laboratory*, July 2020. doi: 10.1101/2020.07.12.199554. URL http://dx.doi.org/10.1101/2020.07.12.199554.

ESM Team. Esm cambrian: Revealing the mysteries of proteins with unsupervised learning, 2024. URL https://evolutionaryscale.ai/blog/esm-cambrian.

Ferreira, J. and Castro, F. Advances in protein solubility and thermodynamics: quantification, instrumentation, and perspectives. *CrystEngComm*, 25(46):6388–6404, 2023. ISSN 1466-8033. doi: 10.1039/d3ce00757j. URL http://dx.doi.org/10.1039/D3CE00757J.

Glaser, M. and Brägelmann, J. Esm-effect: An effective and efficient fine-tuning framework towards accurate prediction of mutation's functional effect. February 2025. doi: 10.1101/2025.02.03.635741. URL http://dx.doi.org/10.1101/2025.02.03.635741.

Heinzinger, M., Weissenow, K., Sanchez, J. G., Henkel, A., Mirdita, M., Steinegger, M., and Rost, B. Bilingual language model for protein sequence and structure. *NAR Genomics and Bioinformatics*, 6(4), September 2024. ISSN 2631-9268. doi: 10.1093/nargab/lqae150. URL http://dx.doi.org/10.1093/nargab/lqae150.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Hu, W. and Ohue, M. Spatialppiv2: Enhancing protein–protein interaction prediction through graph neural networks with protein language models. *Computational and Structural Biotechnology Journal*, 27:508–518, 2025.

Jiang, F., Li, M., Dong, J., Yu, Y., Sun, X., Wu, B., Huang, J., Kang, L., Pei, Y., Zhang, L., Wang, S., Xu, W., Xin, J., Ouyang, W., Fan, G., Zheng, L., Tan, Y., Hu, Z., Xiong, Y., Feng, Y., Yang, G., Liu, Q., Song, J., Liu, J., Hong, L., and Tan, P. A general temperature-guided language model to design proteins of enhanced stability and activity. *Science Advances*, 10(48), November 2024. ISSN 2375-2548. doi: 10.1126/sciadv.adr2641. URL http://dx.doi.org/10.1126/sciadv.adr2641.

Jing, B., Eismann, S., Suriana, P., Townshend, R. J., and Dror, R. Learning from protein structure with geometric vector perceptrons. *arXiv preprint arXiv:2009.01411*, 2020.

Khurana, S., Rawi, R., Kunji, K., Chuang, G.-Y., Bensmail, H., and Mall, R. Deepsol: a deep learning framework for sequence-based protein solubility prediction. *Bioinformatics*, 34(15):2605–2613, March 2018. ISSN 1367-4811. doi: 10.1093/bioinformatics/bty166. URL http://dx.doi.org/10.1093/bioinformatics/bty166.

Li, M., Tan, Y., Ma, X., Zhong, B., Yu, H., Zhou, Z., Ouyang, W., Zhou, B., Tan, P., and Hong, L. Prosst: Protein language modeling with quantized structure and disentangled attention. *Advances in Neural Information Processing Systems*, 37:35700–35726, 2024.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., and Rives, A. Evolutionary-scale prediction of atomic level protein structure with a language model, July 2022. URL http://dx.doi.org/10.1101/2022.07.20.500902.

Lombard, V., Timsit, D., Grudinin, S., and Laine, E. Seamoon: from protein language models to continuous structural heterogeneity. September 2024. doi: 10.1101/2024.09.23.614585. URL http://dx.doi.org/10.1101/2024.09.23.614585.

Marquet, C., Heinzinger, M., Olenyi, T., Dallago, C., Erckert, K., Bernhofer, M., Nechaev, D., and Rost, B. Embeddings from protein language models predict conservation and variant effects. *Human genetics*, 141(10):1629–1647, 2022.

McCloskey, M. and Cohen, N. J. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pp. 109–165. Elsevier, 1989.

Meng, Y., Zhang, Z., Zhou, C., Tang, X., Hu, X., Tian, G., Yang, J., and Yao, Y. Protein structure prediction via deep learning: an in-depth review. *Frontiers in Pharmacology*, 16, April 2025. ISSN 1663-9812. doi: 10.3389/fphar.2025.1498662. URL http://dx.doi.org/10.3389/fphar.2025.1498662.

Nijkamp, E., Ruffolo, J. A., Weinstein, E. N., Naik, N., and Madani, A. Progen2: Exploring the boundaries of protein language models. *Cell Systems*, 14(11): 968–978.e3, November 2023. ISSN 2405-4712. doi: 10.1016/j.cels.2023.10.002. URL http://dx.doi.org/10.1016/j.cels.2023.10.002.

Olsen, T. H., Moal, I. H., and Deane, C. M. Ablang: an antibody language model for completing antibody sequences. *Bioinformatics Advances*, 2(1):vbac046, 2022.

Outeiral, C. and Deane, C. M. Codon language embeddings provide strong signals for use in protein engineering. *Nature Machine Intelligence*, 6(2):170–179, 2024.

Price, W. N., Handelman, S. K., Everett, J. K., Tong, S. N., Bracic, A., Luff, J. D., Naumov, V., Acton, T., Manor, P., Xiao, R., Rost, B., Montelione, G. T., and Hunt, J. F. Large-scale experimental studies show unexpected amino acid effects on protein expression and solubility in vivo in e. coli. *Microbial Informatics and Experimentation*, 1(1), June 2011. ISSN 2042-5783. doi: 10.1186/2042-5783-1-6. URL http://dx.doi.org/10.1186/2042-5783-1-6.

Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, P., Canny, J., Abbeel, P., and Song, Y. Evaluating protein transfer learning with tape. *Advances in neural information processing systems*, 32, 2019.

Rao, R., Meier, J., Sercu, T., Ovchinnikov, S., and Rives, A. Transformer protein language models are unsupervised structure learners. *Cold Spring Harbor Laboratory*, December 2020. doi: 10.1101/2020.12.15.422761. URL http://dx.doi.org/10.1101/2020.12.15.422761.

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

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., and Fergus, R. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences, April 2019. URL http://dx.doi.org/10.1101/622803.

Roche, R., Moussad, B., Shuvo, M. H., Tarafder, S., and Bhattacharya, D. Equipnas: improved protein–nucleic acid binding site prediction using protein-language-model-informed equivariant deep graph neural networks. *Nucleic Acids Research*, 52(5):e27–e27, January 2024. ISSN 1362-4962. doi: 10.1093/nar/gkae039. URL http://dx.doi.org/10.1093/nar/gkae039.

Rocklin, G. J., Chidyausiku, T. M., Goreshnik, I., Ford, A., Houliston, S., Lemak, A., Carter, L., Ravichandran, R., Mulligan, V. K., Chevalier, A., Arrowsmith, C. H., and Baker, D. Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science*, 357(6347):168–175, July 2017. ISSN 1095-9203. doi: 10.1126/science.aan0693. URL http://dx.doi.org/10.1126/science.aan0693.

Satorras, V. G., Hoogeboom, E., and Welling, M. E(n) equivariant graph neural networks, 2021. URL https://arxiv.org/abs/2102.09844.

Schmirler, R., Heinzinger, M., and Rost, B. Fine-tuning protein language models boosts predictions across diverse tasks. *Nature Communications*, 15 (1), August 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-51844-2. URL http://dx.doi.org/10.1038/s41467-024-51844-2.

Stärk, H., Dallago, C., Heinzinger, M., and Rost, B. Light attention predicts protein location from the language of life. *Bioinformatics Advances*, 1(1):vbab035, 2021.

Stark, H., Faltings, F., Choi, M., Xie, Y., Hur, E., O'Donnell, T., Bushuiev, A., Uçar, T., Passaro, S., Mao, W., et al. Boltzgen: Toward universal binder design. *bioRxiv*, pp. 2025–11, 2025.

Su, J., Han, C., Zhou, Y., Shan, J., Zhou, X., and Yuan, F. Saprot: Protein language modeling with structure-aware vocabulary. October 2023. doi: 10.1101/2023.10.01.560349. URL http://dx.doi.org/10.1101/2023.10.01.560349.

Tan, Y., Zheng, J., Hong, L., and Zhou, B. Protsolm: Protein solubility prediction with multi-modal features. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 223–232. IEEE, 2024.

Thumuluri, V., Martiny, H.-M., Almagro Armenteros, J. J., Salomon, J., Nielsen, H., and Johansen, A. R. Netsolp: predicting protein solubility in escherichia coli using language models. *Bioinformatics*, 38(4):941–946, November 2021. ISSN 1367-4811. doi: 10.1093/bioinformatics/btab801. URL http://dx.doi.org/10.1093/bioinformatics/btab801.

Tsuboyama, K., Dauparas, J., Chen, J., Laine, E., Mohseni Behbahani, Y., Weinstein, J. J., Mangan, N. M., Ovchinnikov, S., and Rocklin, G. J. Megascale experimental analysis of protein folding stability in biology and design. *Nature*, 620(7973): 434–444, July 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06328-6. URL http://dx.doi.org/10.1038/s41586-023-06328-6.

van der Weg, K., Merdivan, E., Piraud, M., and Gohlke, H. Topec: prediction of enzyme commission classes by 3d graph neural networks and localized 3d protein descriptor. *Nature Communications*, 16(1):2737, 2025.

Vendruscolo, M., Knowles, T. P. J., and Dobson, C. M. Protein solubility and protein homeostasis: A generic view of protein misfolding disorders. *Cold Spring Harbor Perspectives in Biology*, 3(12):a010454–a010454, August 2011. ISSN 1943-0264. doi: 10.1101/cshperspect.a010454. URL http://dx.doi.org/10.1101/cshperspect.a010454.

Vig, J., Madani, A., Varshney, L. R., Xiong, C., Socher, R., and Rajani, N. F. Bertology meets biology: Interpreting attention in protein language models, 2020. URL https://arxiv.org/abs/2006.15222.

Wang, Z., Combs, S. A., Brand, R., Calvo, M. R., Xu, P., Price, G., Golovach, N., Salawu, E. O., Wise, C. J., Ponnapalli, S. P., and Clark, P. M. Lm-gvp: an extensible sequence and structure informed deep learning framework for protein property prediction. *Scientific Reports*, 12(1), April 2022. ISSN 2045-2322. doi: 10.1038/s41598-022-10775-y. URL http://dx.doi.org/10.1038/s41598-022-10775-y.

Watson, J. L., Juergens, D., Bennett, N. R., Trippe, B. L., Yim, J., Eisenach, H. E., Ahern, W., Borst, A. J., Ragotte, R. J., Milles, L. F., et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976): 1089–1100, 2023.

Weissenow, K., Heinzinger, M., and Rost, B. Protein language-model embeddings for fast, accurate, and alignment-free protein structure prediction. *Structure*, 30(8):1169–1177, 2022.

Yang, K. K., Fusi, N., and Lu, A. X. Convolutions are competitive with transformers for protein sequence pretraining. *Cell Systems*, 15(3):286–294, 2024.

Yu, H., Deng, H., He, J., Keasling, J. D., and Luo, X. Unikp: a unified framework for the prediction of enzyme kinetic parameters. *Nature communications*, 14(1):8211, 2023.

Yuan, Y., Luo, H., and Tian, Y. Lmprotein: a protein language model based framework for protein structural property prediction. *Physical Chemistry Chemical Physics*, 28(2):1747–1758, 2026. ISSN 1463-9084. doi: 10.1039/d5cp01861g. URL http://dx.doi.org/10.1039/D5CP01861G.

Zhang, N., Bi, Z., Liang, X., Cheng, S., Hong, H., Deng, S., Lian, J., Zhang, Q., and Chen, H. Ontoprotein: Protein pretraining with gene ontology embedding, 2022. URL https://arxiv.org/abs/2201.11147.

Zhang, X., Hu, X., Zhang, T., Yang, L., Liu, C., Xu, N., Wang, H., and Sun, W. Plm_sol: predicting protein solubility by benchmarking multiple protein language models with the updated escherichia coli protein solubility dataset. *Briefings in Bioinformatics*, 25(5), July 2024. ISSN 1477-4054. doi: 10.1093/bib/bbae404. URL http://dx.doi.org/10.1093/bib/bbae404.

Zhang, Y., Huang, C., Wang, Y., Li, S., and Sun, S. Clgnn: Contrastive learning and graph neural network for protein–ligand binding affinity prediction. *Journal of Chemical Information and Modeling*, 65(4):1724–1735, 2025.

A. Datasets

We evaluate ProtEGNN on seven datasets spanning three protein task regimes: protein property prediction, mutational-effect prediction, and protein function annotation. Together, these tasks probe global and local sequence–structure relationships under both sequence-diverse and mutation-centric settings making them a comprehensive testbed for assessing the value of explicit 3D geometry.

A.1. Protein Property Prediction

To assess the prediction of intrinsic protein properties that depend on global structure and physicochemical context, we evaluate solubility and stability.

Solubility We use the PSI Biology solubility dataset (Bhandari et al., 2020), curated and cleaned by (Thumuluri et al., 2021), containing 11,226 *E. coli*–expressed proteins labeled as soluble or insoluble. We follow the standard five-fold, label-balanced cross-validation protocol with sequence identity capped at 25%. To evaluate out-of-distribution generalization, we additionally test on the independent Price dataset (Price et al., 2011) which contains 1323 highly expressed proteins. This dataset also does not share any sequences with identity greater than 25% to the PSI Biology dataset, ensured using USEARCH (Edgar, 2010). Solubility is a graph-level binary classification task with ROC_AUC as the metric.

Stability We evaluate protein stability using two complementary datasets that probe distinct notions of thermostability. The first is the Meltome Atlas dataset (referred to as Meltome in results), which contains $\sim$ 23,300 protein sequences with melting temperatures (T_m) measured via mass spectrometry. T_m provides a continuous-valued proxy for intrinsic thermostability and reflects both local packing interactions and long-range structural organization. We follow the train-test splits from the FLIP benchmark (Dallago et al., 2021), which enforces redundancy reduction by clustering sequences at 20% pairwise sequence identity. Predictions on Meltome are formulated as a graph-level regression task and evaluated using SPR.

In contrast, the dataset introduced by Rocklin et al. (2017) (referred to as Stability in results) from the TAPE (Rao et al., 2019) benchmark, containing $\sim$ 69,000 records, measures stability indirectly through protease resistance of de novo–designed mini-proteins, capturing relative fitness within local mutational neighborhoods rather than absolute thermodynamic stability. We adopt the standard TAPE split, where training and validation sets are drawn from four rounds of experimental design, and the test set consists of single-point mutants (Hamming distance 1) around 17 selected high-performing designs. This task is formulated as a graph-level regression problem analogous to learning mutational fitness landscapes evaluated using SPR.

A.1.1. Mutational Effect Prediction

Fine-tuned PLMs have been shown to perform well on mutational effect prediction, as they implicitly capture evolutionary constraints and co-variation patterns that correlate strongly with fitness changes (Glaser & Brägelmann, 2025). Given this strong baseline, our aim was to test whether explicit geometric modeling with EGNNs can be competitive in this regime. In our setup, we provide the ProtEGNN with a single wild-type structure and vary only the residue embeddings for each mutant sequence, keeping the geometry fixed. This reflects a realistic experimental scenario, where typically only one high-quality structure is available and variant-specific structures are either unavailable or indistinguishable due to the limited sensitivity of structure predictors to single or few mutations.

GFP From the TAPE benchmark (Rao et al., 2019), we use the fluorescence dataset (GFP), a regression task where $\sim$ 54,000 sequences are mapped to their log-fluorescence intensity. Training variants are within Hamming distance 3 of the wild type, while test variants contain four or more mutations, explicitly probing generalization to unseen mutation combinations. GFP is a node-level regression task measured by SPR.

GB1 We evaluate on the GB1 landscape from the FLIP benchmark (Dallago et al., 2021), which measures binding affinity of $\sim$8,700 variants of the immunoglobulin-binding domain of Protein G. Mutations are introduced at four positions, producing a highly epistatic landscape. We use the standard 3-vs-rest split, where single, double, and triple mutants are used for training and more distant variants for testing. GB1 is a node-level regression task measured by SPR.

A.1.2. Protein Function Annotation

Subcellular Localization To evaluate functional annotation, we use the harder version of DeepLoc dataset (Almagro Armenteros et al., 2017), SetHard, developed by (Stärk et al., 2021). This dataset, $\sim$ 11,700 records, frames subcellular localization as a 10-class per-protein classification task. Sequences are filtered to remove redundancy above 20% identity

between splits using MMseqs2, ensuring that predictions require generalization beyond close homologs. Subcellular localization (referred to as Sub-loc in results) is a graph-level multi-classification task measured by Accuracy (ACC).

B. Hyper-parameters

Table 6. Hyperparameters in `ProtEGNN`.

Parameter	Default	Possible values	Explanation / where used
batch_size	2	int{2,4,8,16}	PyG DataLoader batch size.
epochs	5	int{5,10,20,30,40,50,60}	Training epochs for cross validated runs.
learning_rate	1e-5	float(log-uniform1e-4..1e-2)	AdamW learning rate.
weight_decay	1e-3	float({1e-6,1e-5,3e-5,1e-4})	AdamW weight decay.
scheduler	CosineAnnealingWarmRestarts	{CosineAnnealingLR,CosineAnnealingWarmRestarts}	LR scheduler choice (eta_min is set to LR/100 in both).
val_maximize	SPR	ACC, ROC_AUC, SPR	Early-stopping/model-selection metric.
early_stopping_delta	0.01	float≥ 0	Minimum improvement threshold for early stopping.
angstroms	10	float{5,10,20})	Distance cutoff (Å) for creating edges
model_layers	3	int({2,3,4,5,6,8})	Number of EGNN message-passing layers.
activation	silu	{silu,relu}	Nonlinearity in EGNN and heads.
egnn_hidden_nf	128	int({64,96,128,142,256,320})	Hidden width inside EGNN layers.
egnn_output_dim	96	int({64,96,128,256})	Output node embedding dim from EGNN backbone.
egnn_norm	None	None or layernorm	If layernorm, inserts LayerNorms in EGNN MLPs.
use_adaptive_pooling	True	{True,False}	If True, learns weights over mean/max/sum pooling (graph tasks).
pool_attention_hidden	[64]	int({32,64,128,256})	Hidden dim of pooling-attention MLP.
fc_head_hidden_dims	[256,128,64]	list[int]	Hidden layer sizes for graph-level head (MLPHead).
node_head_hidden_dims	[64]	list[int]	Hidden layer sizes for node-level head (MLPHead).
head_dropout	0.15	floatin[0,1](0.1--0.3)	Dropout in graph-level head.
node_dropout	0.3	floatin[0,1](0.2--0.4)	Dropout in node-level head.
pos_weight	0.67	{0.5,0.67,0.8})	Positive-class weight for BCEWithLogits (binary graph tasks).
use_focal_loss.enabled	False	{True,False}	Switch focal loss for binary classification.
use_focal_loss.focal_alpha	0.25	{0.25,0.5}	Focal loss α (binary).
use_focal_loss.focal_gamma	2.0	{1.0,2.0}	Focal loss γ (binary).
use_huber_loss.enabled	False	{True,False}	Switch Huber loss for regression.
use_huber_loss.huber_delta	1.0	{0.5,1.0,1.5}	Huber δ threshold.
use_label_smoothing.enabled	True	{True,False}	Enables label smoothing loss for multiclass.
use_label_smoothing.label_smoothing_epsilon	0.1	{0.05,0.1,0.15}	Label smoothing strength ϵ.
use_focal_loss_multiclass.enabled	False	{True,False}	Switch focal loss for multiclass.

Parameter	Default	Possible values	Explanation / where used
`use_focal_loss_multiclass.gamma`	2.0	{`1.5,2.0,2.5,3.0`}	Multiclass focal γ.