
Electro-Prot: Learning Protein Representations from Electrostatics

Paula Sharoubbeem

Dept. of Computational & Systems Biology
University of Pittsburgh
pas195@pitt.edu

Sumeet Kothare

Dept. of Computational & Systems Biology, Pitt
Computational Biology Dept., CMU
skothare@alumni.cmu.edu

Jacky Chen

Joint CMU-Pitt Ph.D. Program
in Computational Biology (CPCB)
jackychen@pitt.edu

David R. Koes

Dept. of Computational & Systems Biology
University of Pittsburgh
dkoes@pitt.edu

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. 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 $r = 0.97$ (MAE = 0.81 kT/e) on a held-out test set. On downstream tasks, EP achieves the highest reported mean Protein-Protein Interaction AUROC among the evaluated standalone representations whereas, SaProt performs better on structural similarity and binding-site prediction, while the binding-affinity comparisons are mixed. PLMs remain stronger on GO, and protein family classification. 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. 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 [5, 11], whereas structure-aware variants like SaProt extend this to spatial protein geometry [8, 16]. 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 information and does that signal encode transferable protein representations?

We use electrostatics as the premise to ask this question since the potential, ϕ , governs binding, recognition, and catalysis [6, 7], and contains complex long-range interactions. The electrostatic

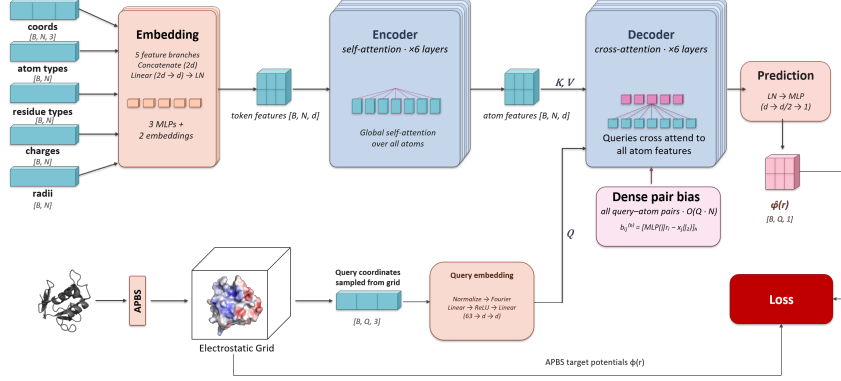


Figure 1: Electro-Prot Encoder-Decoder Pair-Bias Architecture

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 [1, 3, 9]. We treat electrostatic potential prediction as a means to a representation rather than an end in itself, unlike [18]. 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 [12]. 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.

2 Methods

2.1 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 [3] 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%). Resampling the pool keeps coverage of the protein broad while guaranteeing both polarities at every step.

Decoder and query-atom distance pair-bias. The query coordinate q_i is lifted by a Fourier encoding with 10 octave-spaced bands [12, 17], 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:

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

where q_i is the i -th query point, x_j the coordinate of atom j , and k_j its key vector. The bias 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

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

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. A linear head maps each query state to a potential in kT/e.

2.2 Data Preparation

Data Source and Splits. As described in Appendix A, we queried the RCSB Protein Data Bank (PDB) [2] for high-quality structures with at least 30 residues, yielding 190,336 entries. PDB2PQR and APBS [1, 9] were used to produce electrostatic potentials as volumetric grids resulting in 181,089 structure potential pairs. PDBFixer [4] is used to attempt to rescue structures where calculations initially failed. MMseqs2 [15] 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.

Training Augmentation. During training, we sampled points from the APBS field, as described in section 2.1. We then rotated and translated the atom coordinates and the grid query points together around the same APBS grid center. The sampled potential values remained attached to their associated query points. Repeated exposure to these augmented orientations aims to teach the model to predict the same electrostatic potential across the coordinate frames.

2.3 Protein Representation Benchmarking

ProteinShake benchmarks. We use the ProteinShake benchmarking suite, which provides curated protein-structure datasets with standardized tasks, splits, and evaluation metrics [10]. Our eight tasks span protein-level enzyme class, protein family, and GO annotation (EC, GO, and Pfam), structural classification and similarity (SC and SSim), ligand binding affinity (LBA), and residue-level protein-protein interface and binding-site prediction (PPI and BS).

All representations are compared using the same protocol of mean pooling and a linear probe. With pretrained parameters θ frozen, each protein is encoded once. Two representations extracted from the model are considered: final encoder states $\mathbf{h}_i^{\text{enc}}$ and final decoder states $\mathbf{h}_i^{\text{dec}} = D(\mathbf{x}_i; \mathbf{H}^{\text{enc}})$, obtained by querying the encoder representation $\mathbf{H}^{\text{enc}}$ 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 $\mathcal{A}_r$, then over the R residues for protein-level tasks:

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

giving a residue-level representation directly comparable to a PLM. A single affine layer $\hat{\mathbf{y}} = \mathbf{W}\mathbf{x} + \mathbf{b}$ is the only trainable component, to strictly measure linear separability without confounding performance with probe expressivity. For feature-level fusion, we directly concatenate the EP and PLM representations at the task appropriate representation level before applying the same linear probe. BS and PPI are residue-level tasks, so the probe is fit on $\mathbf{z}_r$. The remaining six tasks are protein-level and use $\mathbf{z}$, with the ProteinShake baseline’s ligand fingerprints appended for the affinity task. Task-specific training is further described in Appendix B.

3 Results

3.1 Electrostatic Potential Prediction Quality

On held-out test structures, EP reproduces APBS calculated potentials with a pointwise Pearson correlation of 0.97 and a mean absolute error of 0.81 kT/e; 99.1% of proteins reach a per-protein correlation above 0.90 and 17.9% above 0.98 (Appendix C). Predicted potentials recover the polarity and spatial organization of the ground-truth field.

Impact of Protein Size on Electrostatic Predictions. We completed full-grid spanning evaluations for 112 of the 120 prespecified size-binned proteins with a mean correlation $r = 0.974$ and MAE of $0.845 k_B T/e$. Between the smallest and largest atom-count bands, mean MAE increased from 0.617 to $1.370 k_B T/e$, while mean r decreased from 0.978 to 0.957 (Figure D.2). The largest band

included 14 of 20 selected proteins. Separately, 44 of 47 proteins selected for low correlation from Figure C.1, panel **b** were evaluated with EP; 28 had $r < 0.90$. Their median radius of gyration was 44.3 Å, versus 26.3 Å for the other 16 cases (Figure D.1).

Sensitivity of Electrostatic Potential Predictions to Rotation and Translation. On 118 proteins, rotating or translating the atoms and query coordinates in tandem produced mean absolute prediction changes of 0.086 and $0.003 k_B T/e$, respectively, at uniform queries. Mean MAE against APBS remained approximately $0.893 k_B T/e$; prediction changes were larger at high-potential queries (Appendix E, Figure E.1).

Effect of Protonation on Potential Predictions. On 12 preselected proteins prepared with explicit PROPKA [13, 14] method at pH 5, 7, and 9, mean correlations were 0.962, 0.974, 0.973 respectively and mean MAEs were 0.539, 0.448, $0.469 k_B T/e$ respectively. At corresponding queries, predicted changes relative to pH 7 correlated with APBS changes at 0.841 for pH 5 and 0.881 for pH 9. There were two unchanged pH 9 cases with undefined change correlations (Appendix F, Figure F.1).

3.2 General Protein Representation Tasks

Table 1: Representation performance on the ProteinShake Benchmark Suite. Values are mean $\pm$ sample standard deviation over five seeds. **Bold** marks the model with the highest uncertainty-adjusted score ($\bar{x} - s/\sqrt{5}$), which penalizes high volatility to ensure reliable ranking. Underlining marks the standalone model with the highest uncertainty-adjusted score.

	EC	GO	SC	PFam	SSim	PPI	BS	LBA
Representation	Acc	$F_{\max}$	Acc	Acc	ρ	AUROC	MCC	ρ
<i>Protein language models</i>								
ESM-2 8M	.782 $\pm$.002	.372 $\pm$.003	.622 $\pm$.006	.514 $\pm$.002	.509 $\pm$.015	.692 $\pm$.003	.251 $\pm$.002	.289 $\pm$.008
ESM-2 650M	.706 $\pm$.009	.577 $\pm$.004	.734 $\pm$.001	.624 $\pm$.003	.422 $\pm$.030	.687 $\pm$.005	.451 $\pm$.004	.343 $\pm$.019
ESM-C 600M	.736 $\pm$.016	.530 $\pm$.004	.633 $\pm$.005	.542 $\pm$.006	.359 $\pm$.025	.659 $\pm$.005	.469 $\pm$.001	.336 $\pm$.048
SaProt 35M	.808 $\pm$.010	.471 $\pm$.003	.685 $\pm$.005	.559 $\pm$.004	<u>.626 $\pm$.021</u>	.765 $\pm$.003	.414 $\pm$.002	<u>.357 $\pm$.031</u>
SaProt 650M	.772 $\pm$.019	.556 $\pm$.007	.684 $\pm$.003	.558 $\pm$.001	.567 $\pm$.018	.794 $\pm$.002	<u>.530 $\pm$.001</u>	.333 $\pm$.058
<i>Electro-Prot (EP; ours)</i>								
EP 6.7M (Enc. only)	.641 $\pm$.007	.281 $\pm$.005	.291 $\pm$.005	.277 $\pm$.002	.459 $\pm$.005	.691 $\pm$.008	.240 $\pm$.002	.332 $\pm$.021
EP 6.7M (Enc.+Dec.)	.641 $\pm$.010	.303 $\pm$.003	.315 $\pm$.008	.295 $\pm$.002	.493 $\pm$.006	<u>.826 $\pm$.001</u>	.353 $\pm$.001	.344 $\pm$.024
<i>Feature-level fusion of Electro-Prot (EP) + PLMs</i>								
EP + ESM-2 8M	.777 $\pm$.004	.409 $\pm$.003	.608 $\pm$.006	.488 $\pm$.006	.545 $\pm$.012	.809 $\pm$.004	.374 $\pm$.002	.340 $\pm$.026
EP + ESM-2 650M	.705 $\pm$.006	.576 $\pm$.004	.732 $\pm$.002	.613 $\pm$.003	.451 $\pm$.023	.798 $\pm$.003	.467 $\pm$.002	.344 $\pm$.025
EP + ESM-C 600M	.733 $\pm$.008	.536 $\pm$.004	.637 $\pm$.005	.547 $\pm$.005	.417 $\pm$.023	.782 $\pm$.003	.490 $\pm$.001	.346 $\pm$.022
EP + SaProt 35M	.800 $\pm$.006	.477 $\pm$.003	.665 $\pm$.005	.537 $\pm$.001	.632 $\pm$.006	.831 $\pm$.002	.444 $\pm$.001	.329 $\pm$.027
EP + SaProt 650M	.786 $\pm$.007	.560 $\pm$.006	.676 $\pm$.005	.545 $\pm$.003	.589 $\pm$.015	.836 $\pm$.002	.533 $\pm$.001	.360 $\pm$.024

Table 1 compares EP with 5 PLMs and 5 feature-level fusions across 8 tasks.

Standalone representations. EP is competitive on tasks whose underlying mechanism is interaction-centric and underperforms, relative to sequence models, on tasks driven by sequence homology. On PPI, EP reaches 0.826 AUROC, above every PLM evaluated (best PLM: 0.794 AUROC). On BS, EP reaches an MCC of 0.353, compared to 0.251 for ESM-2 8M. Furthermore, on LBA EP reaches ρ of 0.344, compared with 0.289 for ESM-2 8M. The picture inverts on GO, SC, EC, and PFam, where the PLMs lead by 0.17–0.42. When looking at EP’s encoder-only readouts, we observe a drop across the board, while EC remains relatively unchanged. The largest decreases occur for PPI, from 0.826 to 0.691 AUROC, and BS, from 0.353 to 0.240 MCC.

Feature-level fusion. Fusion of EP with PLMs generally improved results across our interaction-centric tasks at various PLM scales. When comparing EP fusion with ESM-2, we observe improvements in all interaction-centric tasks: LBA, BS, PPI, SSim ($\Delta_{\text{fusion-PLM}} = +0.051/+0.001$, $+0.123/+0.016$, $+0.117/+0.111$, and $+0.036/+0.029$, respectively, for ESM-2 8M/650M). Furthermore, the best performing models for interaction-centric tasks are EP fusion models (SSim: 0.632 ± 0.006 (EP + SaProt 35M); PPI: 0.836 ± 0.002 (EP + SaProt 650M); BS: 0.533 ± 0.001 (EP + SaProt 650M); LBA: 0.360 ± 0.024 (EP + SaProt 650M)). SC, PFam, and EC all remain unchanged or slightly degraded ($\Delta_{\text{fusion-PLM}} = -0.026$ to $+0.014$), whereas GO changes by

+0.037/ − 0.001 for ESM-2 8M/650M. Shuffle experiments (Appendix G) confirm that these gains arise from genuine electrostatic information rather than from increasing the dimensionality of the probe input. This effect is particularly clear for ESM-2 8M, whose parameter count is similar to that of EP ($\Delta(\text{real fusion} - \text{shuffled})$: LBA +0.041, BS +0.111, PPI +0.117, SSim +0.082, GO +0.041).

4 Discussion and Conclusion

Electrostatic pretraining transfers selectively. EP leads on PPI, approaches the best standalone LBA score, and trails the strongest BS baselines and on GO, SC, and Pfam relative to PLMs. It performs best where labels depend on direct physical interactions and worst where they are defined by evolutionary relationships. The electrostatic objective therefore captures additional signal from interaction-centric physics rather than simply recapitulating sequence evolutionary signal. Additionally using decoder states obtained from atom-coordinate queries improves performance over encoder states alone on all tasks except EC. The largest gains are on PPI (AUROC, 0.691 to 0.826) and BS prediction (MCC, 0.240 to 0.353). These interaction-centric gains show that atom-coordinate decoder queries produce a field-oriented representation enriched for the electrostatic determinants of molecular recognition.

Electrostatics and evolutionary representations are complementary. Fusing EP representations with PLM representations improves PPI, SSim, and BS for every PLM partner and improves GO and LBA in four of the five pairings. EC degrades in four of the five pairings, while SC and Pfam are mixed or slightly lower overall, highlighting the interaction-centric versus homology-driven distinction. This suggests EP contributes complementary physical/structural information, but naive feature-level concatenation can dilute or interfere with PLM features on sequence-driven classification tasks. Furthermore, fusion results generally exceeded matched shuffled controls for PLMs on PPI, SC, SSim, and BS, showing that fusion adds genuine complementary electrostatic signal (Appendix G).

Impact on Electrostatic Potential Prediction Quality. Our analysis, as described in section 3.1 and in Appendices D, E, F, shows that protein size attributes, pH for PROPKA preparation, and the orientation and position of atomic coordinates and query points have a quantifiable impact of varying degrees on the overall electrostatic potential prediction quality of our model as measured by Pearson correlation and MAE. In addition to sheer number of atoms, the extent or the spread of the mass of protein, described by the radius of gyration, indicates an association with poorer correlation and errors of predictions compared to the ground truth; however, they do not isolate their causes. Also, low correlations at modest atom counts (Figure D.1) indicate that atom count alone cannot identify unreliable predictions. The nonzero sensitivity of EP predictions to rigid transformations do not establish total invariance due to our training data augmentations however, the nearly unchanged mean prediction accuracy against APBS ground-truth support stability of average sampled-query accuracy despite changes in individual protein predictions. Finally, for different protonation levels, we observe agreement of EP potential predictions with APBS potential changes. However, this responsiveness of EP predictions to altered pH inputs does not establish overall robustness of predictions across several pH values.

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. Our size-binned evaluation shows increasing absolute prediction error with atom count and radius of gyration despite high correlations among predicted and APBS potentials. 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.

References

- [1] 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.
- [2] 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>.
- [3] 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.
- [4] 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.
- [5] ESM Team. ESM Cambrian: Revealing the mysteries of proteins with unsupervised learning. EvolutionaryScale blog, 2024. <https://evolutionaryscale.ai/blog/esm-cambrian>.
- [6] 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.
- [7] Barry Honig and Anthony Nicholls. Classical electrostatics in biology and chemistry. *Science*, 268(5214):1144–1149, 1995.
- [8] 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)*, 2021.
- [9] Elizabeth Jurrus, Dave Engel, Keith Star, et al. Improvements to the APBS biomolecular solvation software suite. *Protein Science*, 27(1):112–128, 2018.
- [10] 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, pages 58277–58289, 2023.
- [11] 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.
- [12] 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.
- [13] Mats H. M. Olsson, Chresten R. Søndergaard, Michal Rostkowski, and Jan H. Jensen. PROPKA3: Consistent treatment of internal and surface residues in empirical pKa predictions. *Journal of Chemical Theory and Computation*, 7(2):525–537, 2011. doi: 10.1021/ct100578z.
- [14] Chresten R. Søndergaard, Mats H. M. Olsson, Michał Rostkowski, and Jan H. Jensen. Improved treatment of ligands and coupling effects in empirical calculation and rationalization of pKa values. *Journal of Chemical Theory and Computation*, 7(7):2284–2295, 2011. doi: 10.1021/ct200133y.
- [15] 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.
- [16] 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.
- [17] 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.
- [18] 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.

A Dataset Construction

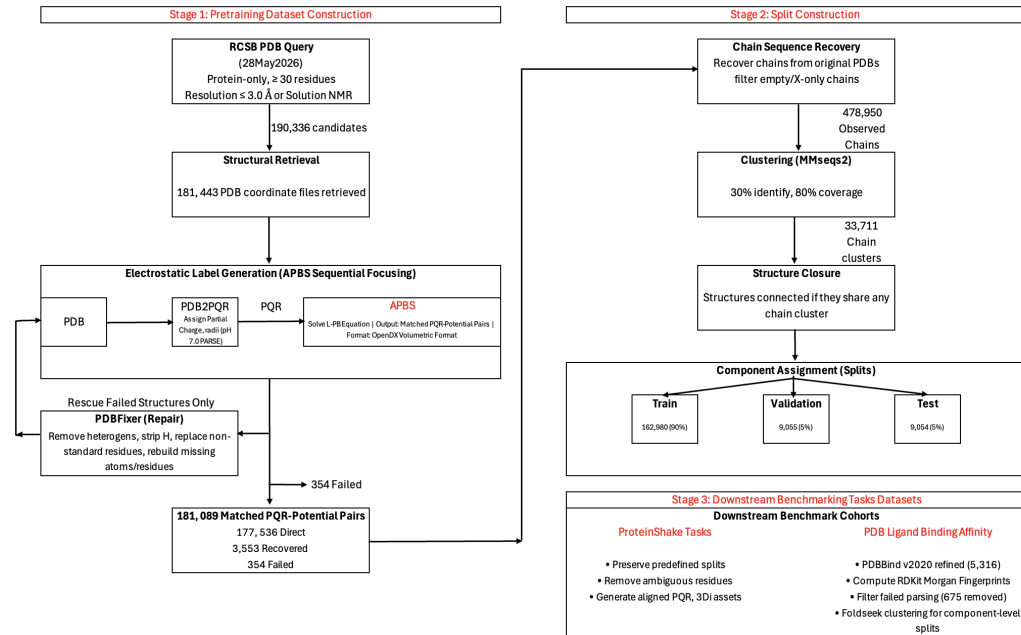


Figure A.1: Dataset Construction Pipeline

A.1 Building the Pretraining Dataset

On 28 May 2026 we queried the RCSB PDB [2] 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 [3] first assigns protonation states, partial charges, and atomic radii at pH 7.0 and 0.15 M monovalent salt. APBS [1, 9] then computes the resulting potential on a 0.5 Å 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 [4], 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.

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 [15] 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.

A.3 Downstream Benchmarking

The downstream benchmark data was separate from the pretraining data and was used only to evaluate frozen representations. For the ProteinShake tasks [10] we kept each task’s published split unchanged, removed structures with ambiguous residue identities, and generated the inputs each representation required.

Different representations failed on slightly different proteins for their own reasons, and allowing each to use whatever subset it could handle would have made the results incomparable. We therefore evaluated every representation on the intersection of the proteins all of them supported, so that every row of Table 1 was computed on an identical set of proteins. Representations were extracted once and cached, all three data splits were standardized using training split statistics, five seeded probes were trained per configuration, checkpoints were selected on validation, and test data were evaluated once.

For ligand binding affinity we obtained the 5,316 complexes of the PDDBind v2020 refined set, discarded 675 whose ligand chemistry could not be parsed, and retained 4,641. These were grouped by structural rather than sequence similarity into 640 groups, which were again assigned whole to splits, giving 3,713 training, 464 validation, and 464 test complexes. Applying the shared-cohort requirement left 3,515, 417, and 415. Because the ProteinShake PDDBind dataset and split manifest were unavailable through the package, we generated a structure-based split at a 70% similarity threshold following the same splitting protocol.

A.4 Settings

Stage	Setting
Charge assignment	PARSE force field, pH 7.0, water removed
Electrostatics	linearized Poisson-Boltzmann, sequential focusing
Dielectric	2.0 solute, 78.54 solvent
Ionic strength	0.15 M monovalent, radii 2.0 and 1.8 Å
Temperature	310 K
Surface	smoothed molecular, 1.4 Å probe
Chain clustering	30% identity, 80% coverage of shorter chain
Split ratio	90 / 5 / 5, seed 42
Ligand fingerprints	Morgan, radius 2, 1024 bits

APBS input. 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
  srfr smol
  chgm spl2
  sdens 10.00
  srdr 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
```

B 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 Electro-Prot Prediction Quality

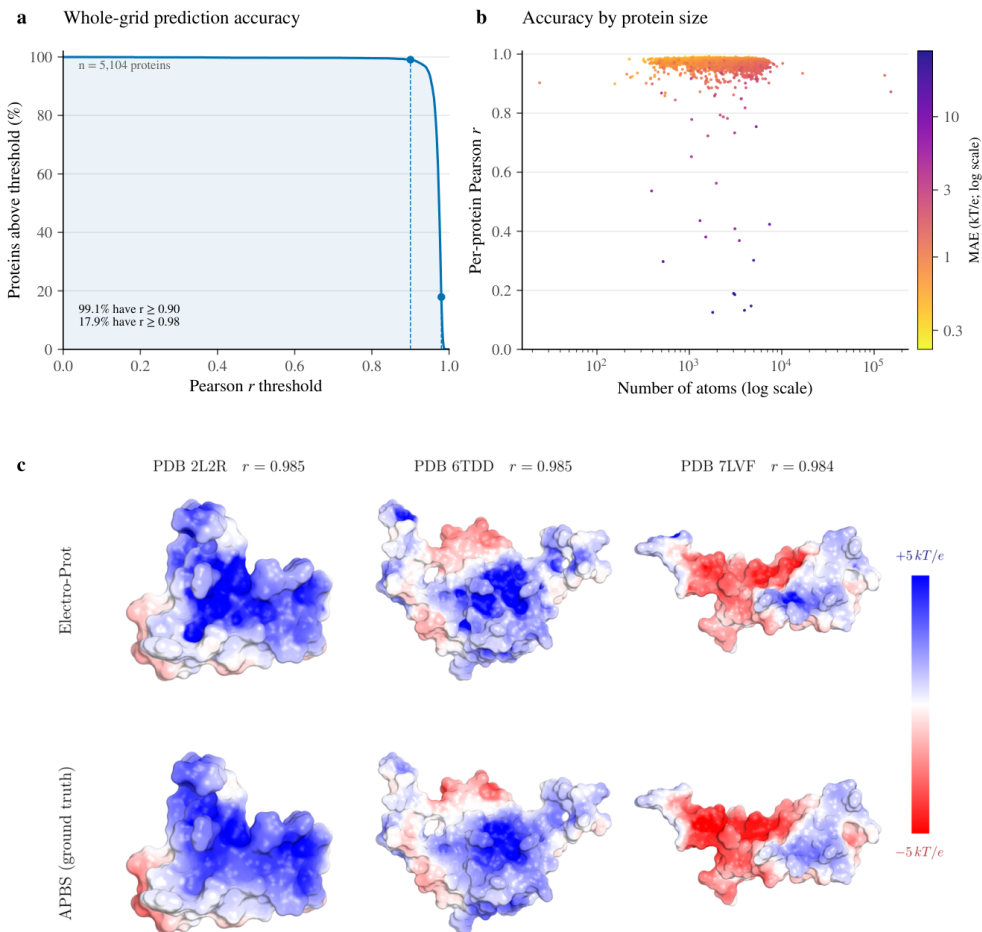


Figure C.1: Electro-Prot prediction quality against APBS potentials on held-out structure split test proteins. Metrics are computed over every finite target voxel of each protein’s native APBS grid. **(a)** Percentage of proteins whose per-protein Pearson correlation meets or exceeds each threshold. **(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. **(c)** Electrostatic potential prediction with EP Compared to APBS Ground truth predictions for 2L2R, 6TDD, and 7LVF.

D Impact of Protein Size on Electrostatic Potential Predictions

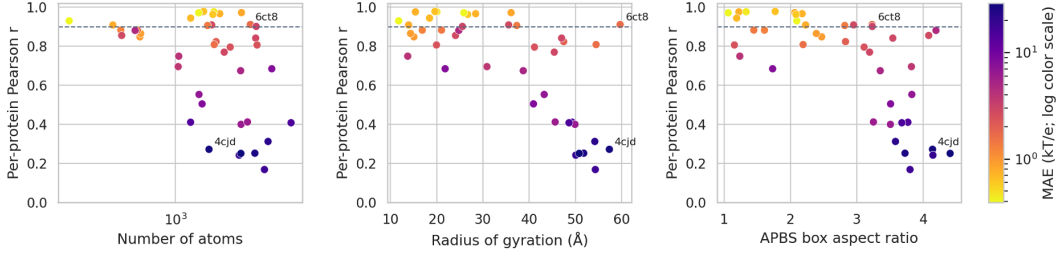


Figure D.1: Each point represents one of 44 completed evaluations of proteins selected for historical $r < 0.90$; 28 currently fall below this threshold (dashed line). Pearson correlation with APBS is plotted against the number of atoms including hydrogens (left, logarithmic axis), radius of gyration (middle panel), and the longest-to-shortest physical edge ratio of the APBS grid box (right). Color shows MAE in kT/e on a logarithmic scale, from yellow (lower error) to purple (higher error). Metrics use all finite stored APBS grid locations. The labeled examples illustrate different accuracies at similar radii of gyration. This selected cohort does not estimate how often low-correlation predictions occur in the test population.

We evaluated frozen EP on 112 held-out proteins by randomly selecting 20 proteins, without replacement, from each of the six training-set-defined atom count bins. Selection was not subject to prediction quality or input availability. Atom counts include all atoms including metals and hydrogens.

For each protein, EP encoded the atomic inputs and predicted the potentials at given query points across the APBS reference grid. For predictions p_j and APBS values t_j at M evaluated locations, we calculated

$$r = \frac{\sum_{j=1}^M (p_j - \bar{p})(t_j - \bar{t})}{\sqrt{\sum_{j=1}^M (p_j - \bar{p})^2} \sqrt{\sum_{j=1}^M (t_j - \bar{t})^2}},$$

$$\text{MAE} = \frac{1}{M} \sum_{j=1}^M |p_j - t_j|,$$

where bars denote query-wise means and MAE is expressed in kT/e .

We also used exploratory radius of gyration ranges of < 15 , $15 - < 20$, $20 - < 25$, $25 - < 30$, $30 - < 40$, and ≥ 40 Å. For N PQR atoms at positions $\mathbf{x}_a$, we calculated

$$\bar{\mathbf{x}} = \frac{1}{N} \sum_{a=1}^N \mathbf{x}_a,$$

$$R_g = \sqrt{\frac{1}{N} \sum_{a=1}^N \|\mathbf{x}_a - \bar{\mathbf{x}}\|^2}.$$

These radius groups were not used for protein selection.

In Figure D.2, bars show the arithmetic means of per-protein metrics within each size bin. Mean MAE increased from 0.617 to 1.370 kT/e between the smallest and largest atom-count bands whereas, mean r decreased from 0.9782 to 0.9574. MAE also increased across the radius groups. Thus, high potential prediction correlation exists despite increasing absolute error.

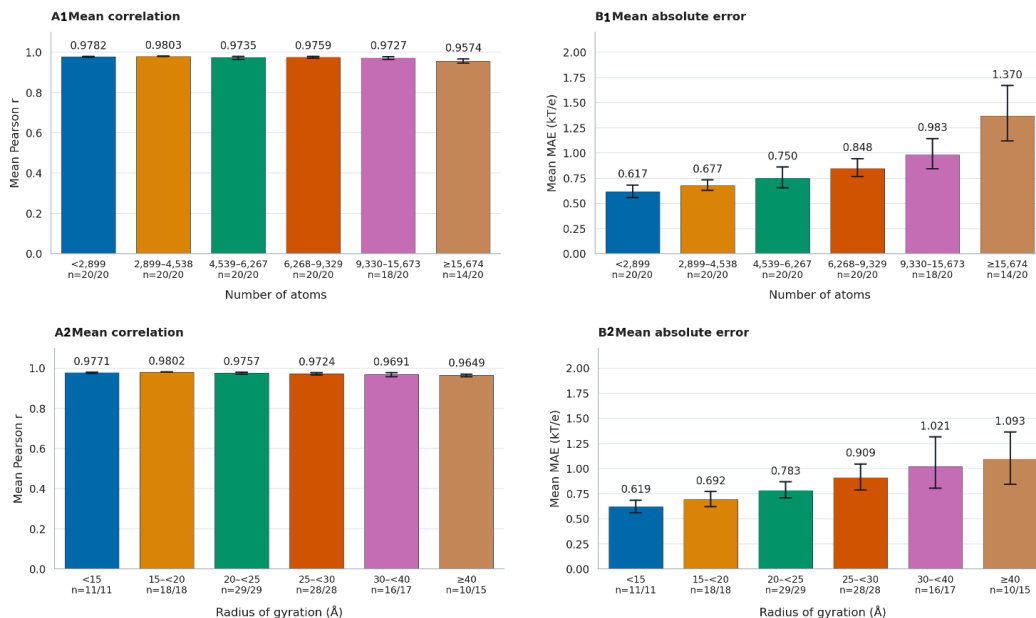


Figure D.2: Assessing the effect of variable number of protein atoms and radius of gyration on predicted potential Mean Correlation and Mean Absolute Error. Bars show mean per-protein Pearson correlation (A1, A2) and mean absolute error (MAE; B1, B2) for 112 completed full grid evaluations of frozen EP. Metrics were calculated over every finite location of each protein’s APBS reference grid, with proteins weighted equally when averaging. The top row groups proteins into atom-count bands defined using the training-set distribution; the bottom row regroups the same evaluated proteins by radius of gyration. Atom counts include hydrogens in the prepared PQR. Whiskers indicate 95% percentile bootstrap intervals for the means, obtained from 2,000 within-band protein resamples. Colors distinguish bins. Mean absolute error increases with size and spatial extent, while mean correlations remain high.

E Evaluating Sensitivity of Electrostatic Potential Predictions to Orientation and Position

We evaluated the frozen EP model on a prespecified panel of 120 held-out proteins selected randomly as 20 proteins from each of the six atom-count bins defined using the training set using a seed of 42. Out of these, 118 had the required inputs. For each protein, we sampled 2048 queries uniformly from finite APBS grid points and a separate 2048 queries sampled from the highest 5% of the absolute APBS electrostatic potentials. The latter provided a targeted analysis of large-magnitude potentials.

Similar to the data augmentation used on the training set proteins, we transformed the atom coordinates of the protein while retaining their corresponding APBS labels. We compared the original coordinates with 20 random rotations about the grid center, six translations of ± 5 along the coordinate axes, and 20 combined transformations. For these combined transformations, translation components were drawn independently from a zero mean Gaussian with a standard deviation of 1 such that it was subsequently clipped to ± 5 . Grid origins moved with translations. We measured absolute prediction changes and changes in MAE against APBS thereby averaging transformations within each protein before averaging across proteins. Confidence intervals used 2000 protein bootstrap resamples.

In panels A and B, we first calculated the absolute difference in potential predictions between transformed (rotated, translated, combined) and the original predictions at the corresponding queries, averaged over the queries (2048), and then averaged over transformations (20 rotations, 6 translations, 20 combined) within each family (x-axis). In these panels, exact invariance would correspond to 0 mean absolute prediction change hence, lower y-axis value is better and leans towards invariance. In panels C and D, we instead calculated the MAE against APBS, over 2048 queries for each transformation, averaged these MAEs within each transformation family, and subtracted the protein’s original coordinate (identity) MAE against APBS (averaged over 2048 queries). For these panels,

negative values indicate reduced error whereas positive values indicate increased error. Each dot in all the panels summarizes one protein, diamonds and error bars show means across 118 proteins, and the 95% protein-bootstrap confidence intervals.

As in Figure E.1, mean prediction changes under rotation, translation, and combined transformations were 0.0862, 0.0033, and 0.0834 kT/e for uniform grid queries. Mean MAE changes remained near zero, but the wider spread in panel D shows that this average conceals greater variation in protein-level accuracy changes. Hence, despite a nonzero prediction sensitivity, average accuracy changed little under the tested transformations.

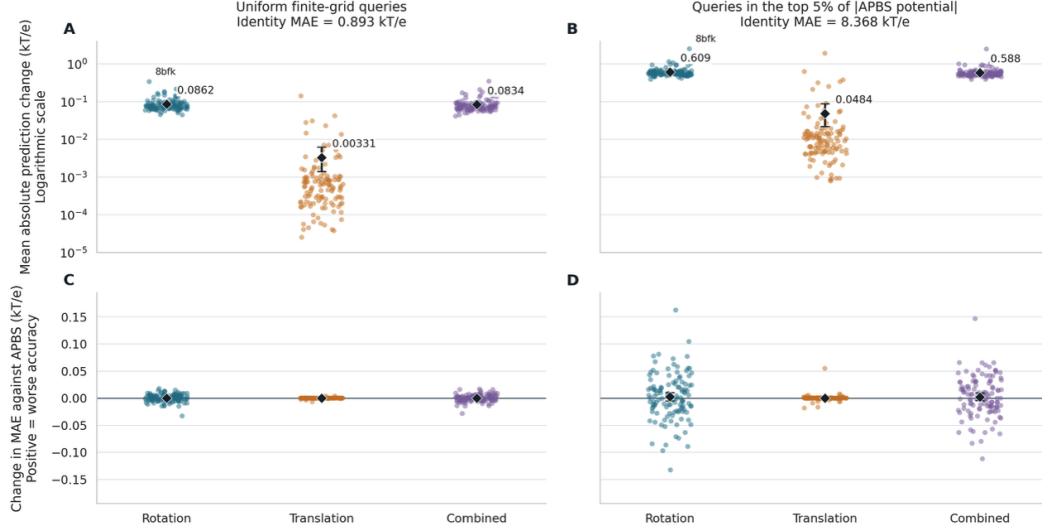


Figure E.1: Evaluating the sensitivity of electrostatic potential predictions to rotations and translations. A–B: Mean absolute prediction change relative to the original coordinates, on logarithmic axes. C–D: Change in MAE against APBS. Left panels use uniformly sampled queries; right panels use queries from the highest 5% of absolute APBS potential. Each dot represents one of 118 proteins, averaged over transformations within its family. Diamonds indicate protein means, and error bars show 95% bootstrap confidence intervals. Horizontal jitter separates overlapping dots. “Identity MAE” denotes prediction error before transformation.

F Evaluating the Impact of Protonation on Potential Predictions

We evaluated frozen EP on 12 preselected test proteins. PDB2PQR used PARSE parameters and explicit PROPKA calculation of pKa values of ionizable groups from the protein’s 3D structure at pH 5, 7, and 9. APBS grid geometry (one common grid) and solver settings (salt concentration, temperature, and dielectric settings: 0.15 M, 298.15 K, and 2.0/78.54) remained unchanged whereas the number of hydrogens (protonation) were free to change.

For each protein i , we sampled $N = 2048$ finite pH-7 APBS grid locations uniformly without replacement and reused them at pH 5 and 9. Predictions used the complete encoder–decoder model. Let $\hat{\phi}_{ihj}$ and ϕ_{ihj} denote predicted and APBS potentials, respectively, at pH h and query j . Per-protein metrics were

$$r_{ih} = \text{corr}_j(\hat{\phi}_{ihj}, \phi_{ihj}), \quad \text{MAE}_{ih} = \frac{1}{N} \sum_{j=1}^N |\hat{\phi}_{ihj} - \phi_{ihj}|,$$

where corr_j denotes Pearson correlation across query locations. We also defined

$$\Delta \hat{\phi}_{ihj} = \hat{\phi}_{ihj} - \hat{\phi}_{i7j}, \quad \Delta \phi_{ihj} = \phi_{ihj} - \phi_{i7j},$$

and applied the same correlation and MAE definitions to these change vectors, yielding r_{ih}^{Δ} and MAE_{ih}^{Δ} . Metrics were averaged with equal protein weights. For each plotted mean, we generated

2000 bootstrap resamples by sampling the available protein level metrics with replacement using the original set’s sample size. The error bars span the 2.5 and 97.5 percentiles of the resulting means of the resamples.

At pH 5/7/9, mean r was 0.962/0.974/0.973 and mean MAE was 0.539/0.448/0.469 $k_B T/e$. For changes from pH 7 to pH 5/9, mean r^Δ was 0.841/0.881 and mean MAE^Δ was 0.252/0.088 $k_B T/e$. Two proteins had zero changes at pH 9: their undefined r^Δ values were excluded ($n = 10$), while their zero change errors were retained ($n = 12$). At the pH-7 reference, change errors are zero by construction and change correlations are undefined.

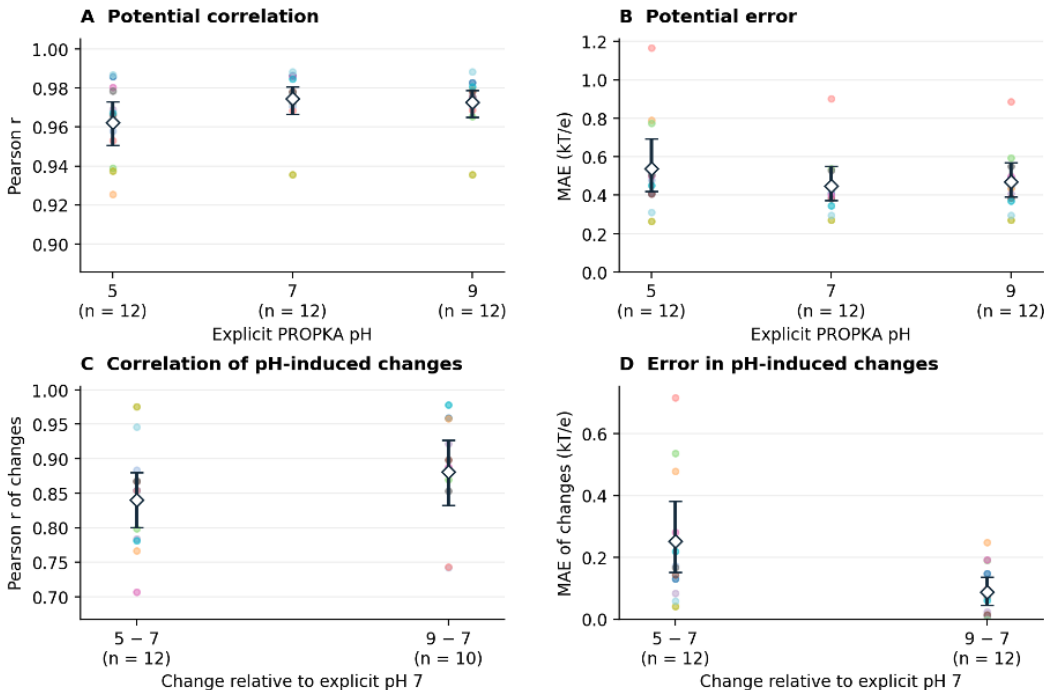


Figure F.1: Panels A and B: Pearson correlation and MAE of predicted versus APBS potentials. Panels C and D: Corresponding metrics for potential changes relative to explicit PROPKA pH 7. Colored circles represent individual proteins; diamonds and error bars show protein means and 95% bootstrap intervals respectively. All metrics use 2048 corresponding uniform queries per protein

G Shuffled-Embedding Control: Electrostatic Information versus Probe Capacity

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 as seen in Table G.1. 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

Table G.1: Dimension-matched shuffled-feature control. Upper panel: absolute performance with permuted PB embeddings, mean $\pm$ sample standard deviation over three 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
<i>Shuffled Electro-Prot embeddings</i>								
ESM-2 8M	.781 $\pm$.002	.368 $\pm$.001	.546 $\pm$.014	.475 $\pm$.005	.463 $\pm$.023	.692 $\pm$.001	.263 $\pm$.002	.299 $\pm$.035
SaProt 35M	.807 $\pm$.009	.477 $\pm$.004	.627 $\pm$.011	.524 $\pm$.004	.617 $\pm$.018	.768 $\pm$.003	.424 $\pm$.002	.334 $\pm$.018
ESM-2 650M	.712 $\pm$.003	.579 $\pm$.003	.721 $\pm$.003	.622 $\pm$.001	.429 $\pm$.037	.688 $\pm$.004	.453 $\pm$.002	.326 $\pm$.014
ESM-C 600M	.751 $\pm$.005	.533 $\pm$.003	.611 $\pm$.010	.542 $\pm$.017	.366 $\pm$.023	.668 $\pm$.003	.477 $\pm$.002	.379 $\pm$.020
SaProt 650M	.785 $\pm$.014	.554 $\pm$.005	.655 $\pm$.009	.548 $\pm$.006	.576 $\pm$.008	.795 $\pm$.002	.534 $\pm$.001	.352 $\pm$.047
Δ (real fusion – shuffled)								
ESM-2 8M	–.004	+.040	+.062	+.013	+.082	+.117	+.111	+.041
SaProt 35M	–.007	.000	+.037	+.013	+.015	+.063	+.020	–.004
ESM-2 650M	–.006	–.003	+.012	–.010	+.022	+.109	+.014	+.018
ESM-C 600M	–.019	+.003	+.025	+.005	+.051	+.114	+.013	–.033
SaProt 650M	+.001	+.006	+.020	–.004	+.013	+.041	–.001	+.009

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.