JUST ADD STRUCTURE: PROTEIN LANGUAGE MODELS COMBINED WITH STRUCTURAL EQUIVARIANCE EXCEL AT PROTEIN TASKS

Anonymous authors
Paper under double-blind review

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 INTRODUCTION

Protein Language Models trained on large-scale sequence data are widely used for protein analysis and 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). PLMs capture evolutionary and functional constraints from sequence and have even been shown to implicitly reflect aspects of protein structure, including fold classes and residue–residue contacts (Rao et al., 2020; Vig et al., 2020; Lombard et al., 2024). However, this structural signal is indirect, arising from sequence co-variation rather than explicit geometric reasoning, even though many protein properties, such as solubility and thermodynamic stability, are fundamentally governed by spatial interactions (Meng et al., 2025).

State-of-the-art protein models typically rely on sophisticated fine-tuning or parameter-efficient adaptation of increasingly large protein language models (Jiang et al., 2024; Schmirler et al., 2024). Many approaches further employ ensembling (Thumuluri et al., 2021) or mix-and-match architectures that combine multiple PLMs and downstream heads (Yuan et al., 2026; Zhang et al., 2024), often requiring thousands of experimental configurations to perform well (Bikias et al., 2025). Such adaptation-heavy pipelines are costly, difficult to reproduce, and can induce catastrophic forgetting, degrading the general representations learned during pretraining (Heinzinger et al., 2024). Current approaches emphasize elaborate adaptation of Protein Language Models (PLMs), treating sequence representations as the primary lever for improvement while underutilizing principled geometric modeling. Even when structure is incorporated, it is typically done indirectly via handcrafted features or discretized 3D tokens rather than continuous, symmetry-aware geometry (Su et al., 2023a; Li et al., 2024). As a result, many methods are costly to train yet still lack explicit

structural inductive biases. We challenge this fine-tune-first paradigm and argue that combining frozen PLM representations with explicit geometric reasoning provides a more direct and efficient path to improved protein task performance. Please see A.1 for an overview of related approaches.

In this work, we introduce ProtEGNN, a simple multi-modal sequence and structure model that pairs static PLM residue embeddings with an Equivariant Graph Neural Network (EGNN) (Satorras et al., 2021). By integrating sequence-derived features with explicit 3D geometry, ProtEGNN achieves competitive or superior performance across diverse benchmarks while remaining substantially more parameter efficient than leading baselines.

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 match 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 METHODOLOGY

2.1 EQUIVARIANT GRAPH CONSTRUCTION

Details on how sequence and structure representations are created can be found in A.5. 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 $h^t = \{h_0^t, \ldots, h_{n-1}^t\}$, 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:

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

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

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

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

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 Equation 2 preserves permutation invariance. Equation 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 $\phi_x(m_{ij})$.

3 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. Full dataset descriptions, preprocessing steps, splits, and evaluation protocols are provided in the Appendix A.2. For detailed explanation of results, refer to A.3.

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	**60.85 ± 0.48**	**77.13 ± 0.03**	**72.62 ± 2.05**	**69.48 ± 0.12**	**89.95 ± 0.58**	**57.21 ± 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	**83.80 ± 0.36**	75.90 ± 0.05	68.01 ± 0.32	89.44 ± 0.07	65.98 ± 0.80
	ProtEGNN	**78.65 ± 0.08**	82.40 ± 0.02	**78.31 ± 2.03**	**69.13 ± 0.02**	**89.51 ± 0.30**	**66.81 ± 0.09**

Baseline Experiments Table 1 demonstrates that adding explicit geometric modeling consistently improves performance over pre-trained and fine-tuned PLM 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.

Solubility Table 2a reports solubility prediction results on the PSI-Biology dataset along with model sizes. We compare ProtEGNN against Prot-T5 (Elnaggar et al., 2020), the MSA-based ESM-MSA (Rao et al., 2021), and NetSolP (Thumuluri et al., 2021), the previous state-of-the-art solubility predictor trained specifically on PSI-Biology. Despite being orders of magnitude smaller, both ProtEGNN variants perform competitively, with ProtEGNN(ESMc) achieving the best ROC-AUC (0.78) while using over 100× fewer trainable parameters than large PLM-based baselines. To evaluate generalization, we test on the independent Price dataset (Price et al., 2011). As shown in Table 2b, ProtEGNN(ESMc) again achieves the best zero-shot performance (ROC-AUC = 0.77) with only 1.2M parameters, outperforming substantially larger and more complex methods. The smaller ProtEGNN(t6) remains competitive (0.74) and surpasses multimodal approaches such as ProtSolM (Tan et al., 2024) and PLMSol (Zhang et al., 2024), which rely on end-to-end PLM training, handcrafted features, or ensemble architectures.

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

Stability On the Meltome thermostability benchmark (Table 3a), performance depends strongly on PLM capacity: ProtEGNN(t6) underperforms (SPR = 0.61), while scaling to ESMc yields state-of-the-art performance (SPR = 0.79). Notably, ProtEGNN(ESMc) outperforms resource-intensive baselines that require extensive architectural adaptation (Bikias et al., 2025), specialized pre-training (Jiang et al., 2024), and structure-aware PLMs (Li et al., 2024; Su et al., 2023b), while training only a lightweight EGNN head on a frozen PLM, highlighting the efficiency of explicit geometric modeling. On the second stability benchmark (Table 3b) ProtEGNN(t6) achieves strong performance (SPR = 0.77), matching fine-tuned billion-parameter models despite using only 400K trainable parameters. Scaling to ProtEGNN(ESMc) further improves results (SPR = 0.82), surpassing hybrid sequence models (Yang et al., 2024; Yuan et al., 2026) and approaching the best LoRA-tuned baseline (Schmirler et al., 2024), which requires orders-of-magnitude more parameters. Together, these results show that while PLM scale benefits global thermostability prediction, explicit geometric inductive biases can substantially reduce the need for large-scale adaptation in local stability tasks.

GFP & GB1 For mutational-effect prediction, we use a single wild-type structure for all variants, fixing the underlying geometry across the landscape. Despite not including mutation-induced structural rearrangements, our equivariant geometric model captures local physical constraints that remain informative under sequence perturbations. On GFP (Table 4a), ProtEGNN(T6) achieves state-of-the-art performance (SPR = 0.70), matching LoRA-finetuned Ankh-FT (Schmirler et al., 2024) while training only 500K parameters and using features from a small 8M-parameter PLM backbone, outperforming substantially larger models such as Prot-T5 (Elnaggar et al., 2020) and Ankh3 (Alsamkary et al., 2025). On GB1 (Table 4b), both ProtEGNN(T6) and ProtEGNN(ESMc) reach SPR = 0.90, matching the best-performing 5.7B-parameter Ankh3 (Alsamkary et al., 2025) model with as few as 400K trainable parameters. In contrast, parameter-efficient fine-tuning of multi-billion–parameter ESM2-T36-FT (Schmirler et al., 2024) still underperforms without explicit geometry. Together, these results show that even incorporating a single wild-type structure enables

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

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.

Subcelluar 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

small PLMs to match the predictive power of much larger sequence-only models, without extensive fine-tuning or variant-specific structure prediction.

Subcelluar Localization Table 5 shows that ProtEGNN(t6) matches ProSST-T5 (Heinzinger et al., 2024) while training only 400K parameters and an 8M parameter small PLM backbone, whereas ProSST, a fine-tuned ProtT5 (Elnaggar et al., 2020) variant, exhibits degraded performance relative to its base model, Prot-T5 (Elnaggar et al., 2020) highlighting catestrophic forgetting through fine-tuning. Scaling to ProtEGNN(ESMc) yields ACC = 0.67 with just 500K parameters, closely approaching the best LoRA-tuned ESM2-T36 result (0.68) reported by (Schmirler et al., 2024). Overall, these results show that adding a lightweight EGNN head to frozen PLM representations is a more efficient and robust alternative to fine-tuning large PLMs, even with parameter-efficient methods.

4 Discussion

Our results show that explicit, equivariant geometric modeling can replace, and often outperform, large-scale PLM fine-tuning while being 100–1000$\times$ more parameter-efficient. This highlights a fundamental limitation of fine-tune-first strategies, which are computationally costly and can reduce model generality through task specialization. While our approach relies on predicted structures, future work should incorporate uncertainty-aware modeling for low-confidence regions and explore multi-task EGNN heads and interpretability tools. 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. More detailed discussion can be found in A.4

REFERENCES

José Juan Almagro Armenteros, Casper Kaae Sønderby, Søren Kaae Sønderby, Henrik Nielsen, and Ole Winther. Deeploc: prediction of protein subcellular localization using deep learning. *Bioinformatics*, 33(21):3387–3395, 2017.

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

Bikash K Bhandari, Paul P Gardner, and Chun Shen Lim. 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.

Thomas Bikias, Evangelos Stamkopoulos, and Sai T Reddy. 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.

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, February 2022. ISSN 1367-4811. doi: 10.1093/bioinformatics/btac020. URL http://dx.doi.org/10.1093/bioinformatics/btac020.

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. November 2021. doi: 10.1101/2021.11.09.467890. URL http://dx.doi.org/10.1101/2021.11.09.467890.

Alexandre Duval, Simon V. Mathis, Chaitanya K. Joshi, Victor Schmidt, Santiago Miret, Fragkiskos D. Malliaros, Taco Cohen, Pietro Liò, Yoshua Bengio, and Michael Bronstein. A hitchhiker's guide to geometric gnns for 3d atomic systems, 2023. URL https://arxiv.org/abs/2312.07511.

Robert C. Edgar. 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.

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. 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.

Moritz Glaser and Johannes Brägelmann. 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.

Michael Heinzinger, Konstantin Weissenow, Joaquin Gomez Sanchez, Adrian Henkel, Milot Mirdita, Martin Steinegger, and Burkhard Rost. 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.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

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

Fan Jiang, Mingchen Li, Jiajun Dong, Yuanxi Yu, Xinyu Sun, Banghao Wu, Jin Huang, Liqi Kang, Yufeng Pei, Liang Zhang, Shaojie Wang, Wenxue Xu, Jingyao Xin, Wanli Ouyang, Guisheng Fan, Lirong Zheng, Yang Tan, Zhiqiang Hu, Yi Xiong, Yan Feng, Guangyu Yang, Qian Liu, Jie Song, Jia Liu, Liang Hong, and Pan Tan. 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.

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

Sameer Khurana, Reda Rawi, Khalid Kunji, Gwo-Yu Chuang, Halima Bensmail, and Raghvendra Mall. Deepsol: a deep learning framework for sequence-based protein solubility prediction. *Bioinformatics*, 34(15):2605–2613, March 2018. ISSN 1367-4811. doi: 10.1093/bioinformatics/bty166. URL http://dx.doi.org/10.1093/bioinformatics/bty166.

Mingchen Li, Yang Tan, Xinzhu Ma, Bozitao Zhong, Huiqun Yu, Ziyi Zhou, Wanli Ouyang, Bingxin Zhou, Pan Tan, and Liang Hong. Prosst: Protein language modeling with quantized structure and disentangled attention. *Advances in Neural Information Processing Systems*, 37:35700–35726, 2024.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. 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.

Valentin Lombard, Dan Timsit, Sergei Grudinin, and Elodie Laine. 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.

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.

Yajie Meng, Zhuang Zhang, Chang Zhou, Xianfang Tang, Xinrong Hu, Geng Tian, Jialiang Yang, and Yuhua Yao. 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.

Erik Nijkamp, Jeffrey A. Ruffolo, Eli N. Weinstein, Nikhil Naik, and Ali Madani. 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.

W Nicholson Price, Samuel K Handelman, John K Everett, Saichiu N Tong, Ana Bracic, Jon D Luff, Victor Naumov, Thomas Acton, Philip Manor, Rong Xiao, Burkhard Rost, Gaetano T Montelione, and John F Hunt. 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.

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.

Roshan Rao, Joshua Meier, Tom Sercu, Sergey Ovchinnikov, and Alexander Rives. 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.

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*, pp. 8844–8856. PMLR, 2021.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences, April 2019. URL http://dx.doi.org/10.1101/622803.

Rahmatullah Roche, Bernard Moussad, Md Hossain Shuvo, Sumit Tarafder, and Debswapna Bhattacharya. 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.

Gabriel J. Rocklin, Tamuka M. Chidyausiku, Inna Goreshnik, Alex Ford, Scott Houliston, Alexander Lemak, Lauren Carter, Rashmi Ravichandran, Vikram K. Mulligan, Aaron Chevalier, Cheryl H. Arrowsmith, and David Baker. 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.

Victor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks, 2021. URL https://arxiv.org/abs/2102.09844.

Robert Schmirler, Michael Heinzinger, and Burkhard Rost. 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.

Hannes Stärk, Christian Dallago, Michael Heinzinger, and Burkhard Rost. Light attention predicts protein location from the language of life. *Bioinformatics Advances*, 1(1):vbab035, 2021.

Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. Saprot: Protein language modeling with structure-aware vocabulary. October 2023a. doi: 10.1101/2023.10.01.560349. URL http://dx.doi.org/10.1101/2023.10.01.560349.

Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. Saprot: Protein language modeling with structure-aware vocabulary. *BioRxiv*, pp. 2023–10, 2023b.

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

Vineet Thumuluri, Hannah-Marie Martiny, Jose J Almagro Armenteros, Jesper Salomon, Henrik Nielsen, and Alexander Rosenberg Johansen. 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.

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

Jesse Vig, Ali Madani, Lav R. Varshney, Caiming Xiong, Richard Socher, and Nazneen Fatema Rajani. Bertology meets biology: Interpreting attention in protein language models, 2020. URL https://arxiv.org/abs/2006.15222.

Zichen Wang, Steven A. Combs, Ryan Brand, Miguel Romero Calvo, Panpan Xu, George Price, Nataliya Golovach, Emmanuel O. Salawu, Colby J. Wise, Sri Priya Ponnapalli, and Peter M. Clark. 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.

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

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

Yongna Yuan, Hui Luo, and Yaojie Tian. 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.

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, 2022. URL https://arxiv.org/abs/2201.11147.

Xuechun Zhang, Xiaoxuan Hu, Tongtong Zhang, Ling Yang, Chunhong Liu, Ning Xu, Haoyi Wang, and Wen Sun. 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.

Yunjiang Zhang, Chenyu Huang, Yaxin Wang, Shuyuan Li, and Shaorui Sun. Cl-gnn: Contrastive learning and graph neural network for protein–ligand binding affinity prediction. *Journal of Chemical Information and Modeling*, 65(4):1724–1735, 2025.

A Appendix

A.1 Related Works

Protein Large Language Models Many predictive methods build on pretrained protein language models 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., 2023a) 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.

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).

A.2 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. 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).

A.2.1 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 Spearman's rho (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.2.2 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 ∼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.2.3 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).

A.3 Results - detailed

A.3.1 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 summarizes the effect of explicit geometric modeling across tasks and PLM scales. Overall, incorporating structure via an equivariant graph head improves performance over both pretrained and fine-tuned sequence-only baselines across all tasks and PLM backbones sizes, with the sole exception of ProtEGNN(ESMc) on the Stability dataset, where full fine-tuning achieves a higher score. For the small ESM2-T6 backbone, ProtEGNN yields consistent gains across every benchmark, demonstrating that explicit geometry can substantially enhance the predictive power of sequence representations. For the larger ESMc backbone, ProtEGNN excels over fine-tuning, achieving the best overall performance on five out of six tasks.

At the same time, the results highlight an important interaction between PLM scale and task type. Increasing PLM scale from ESM2-T6 to ESMc leads to consistent performance improvements on global protein property and function tasks such as thermostability, solubility, and subcellular localization, indicating that larger language models capture richer global and evolutionary context. In contrast, for mutational-effect prediction tasks (GFP and GB1), the smaller backbone often matches the performance of the larger model once explicit structure is introduced. This suggests that, in mutation-centric settings where the protein fold is fixed and sequence changes are local, explicit geometric modeling can largely compensate for reduced PLM capacity. Taken together, these findings indicate that while PLM scale remains beneficial, principled geometric inductive biases can substantially narrow, or even close, the performance gap.

A.3.2 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.

A.3.3 THERMOSTABILITY

On the first thermostability benchmark, Meltome, (Dallago et al., 2021), we observe that PLM capacity plays a larger role than in other tasks. Table 3a shows that the ProtEGNN(t6) variant achieves only 0.61 SPR, trailing substantially behind approaches built on large pretrained backbones. Scaling the PLM size within our framework yields a large jump: ProtEGNN(ESMc) reaches 0.79 SPR, outperforming all compared methods by a significant margin. Prot-T5-FT (Schmirler et al., 2024) attains strong performance after adapting only 3.5M out of the 1.2B parameters of Prot-T5 (Elnaggar et al., 2020), yet lacks structural bias. However, several baselines rely on resource-intensive pretraining and tuning pipelines: PLM-Fit Bikias et al. (2025) benchmarks combinations of multiple PLMs with methods such as feature extraction and bottleneck adapters across five protein engineering datasets, requiring 3,000+ experiments and PRIME Jiang et al. (2024) is pre-trained on 96M bacterial sequences annotated with optimal growth temperatures and then fine-tuned on downstream tasks.

Several methods additionally incorporate structural signals: ProSST Li et al. (2024) learnt a structure-quantized representation with disentangled sequence–structure attention, and SaProt (Su et al., 2023a) augmented PLMs with quantized 3D structural tokens to inject geometry directly into the sequence model. In comparison, our approach keeps the PLM frozen and uses its embeddings as node representations for an EGNN, achieving state-of-the-art SPR with orders-of-magnitude fewer parameters than millioins-to-billions parameter PLM based pipelines. Overall, the weak performance of ProtEGNN(t6) relative to these strong baselines, paired with the clear gains from ProtEGNN(ESMc), suggests that thermostability prediction depends heavily on rich pretrained representations and that the way structure is integrated into PLMs is consequential, motivating a shift from scale-first to geometry-aware design.

On the second stability benchmark, which measures protease susceptibility for de novo–designed mini-proteins, a different pattern emerges (Table 3b). In contrast to Meltome, the small-backbone variant ProtEGNN(t6) achieves strong performance (SPR = 0.77), matching Ankh-FT and outperforming sequence-only baselines such as CARP (Yang et al., 2024). This suggests that for this dataset, where generalization is evaluated locally around optimized designs, explicit geometric modeling can compensate for limited PLM capacity.

Scaling the backbone further improves performance: ProtEGNN(ESMc) attains an SPR of 0.82 while training fewer than one million task-specific parameters, approaching the best-performing baseline, ESM2-T36-FT (SPR = 0.84), which relies on parameter-efficient fine-tuning of a multi-billion–parameter model. Notably, the remaining gap highlights that while PEFT on very large PLMs can be beneficial, explicit geometric inductive biases allow comparable performance with orders-of-magnitude fewer trainable parameters.

Together with the Meltome results, this comparison indicates that thermostability prediction benefits from both rich pretrained representations and explicit structure, but that the relative importance of PLM scale versus geometry depends on the nature of the stability task—global across diverse proteins versus local around optimized designs.

A.3.4 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.

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 and significantly higher computational cost. 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.

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.

A.3.5 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.

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 efficient fine tuning techniques approaches incur significantly higher computational costs compared to our method, which simply adds a lightweight geometric head to frozen PLM representations. These results reinforce 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.

A.4 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$\times$ more parameter-efficient. By pairing frozen PLM residue embeddings with a lightweight $E(3)$-equivariant GNN, ProtEGNN achieves strong performance across solubility, thermostability, and mutational-effect prediction while training fewer than 1M task-specific parameters. 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

The ProtEGNN design offers two practical advantages. First, it substantially reduces training, tuning, and replication costs compared to adaptation of billion-parameter PLMs or large ensemble pipelines, both in terms of overall computational cost and memory requirements. Second, it provides a principled mechanism for injecting structural inductive biases, equivariance and continuous 3D geometry, that are difficult to recover through fine-tuning alone and appear particularly valuable on benchmarks with genuine structural diversity, such as solubility and thermostability. 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.

At the same time, our results highlight important task-dependent limits. The performance gap between ProtEGNN(T6) and ProtEGNN(ESMc) on thermostability and subcellular localization, contrasted with ProtEGNN(T6) matching state-of-the-art accuracy on mutational fitness tasks (GFP, GB1), indicates that PLM capacity matters more for some tasks than others. In particular, small frozen embeddings appear to lack certain global or evolutionary signals, such as long-range coevolution or organism-level priors, needed for fine-grained thermostability regression, whereas dense mutational landscapes benefit more from strong inductive biases and local reasoning. This observation suggests that scaling the PLM is not uniformly beneficial, and that architectural bias and task dynamics play a decisive role.

Finally, several experimental choices bound the scope of our conclusions. For mutational datasets (GFP, GB1), we reuse a single wild-type structure across all variants, reflecting the limited backbone changes produced by current structure predictors, like ESMFold, for near-neighbor mutations; however, this setup limits insight into cases where mutations induce genuine conformational change. 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.

A.5 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.

A.6 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.

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

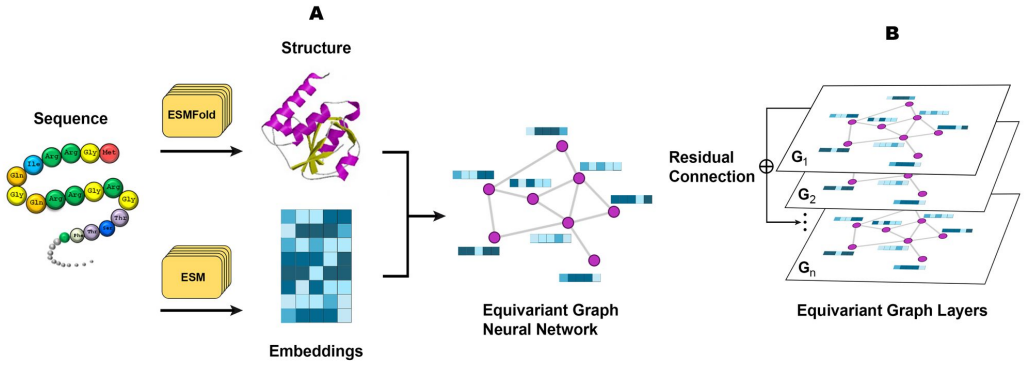

Figure 1: Integration of protein sequence and structure data using a EGNN architecture in ProtEGNN. In **Panel A**, a protein sequence is processed through two distinct pathways: ESMFold predicts the 3D structure of the protein, while ESM (ESMc or ESM2-T6) generates sequence embeddings. The predicted 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.