ELECTRO-PROT: LEARNING PROTEIN REPRESENTATIONS THROUGH ELECTROSTATIC POTENTIALS

Anonymous authors
Paper under double-blind review

ABSTRACT

Protein Language Models (PLMs) trained on sequence, as well as structure-aware variants that integrate structural information, have emerged as powerful tools to learn protein representations applicable to a wide range of downstream tasks. However, neither of these frameworks explicitly accounts for the electrostatic interactions that govern protein functions. Electrostatic potentials are the primary determinants of binding, recognition, long-range interactions, and catalysis, and serve as a cornerstone of our understanding of biophysical interactions in proteins. We investigate whether a continuous electrostatic field can serve as a pretraining target to encode general protein representations. We present Electro-Prot, a 6.7 million-parameter encoder-decoder transformer model trained to predict Poisson-Boltzmann electrostatic potentials at arbitrary points around a protein. Coordinate queries cross-attend to atoms to predict the electrostatic potential as a continuous field. Electro-Prot attains $r = 0.969$ (MAE = 0.81 kT/e) on a held-out test set. On downstream tasks, our embeddings match or outperform PLMs in interaction-centric tasks such as protein-protein interfacing and mutation induced change in free binding predictions, despite utilizing a dataset that is orders of magnitude smaller. Combining Electro-Prot and PLM embeddings improves results in structure similarity, protein-protein interface, Gene Ontology, and binding site detection. suggesting that Electro-Prot learns a complementary signal not captured in evolutionary statistics. Furthermore, we investigated protein-protein binding free energy predictions, where we outperformed PLMs in mutation-induced changes in binding affinity. Our results suggest that electrostatic pretraining provides a viable, parameter- and data- efficient signal for learning representations relevant to interaction-centric tasks.

1 INTRODUCTION

Protein representation models encode protein sequences or structures into embeddings that transfer across a wide variety of biological tasks. Protein language models (PLMs) such as ESM-2 and ESM-C learn transferable representations from evolutionary sequence statistics (Lin et al., 2023; ESM Team, 2024), whereas structure-aware variants like SaProt extend this to spatial protein geometry (Jing et al., 2021b; Su et al., 2024). The success of these models raises a question: is there a physical learning signal that can add information that is not already captured by evolutionary and structural information and can that signal be used to encode transferable protein representations?

We use electrostatics as the premise to ask this question since the electrostatic potential, ϕ, governs binding, recognition, catalysis, and contains complex long-range interactions (Honig & Nicholls, 1995; Fogolari et al., 2002). The electrostatic potential can be calculated with Poisson-Boltzmann theory, which integrates the protein charge distribution and geometry with the dielectric boundary of the solvent and the ionic screening of the solute (Dolinsky et al., 2004; Baker et al., 2001; Jurrus et al., 2018). We treat electrostatic potential prediction as a means to a representation rather than an end in itself, unlike Wu & Luo (2025). We present Electro-Prot (EP), a 6.7M-parameter encoder-decoder transformer that predicts ϕ at arbitrary coordinates, in the manner of neural field representations of continuous signals (Mildenhall et al., 2020). We report accuracy against ground truth electrostatics, and, more importantly, demonstrate the learned representations transfer to standard protein benchmarks and show that EP embeddings provide complementary information to traditional PLM embeddings.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

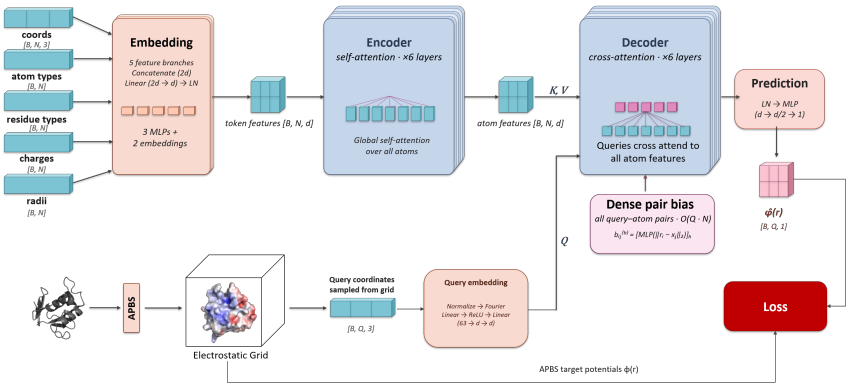

Figure 1: Electro-Prot Encoder-Decoder Pair-Bias Architecture

2 Related Work

2.1 Electrostatic Properties and Potential Prediction

Continuum electrostatics methods estimate biomolecular potentials from a molecular charge distribution and a dielectric model of the solute and solvent. PDB2PQR (Dolinsky et al., 2004) pepares structures with assigned charges and radii, and APBS (Baker et al., 2001; Jurrus et al., 2018) numerically solves the Poisson-Boltzmann problems. EP is trained using these APBS calculated potentials.

2.2 Sequence and Structure-Based Protein Representations

Protein representation models encode protein sequences or structures into embeddings that transfer across a wide variety of biological tasks. Protein language models (PLMs) such as ESM-2 and ESM-C learn transferable representations from evolutionary sequence statistics (Lin et al., 2023; ESM Team, 2024; Candido et al., 2026), whereas structure-aware variants like SaProt extend PLMs by incorporating structural information into the token vocabulary (Jing et al., 2021b; Su et al., 2024). These approaches motivate our comparisons with PLM representations and our evaluation of EP-PLM fusion to assess whether EP can encode information not already captured by evolutionary information.

3 Methods

3.1 Architecture

As shown in Figure 1, EP is an encoder-decoder transformer with a model dimension (d_{model}) of 256, 4 heads, 6 encoder layers, and 6 decoder layers. **Encoder.** Each atom in the input structure is treated as a token. The element and residue type are categorically embedded, whereas the coordinates, PDB2PQR (Dolinsky et al., 2004) partial charge, and atomic radius are encoded by small MLPs. The five encodings are concatenated, projected to d_{model}, and passed through six pre-LayerNorm layers of self-attention. Within each layer, self-attention is followed by a two layer feed-forward MLP. The encoder never sees the query, so its output, one 256-dimensional vector per atom, depends on the protein structure alone.

Query point sampling. Since the majority of calculated electrostatic potentials are in bulk solvent and so are near-zero, uniform random sampling provides a low information signal for training. As an alternative, we use dual polarity stratified sampling (DPSS), which samples 512×32 points at each training step and selects 512 queries from it: positive tail of the distribution (10%), negative tail of the distribution (10%), and the rest uniformly at random (80%).

Decoder and query-atom distance pair-bias. The query coordinate q_i is lifted by a Fourier encoding with 10 octave-spaced bands (Mildenhall et al., 2020; Tancik et al., 2020), projected to d_{model}, and passed through six cross-attention layers over the encoder's atom states. Attention is not conditioned

on pairwise geometry and is therefore blind to the query-atom separation that determines ϕ, so we add a learned scalar per pair and head inside the softmax:

$$e_{ij} = \frac{q_i \cdot k_j}{\sqrt{d_h}} + b_{ij}, \qquad b_{ij} = \mathrm{MLP}(\|q_i - x_j\|_2) \in \mathbb{R}^H, \tag{1}$$

where q_i is the i-th query point, x_j the coordinate of atom j, and k_j its key vector. The bias, b_{ij}, acts as a multiplicative, distance-dependent gate on attention. This matches the form of the target, which, to the first order, is the Coulomb sum

$$\phi(q_i) \;\approx\; \sum_j \frac{c_j}{\varepsilon \, \|q_i - x_j\|_2}, \tag{2}$$

where ϕ is the electrostatic potential in kT/e, c_j the partial charge on atom j, and ε the dielectric constant of the medium; the full Poisson-Boltzmann target adds ionic screening and dielectric-boundary terms. Attention with a learned distance kernel therefore has the same functional form as the quantity being predicted. An MLP maps the decoder output to an electrostatic potential in units of (kT/e), and is trained by minimizing an L_2 loss between the predicted and target potentials.

3.2 Electro-Prot Training Data Preparation

As described in Appendix A, we queried the RCSB Protein Data Bank (PDB) (Berman et al., 2000) for high-quality structures with at least 30 residues, yielding 190,336 entries. PDB2PQR and APBS (Baker et al., 2001; Jurrus et al., 2018) were used to produce electrostatic potentials as volumetric grids resulting in 181,089 structure potential pairs. PDBFixer (Eastman et al., 2017) is used to attempt to rescue structures where calculations initially failed. MMseqs2 (Steinegger & Söding, 2017) was then used to cluster observed chains at 30% sequence identity and 80% shorter-chain coverage, grouping connected structures into 90/5/5 train/val/test splits to prevent homolog leakage.

3.3 Protein Representations

With pretrained parameters θ frozen, each protein is encoded once. Two representations extracted from the model are considered: final encoder states $\mathbf{h}_i^{\mathrm{enc}}$ and final decoder states $\mathbf{h}_i^{\mathrm{dec}} = D(\mathbf{x}_i; \mathbf{H}^{\mathrm{enc}})$, obtained by querying the encoder representation $\mathbf{H}^{\mathrm{enc}}$ at each atom coordinate $\mathbf{x}_i$. For either representation, the per-atom states $\mathbf{h}_i$ are pooled uniformly within each residue r over its atoms $\mathcal{A}_r$ for residue-level representations, $\mathbf{z}_r$, and over the R residues for protein-level representations, $\mathbf{z}$:

$$\mathbf{z}_r = \frac{1}{|\mathcal{A}_r|} \sum_{i \in \mathcal{A}_r} \mathbf{h}_i, \qquad \mathbf{z} = \frac{1}{R} \sum_{r=1}^{R} \mathbf{z}_r \tag{3}$$

3.4 EP-PLM Fusion Techniques

EP-PLMs uses three feature level fusion strategies for combining their representations. Feature-level fusion by concatenation aligns residue embeddings from EP and PLMs through the feature dimension, so each respective model retains their original dimensions. We also utilize two learned residual fusion modules. Low-Rank Bilinear Fusion projects representations into a shared low dimensional space, fuses projections through elementwise multiplications, and maps the interaction as an additive correction to each PLM representation Kim et al. (2017). Moreover, Gated Residual Fusions utilizes a EP conditioned gate to modulate residual corrections to ESM-2, ESM-C, and SaProt representations Arevalo et al. (2017); Savarese et al. (2016). Fused representations then utilize the probe described in Appendix C to predict downstream tasks. For tasks with ligand information, ligand fingerprints are additionally concatenated.

3.5 GVP-Electro-Prot Fusion

We augment an atomic GVP-GNN (Jing et al., 2021a;b) by concatenating frozen EP decoder states with the GVP's scalar element embeddings ($9 + 256 = 265$ features per atom node). Ligand nodes receive zero EP states. We consequently widen the input normalization and projection, retain the baseline graph and the downstream architecture and training methodology, and train the GVP on affinity labels with EP remaining frozen. Feature extraction, architectural details, and experimental arms are described in Appendix H and in Fig 6.

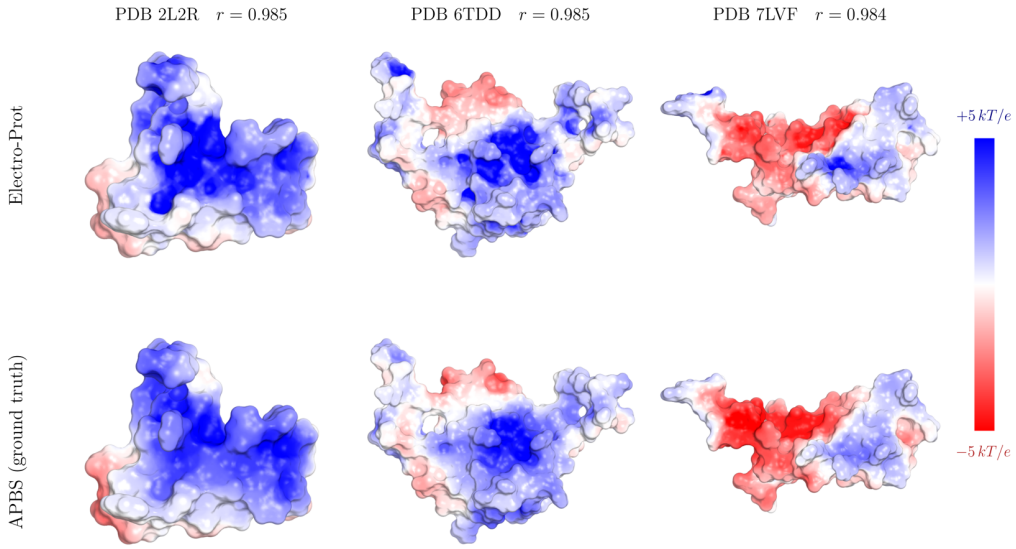

Figure 2: Electrostatic potential prediction with EP Compared to APBS Ground truth predictions.

4 Experiments and Results

We examine how EP performs electrostatic prediction and on a variety of residue and protein level representations tasks that are driven by structure, evolutionary statistics, and biophysical interactions.

4.1 Electrostatic Predictions with Electro-Prot

On 5,104 completed held-out test structures, EP reproduces APBS calculated potentials with a pointwise Pearson correlation of 0.969 and a mean per-protein absolute error of $0.810\,kT/e$; 99.1% of proteins reach a per-protein correlation above 0.90 and 17.9% above 0.98 (Appendix J). Predicted potentials recover the polarity and spatial organization of the ground-truth field Fig. 2.

4.1.1 Dataset and Task Preparations

ProteinShake. We evaluate frozen EP and PLM representations along with their fusion on eight ProteinShake (Kucera et al., 2023) tasks spanning protein function, structure, and molecular interactions. Within each comparison, we subject the models to fixed splits and common task cohorts. A single linear layer $\hat{\mathbf{y}} = \mathbf{W}\mathbf{x} + \mathbf{b}$ is trained on the fixed representations. Further information is detailed in Appendix C.1.

Pretraining study with shared corpus. We conduct additional experiments where we train EP, ESM-2 8M, and SaProt 35M with a random initialization on a common corpus of 22,669 training PDB samples to control for differences in training data during evaluation. We evaluate the frozen representations at 4, 8, 16, and 24 hours of training-loop walltime in a time-controlled study. We detail in Appendix C.2 the corpus dataset construction, held-out sets, and timing.

PPB Affinity. We evaluate frozen EP and PLM representations on PPB affinity, which combines measurements of protein–protein binding free-energy (ΔG) from ATLAS, Affinity Benchmark v5.5, PDBbind v2020, SAbDab, and SKEMPI v2.0. (Liu et al., 2024; Borrman et al., 2017; Vreven et al., 2015; Guest et al., 2021; Liu et al., 2015; Dunbar et al., 2014; Jankauskaitė et al., 2019). We used Foldseek to create structural train/validation/test splits for downstream assessment. A single linear layer is trained on the concatenated protein-level representations. Further information is detailed in Appendix C.3

Table 1: Electro-Prot and PLMs representation evaluation on the ProteinShake Benchmark Suite. Values are mean ± sample standard deviation over five seeds. **Bold** marks the best available mean and underline is second best.

	EC	GO	SC	PFam	SSim	PPI	BS	LBA
Representation	Acc	$F_{\max}$	Acc	Acc	ρ	AUROC	MCC	ρ
Protein language models								
ESM-2 8M	.782 ±.002	.372 ±.003	.622 ±.006	.514 ±.002	.509 ±.015	.692 ±.003	.251 ±.002	.289 ±.008
ESM-2 650M	.706 ±.009	**.577 ±.004**	**.734 ±.001**	**.624 ±.003**	.422 ±.030	.687 ±.005	.451 ±.004	.343 ±.019
ESM-C 600M	.736 ±.016	.530 ±.004	.633 ±.005	.543 ±.007	.359 ±.025	.659 ±.005	.469 ±.001	.336 ±.048
SaProt 35M	**.808 ±.010**	.471 ±.003	.685 ±.005	.559 ±.004	**.626 ±.021**	.765 ±.003	.414 ±.002	**.357 ±.031**
SaProt 650M	.772 ±.019	.556 ±.007	.684 ±.003	.558 ±.001	.567 ±.018	.794 ±.002	**.530 ±.001**	.333 ±.058
Electro-Prot (EP; ours)								
EP 6.7M	.632 ±.009	.282 ±.003	.295 ±.003	.249 ±.006	.559 ±.005	**.812 ±.001**	.323 ±.002	.350 ±.022

SKEMPI 2.0. We predict the changes in protein-protein binding free energy induced by mutations ($\Delta\Delta G$, kcal/mol) using the SKEMPI 2.0 dataset (Jankauskaitė et al., 2019) along with three-fold cross-validation using PDB identifier groupings. A single linear layer is trained on the concatenated protein-level representations. FoldX RepairPDB and BuildModel prepare paired wild-type and mutant structures. (Delgado et al., 2019)

ATOM3D. We used the PDBBind-derived ATOM3D ligand binding affinity task and its stricter 30% protein sequence-identity split containing 3,507 training, 466 validation, and 490 test complexes(Townshend et al., 2021; Wang et al., 2004; Liu et al., 2015). The GVP model receives the supplied pocket and ligand atoms and predicts the affinity, $pK = -\log_{10}(K)$ with K being the binding affinity in molar units. RMSE is reported in pK units.

4.1.2 ProteinShake Evaluation

We examine how EP performs on a variety of residue and protein level representation tasks and datasets provided by ProteinShake in Table 1. Our eight tasks span protein-level representations such as enzyme class, protein family, and GO annotation (EC, GO, and PFam), structural classification and similarity (SC and SSim), ligand binding affinity (LBA), as well as residue-level protein-protein interface and binding-site prediction (PPI and BS).

Result. EP is competitive on tasks whose underlying mechanisms are structural or interaction-centric. On interaction-centric tasks in particular, EP outperforms other models despite being substantially more data- and parameter-efficient. On PPI, EP achieves an AUROC of 0.81, compared with 0.79 for the best PLM. On BS, EP is again stronger with an MCC of 0.32 versus 0.25 for the best PLM. On LBA, EP attains a Spearman correlation of $\rho = 0.56$, outperforming ESM at 0.51, although SaProt remains the strongest model on this task at 0.63.

In contrast, on tasks driven primarily by evolutionary statistics, PLMs continue to outperform EP. This distinction supports our hypothesis that EP captures structural and interaction-related signals that are complementary to, rather than redundant with, the evolutionary information learned by PLMs. This result may be confounded by the fact that the PLM baselines are trained on a vast corpus of data, which includes test set leakage, where as EP does not.

4.1.3 Pretraining Efficiency on a Shared Protein Corpus.

Previous PLMs have been trained on a wide corpus of data, including data very similar used in the test set of our downstream tasks. In order test the effects of potential data leakage. EP was trained on 160k proteins while ESM-2 8M and Saprot 35M were trained on 65m and 40m proteins respectively. In order to remove effects of data leakage, we conduct a set of experiments where we retrain ESM-2 8M and SaProt 35M with the same data and training time constraits as EP.

Results. Under these controls, Electro-Prot strongly outperforms the baseline models on all two-body interaction-based tasks. For binding-site prediction, EP achieves an MCC of 0.281, compared with 0.161 and 0.240 for ESM-2 8M and SaProt 35M, respectively. A similar pattern holds for LBA

Table 2: PPB-Affinity binding free-energy (ΔG) prediction. Overall evaluates a probe trained on all sources, each dataset column evaluates a separate probe trained only on that source. Values are mean $\pm$ sample standard deviation over five seeds. **Bold** marks the best mean and underlining the second-best mean within each column and metric. Pearson, Spearman, and RMSE are reported in Appendix Table 7.

Model	Overall	ATLAS	Affinity Benchmark v5.5	PDBbind v2020	SAbDab	SKEMPI v2.0
Test measurements	2,251	108	30	642	185	1,286
(MAE ↓ (kcal/mol)						
ESM-2 8M	1.874 ±.084	1.562 ±.602	1.962 ±.460	1.764 ±.021	1.570 ±.062	2.669 ±.454
ESM-2 650M	1.980 ±.247	1.438 ±.236	**1.703 ±.253**	1.732 ±.035	1.661 ±.140	2.957 ±.622
ESM-C 600M	2.391 ±.169	1.972 ±.460	1.757 ±.108	1.808 ±.017	1.692 ±.212	2.422 ±.266
SaProt 35M	1.780 ±.051	1.297 ±.381	1.955 ±.130	1.751 ±.039	2.313 ±.049	2.574 ±.770
SaProt 650M	**1.770 ±.024**	1.411 ±.122	2.032 ±.251	**1.730 ±.023**	1.589 ±.206	**2.260 ±.171**
Electro-Prot 6.7M	2.174 ±.040	**1.043 ±.164**	2.378 ±.126	1.840 ±.010	**1.442 ±.068**	2.487 ±.320

and PPI, where EP outperforms the second-best model ($0.35/0.80$ vs. $0.32/0.71$)(Fig. 4). For tasks such as EC and GO, matching the pretraining data and training budget substantially reduces the performance gap between EP and the PLM baselines shown in the above section. For instance the gap between ESM-2 and EP was changed form 0.150 (EC accuracy) to -0.068. The same is shown for GO task, where the gap changed form 0.090 to 0.012 ($F_{\max}$). Together, these results suggest that EP is particularly effective at learning interaction- and structure-related signals, and that its competitiveness relative to PLMs becomes more apparent when pretraining data and compute are controlled.

4.1.4 Predicting Protein-Protein Binding Free Energy

To further assess the interaction-centric information EP learns through electrostatics, we evaluate how our model performs on protein-protein binding free energy (ΔG) predictions using the PPB-Affinity dataset. We evaluate a linear probe trained on the entire PPB Affinity dataset and individual probes trained on each distinct dataset, using the same structural train/validation/test splits across the representations.

Results. For binding free-energy (ΔG) prediction, EP achieves the lowest MAE on ATLAS and SAbDab, but PLMs remain stronger on the pooled PPB-Affinity task (Table 2). On ATLAS and SAbDab, EP achieves the lowest MAE at 1.043 and 1.442 kcal/mol respectively, compared with 1.297 for SaProt 35M and 1.570 for ESM-2 8M, the strongest PLMs on each dataset. The picture flips on Affinity Benchmark, PDBbind, and SKEMPI absolute binding affinity, where EP scores an MAE of 2.378, 1.840, and 2.487 kcal/mol respectively, compared with 1.703 for ESM-2 650M and $1.730/2.260$ for SaProt 650M, although EP outperforms both ESM-2 models and SaProt 35M on SKEMPI. On the overall dataset, EP reaches an MAE of 2.174 kcal/mol, which is higher than all PLMs except ESM-C 600M, which scored 2.391 kcal/mol. Appendix Fig 5 compares predicted and measured ΔG across all five datasets for EP and PLMs. The results show EP's advantages in absolute binding free-energy prediction are dataset dependent rather than consistent across PPB Affinity.

4.1.5 Predicting Mutation Induced Changes of Protein-Protein Binding

Predicting mutation-induced changes in protein-protein free binding energy connects protein representations to experimentally measured binding affinities. We evaluate whether Electro-Prot representations captures information about the energetic consequences of mutations by predicting experimentally measured $\Delta\Delta G$,

Results. As shown in Table 3, EP achieves the lowest prediction error on mutation-induced $\Delta\Delta G$ predictions (Table 3). EP achieves an MAE of 1.348, RMSE of 1.889 kcal/mol, and 0.369 Spearman ρ compared to 1.445, 2.046, and 0.359 for ESM-C 600M, the strongest PLM on these metrics. For Pearson r, EP scores 0.446, slightly below ESM-C 600M at 0.461, but above both ESM-2 models ($0.375/0.385$ for 8M/650M) and both SaProt models ($0.305/0.373$ for 35M/650M). Our results

Table 3: EP performance against ESM-2 8M and SaProt 35M on SKEMPI 2.0 mutation-induced $\Delta\Delta G$ prediction. The metrics are computed from pooled held-out predictions across three folds for each seed. Values are mean ± sample standard deviation over five seeds. **Bold** marks the best mean and underlining the second-best mean in each column. MAE and RMSE are reported in kcal/mol.

Representation	Pearson r ↑	Spearman ρ ↑	MAE ↓	RMSE ↓
Protein language models				
ESM-2 8M	.375 ±.011	.306 ±.005	1.467 ±.015	2.084 ±.038
ESM-2 650M	.385 ±.030	.344 ±.023	1.523 ±.060	2.247 ±.167
ESM-C 600M	**.461 ±.007**	.359 ±.021	1.445 ±.018	2.046 ±.043
SaProt 35M	.305 ±.006	.281 ±.014	1.484 ±.022	2.142 ±.053
SaProt 650M	.373 ±.018	.302 ±.011	1.521 ±.030	2.242 ±.064
Electro-Prot (EP; ours)				
EP 6.7M	.446 ±.017	**.369 ±.009**	**1.348 ±.010**	**1.889 ±.017**

Table 4: Electro-Prot fusions with PLMs on the ProteinShake benchmark. Each cell shows the fusion score averaged over five seeds, followed by the mean signed change from the matching standalone PLM score on the same seeds. Positive values + indicate an increase in the metric. Bold marks the best available fusion score within each task.

PLM	EC	GO	SC	PFam	SSim	PPI	BS	LBA
	Acc	$F_{\max}$	Acc	Acc	ρ	AUROC	MCC	ρ
Feature-level fusion by concatenation								
ESM-2 8M	.776 (−0.007)	.398 (+0.026)	.598 (−0.024)	.486 (−0.029)	.578 (+0.069)	.802 (+0.109)	.356 (+0.105)	**.361** (+0.072)
ESM-2 650M	.703 (−0.002)	.577 (+0.000)	**.729** (−0.005)	**.614** (−0.010)	.502 (+0.080)	.785 (+0.098)	.471 (+0.019)	.343 (+0.000)
ESM-C 600M	.738 (+0.002)	.528 (−0.001)	.631 (−0.002)	.518 (−0.025)	.457 (+0.098)	.777 (+0.118)	.490 (+0.020)	.350 (+0.014)
SaProt 35M	.787 (−0.021)	.468 (−0.003)	.663 (−0.022)	.542 (−0.017)	.628 (+0.002)	.829 (+0.064)	.444 (+0.029)	.336 (−0.021)
SaProt 650M	.794 (+0.022)	.559 (+0.002)	.673 (−0.011)	.553 (−0.005)	.588 (+0.021)	.834 (+0.040)	**.533** (+0.003)	.358 (+0.025)

suggest that EP captures mutation-sensitive physical energetic information, further supporting that electrostatic pretraining learns interaction centric representations.

4.2 Representation Fusions and Complementarity with PLMs

Combining learned sequence features with structural predictors has precedent in DeepFRI and studies integrating pretrained PLM representations into geometric networks (Gligorijević et al., 2021; Wu et al., 2023). Our comparisons examine EP and PLM representations separately, their combination through varying aligned feature concatenation methods, and an atomic GVP with and without frozen, atom-aligned EP decoder states. These comparisons test predictive utility and complementarity.

4.2.1 Evaluating EP and PLMs Fusions on ProteinShake

Result. As shown in Table 4, fusion of EP with PLMs generally improved results across the interaction-centric ProteinShake tasks at various PLM scales. When comparing EP fusion with ESM-2, we observe improvements in all interaction-centric tasks: LBA, BS, PPI, SSim ($\Delta_fusion - PLM = +0.072/+0.000, +0.105/+0.019, +0.109/+0.098$, and $+0.069/+0.080$, respectively, for ESM-2 8M/650M). Furthermore, the best performing models for interaction-centric tasks are EP fusion models (SSim: 0.628 ± 0.017 (EP + SaProt 35M); PPI: 0.834 ± 0.002 (EP + SaProt 650M); BS: 0.533 ± 0.001 (EP + SaProt 650M); LBA: 0.361 ± 0.011 (EP + ESM-2 8M)). SC and PFam degrade, while EC shows mixed changes ($\Delta_fusion - PLM = -0.021$ to $+0.022$), whereas GO changes by $+0.026/+0.000$ for ESM-2 8M/650M. Shuffle experiments (Appendix E.2) support a contribution from EP representations beyond increasing the dimensionality of the probe input for several tasks and PLM partners. This effect is particularly clear for ESM-2 8M, whose parameter count is similar to that of EP (Δ(real fusion − shuffled): LBA $+0.037$, BS $+0.107$, PPI $+0.112$, SSim $+0.020$, GO $+0.037$). Low-rank bilinear and PLM-conditioned gated residual corrections achieve a higher GO and SSim scores than concatenation across all PLMs, with the largest SSim gain observed for ESM-C 600M ($\Delta_{fusion-concat} = +0.061$, gated fusion)(Appendix Table 5).

For ESM-2 650M, gated fusion increases mean PPI performance by $+0.042$, while bilinear fusion increases mean EC accuracy by $+0.061$ relative to concatenation. These gains suggest that a learned fusion utilizes interactions between EP and PLM representations that concatenation doesn't capture.

4.2.2 Evaluating Performance of GVP-GNN Fusion with Electro-Prot Representations on ATOM3D Ligand Binding Affinity Task

We evaluate whether frozen EP representations improve an atomic geometric predictor (Geometric Vector Perceptron or GVP-GNN) of protein-ligand binding affinity. The comparison includes an established GVP-GNN baseline, its augmentation with EP features (5-epoch and 20-epoch fusion models), an EP-only ridge regression readout, and controls addressing direct chemical inputs, pretraining, and feature randomization to confirm EP feature importance as further described in Appendix H.2, Table 9. Using the exact architecture and training methodology from the original GVP-GNN model release, we reproduced the GVP baseline RMSE (1.5717 ± 0.0960 vs. the literature's 1.594 ± 0.073) for ATOM3D (Jing et al., 2021b;a).

GVP-EP Fusion Results (Fig. 7a). Frozen EP protein representations improve the a GVP-GNN model's performance on ATOM3D LBA binding affinity predictions dependent on the checkpoint's training duration. Compared to the GVP baseline, a GVP fusion model using the EP representations from 5-epochs of pretraining (GVP + EP-5) improves the mean prediction error (1.5717 to 1.4888 RMSE), mean Pearson correlation (0.3878 to 0.5054), and Spearman ranking (0.3851 to 0.4876). However, the 20-epoch (GVP + EP-20) fusion model improves correlations without reducing RMSE suggesting that overtraining may result in less informative representations. Both GVP fusion models reduce the standard deviation across seeds compared to the GVP-only baseline.

GVP-EP 20 epoch Model Comparisons (Fig. 7b). Across the 10 training seeds, the GVP + EP-20 model achieved a higher mean Pearson correlation than GVP-only baseline (0.4381 versus 0.3853) however, it also scored a higher RMSE (1.6002 versus 1.5612). Direct charge/radii inputs yielded comparable performance (RMSE 1.5839; Pearson 0.4109). Pretrained EP-20 outperformed the tested frozen random EP initialization (RMSE 1.7663; Pearson 0.2719). Thus, pretraining improved performance relative to this random-feature control, but EP-20 did not establish an advantage over GVP or direct chemical inputs across all metrics.

Pocket-only (EP-5 and EP-20) linear readouts (Fig. 7c). We averaged the 256 embeddings for each atom within the residue and then averaged the 256 embeddings for each residue to obtain a 256-dimensional pocket vector per protein complex. We implemented training-set standardization and validation MS selected regularization for the Ridge readouts ($\alpha = 100$ for EP-5 and 0.1 for EP-20) as described in detail in Appendix H.3. EP-5 achieved test RMSE 1.5311 and Pearson correlation 0.4203, compared with 1.5717 and 0.3878 for the five-seed GVP baseline. EP-20 ridge regression performed worse (RMSE 1.6745; Pearson 0.3096). These results suggest that the affinity related information from this linear readout is dependent on the checkpoint training duration. Ridge results represent single deterministic fits, rather than seed-averaged estimates. In these linear readouts, no ligand descriptor or context was provided however, the pocket definition depends on the complex containing the ligand. Hence, these ridge regression evaluations were not ligand-independent.

4.3 Representations Ablations

Results. EP's Encoder-Decoder representation (final decoder states $\mathbf{h}_i^{\text{dec}} = D(\mathbf{x}_i; \mathbf{H}^{\text{enc}})$) improve, compared to the Encoder-only representation, at all metrics across all tasks within the ProteinShake and SKEMPI experiments (Table 8a,b). For the PPB-Affinity task, Encoder-Decoder once again dominates by producing the least MAE for all datasets except Affinity Benchmark v5.5. For the Overall dataset that is a combination of all the datasets in this task, Encoder only representation performs comparably with the Encoder-Decoder representation with a slight edge by producing the least MAE (Table 8c). These results suggest that the atom-coordinate decoder queries can improve the accessibility of these interaction centric information.

5 DISCUSSION

Electrostatic pretraining transfers selectively. EP leads the evaluated standalone PLMs on PPI, beats ESM-2 and C on SSim and LBA (Table 1, Appendix Fig. 4). Moreover, achieves the lowest MAE and RMSE for mutation-induced $\Delta\Delta G$ (Table 3). However, PLMs generally remain stronger on EC, GO, SC, and PFam. PPB-affinity gains for EP are dependent on the dataset; EP leads on ATLAS and SAbDab datasets but trails the PLMs on the others (Table 2). The encoder only states degrade Electro-prot performance on all ProteinShake tasks except EC with noticeable losses in PPI and BS. These results suggest that EP generally performs best when labels depend on biophysical interactions and under-performs when they are defined by evolutionary relationships. The electrostatic objective therefore captures additional signal from interaction-centric physics rather than simply recapitulating sequence evolutionary signal.

Task-dependent complementarity of Electro-Prot features for fusion models. Feature-level fusion by concatenation improves PPI, SSim, and BS for every PLM partner. GO, EC, and LBA gains depend on the partner and fusion method whereas SC and PFam consistently decline (Tables 4 and 1). Shuffled-feature controls support a contribution beyond added input dimensions, especially on PPI (Appendix E.2). These controls establish the predictive informativeness of aligned EP features but do not yet isolate electrostatic supervision from structural and chemical inputs. Furthermore, the EP-5 gains with GVP indicate that EP features can also benefit an atomic geometric predictor at protein-ligand binding affinity tasks. However, EP-20 improves the correlations without reducing the RMSE and offers no consistent advantage over direct charges/radii (Figure 7). Hence, complementarity and informative representations may not always be achieved with longer pretraining.

6 CONCLUSION

Overall, we show that pretraining protein representation models on an electrostatic-based task provides strong signal not captured by previous pretraining objectives. In interaction type downstream tasks, we show that Electro-Prot can outperform other models despite being orders of magnitude more parameter efficient. In addition, we explore strategies to fuse our embeddings with other PLMs and show additional gains in performance, suggesting that electrostatics provide strong complimentary signal not captured otherwise. Because our model is small and parameter-efficient, it can be easily integrated into existing methods and pipelines with little additional cost. Our model and its success may unlock additional performance gains in downstream tasks commonly used in drug discovery and biophysics.

Limitations and Outlook. EP is pretrained on only electrostatics, so it complements sequence models rather than replacing them outright on tasks that derive labels from homology (EC, GO, SC, Pfam). The quadratic scaling of the transformer architecture poses challenges for training on large structures. Limitations in both data and compute may contribute to the reduced electrostatic potential prediction performance observed for larger structures, an effect that warrants further investigation. In addition further, pretraining on electrostatics may benefit significantly from scaling up both data and compute, but are not explored in current work. Despite these limitations, the fusion results suggest a promising path toward more compact and informative protein representations through joint training on electrostatics, structure reconstruction, and sequence reconstruction.

REPRODUCIBILITY STATEMENT

The Methods section describes the EP architecture, pretraining objective, representation extraction, and fusion strategies. Appendix A documents electrostatic target generation and pretraining dataset construction. Appendix B describes downstream benchmarking dataset preparation, the splits for evaluation, shared corpus pretraining protocol, linear probe training, and model selection. Appendix H, Appendix E.2, and other appendices describe additional experiments and experimental controls. Electro-rot anonymized repository: https://github.com/PAS195Pitt/electro-prot-iclr2027-review

AI USE STATEMENT

We used ChatGPT-Codex and Claude to assist with software implementation and debugging, test development, provenance and reproducibility checks. AI tools also assisted with identifying relevant literature, providing feedback on experimental design, and preparing explanatory figures and tables. The authors directed these activities, critically evaluated the suggestions against the implementation and source literature, and retained control over scientific and editorial decisions.

REFERENCES

John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A. González. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*, 2017. URL https://arxiv.org/abs/1702.01992.

Nathan A. Baker, David Sept, Simpson Joseph, Michael J. Holst, and J. Andrew McCammon. Electrostatics of nanosystems: Application to microtubules and the ribosome. *Proceedings of the National Academy of Sciences*, 98(18):10037–10041, 2001.

Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, T Nand Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic Acids Research*, 28(1): 235–242, 2000. doi: 10.1093/nar/28.1.235. URL https://www.rcsb.org.

Tyler Borrman, Jennifer Cimons, Michael Cosiano, Michael Purcaro, Brian G. Pierce, Brian M. Baker, and Zhiping Weng. ATLAS: A database linking binding affinities with structures for wild-type and mutant TCR-pMHC complexes. *Proteins: Structure, Function, and Bioinformatics*, 85(5): 908–916, 2017. doi: 10.1002/prot.25260.

Salvatore Candido, Thomas Hayes, Alexander Derry, Roshan Rao, Zeming Lin, Robert Verkuil, Bryan Z. Wu, Jin Sub Lee, Elise S. Bruguera, Jehan A. Keval, et al. Language modeling materializes a world model of protein biology. *bioRxiv*, 2026. doi: 10.64898/2026.06.03.729735.

Javier Delgado, Leandro G. Radusky, Damiano Cianferoni, and Luis Serrano. FoldX 5.0: working with RNA, small molecules and a new graphical interface. *Bioinformatics*, 35(20):4168–4169, 2019. doi: 10.1093/bioinformatics/btz184.

Todd J. Dolinsky, Jens E. Nielsen, J. Andrew McCammon, and Nathan A. Baker. PDB2PQR: An automated pipeline for the setup of poisson–boltzmann electrostatics calculations. *Nucleic Acids Research*, 32:W665–W667, 2004.

James Dunbar, Konrad Krawczyk, Jinwoo Leem, Terry Baker, Angelika Fuchs, Guy Georges, Jiye Shi, and Charlotte M. Deane. SAbDab: the structural antibody database. *Nucleic Acids Research*, 42(D1):D1140–D1146, 2014. doi: 10.1093/nar/gkt1043.

Peter Eastman, Jason Swails, John D. Chodera, Robert T. McGibbon, Yutong Zhao, Kyle A. Beauchamp, Lee-Ping Wang, Andrew C. Simmonett, Matthew P. Harrigan, Chaya D. Stern, Rafal P. Wiewiora, Bernard R. Brooks, and Vijay S. Pande. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Computational Biology*, 13(7): e1005659, 2017. doi: 10.1371/journal.pcbi.1005659.

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

Federico Fogolari, Alessandro Brigo, and Henriette Molinari. The Poisson–Boltzmann equation for biomolecular electrostatics: a tool for structural biology. *Journal of Molecular Recognition*, 15(6): 377–392, 2002.

Vladimir Gligorijević, P. Douglas Renfrew, Tomasz Kosciolek, et al. Structure-based protein function prediction using graph convolutional networks. *Nature Communications*, 12:3168, 2021. doi: 10.1038/s41467-021-23303-9. URL https://www.nature.com/articles/s41467-021-23303-9.

Johnathan D. Guest, Thom Vreven, Jing Zhou, Iain Moal, Jeliazko R. Jeliazkov, Jeffrey J. Gray, Zhiping Weng, and Brian G. Pierce. An expanded benchmark for antibody-antigen docking and affinity prediction reveals insights into antibody recognition determinants. *Structure*, 29(6): 606–621.e5, 2021. doi: 10.1016/j.str.2021.01.005.

Barry Honig and Anthony Nicholls. Classical electrostatics in biology and chemistry. *Science*, 268 (5214):1144–1149, 1995.

Justina Jankauskaitė, Brian Jiménez-García, Justas Dapkūnas, Juan Fernández-Recio, and Iain H. Moal. SKEMPI 2.0: an updated benchmark of changes in protein–protein binding energy, kinetics and thermodynamics upon mutation. *Bioinformatics*, 35(3):462–469, 2019. doi: 10.1093/bioinformatics/bty635. URL https://doi.org/10.1093/bioinformatics/bty635.

Bowen Jing, Stephan Eismann, Pratham N. Soni, and Ron O. Dror. Equivariant graph neural networks for 3D macromolecular structure, 2021a. URL https://arxiv.org/abs/2106.03843. Workshop on Computational Biology at ICML.

Bowen Jing, Stephan Eismann, Patricia Suriana, Raphael J. L. Townshend, and Ron Dror. Learning from protein structure with geometric vector perceptrons. In *International Conference on Learning Representations (ICLR)*, 2021b.

Elizabeth Jurrus, Dave Engel, Keith Star, et al. Improvements to the APBS biomolecular solvation software suite. *Protein Science*, 27(1):112–128, 2018.

Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Hadamard product for low-rank bilinear pooling. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=r1rhWnZkg.

Tim Kucera, Carlos Oliver, Dexiong Chen, and Karsten Borgwardt. Proteinshake: Building datasets and benchmarks for deep learning on protein structures. In *Advances in Neural Information Processing Systems*, volume 36, pp. 58277–58289, 2023.

Zeming Lin, Halil Akin, Roshan Rao, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574.

Huaqing Liu, Peiyi Chen, Xiaochen Zhai, Ku-Geng Huo, Shuxian Zhou, Lanqing Han, and Guoxin Fan. PPB-Affinity: Protein-protein binding affinity dataset for AI-based protein drug discovery. *Scientific Data*, 11(1):1316, 2024. doi: 10.1038/s41597-024-03997-4.

Zhihai Liu, Yan Li, Li Han, Jie Li, Jie Liu, Zhixiong Zhao, Wei Nie, Yuchen Liu, and Renxiao Wang. PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics*, 31 (3):405–412, 2015. doi: 10.1093/bioinformatics/btu626. URL https://doi.org/10.1093/bioinformatics/btu626.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision (ECCV)*, 2020.

Pedro H. P. Savarese, Leonardo O. Mazza, and Daniel R. Figueiredo. Learning identity mappings with residual gates. *arXiv preprint arXiv:1611.01260*, 2016. URL https://arxiv.org/abs/1611.01260.

Martin Steinegger and Johannes Söding. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11):1026–1028, 2017. doi: 10.1038/nbt.3988.

Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. SaProt: Protein language modeling with structure-aware vocabulary. In *International Conference on Learning Representations (ICLR)*, 2024.

Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Raphael J. L. Townshend, Martin Vögele, Patricia Suriana, Alexander Derry, Alexander Powers, Yianni Laloudakis, Sidhika Balachandar, Bowen Jing, Brandon Anderson, Stephan Eismann, Risi Kondor, Russ B. Altman, and Ron O. Dror. ATOM3D: Tasks on molecules in three dimensions. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021. URL https://arxiv.org/abs/2012.04035.

Thom Vreven, Iain H. Moal, Anna Vangone, Brian G. Pierce, Panagiotis L. Kastritis, Mieczyslaw Torchala, Raphael Chaleil, Brian Jiménez-García, Paul A. Bates, Juan Fernández-Recio, Alexandre M. J. J. Bonvin, and Zhiping Weng. Updates to the integrated protein–protein interaction benchmarks: Docking benchmark version 5 and affinity benchmark version 2. *Journal of Molecular Biology*, 427(19):3031–3041, 2015. doi: 10.1016/j.jmb.2015.07.016.

Renxiao Wang, Xueliang Fang, Yipin Lu, and Shaomeng Wang. The PDBbind database: Collection of binding affinities for protein–ligand complexes with known three-dimensional structures. *Journal of Medicinal Chemistry*, 47(12):2977–2980, 2004. doi: 10.1021/jm030580l. URL https://doi.org/10.1021/jm030580l.

Fang Wu, Lirong Wu, Dragomir Radev, Jinbo Xu, and Stan Z. Li. Integration of pre-trained protein language models into geometric deep learning networks, 2023. URL https://arxiv.org/abs/2212.03447v2. Version 2.

Yongxian Wu and Ray Luo. Spacier: A dataset for modeling electrostatic poisson–boltzmann atomic solvation potentials. In *NeurIPS Workshop on Machine Learning and the Physical Sciences*, 2025.

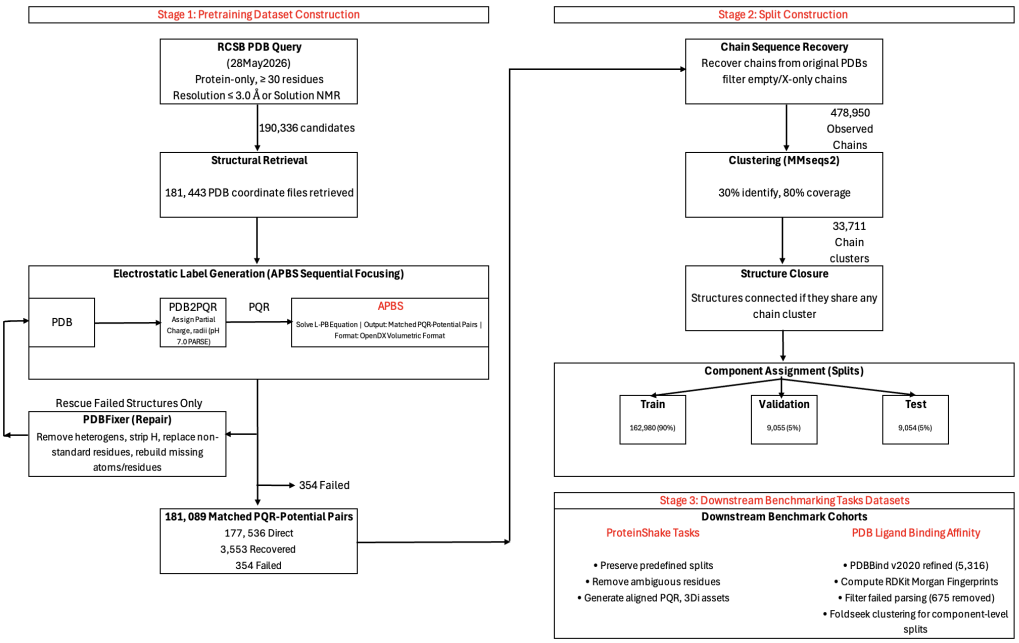

Figure 3: Dataset Construction Pipeline

A Electrostatic Target Generation and Dataset Construction

A.1 Building the Pretraining Dataset

On 28 May 2026 we queried the RCSB PDB (Berman et al., 2000) for experimental, protein-only entries with at least 30 residues, requiring either a resolution of 3.0 Å or better or that the structure was solved by solution NMR. This returned 190,336 entries, of which 181,443 were retrieved as coordinate files.

Each structure then needs an electrostatic field to serve as a training target. A deposited structure carries no charges, so PDB2PQR (Dolinsky et al., 2004) first assigns protonation states, partial charges, and atomic radii at pH 7.0 and 0.15 M monovalent salt. APBS (Baker et al., 2001; Jurrus et al., 2018) then computes the resulting potential on a 0.5Å resolution 3D grid by numerically solving the Poisson-Boltzmann equation. Grid dimensions are set from the protein size using PDB2PQR defaults. The output is a volumetric map giving the potential at every grid point, which we store in single precision together with its origin and spacing. Training queries and their labels are drawn from these grid points.

Some deposited structures are incomplete or contain unusual chemistry and fail this calculation. These are repaired once with PDBFixer (Eastman et al., 2017), which removes non-protein components, strips existing hydrogens, substitutes nonstandard residues, and rebuilds missing atoms, and are then retried. Protonation is always left to PDB2PQR so that pH is treated identically across the dataset. In total 177,536 structures succeeded directly, 3,553 more were recovered after repair, and 354 failed both attempts, giving 181,089 structures paired with an electrostatic map.

APBS input file used for electrostatic potential generation:

```
read
    mol pqr PQR_FILE
end

elec
    mg-auto
    dime GRID_DIMENSIONS
```

```
    cglen COARSE_GRID_LENGTH
    fglen FINE_GRID_LENGTH
    cgcent mol 1
    fgcent mol 1
    mol 1
    lpbe
    bcfl sdh
    pdie 2.0
    sdie 78.54
    srfm smol
    chgm spl2
    sdens 10.00
    srad 1.40
    swin 0.30
    temp 310.0
    calcenergy total
    calcforce no
    ion charge +1 conc 0.15 radius 2.0
    ion charge -1 conc 0.15 radius 1.8
    write pot dx OUTPUT_DIR_POT
end

print elecEnergy 1 end
quit
```

A.2 Splitting the Dataset

Leakage is a concern when splitting the dataset. If a test protein has a close relative in the training set, the model can score well due to memorization. We split at the level of related groups rather than individual structures. The chains were clustered with MMseqs2 (Steinegger & Söding, 2017) and similar chains were clustered together yielding 33,711 clusters from 478, 950 chains. Any two structures that share any cluster are linked together in the same group. Since chains are the clustering unit, a multi-chain structure may belong to several clusters at once. Overlapping chain clusters are merged into separate structure groups giving us a total of 17,286 groups and assign each group as a whole into one split. No cluster is divided as a result. Groups were placed largest first into whichever split was furthest from its target size, giving 162,980 training, 9,055 validation, and 9,054 test structures.

B Dataset and Task Preparation Details

C Probe configuration

Probe. Each probe is a single `nn.Linear` layer trained with AdamW at learning rate 10^{-3} and weight decay 0, for at most 300 epochs with early stopping on the task's validation metric (patience 50). Batch size is 256 for protein-level tasks and 8192 for residue-level tasks. Features are standardized to zero mean and unit variance using training-split statistics; regression targets are likewise standardized on the training split and inverse-transformed before scoring. Losses are cross-entropy (multiclass), binary cross-entropy with positive-class weighting capped at 50 for gene ontology (binary and multi-label), and mean squared error (regression). For each probe, we selected the best validation checkpoint and used a validation selected probability cutoff for binary binding-site test predictions. Runs are deterministic and are evaluated over 5 seeds.

C.1 ProteinShake Tasks and Splits

We use ProteinShake which contains datasets for enzyme classification, Gene Ontology prediction, structural classification, protein-family classification, structure-similarity prediction, binding-site detection, ligand-affinity prediction, and protein-protein interface prediction (Kucera et al., 2023). Enzyme classification uses the top-level EC category, Gene Ontology uses the molecular-function branch, and structural classification uses the SCOP family level. 6 Tasks retain ProteinShake supplied structure splits at 0.7 threshold, PDBBind was unavailable through the package so it was obtained manually. Foldseek was then utilized to split the dataset at a TM-score of 0.7 and coverage of 0.8. Proteinshake official split utilized a LDDT methodology. Morgan fingerprints at 1024 radius bits and Maccs ligand featurization were utilized for ligand featurization. A linear probe was utilized for downstream prediction for each task.

C.2 Shared Corpus and Pretraining Time Points

We constructed a pretraining corpus from the union of raw training PDB enteries obtained from the 8 ProteinShake tasks. Any PDB that was assigned to a validation or test split in any of the proteinshake tasks was excluded from the final corpus, including training split examples from other tasks. Furthermore, to prepare the dataset for PLMs, we excluded enteries with identical chain sequences to held out ProteinShake datasets, and require PLM chains to contain 1-1022 chain sequences, preserving their published setup. The corpus final count is 25,202 PDB enteries which is then separated into 22,669 training and 2533 validation examples (90/10 split) for pretraining. Each timed run ran on identical hardware utilizing one L40 for 24 hours and retained their checkpoint at 4, 8, 16, and 24 hours for downstream evaluation. Each model was pretrained using their expected input representations, Fasta for ESM models, Foldseek tokens for SaProt, and PDB2PQR charge and radii assignment for PQR files was done for EP. Proteinshake downstream were evaluated in the same was as the headline results on 5 seeds. We then evaluate on the dfownstream tasks using our common linear probe set up.

C.3 PPB Affinity.

Foldseek was used to create structural train(7880), val(1126), and test(2251) sets using a TM-score 0.7 and a shorter-to-longer chain length ratio 0.8 for downstream assessment. We concatenate the representations of the two intersecting protein partners, retaining 11,257 measurements after PQR PARSE algorithm preparation. The residue representations were averaged across interacting protein partners, and then the two partner representations are concatenated. A linear probe is trained on the concatenated protein-level representations first on the full combined dataset, then on each dataset separately.

C.4 SKEMPI 2.0 Preparation

Skempi 2.0 contains 7085 measurements across 345 pdb enteries, we excluded 287 and 186 measurements for lacking mutation dissociation constants and inequality valued affinities.FoldX 5.1 RepairPDB and BuildModel, using a single BuildModel run per variant, were utilized to construct our wildtype and mutation pairs. PDB2PQR with PH 7 and temperature of 310, hydrogen bonding network optimization and water removal. 203 structural entries were excluded due to PDB2PQR failure. final cohort consisted of 6,359 measurements across 329 PDB identifiers

D MATCHED CORPUS PRE-TRAINING DOWNSTREAM EVALUATION ON PROTEINSHAKE

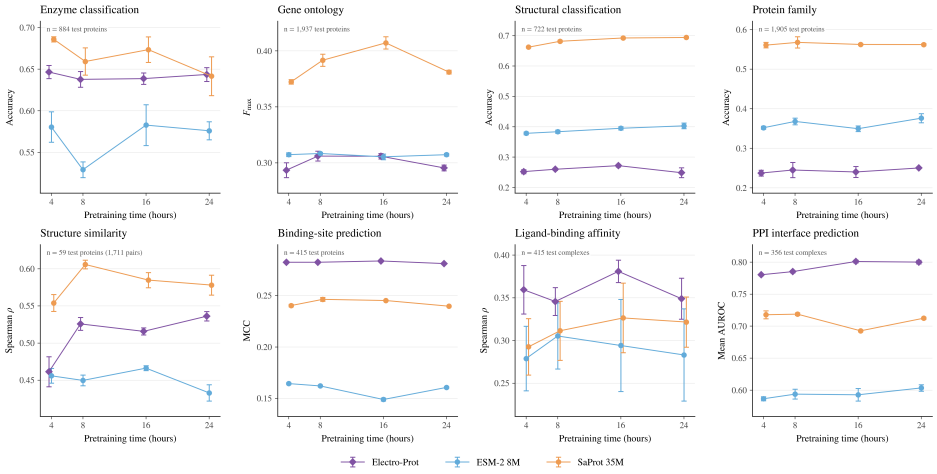

Figure 4: EP, ESM-2 8M, and SaProt 35M representation evaluation on ProteinShake Benchmark Suite at different pretraining durations. Points show means over five probe-training seeds and whiskers indicate ± one sample standard deviation.

E FUSION EXPERIMENTS

E.1 EP-PLM CONCATENATION, BILINEAR, AND GATED FUSION RESULTS

Table 5 compares the three fusion strategies across the eight ProteinShake tasks.

Table 5: Electro-Prot fusions with PLMs on ProteinShake. Each cell shows the mean score over five seeds, with the mean signed change from the matching standalone PLM below it in parentheses. Bold marks the best available fusion score within each task.

(a) Feature-level fusion by concatenation

PLM	EC	GO	SC	PFam	SSim	PPI	BS	LBA
	Acc	$F_{\max}$	Acc	Acc	ρ	AUROC	MCC	ρ
ESM-2 8M	.776 (−0.007)	.398 (+0.026)	.598 (−0.024)	.486 (−0.029)	.578 (+0.069)	.802 (+0.109)	.356 (+0.105)	**.361** (+0.072)
ESM-2 650M	.703 (−0.002)	.577 (+0.000)	**.729** (−0.005)	**.614** (−0.010)	.502 (+0.080)	.785 (+0.098)	.471 (+0.019)	.343 (+0.000)
ESM-C 600M	.738 (+0.002)	.528 (−0.001)	.631 (−0.002)	.518 (−0.025)	.457 (+0.098)	.777 (+0.118)	.490 (+0.020)	.350 (+0.014)
SaProt 35M	.787 (−0.021)	.468 (−0.003)	.663 (−0.022)	.542 (−0.017)	.628 (+0.002)	.829 (+0.064)	.444 (+0.029)	.336 (−0.021)
SaProt 650M	.794 (+0.022)	.559 (+0.002)	.673 (−0.011)	.553 (−0.005)	.588 (+0.021)	.834 (+0.040)	**.533** (+0.003)	.358 (+0.025)

(b) Low-rank bilinear residual fusion

PLM	EC	GO	SC	PFam	SSim	PPI	BS	LBA
	Acc	$F_{\max}$	Acc	Acc	ρ	AUROC	MCC	ρ
ESM-2 8M	.753 (−0.030)	.430 (+0.058)	.561 (−0.061)	.469 (−0.045)	.611 (+0.103)	.787 (+0.095)	.335 (+0.084)	.317 (+0.028)
ESM-2 650M	.764 (+0.059)	**.598** (+0.021)	.712 (−0.022)	.608 (−0.015)	.509 (+0.088)	.787 (+0.099)	.437 (−0.014)	.339 (−0.004)
ESM-C 600M	.717 (−0.019)	.553 (+0.023)	.615 (−0.018)	.521 (−0.022)	.510 (+0.151)	.753 (+0.094)	.455 (−0.014)	.306 (−0.030)
SaProt 35M	.779 (−0.029)	.507 (+0.036)	.607 (−0.078)	.517 (−0.042)	**.646** (+0.020)	.829 (+0.064)	.409 (−0.005)	.338 (−0.019)
SaProt 650M	**.821** (+0.049)	.578 (+0.022)	.641 (−0.043)	.505 (−0.052)	.611 (+0.044)	**.841** (+0.047)	.490 (−0.040)	.337 (+0.004)

(c) PLM-conditioned gated residual correction

PLM	EC	GO	SC	PFam	SSim	PPI	BS	LBA
	Acc	$F_{\max}$	Acc	Acc	ρ	AUROC	MCC	ρ
ESM-2 8M	.754 (−0.028)	.425 (+0.053)	.579 (−0.043)	.456 (−0.058)	.626 (+0.117)	.806 (+0.113)	.347 (+0.096)	.327 (+0.038)
ESM-2 650M	.722 (+0.016)	.597 (+0.021)	.718 (−0.016)	.613 (−0.011)	.528 (+0.106)	.827 (+0.139)	.461 (+0.010)	.340 (−0.003)
ESM-C 600M	.729 (−0.007)	.556 (+0.026)	.628 (−0.005)	.512 (−0.031)	.518 (+0.159)	.813 (+0.154)	.472 (+0.003)	.315 (−0.021)
SaProt 35M	.805 (−0.003)	.508 (+0.037)	.666 (−0.020)	.529 (−0.030)	.642 (+0.016)	.838 (+0.073)	.427 (+0.012)	.349 (−0.008)
SaProt 650M	.815 (+0.043)	.585 (+0.029)	.665 (−0.019)	.518 (−0.039)	.613 (+0.046)	—	.509 (−0.022)	.343 (+0.010)

E.2 FUSION EXPERIMENTS CONTROLS

To separate genuine electrostatic signal from the effect of widening the probe input, we repeated each fusion arm with EP embeddings permuted across proteins within each split. The permuted features preserve dimensionality, per-dimension statistics, and the probe capacity exactly. However, we purposefully remove only the correspondence between a protein and its own representation. Comparing the shuffled arm to the standalone PLM therefore isolates the effect of added capacity whereas, comparing the real fusion of the PLM and EP to the shuffled arm isolates the effect of electrostatic content.

Electrostatics are genuinely informative, when fused with PLMs, for PPI tasks. The experiment most sharply resolved the question of capacity versus genuine electrostatics informativeness on the PPI task

Table 6: Dimension-matched shuffled-feature control. Upper panel: absolute performance with permuted PB embeddings, mean ± sample standard deviation over five seeds. Lower panel: Δ between the real fusion of Table 1 and the shuffled arm; positive values indicate signal attributable to electrostatic content rather than to feature dimensionality.

	EC	GO	SC	PFam	SSim	PPI	BS	LBA
PLM partner	Acc	$F_{\max}$	Acc	Acc	ρ	AUROC	MCC	ρ
Shuffled Electro-Prot embeddings								
ESM-2 8M	.771 ±.007	.361 ±.011	.536 ±.003	.452 ±.002	.558 ±.010	.689 ±.003	.248 ±.002	.323 ±.046
SaProt 35M	.792 ±.002	.473 ±.004	.614 ±.005	.525 ±.002	.633 ±.014	.763 ±.004	.414 ±.001	.336 ±.027
ESM-2 650M	.722 ±.004	.573 ±.005	.711 ±.004	.627 ±.007	.441 ±.036	.687 ±.004	.450 ±.002	.340 ±.030
ESM-C 600M	.731 ±.016	.522 ±.003	.590 ±.006	.549 ±.002	.404 ±.026	.663 ±.003	.470 ±.001	.342 ±.025
SaProt 650M	.787 ±.006	.549 ±.004	.648 ±.007	.541 ±.004	.569 ±.016	—	.528 ±.002	.351 ±.043
Δ *(real fusion − shuffled)*								
ESM-2 8M	+.004	+.037	+.062	+.034	+.020	+.112	+.107	+.037
SaProt 35M	−.005	−.005	+.049	+.017	−.005	+.066	+.029	.000
ESM-2 650M	−.018	+.004	+.017	−.013	+.061	+.098	+.021	+.003
ESM-C 600M	+.006	+.007	+.041	−.031	+.053	+.114	+.019	+.008
SaProt 650M	+.007	+.009	+.024	+.012	+.020	—	+.005	+.006

as seen in Table 6. The performance of shuffling, compared to the standalone PLM (Table 1), is almost exactly equal (0.692 AUROC for shuffled versus 0.692 AUROC for standalone ESM-2 8M; 0.795 AUROC for shuffled versus 0.794 AUROC for standalone SaProt 650M). In contrast, real fusion leads to a positive increase in AUROC of 0.117 and 0.041 for ESM-2 8M and SaProt 650M respectively. Hence, additional dimensions alone contribute nothing and the entire PPI gain is attributable to electrostatic content. Structure similarity shows the same ordering for every partner, while binding site does so for four of five with smaller margins.

F PROTEIN-PROTEIN FREE BINDING ENERGY PREDICTION ON PPB-AFFINITY DATASET

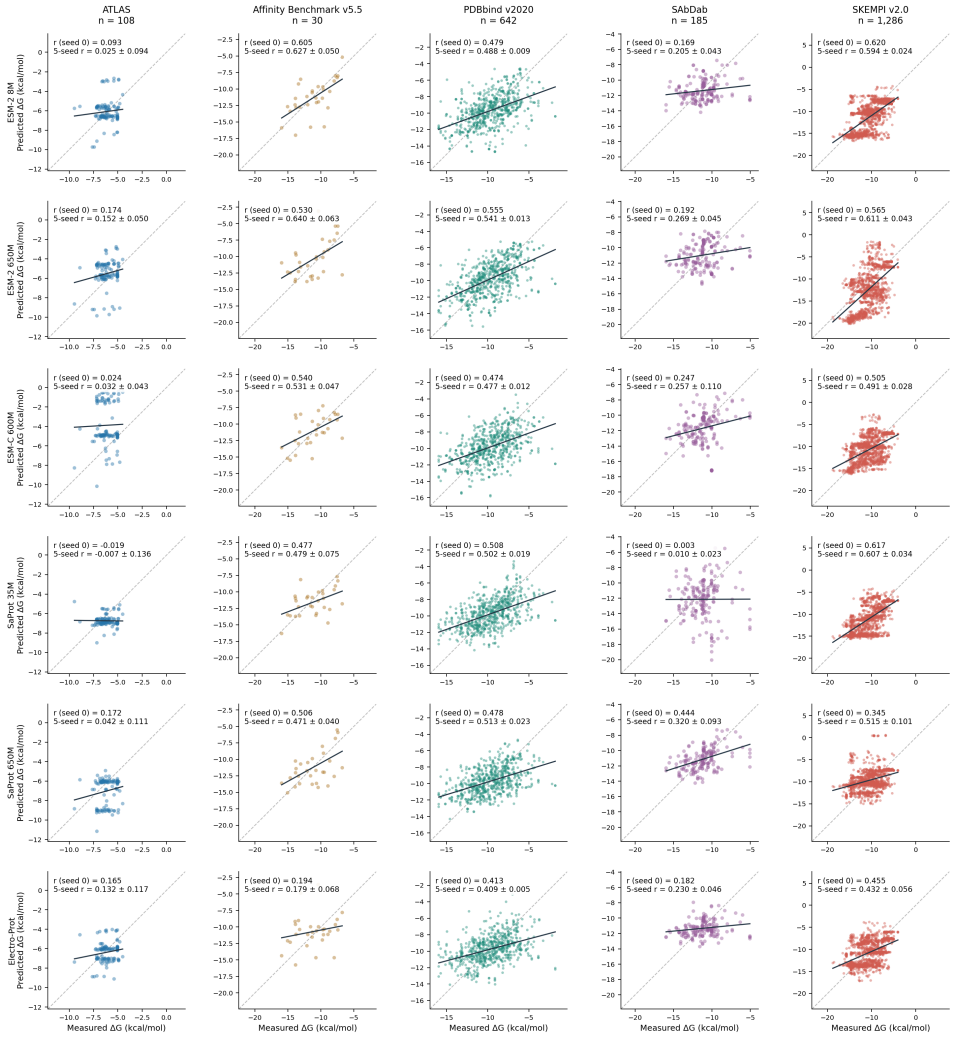

Figure 5: PPB-Affinity binding free-energy (ΔG) prediction scatterplot for probes trained separately on each dataset. Points show held-out predictions from seed 0. Solid lines show linear fits to the displayed points. Dashed lines indicate predicted = measured. Annotations report Pearson r for seed 0 and the mean $\pm$ sample standard deviation across five seeds.

Table 7: PPB-Affinity binding free-energy (ΔG) prediction. Overall evaluates a probe trained on all sources, each dataset column evaluates a separate probe trained only on that source. Values are mean ± sample standard deviation over five seeds. Metrics are Pearson, Spearman, and RMSE. **Bold** marks the best mean and underlining the second-best mean within each column and metric.

Model	Overall	ATLAS	Affinity Benchmark v5.5	PDBbind v2020	SAbDab	SKEMPI v2.0
Test measurements	2,251	108	30	642	185	1,286
(a) Pearson r ↑						
ESM-2 8M	.618 ±.008	.025 ±.094	.627 ±.050	.488 ±.009	.205 ±.043	.594 ±.024
ESM-2 650M	**.653 ±.018**	**.152 ±.050**	**.640 ±.063**	**.541 ±.013**	.269 ±.045	**.611 ±.043**
ESM-C 600M	.625 ±.019	.032 ±.043	.531 ±.047	.477 ±.012	.257 ±.110	.491 ±.028
SaProt 35M	.606 ±.023	-.007 ±.136	.479 ±.075	.502 ±.019	.010 ±.023	.607 ±.034
SaProt 650M	.648 ±.015	.042 ±.111	.471 ±.040	.513 ±.023	**.320 ±.093**	.515 ±.101
Electro-Prot	.444 ±.028	.132 ±.117	.179 ±.068	.409 ±.005	.230 ±.046	.432 ±.056
(b) Spearman ρ ↑						
ESM-2 8M	.640 ±.006	-.025 ±.038	.611 ±.068	.513 ±.009	.218 ±.070	.584 ±.040
ESM-2 650M	**.673 ±.011**	**.109 ±.069**	**.645 ±.083**	**.547 ±.016**	.286 ±.051	**.634 ±.059**
ESM-C 600M	.637 ±.018	-.002 ±.029	.523 ±.063	.494 ±.010	.287 ±.090	.492 ±.026
SaProt 35M	.612 ±.026	-.011 ±.112	.445 ±.075	.526 ±.022	.090 ±.018	.612 ±.048
SaProt 650M	.657 ±.015	.009 ±.150	.430 ±.049	.523 ±.025	**.339 ±.124**	.532 ±.083
Electro-Prot	.455 ±.023	.094 ±.116	.237 ±.090	.445 ±.003	.283 ±.052	.423 ±.051
(d) RMSE ↓ (kcal/mol)						
ESM-2 8M	2.475 ±.104	1.895 ±.667	2.532 ±.534	2.303 ±.031	2.087 ±.039	3.394 ±.563
ESM-2 650M	2.557 ±.317	1.825 ±.260	**2.196 ±.313**	**2.211 ±.021**	2.132 ±.131	3.697 ±.776
ESM-C 600M	3.070 ±.246	2.466 ±.697	2.325 ±.059	2.321 ±.019	2.269 ±.345	3.073 ±.332
SaProt 35M	**2.300 ±.081**	1.578 ±.440	2.547 ±.280	2.235 ±.028	3.266 ±.075	3.287 ±.917
SaProt 650M	2.331 ±.026	1.708 ±.085	2.542 ±.259	2.219 ±.027	2.052 ±.197	**2.843 ±.191**
Electro-Prot	2.755 ±.068	**1.302 ±.214**	2.949 ±.266	2.382 ±.010	**1.970 ±.051**	3.091 ±.317

G Encoder-Decoder Representation Ablations

Table 8: (a) Electro-Prot representation ablation on ProteinShake. Values are mean ± sample standard deviation over five seeds. Bold marks the higher mean within each task. (b) Electro-Prot representation ablation on SKEMPI mutation-induced $\Delta\Delta G$ prediction. Metrics are computed from pooled held-out predictions across the same three PDB-grouped folds for each seed, covering 6,359 measurements. Values are mean ± sample standard deviation over five seeds. Bold marks the better mean for each metric.

(a) ProteinShake

	EC	GO	SC	PFam	SSim	PPI	BS	LBA
Representation	Acc	$F_{\max}$	Acc	Acc	ρ	AUROC	MCC	ρ
Encoder only	.596 ±.009	.261 ±.005	.227 ±.013	.225 ±.004	.482 ±.009	.663 ±.002	.244 ±.001	**.351 ±.018**
Encoder–decoder	**.632 ±.009**	**.282 ±.003**	**.295 ±.003**	**.249 ±.006**	**.559 ±.005**	**.812 ±.001**	**.323 ±.002**	.350 ±.022

(b) SKEMPI mutation-induced $\Delta\Delta G$

	Pearson	Spearman	MAE	RMSE
Representation	r ↑	ρ ↑	kcal/mol ↓	kcal/mol ↓
Encoder only	.295 ±.014	.256 ±.007	1.526 ±.019	2.136 ±.057
Encoder–decoder	**.446 ±.017**	**.369 ±.009**	**1.348 ±.010**	**1.889 ±.017**

(c) PPB-Affinity ΔG: MAE (kcal/mol ↓)

Representation	Overall	ATLAS	Affinity Benchmark v5.5	PDBbind v2020	SAbDab	SKEMPI v2.0
Encoder only	**2.060 ±.012**	1.455 ±1.085	**1.990 ±.089**	1.841 ±.004	1.538 ±.083	2.991 ±.391
Encoder–decoder	2.174 ±.040	**1.043 ±.164**	2.378 ±.126	**1.840 ±.010**	**1.442 ±.068**	**2.487 ±.320**

H GVP-GNN and Electro-Prot Feature Fusion and Experimental Details

H.1 Architecture and Fusion Details

GVP-GNN passes messages along a geometric graph using scalar channels and three-dimensional vector channels (Jing et al., 2021b;a). Scalars encode features such as atom identity and distance, while vector channels carry directional information. We use the vector-gated atomic implementation: a radius graph with a 4.5 Å cutoff and the library's default neighbor cap, 16 radial-basis distance features and a unit-displacement vector per edge, and five message-passing layers with 100 scalar and 16 vector hidden channels. A final scalar projection and mean pooling over pocket and ligand nodes feed an affinity prediction head.

For pocket atom i, EP is queried at its original coordinate and supplies a frozen decoder state $\mathbf{e}_i \in \mathbb{R}^{256}$ immediately before the potential-prediction head. This state is a learned feature vector instead of the scalar predicted potential. We concatenate it with the GVP element embedding $\mathbf{a}_i \in \mathbb{R}^9$:

$$\mathbf{s}_i = [\mathbf{a}_i \,\|\, \widetilde{\mathbf{e}}_i], \qquad \widetilde{\mathbf{e}}_i = \begin{cases} \mathbf{e}_i, & i \text{ is a pocket atom,} \\ \mathbf{0}_{256}, & i \text{ is a ligand atom.} \end{cases} \tag{4}$$

Here $\|$ means concatenation along the feature dimension wherein each node receives 265 scalar input channels instead of 9. We widen the input normalization and projection accordingly; subsequent message-passing dimensions, pooling, and prediction head match the baseline. EP remains frozen and its features are cached, while GVP is trained on affinity labels. The zero ligand block indicates unavailable EP features instead of a zero physical ligand charge. It can also provide a pocket-ligand cue absent from the baseline's consumed node features. EP features are supplied as scalar channels, but the EP coordinate encoding does not guarantee rotation invariance; the complete fused pipeline therefore does not inherit an equivariance guarantee merely from its GVP component.

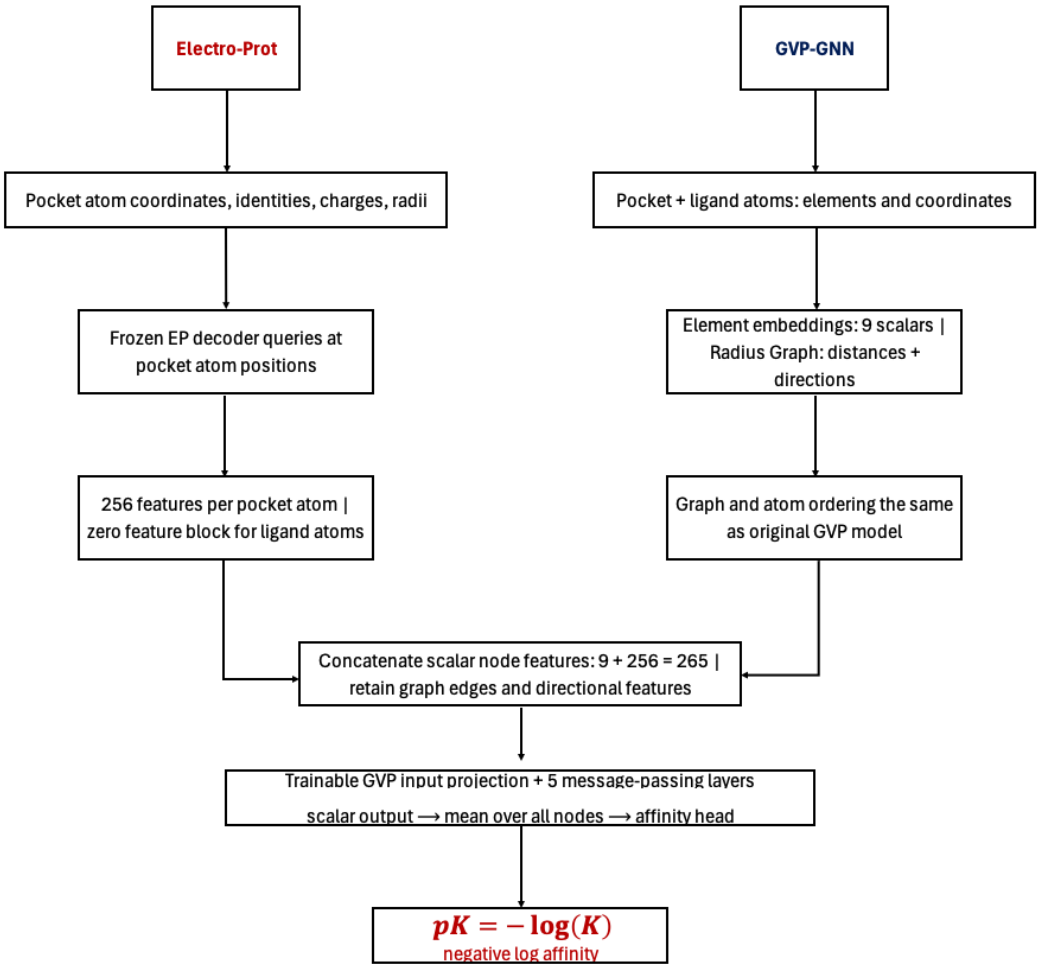

Figure 6: **GVP-EP fusion.** Frozen, atom-aligned EP decoder states augment scalar node inputs. Coordinates and graph edges follow the same construction as in the baseline; only the initial scalar feature width and its projection change. Affinity-loss gradients update only GVP.

H.2 Experimental Arms and Training Protocol

Table 9 summarizes the six experimental arms evaluated. All GVP arms use the same pocket-ligand graph and added features are supplied only for pocket atoms with zeros supplied for ligand atoms. EP and random-EP feature embeddings remain frozen.

Table 9: Experimental arms. Feature counts refer to GVP scalar inputs per atom, except for the pooled EP-only representation.

Arm	Input representation	Question addressed
GVP baseline	Element features and geometry: 9 scalar inputs	Reference geometric predictor
GVP + EP	Element features + 256 pretrained EP features: 265 inputs	Does pretrained EP help GVP?
EP-only ridge	256 pooled pocket features: linear ridge readout	What can a pocket-only linear predictor recover?
GVP + charges/radii	Element features + standardized charge and radius: 11 inputs	Are direct chemical inputs sufficient?
GVP + random EP	Element features + 256 features from an untrained EP: 265 inputs	Does EP pretraining matter?

The original GVP comparison and controls specify ten training seeds, 50 epochs, Adam with learning rate 10^{-4}, batch size 8, and selection by validation MSE. The EP-only ridge predictor is fit once, with regularization selected on validation data. We report test RMSE, Pearson and Spearman correlations, and R^2, summarizing neural-network runs by mean and population standard deviation.

H.3 Pocket-only ridge regression.

For each ATOM3D LBA-30 complex, we used cached, frozen EP-5 or EP-20 decoder hidden states immediately before the electrostatic-potential prediction head. The decoder was queried at the original pocket atom coordinates using real charges and radii. Each atom supplied a feature vector $h_{ia} \in \mathbb{R}^{256}$. For complex i, we constructed

$$z_i = \frac{1}{|\mathcal{R}_i|} \sum_{r \in \mathcal{R}_i} \frac{1}{|\mathcal{A}_{ir}|} \sum_{a \in \mathcal{A}_{ir}} h_{ia},$$

where $\mathcal{R}_i$ contains residues represented in the pocket and $\mathcal{A}_{ir}$ contains their selected pocket atoms. This pooling weights represented residues equally.

Using the fixed training, validation, and test splits (3,507/466/490 complexes), we standardized each feature using training-set statistics only. We fitted an intercept and 256 coefficients to the ATOM3D negative log affinity targets by minimizing

$$\sum_{i \in \text{train}} (y_i - b - w^\top \tilde{z}_i)^2 + \alpha \|w\|_2^2,$$

with an unpenalized intercept. We used scikit-learn `StandardScaler` and `Ridge` with the `cholesky` solver.

For each checkpoint independently, we searched $\alpha \in \{0.01, 0.1, 1, 10, 100, 1000, 10000\}$ and selected the model with the lowest validation MSE. The selected values were 100 for EP-5 and 0.1 for EP-20. We evaluated the selected training-fitted model on the test split without refitting on training plus validation data. ElectroProt remained frozen, and no GVP network or ligand descriptor was used. Each reported ridge result is one selected deterministic fit; no across-seed standard deviation was estimated. The representations are pocket-only but not ligand-independent, because pocket selection uses the ligand-containing complex.



Figure 7: **ElectroProt representations for ATOM3D ligand binding affinity.** All models are evaluated on the same 490 test complexes from the 30% sequence-identity split. **(a)** Controlled five-seed comparison of GVP-only baseline and GVP augmented with frozen EP-5 or EP-20 decoder features (GVP fusion models). **(b)** Separate ten-seed comparison of GVP-only baseline, pretrained EP-20 features, direct charges/radii, and features from one frozen random EP initialization. **(c)** Deterministic ridge readouts of pooled EP pocket features, compared with the same five-seed GVP-only baseline shown in (a). EP-5 and EP-20 denote checkpoints after five and twenty pretraining epochs. Columns report RMSE (lower is better), Pearson correlation, Spearman correlation, and R^2 (higher is better). Small points denote individual training seeds; filled circles and whiskers denote means and population standard deviations. Open diamonds denote single ridge fits without across-seed uncertainty estimates. Numeric annotations report means or individual ridge results.

I GVP-GNN Fusion with Electro-Prot Results on ATOM3D LBA

J Electro-Prot Prediction Quality

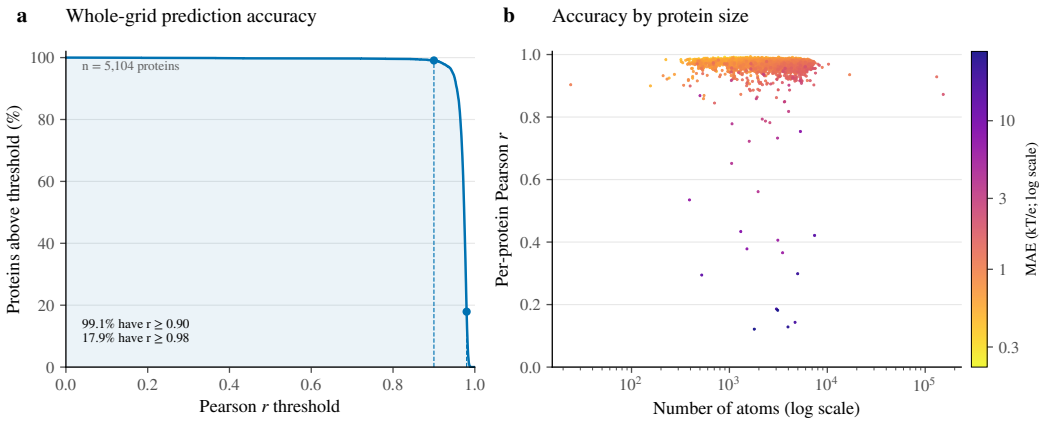

Figure 8: Electro-Prot prediction quality against APBS potentials on 5,104 completed held-out structure split test proteins. Metrics are computed over every finite target voxel of each protein's native APBS grid. **(a)** Percentage of proteins whose per-protein Pearson correlation meets or exceeds each threshold. **(b)** Per-protein Pearson correlation against input atom count, colored by per-protein MAE in kT/e. Atom count and the MAE color scale are logarithmic; each point represents one protein.