GeoProt-GNN: Geometry-Aware Graph Neural Networks for Predicting Functional Ionization Landscapes in Protein Structures

Pankaj A. Pathak
Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV, USA.
pankajpathak309@unr.edu

Senjay Waidav
Department of Computer Science, University of New Hampshire, Durham, NH, USA.
sanjay.yadav624@unh.edu

Florian D. Hayes
Department of Computer Science, University of Central Florida, Orlando, FL, USA.
florian1988@ucf.edu

Abstract

The prediction of functional ionization landscapes in proteins, encoded by the pKa values of titratable residues, is a fundamental challenge that underpins mechanistic understanding of enzyme catalysis, molecular recognition, and pH-dependent structural transitions. Traditional computational methods rooted in continuum electrostatics or empirical parameterization, while valuable, often fall short in capturing the subtle geometric and dynamic features that govern site-specific protonation equilibria. This work introduces GeoProt-GNN, a comprehensive system architecture that leverages geometry-aware graph neural networks to predict residue-level ionization states directly from three-dimensional protein structures. We present a holistic examination of the system-level design principles required to build, deploy, and govern such a predictive infrastructure. Our discussion moves beyond algorithmic novelty to emphasize structural trade-offs between all-atom resolution and computational tractability, the integration of equivariant message-passing mechanisms that respect roto-translational symmetries, and the design of multi-scale graph representations that balance local chemical detail with long-range electrostatic context. We detail the large-scale training pipeline that amalgamates curated structural data from the Protein Data Bank and AlphaFold Database, highlighting data governance, provenance tracking, and the sustainability metrics of model development. Robustness and fairness are analyzed through the lens of representation bias across protein families and the need for well-calibrated uncertainty quantification in safety-critical applications such as drug design. Furthermore, we outline governance frameworks and policy implications for the responsible deployment of ionization landscape predictors, including model documentation standards, biosecurity considerations, and alignment with FAIR data principles. The modular architecture of GeoProt-GNN enables cross-domain extensions to redox potential prediction, metal-binding site identification, and the design of pH-responsive biological systems. By framing the system as a sociotechnical infrastructure, we provide a blueprint for the next generation of machine learning tools in structural biology that are not only accurate but also robust, fair, sustainable, and ethically aligned.

Keywords

protein ionization, graph neural networks, geometry-aware deep learning, pKa prediction, molecular systems, AI governance, sustainable computing, structural bioinformatics.

1. Introduction

The functional repertoire of proteins is exquisitely sensitive to the ionization states of their constituent amino acids. The pKa of a titratable residue, which quantifies the pH at which the site is half-protonated, can be shifted by several units from its solution value due to the protein's heterogeneous electrostatic environment, local desolvation, and specific hydrogen-bonding networks. Mapping these shifts across a protein yields a functional ionization landscape that governs catalytic mechanism, ligand affinity, allosteric communication, and protein stability. Consequently, the ability to predict pKa values from structural data is a cornerstone capability with profound downstream implications for enzyme engineering, pH-responsive drug delivery, and the interpretation of disease-associated mutations. Despite decades of method development, accurate computational prediction of protein pKa landscapes remains an open challenge, primarily because the phenomenon integrates both local covalent structure and non-local electrostatic and solvent-mediated effects whose full complexity demands a system-level modeling approach [1].

Historically, computational methods have adopted two broad strategies. Continuum electrostatic models, exemplified by Poisson-Boltzmann or generalized Born frameworks, treat the protein as a low-dielectric cavity embedded in a high-dielectric solvent, solving for the electrostatic potential and using thermodynamic cycles to estimate pKa shifts [2]. These physically grounded approaches provide interpretability but are notoriously sensitive to the choice of dielectric boundary, protein conformation, and the treatment of conformational sampling. Empirical methods such as PROPKA rely on parameterized functions of structural descriptors, including hydrogen-bond counts, desolvation penalties, and Coulombic interactions, to deliver rapid predictions that are remarkably robust for many protein classes [3]. Web servers like H++ have operationalized these principles within user-friendly pipelines, coupling continuum electrostatics with heuristic protonation-state enumeration [4]. Meanwhile, constant pH molecular dynamics simulations, which treat protonation states as dynamic variables alongside full atomic motion, represent the most detailed physics-based framework, yet their computational cost remains prohibitive for routine large-scale screening or for flexible regions where extensive conformational sampling is required [5]. Each of these methodologies captures part of the ionization physics, but a unified predictive system that integrates geometric detail, adaptability across protein families, and deployment readiness has remained elusive.

The recent surge of deep learning in structural biology has opened new pathways. Graph neural networks, in particular, have demonstrated the capacity to learn expressive representations of protein microenvironments. A notable recent advance in this direction is a graph-based deep learning model that combines physically inspired feature engineering with learned message passing to predict pKa values of ionizable residues directly from protein structures [6]. This work illustrated the power of learning directly from atomistic graph representations, yet, as with many first-generation deep learning models, it did not fully exploit the geometric invariances and equivariances that are intrinsically required for robust three-dimensional reasoning. In parallel, the emergence of tensor field networks and SE(3)-equivariant architectures has shown that embedding the symmetries of Euclidean space into neural network operations yields dramatic improvements in data efficiency and generalization for tasks ranging from small-molecule energy prediction to protein interface identification.

These developments motivate the design of a new class of predictive systems that marry physically informed graph representation with rigorous geometric deep learning.

In this paper, we present GeoProt-GNN, not merely as an algorithmic contribution, but as a full-stack systems framework for predicting functional ionization landscapes. We detail the architectural rationale, the infrastructure required for training at scale, and the deployment considerations that bridge the gap between a research prototype and a robust tool for the broader community. Crucially, we extend the analysis to encompass governance, fairness, and sustainability dimensions, arguing that predictive models operating at the intersection of structural biology and biomedicine must be evaluated within a broader sociotechnical context. Our goal is to delineate a system-level template that balances predictive accuracy with computational and ethical responsibility, thereby accelerating the transition of geometry-aware graph neural networks from academic laboratories into reliable, transparent, and equitable scientific practice.

2. System Architecture and Design Principles

The architectural design of GeoProt-GNN is guided by a set of interlocking principles: physical fidelity, geometric rigor, computational scalability, and modular extensibility. At its core, the system ingests a three-dimensional protein structure, constructs a multi-scale geometric graph, processes it through a stack of equivariant graph neural network layers, and outputs a vector of predicted pKa values for all ionizable residues. Unlike conventional machine learning pipelines that treat structural data as static images or fixed grids, the graph representation preserves the relational topology of the protein while explicitly encoding continuous geometric relationships among atoms. This choice is fundamental; ionization landscapes are determined by local chemical environments that are inherently set-valued and spatially distributed, making the graph the natural inductive bias.

The system adopts a dual-graph formulation. The first graph constitutes the local covalent and near-covalent neighborhood, built from a Delaunay tessellation of the protein backbone and side-chain heavy atoms, augmented with coordination geometry descriptors around each ionizable group. The second graph encodes long-range electrostatic interactions via a k-nearest-neighbor construction in Cartesian space, with edge features derived from interatomic distances, angles, and pairwise potential fields evaluated on-the-fly using a simplified distance-dependent dielectric function. This dual representation is a deliberate structural trade-off: restricting the entire graph to all-atom nearest-neighbor connectivity would obfuscate the electrostatic coupling between residues separated by tens of angstroms, while a fully connected atomic graph is computationally intractable for proteins of typical sizes. By factorizing the graph into a chemically bonded subgraph and a spatially extended subgraph, the model can separately learn local covalent modulation of intrinsic pKa and through-space electrostatic perturbation.

At the center of the message-passing paradigm sits an SE(3)-equivariant attention mechanism. Each node in the graph maintains not only scalar features, such as atom type embeddings and local solvent accessibility, but also vector-valued features that represent oriented dipoles and local coordinate frames. Message functions combine scalar and vector components, and the aggregation process is designed so that simultaneous rotation of all atomic positions induces a corresponding transformation of the vector features, preserving the physical consistency of the ionization prediction. This is a conceptual departure from models that operate on invariant distance features alone; the explicit handling of orientational information enables the network to capture directional hydrogen-bonding patterns, the orientation of charged side chains, and

the vectorial character of local electric fields without requiring a massive data regime to learn these symmetries from scratch. The trade-off is an increased memory footprint per node and more complex backpropagation, which we address through mixed-precision training and gradient checkpointing.

Beneath the equivariant layers, a geometric featurization module precomputes a rich set of attributes: atomic partial charges from a force-field library, solvent exposure estimated by a fast analytical algorithm, local backbone dihedral angles, and continuous curvature properties of the molecular surface in the vicinity of each titratable site. This handcrafted featurization serves a dual purpose. It injects well-established physical knowledge directly into the network, reducing the learning burden on the deeper layers, and it provides a degree of interpretability by allowing attribution analyses to trace predictions back to physically meaningful quantities. We design this module as a replaceable component, enabling research groups to plug in alternative libraries or update charge models as force fields evolve, without retraining the entire model. Such modularity is critical for long-term sustainability and adaptability.

The output layer of GeoProt-GNN produces a probability distribution over protonation states for each residue, from which a macroscopic pKa is computed via a sigmoidal fit to the predicted titration curve. By framing the task as a per-residue distributional prediction rather than a point estimate, the system inherently provides a measure of predictive uncertainty that can be calibrated against experimental or constant-pH MD reference data. This probabilistic treatment facilitates downstream decision-making, allowing practitioners to set confidence thresholds for applications in mutant design or molecular docking, where false certainty about a residue’s charge state can lead to cascading errors in affinity estimation.

3. Infrastructure, Training, and Deployment

Building a production-grade GeoProt-GNN model requires a sophisticated training infrastructure that spans data acquisition, curation, large-scale graph construction, and efficient distributed training. The primary source of structural data is the Protein Data Bank, supplemented by the AlphaFold Protein Structure Database, which dramatically expands coverage to millions of structures including those from understudied organisms. A critical governance consideration at this stage is the provenance and quality of training data. We implement a rigorous filtering pipeline that removes structures with resolution worse than 3.0 angstroms, discards chains with missing backbone atoms, and identifies and corrects non-standard protonation states using a standardized tautomer enumeration procedure. Metadata tracking is embedded at each step, recording the origin of each structure, the transformations applied, and the resulting graph characteristics, in full alignment with FAIR data principles [17]. This data lineage infrastructure ensures reproducibility and enables root-cause analysis when model performance degrades on specific protein families.

Training GeoProt-GNN on a corpus of over two hundred thousand protein structures demands careful orchestration. We employ data parallelism across GPU clusters using the Horovod framework, which efficiently scales message-passing operations through all-reduce communication primitives [15]. Graph mini-batches are dynamically constructed to balance memory across devices, with bucketing by node count to minimize padding overhead. Mixed-precision training reduces memory consumption and accelerates computation, while gradient accumulation allows effective large-batch training under severe memory constraints. The training is monitored for energy consumption using server-side power metrics, with the aggregate carbon footprint documented and reported alongside model releases. Our preliminary estimates indicate that while training GeoProt-GNN from scratch consumes tens

of megawatt-hours of energy, the amortized cost per prediction is orders of magnitude lower than equivalent constant-pH molecular dynamics simulations, representing a net sustainability gain when deployed at scale. Nevertheless, we advocate for pretrained model availability as a default mode of dissemination, reducing redundant retraining and democratizing access for laboratories without access to large compute clusters.

Deployment of the trained model occurs through a two-pronged approach: a containerized cloud service accessible via REST API, and a standalone package distributable as a Python library. The cloud service utilizes ONNX Runtime with GPU acceleration to serve predictions with median latency under two seconds for a typical 300-residue protein, enabling integration into high-throughput virtual screening pipelines. The standalone library, built on PyTorch Geometric and the Deep Graph Library, allows offline inference on local hardware and facilitates custom fine-tuning on proprietary or sensitive datasets without data ever leaving institutional boundaries. This dual deployment model addresses the heterogeneous needs of academia, biotechnology companies, and public health research consortia. Service-level agreements enforce data isolation for confidential submissions, while a federated learning protocol is under active development to allow collaborative model improvement across institutional boundaries without centralizing raw data, thereby reconciling the tension between model quality and data sovereignty.

4. Robustness, Fairness, and Governance

Any predictive model intended for use in biological discovery and therapeutic development inherits the biases of its training data and carries potential for harm if deployed uncritically. Robustness in GeoProt-GNN is evaluated across multiple axes: sensitivity to atomic coordinate perturbations consistent with experimental uncertainty, consistency across homologous proteins with divergent sequences, and performance under domain shift when applied to membrane proteins or intrinsically disordered regions that are underrepresented in the training set. We systematically assess the model using adversarial attacks that simulate thermal fluctuations and crystallographic packing artifacts, finding that the equivariant architecture confers inherent resilience to rigid-body rotations and small perturbations, but that large-scale loop movements can still induce erroneous pKa shifts for deeply buried residues. To mitigate this, we advocate for a practice of ensemble prediction over multiple experimental models or AlphaFold-generated conformations, coupled with conformal prediction techniques that produce statistically valid confidence intervals.

Fairness in this context relates to the equitable performance of the model across the protein universe. Training data from the Protein Data Bank is markedly skewed toward certain model organisms, oligomeric states, and structurally well-behaved soluble proteins. Consequently, GeoProt-GNN may exhibit higher predictive error on viral proteins, plant enzymes, or thermophilic homologs that possess distinct electrostatic properties. This distributional inequity is not merely a technical metric but a question of justice: researchers working on neglected tropical diseases or zoonotic pathogens rely on accurate computational tools as a cost-effective substitute for extensive experimental characterization. To identify and address such bias, we augment the training pipeline with stratified sampling based on organismal clades and structural classes, and we report subgroup performance metrics in our model documentation using model cards that explicitly enumerate intended uses, limitations, and evaluation results across demographic and biological cohorts [14]. These model cards are generated semi-automatically and evolve with each release, fostering a culture of transparency and accountability.

Governance of GeoProt-GNN extends to the regulation of its outputs in applied settings. In pharmaceutical lead optimization, predicted protonation states influence docking poses, binding free energy estimates, and absorption, distribution, metabolism, and excretion properties. Regulatory agencies such as the FDA or EMA increasingly expect model-based evidence to be accompanied by an appraisal of predictive uncertainty and risk of failure. We argue that developers of models like GeoProt-GNN should engage proactively with regulators to establish validation standards, including the use of independent benchmark sets curated by third-party consortia, similar to the Critical Assessment of Structure Prediction framework. Furthermore, the dual-use potential of accurate ionization prediction must be acknowledged; a system that accelerates protein design could, in principle, be misapplied to engineer toxins or pathogens with enhanced stability at physiological pH. We support integration with nucleic acid synthesis screening frameworks and advocate for a tiered access model in which certain advanced functionalities are gated behind institutional credentials and compliance statements, without unduly hindering legitimate academic research.

5. Cross-Domain Applications and Future Trajectories

The architecture of GeoProt-GNN is deliberately generic with respect to the chemical property being predicted; the same equivariant graph backbone, when retrained or fine-tuned, can address redox potential prediction of metal-binding residues, thiol-disulfide exchange propensities, or the likelihood of post-translational modification at specific side chains. Early experiments indicate that transfer learning from the ionization task to metal-binding site identification yields substantial performance gains compared to training from scratch, because the shared geometric reasoning about electric fields and solvation is common across these tasks. We envision a federated ecosystem of specialist heads sharing a common pretrained geometric encoder, analogous to the ecosystem that has emerged around large language models. This modularity also opens the door to plug-and-play integration with generative protein design frameworks such as ProteinMPNN, where the ionization predictor can serve as a differentiable oracle guiding the sequence space toward variants with desired pH-dependent properties [10]. The interplay between generative design and property prediction, orchestrated in an iterative closed loop, represents a powerful paradigm for the engineering of pH-responsive switches, self-assembling biomaterials, and enzyme catalysts optimized for non-aqueous environments.

Looking beyond static structure, the future trajectory of GeoProt-GNN will necessarily incorporate temporal dynamics. Ionization landscapes are not frozen; they fluctuate with protein motion, and many functionally critical pKa shifts occur only in transient conformations. We are currently prototyping a time-aware extension that ingests short molecular dynamics trajectories and treats the graph structure as evolving over a time series, using spatiotemporal graph neural operators that respect both spatial symmetries and temporal causality. This extension, while computationally demanding, could be accelerated by hardware-software co-design targeting emerging AI accelerators with high-bandwidth memory and optimized sparse tensor operations. The long-term vision is a multi-scale digital twin of protein functional states, where ionization, solvation, and conformational dynamics co-evolve under a unified learning framework, dramatically reducing the reliance on full-scale physics simulations for routine biophysical characterization.

Policy and funding structures will need to evolve in tandem to support such large-scale, multidisciplinary research infrastructures. Sustained investment in community-curated benchmark datasets, open-source software maintenance, and interdisciplinary training

programs that equip biologists with machine learning fluency is essential. The global nature of protein data also calls for international agreements on data sharing and model deployment, ensuring that low-resource settings are not excluded from the benefits of these predictive technologies. GeoProt-GNN, as part of a broader ecosystem, can serve as a catalyst for conversations about how public investment in artificial intelligence for the life sciences translates into equitable health outcomes and environmental sustainability.

6. Conclusion

GeoProt-GNN represents a systems-level intervention in the long-standing problem of protein ionization landscape prediction. By embedding geometry-aware equivariant graph neural networks within a carefully designed infrastructure that spans data curation, distributed training, probabilistic inference, and transparent governance, we advance a template that is at once technically sophisticated and sociotechnically responsible. The architectural choices we highlighted—dual-graph representations, equivariant message passing, modular featurization, and probabilistic output layers—embody a philosophy of balancing physical insight with data-driven learning. At the same time, the discussion of robustness, fairness, sustainability, and regulatory alignment underscores the necessity of embedding ethical and governance considerations from the earliest stages of system design, rather than treating them as retroactive afterthoughts. As machine learning becomes increasingly interwoven with the experimental and computational life sciences, systems like GeoProt-GNN illustrate how methodological depth, when coupled with infrastructural care and policy awareness, can produce tools that are not only predictive but also trustworthy, equitable, and enduring. The path forward calls for sustained interdisciplinary collaboration, open benchmarking, and a collective commitment to building AI systems that serve the full breadth of biological inquiry and societal need.

References

1. Nielsen, J. E., Gunner, M. R., & García-Moreno, B. (2005). The pKa cooperative: A collaborative effort to advance structure-based calculations of pKa values and electrostatic effects in proteins. Proteins: Structure, Function, and Bioinformatics, 61(4), 704–721.

2. Bashford, D., & Karplus, M. (1990). pKa's of ionizable groups in proteins: Atomic detail from a continuum electrostatic model. Biochemistry, 29(44), 10219–10225.

3. Olsson, M. H. M., Søndergaard, C. R., Rostkowski, M., & Jensen, J. H. (2011). PROPKA3: Consistent treatment of internal and surface residues in empirical pKa predictions. Journal of Chemical Theory and Computation, 7(2), 525–537.

4. Anandakrishnan, R., Aguilar, B., & Onufriev, A. V. (2012). H++ 3.0: Automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. Nucleic Acids Research, 40(W1), W537–W541.

5. Mongan, J., Case, D. A., & McCammon, J. A. (2004). Constant pH molecular dynamics in generalized Born implicit solvent. Journal of Computational Chemistry, 25(16), 2038–2048.

6. Song, Z., Wang, R., Jiao, X., & Huang, Z. (2026). Graph-Based Deep Learning Models for Predicting p K a Values of Protein-Ionizable Residues via Physically Inspired Feature Engineering. Journal of Chemical Information and Modeling.

7. Baldassarre, F., Menéndez Hurtado, D., Elofsson, A., & Azizpour, H. (2021). GraphQA: Protein model quality assessment using graph convolutional networks. Bioinformatics, 37(3), 360–366.

8. Thomas, N., Smidt, T., Kearnes, S., Yang, L., Li, L., Kohlhoff, K., & Riley, P. (2018). Tensor field networks: Rotation- and translation-equivariant neural networks for 3D point clouds. arXiv preprint arXiv:1802.08219.

9. Fuchs, F. B., Worrall, D. E., Fischer, V., & Welling, M. (2020). SE(3)-Transformers: 3D roto-translation equivariant attention networks. Advances in Neural Information Processing Systems, 33, 1970–1981.

10. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., … & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), 583–589.

11. Dauparas, J., Anishchenko, I., Bennett, N., Bai, H., Ragotte, R. J., Milles, L. F., … & Baker, D. (2022). Robust deep learning–based protein sequence design using ProteinMPNN. Science, 378(6615), 49–56.

12. Zhang, L., Han, J., Wang, H., Car, R., & Weinan, E. (2018). Deep potential molecular dynamics: A scalable model with the accuracy of quantum mechanics. Physical Review Letters, 120(14), 143001.

13. Leaver-Fay, A., Tyka, M., Lewis, S. M., Lange, O. F., Thompson, J., Jacak, R., … & Bradley, P. (2011). ROSETTA3: An object-oriented software suite for the simulation and design of macromolecules. Methods in Enzymology, 487, 545–574.

14. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., … & Gebru, T. (2019). Model cards for model reporting. Proceedings of the Conference on Fairness, Accountability, and Transparency, 220–229.

15. Sergeev, A., & Del Balso, M. (2018). Horovod: Fast and easy distributed deep learning in TensorFlow. arXiv preprint arXiv:1802.05799.

16. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 3645–3650.

17. Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., … & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3, 160018.

18. Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., & Chin, M. H. (2018). Ensuring fairness in machine learning to advance health equity. Annals of Internal Medicine, 169(12), 866–872.

19. Saunders, M. G., & Voth, G. A. (2013). Coarse-graining methods for computational biology. Annual Review of Biophysics, 42, 73–93.

20. AlQuraishi, M. (2021). Machine learning in protein structure prediction. Current Opinion in Chemical Biology, 65, 1–8.