Dynamic Conformation-Aware Deep Learning for Residue Ionization Prediction in Flexible Protein Systems

Aakash Gaptea
School of Information Technology, University of Cincinnati, Cincinnati, OH, USA.
aakash.gupta@uc.edu

Felix Jarvinen
Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO, USA.
felixjarvinen167@missouri.edu

Abstract

The accurate prediction of ionizable residue pKa values in proteins is fundamental to understanding pH-dependent structure, function, and molecular recognition. Traditional physics-based methods, while interpretable, suffer from systematic errors when applied to flexible systems where conformational dynamics modulate local electrostatic environments. This work presents a dynamic conformation-aware deep learning framework designed to predict residue ionization states across an ensemble of protein conformations, moving decisively beyond the static single-structure paradigm. The system integrates three-dimensional geometric representations with multi-conformer sampling through an architecture that combines graph neural networks, time-aware transformers, and equivariant message passing to capture pH-dependent protonation probabilities in a conformationally resolved manner. We examine the entire lifecycle of such a system, from data infrastructure and featurization to training strategies, deployment scalability, and the broader governance considerations essential for its adoption in drug discovery, enzyme engineering, and synthetic biology. A detailed analysis of structural trade-offs reveals the balance between conformational granularity and computational tractability, the robustness of the model under input perturbations, and the fairness implications arising from training data biases across protein families and organismal sources. The paper argues that sustainable, equitable deployment of dynamic deep learning models for residue ionization requires not only architectural innovation but also transparent data curation practices, continuous monitoring of performance drifts, and the articulation of policy frameworks that guide the use of predictive biochemical models in high-stakes biomedical decision-making.

Keywords

protein ionization, pKa prediction, deep learning, conformational dynamics, graph neural networks, structural biology, molecular simulations, system architecture, robustness, algorithmic fairness.

1. Introduction

The protonation states of ionizable amino acid residues—aspartic acid, glutamic acid, histidine, lysine, arginine, cysteine, and tyrosine—govern the electrostatic landscape of proteins and underpin catalytic activity, allosteric transitions, protein–protein interactions, and

pH-dependent stability. Classical biochemical studies established that the apparent pKa of a residue can shift dramatically relative to its model compound value due to desolvation, hydrogen bonding, and charge–charge interactions within the folded protein [1]. In flexible protein systems, such as intrinsically disordered regions, multi-domain assemblies, and enzymes undergoing large-scale conformational changes, these microenvironments are not static; they fluctuate on timescales ranging from picoseconds to milliseconds, sampling multiple local geometries that each present a distinct energetic context for ionization [2]. Consequently, a single static crystal structure or an AlphaFold-predicted model often fails to capture the full conformational diversity that dictates the ensemble-averaged pKa values observed experimentally.

Computational prediction of protein pKa values has evolved from continuum electrostatics models [3] through empirical structure-based heuristics [4] to machine learning methods that leverage either handcrafted physicochemical descriptors or learned representations from three-dimensional coordinates. Despite these advances, the vast majority of existing deep learning approaches embed a static snapshot as input, implicitly assuming that a single conformation is sufficiently representative. This assumption breaks down with growing evidence that sidechain rotameric states, backbone fluctuations, and water penetration events alter local dielectric environments in ways that static models cannot encode. A truly conformation-aware prediction system must ingest, process, and integrate information from multiple conformers, learning a mapping from dynamic structural ensembles to pH-dependent ionization probabilities.

This paper presents a systems-level analysis of dynamic conformation-aware deep learning for residue ionization prediction. We define the architectural components required to handle multimodal conformational data, discuss the infrastructure necessary to generate and curate large-scale training sets from molecular dynamics simulations, and illuminate the trade-offs among representational fidelity, computational cost, and predictive robustness. Beyond technical architecture, we examine fairness and bias issues, sustainability concerns, and the governance policies that must accompany the deployment of such models in pharmaceutical research and biotechnology. By foregrounding the system as a whole rather than an isolated algorithmic artifact, we aim to guide the design of next-generation computational tools that respect the dynamic, context-dependent nature of protein biophysics.

2. Background and Related Work

Early quantitative frameworks for interpreting protein titration curves were rooted in the linkage between proton binding and conformational equilibria, formalized through the concept of microscopic pKa values [1]. Building on this thermodynamic foundation, continuum electrostatic methods such as those solving the Poisson–Boltzmann equation became the workhorse of pKa prediction for several decades [3]. These approaches rely on a single protein structure with assigned dielectric constants and a finite-difference or boundary-element solver to estimate the electrostatic free energy of charging each titratable group. Although physically intuitive, they are sensitive to the choice of protein dielectric, the treatment of conformational relaxation, and the definition of the solute–solvent boundary, leading to root-mean-square errors often exceeding 1.0 pKa unit on challenging internal residues. Empirical methods like PROPKA [4] addressed some of these limitations by introducing parameterized terms for hydrogen bonding, desolvation, and charge–charge interactions calibrated on experimental datasets, achieving improved accuracy for surface

residues while remaining computationally efficient enough for high-throughput virtual screening.

With the rise of deep learning in the physical sciences, researchers began to formulate pKa prediction as a supervised learning problem on three-dimensional protein graphs. Early graph neural network models featurized residue-level environments using distance-based edge attributes and atomistic node properties, learning directly from experimental pKa data without explicit electrostatics. The performance of such data-driven models was encouraging but often plateaued when tested on blind sets comprising flexible loops or active sites, where static geometries misrepresented the true conformational ensemble. More recent work has sought to incorporate protein dynamics by using short molecular dynamics simulations to generate multiple conformers, then averaging predictions across frames. The field has also witnessed the introduction of physically inspired feature engineering strategies that embed Coulombic and van der Waals energy terms, solvent accessibility metrics, and pKa shifts of model compounds into graph-based architectures [9]. These hybrid representations offer a bridge between first-principles physics and data-driven learning, yet they remain largely confined to single-structure inputs.

Parallel advances in geometric deep learning have yielded rotationally and translationally equivariant neural networks capable of operating on three-dimensional point clouds and molecular graphs without requiring pre-aligned coordinate frames. Such architectures have been applied to protein–ligand binding affinity prediction, mutational effect estimation, and protein–protein interface quality assessment, demonstrating that SE(3)-equivariant message passing can learn rich representations of local chemical environments. In conformation-aware applications, Boltzmann generators [8] and normalizing flows on torsion angles have enabled the generation of physically plausible ensembles from a single structure, circumventing the high cost of exhaustive molecular dynamics sampling. Together, these developments provide the building blocks for a dynamic pKa prediction system that respects symmetry, integrates temporal information, and learns directly from conformational heterogeneity.

3. System Architecture of Dynamic Conformation-Aware Models

A dynamic conformation-aware deep learning pipeline for residue ionization prediction can be decomposed into four principal modules: a conformational ensemble encoder, a spatiotemporal aggregator, a pH-conditioned ionization decoder, and an ensemble inference harmonizer. The ensemble encoder ingests a set of protein conformations, each represented as a graph whose nodes correspond to ionizable and neighboring residues, with edges encoding spatial proximity, hydrogen bonding patterns, and sequence separation. Three-dimensional coordinates are processed by an equivariant message-passing backbone that updates node features while preserving rotational and translational symmetries, ensuring that the learned representations are invariant to the orientation of individual conformers. Conformational identity is encoded through positional or velocity information from the source molecular dynamics trajectory, allowing the model to distinguish whether a given frame corresponds to a rare high-energy state or a highly populated basin. To handle variable ensemble sizes and protein lengths, the architecture employs attention-based pooling mechanisms that can compress an arbitrarily sized set of conformations into a fixed-dimensional latent representation without losing resolution on functionally critical residues.

The spatiotemporal aggregator fuses information across the conformational ensemble over time. While a naive approach would average predictions after independent processing of each frame, we argue that a learned cross-conformation attention module provides a richer

inductive bias. This module applies multi-head attention across the ensemble dimension for each residue, enabling the model to compare sidechain environments in different conformations and weight frames according to their predicted relevance to the titration behavior. For example, a buried aspartic acid that is solvent-exposed only in rare loop-opening fluctuations may have its pKa dominated by those rare events; a cross-conformation attention mechanism can learn to upweight such frames dynamically. The temporal order of conformations is modeled using sinusoidal positional encodings or recurrent gating, depending on whether the simulation trajectory is continuous or represents an uncorrelated ensemble from enhanced sampling. Importantly, the aggregator supports missing modalities: when only a single structure is available, the system gracefully degrades to a static predictor by bypassing cross-conformation attention, with a measurable yet controlled penalty in accuracy.

The decoder head predicts a residue-specific pKa value or, more expressively, a pH-dependent protonation probability curve. Conditioning on pH is achieved by appending the target pH to the aggregated latent vector, allowing a single model to generalize across the entire physiological and experimental pH range. A calibration layer based on Platt scaling or isotonic regression is integrated into the decoder to align predicted probabilities with experimental titration curves observed in NMR or mass spectrometry experiments. The ensemble inference harmonizer then combines decoder outputs into final, interpretable metrics such as the macroscopic pKa, Hill coefficient, and confidence intervals derived from conformer-level variance. This layered architecture not only captures the multimodality of residue ionization behavior but also provides built-in uncertainty quantification, a critical feature for downstream applications in lead optimization where false confidence in an erroneous pKa can derail molecular design campaigns.

4. Data Infrastructure and Conformational Sampling

Training a dynamic conformation-aware model demands a curated, high-quality dataset of experimentally measured microscopic pKa values paired with structural ensembles. While several static pKa benchmarks exist, the construction of a dynamic benchmark involves significant infrastructural challenges. Molecular dynamics simulations must be run for thousands of proteins under varying pH conditions, typically using constant-pH replica exchange or discrete protonation state sampling, which multiplies the required compute by orders of magnitude compared to fixed-protonation equilibration. Furthermore, simulation trajectories must be annotated with residue-wise protonation states extracted at each saved frame through retrospective reweighting or Hamiltonian exchange, a process that is error-prone and requires careful validation against experimental NMR or potentiometric data. We propose a federated data infrastructure in which multiple academic and industrial groups contribute simulation trajectories to a centralized, version-controlled repository with standardized metadata formats, similar to the Protein Data Bank but extended to dynamic ensembles. Such an infrastructure can reduce the burden on individual laboratories and accelerate the development of robust models from diverse protein families, including membrane proteins, thermophilic enzymes, and viral capsids.

The choice of conformer sampling strategy has profound architectural implications. Exhaustive explicit-solvent simulations provide the most physically realistic ensembles but are computationally prohibitive for large-scale training. Coarse-grained models and elastic network normal mode analysis offer a lightweight alternative; however, their ability to capture localized sidechain rearrangements that critically affect pKa values is limited. Deep

generative models trained on structural databases—such as variational autoencoders over torsion angles or diffusion models operating on backbone coordinates—can produce physically plausible conformers at inference time with minimal overhead. Yet these surrogate ensembles may introduce biases if the generator fails to reproduce rare but functionally important states. A hybrid approach that combines a small number of high-quality all-atom frames with many augmented conformers from a learned generative model can balance fidelity and coverage, especially when the generative model is conditioned on local sequence and structure features and trained with a physics-informed loss that penalizes steric clashes and unrealistic electrostatic configurations. Ensuring that the infrastructure supports efficient, streaming access to these ensembles during training is a non-trivial systems engineering task, requiring optimized storage formats, distributed data loading, and on-the-fly featurization pipelines to keep accelerator utilization high.

5. Training Strategies, Robustness, and Generalization

Training conformation-aware models presents unique challenges beyond those encountered in static protein learning. The loss landscape is rugged because small changes in a conformation's sidechain dihedral angles can produce large fluctuations in predicted pKa, leading to noisy gradient updates. To mitigate this, we employ curriculum learning where the model is first pretrained on static structures to build a stable baseline representation, then fine-tuned on increasingly larger and more heterogeneous ensembles. Multi-task auxiliary objectives, such as predicting solvent accessible surface area or hydrogen bond counts, act as regularizers that steer the latent space toward physically meaningful dimensions. Transfer learning from general-purpose protein language models—such as those trained on millions of sequences or on structure prediction tasks—can further bootstrap performance on pKa-specific data, which remains relatively scarce and imbalanced across residue types. Histidine residues, for example, are substantially underrepresented in experimental pKa databases, a data imbalance that a suitably weighted loss function and controlled oversampling can partially address.

Robustness must be evaluated against multiple axes of perturbation. Conformational noise—whether from simulation inaccuracies or generative model artifacts—should not cause catastrophic prediction shifts unless the underlying physicochemical environment genuinely changes. We advocate for adversarial robustness testing in which small, physically plausible perturbations are applied to sidechain torsions or hydrogen positions, and the resulting pKa predictions are compared against the unperturbed ensemble. A robust system will exhibit smooth, bounded variations consistent with thermodynamic expectations, while a brittle model may show step-function discontinuities that erode trust in virtual screening pipelines. Generalization across protein folds and functional classes is equally critical. A model trained predominantly on globular soluble proteins will frequently mispredict pKa values in transmembrane helices, where the low-dielectric lipid bilayer creates radically different electrostatic screening effects. Domain adaptation techniques—either through feature alignment in latent space or through physically informed data augmentation that simulates membrane-like environments—are necessary to extend the operational envelope of such systems. Furthermore, continual learning paradigms should be considered so that models can ingest newly generated experimental and simulation data without catastrophic forgetting, maintaining accuracy across a growing and diversifying protein universe.

6. Deployment Considerations and Scalability

Transitioning from a research-grade prototype to a production-ready service for the broader biochemical community requires careful attention to scalability, latency, and interoperability. A single inference request for a 500-residue protein with 100 conformers may require evaluating tens of thousands of message-passing operations and attention computations, a workload that can be served through optimized batched inference on graphical processing units or specialized tensor processing units. Serving infrastructure must accommodate bursty usage patterns, as medicinal chemistry campaigns often screen hundreds of protein variants overnight. We propose a microservice architecture in which the conformational encoder, aggregator, and decoder are deployed as containerized modules with autoscaling capabilities, connected via a lightweight message broker. On-demand conformer generation can be offloaded to specialized generative services that return ensembles either from precomputed databases or through fast surrogate models, reducing the need for users to run their own simulations.

A critical deployment trade-off lies in the decision between cloud-based centralized inference and on-premises deployment in pharmaceutical environments with strict data sovereignty requirements. Federated inference, where model weights are distributed to local clusters, could resolve privacy concerns while maintaining identical predictive performance, though it demands careful versioning and drift monitoring across heterogeneous hardware. The environmental sustainability of large-scale deep learning infrastructure must also be addressed. Carbon-aware scheduling of training jobs, application of mixed-precision quantization, and model distillation into compact student networks can significantly reduce the energy footprint without degrading predictive accuracy. A lightweight distilled model that approximates the full conformation-aware ensemble prediction within 0.1 pKa units may be sufficient for routine triage, while the full high-fidelity model is reserved for lead optimization stages, realizing a tiered approach that optimizes the resource–accuracy Pareto frontier.

7. Fairness, Bias, and Ethical Implications

Algorithmic fairness in the context of protein science is a multifaceted concept that extends beyond individual-level discrimination to encompass representational imbalances across biological space. Current protein databases used for training pKa models are enriched in well-characterized, soluble, mesophilic proteins from a narrow set of model organisms. Consequently, prediction performance on extremophilic enzymes, disordered regions, or viral proteins may be systematically degraded. This representational harm can propagate into downstream inequities: drug discovery programs targeting neglected tropical diseases, which often rely on homology models of pathogen proteins, may be disproportionately affected by inaccurate ionization predictions, biasing lead optimization away from viable compounds [10]. A fairness audit of conformation-aware pKa models should involve stratified evaluation across biological kingdoms, functional classes, and structural environments, with transparent reporting of performance disparities and efforts to expand the diversity of training ensembles through targeted molecular dynamics campaigns and community data challenges.

Bias can also emerge from the choice of simulation force fields and sampling protocols, which embed theoretical approximations about protein dynamics that may not hold uniformly across chemical space. A model trained exclusively on simulations using a particular water model and fixed-charge force field may inherit that force field's systematic errors in representing salt bridges or metal coordination, resulting in overconfident but incorrect predictions for zinc-finger domains or calcium-binding loops. Mitigating this form of epistemic bias requires multi-fidelity training strategies that blend data from different force

fields, quantum mechanics/molecular mechanics reference calculations, and experimental measurements, with explicit uncertainty quantification that propagates force field disagreements into the final pKa credible intervals. The ethical imperative extends to communication: users of these tools—medicinal chemists, structural biologists, and regulators—must be educated on the limitations and provenance of predictions, resisting the temptation to treat a single pKa value generated by an opaque deep learning model as an unassailable ground truth. Transparent model cards that document training data composition, known failure modes, and domain applicability can serve as a governance instrument for responsible deployment.

8. Policy and Governance in Computational Structural Biology

The integration of dynamic deep learning models into industrial drug development pipelines and publicly funded research infrastructures raises policy questions about validation, reproducibility, and intellectual property. Unlike traditional physics-based software packages whose algorithms are published and independently reimplementable, deep learning models often rely on proprietary weights, training configurations, and data curation processes that are not fully disclosed. Regulatory agencies such as the European Medicines Agency and the U.S. Food and Drug Administration are beginning to develop frameworks for the qualification of computational models used in drug development, but these frameworks were largely designed around mechanistic pharmacokinetic models and have yet to address the specific challenges of data-driven predictors of molecular properties. A consensus is needed on the acceptable validation protocols for pKa prediction models, including requirements for prospective benchmarking against blinded experimental datasets, sensitivity analyses to input conformational ensembles, and comparison with established physics-based reference methods.

Governance must also encompass data governance, given that sensitive protein structures may originate from proprietary pharmaceutical projects or biodefense-related research. Federated learning protocols and differential privacy guarantees can enable collaborative model improvement without exposing raw structural data. However, these privacy-preserving techniques introduce trade-offs with model accuracy and interpretability, requiring careful policy guidance on the acceptable level of noise injection in predictions that influence patient safety through drug design decisions. Open science initiatives that mandate the deposition of model architecture, training codes, and curated benchmark datasets alongside publications will accelerate reproducibility, but they must be balanced with incentives for industry participation. Finally, the spectre of dual-use must be acknowledged: high-accuracy pKa models for flexible proteins could, in principle, be misapplied to engineer pH-sensitive toxins or pathogens. While the risk is currently low, proactive engagement between the computational biology community and biosecurity experts is advisable to establish norms and oversight mechanisms that do not stifle beneficial research while safeguarding against malicious exploitation.

9. Conclusion

Dynamic conformation-aware deep learning represents a paradigm shift in the computational prediction of residue ionization states, moving from rigid, static views of proteins to representations that embrace the inherent flexibility of biomolecular systems. The system we have described brings together equivariant geometric learning, cross-conformation attention, and principled ensemble inference to produce calibrated pKa predictions that capture the breadth of conformational space. Yet the value of such a system extends only as far as the infrastructure that sustains it. Data curation, simulation provenance tracking, robustness

verification, and fairness auditing are not peripheral afterthoughts but core components that determine whether the technology can be safely and equitably deployed across the diverse landscape of protein science. The interplay between architecture design and system governance must be treated as a unified design problem. Without transparent benchmarking protocols, inclusive training data, and environmentally conscious deployment strategies, even the most elegantly designed model risks perpetuating the biases and blind spots of the static era. As we advance toward conformation-aware artificial intelligence for biophysics, the community must build not only more powerful models but also the sociotechnical scaffolding that ensures those models serve the entire breadth of biological inquiry with integrity, reproducibility, and fairness.

References

1. Tanford, C., & Roxby, R. (1972). Interpretation of protein titration curves: Application to lysozyme. Biochemistry, 11(11), 2192–2198.

2. Swenson, J. M., & Bolen, D. W. (1999). A thermodynamic linkage between protein stability and proton binding. Biophysical Chemistry, 80(3), 157–167.

3. Bashford, D., & Karplus, M. (1990). pKa's of ionizable groups in proteins: Atomic detail from a continuum electrostatic model. Biochemistry, 29(45), 10219–10225.

4. 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.

5. 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.

6. Ingraham, J., Garg, V., Barzilay, R., & Jaakkola, T. (2019). Generative models for graph-based protein design. Advances in Neural Information Processing Systems, 32.

7. Chen, Y., Lu, C., Li, M., Wang, S., & Gao, X. (2021). DeepKa: A deep learning based approach to predict pKa values of ionizable residues in proteins. Bioinformatics, 37(16), 2646–2653.

8. Noé, F., Olsson, S., Köhler, J., & Wu, H. (2019). Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. Science, 365(6457), eaaw1147.

9. 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.

10. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447–453.

11. Hie, B. L., Yang, K. K., & Kim, P. S. (2022). Evolutionary velocity with protein language models predicts mutational effects. Nature Biotechnology, 40(7), 1114–1122.

12. Towns, J., Cockerill, T., Dahan, M., Foster, I., Gaither, K., Grimshaw, A., ... & Wilkins-Diehr, N. (2014). XSEDE: Accelerating scientific discovery. Computing in Science & Engineering, 16(5), 62–74.

13. Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. Communications of the ACM, 63(12), 54–63.

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. Anthis, N. J., & Clore, G. M. (2015). Visualizing transient dark states by NMR spectroscopy. Quarterly Reviews of Biophysics, 48(1), 35–116.

16. Gapsys, V., Michielssens, S., Seeliger, D., & de Groot, B. L. (2015). pmx: Automated protein structure and topology generation for alchemical perturbations. Journal of Computational Chemistry, 36(5), 348–354.

17. Beuming, T., & Sherman, W. (2012). Current assessment of scoring functions. Journal of Chemical Information and Modeling, 52(10), 2473–2480.

18. Lu, C., Liu, C., Zhang, Z., & Maragakis, P. (2022). Machine learning prediction of pKa. Current Opinion in Structural Biology, 74, 102377.

19. Riniker, S. (2017). Fixed-charge atomistic force fields for molecular dynamics simulations in the condensed phase: An overview. Journal of Chemical Information and Modeling, 58(3), 565–578.

20. Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a “right to explanation.” AI Magazine, 38(3), 50–57.