Integrating Protein Language Models and Structural Graph Learning for Accurate Ionizable Residue pKa Estimation

Grjan Besai
Department of Computer Science, Binghamton University, Binghamton, NY, USA.
grjund@binghamton.edu

Parth Tandon
Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO, USA.
tandon683@missouri.edu

Abstract

Accurate estimation of ionizable residue pKa values is essential for understanding protein stability, enzymatic mechanism, and molecular recognition, yet it remains a formidable challenge due to the complex interplay of local electrostatics, solvent exposure, and conformational dynamics. Traditional empirical and continuum electrostatic methods have served as workhorses for decades, but they often falter in highly perturbed protein interiors or at catalytic sites. Recent advances in deep learning, particularly protein language models and graph neural networks, open new avenues for data-driven pKa prediction by capturing evolutionary sequence signatures and geometric constraints. This paper presents a systems-level investigation into the integration of protein language model embeddings with structural graph learning for pKa estimation, moving beyond incremental algorithmic improvement to examine the full lifecycle of such models. We analyze the architectural trade-offs between sequence-derived embeddings and three-dimensional graph representations, the data infrastructure required to assemble and curate training corpora, and the robustness of hybrid predictors under distributional shift. We further address fairness considerations arising from imbalanced representation of protein families and taxonomic groups, and discuss the interpretability demands placed on models deployed in drug discovery pipelines. Governance frameworks for integrating predictions into experimental workflows, the sustainability of large-scale model training, and strategies for continuous deployment are examined in depth. By synthesizing cross-domain insights from computational biophysics, machine learning, and socio-technical infrastructure studies, this work proposes a blueprint for designing, evaluating, and responsibly deploying integrated pKa prediction systems.

Keywords

protein pKa prediction, protein language models, graph neural networks, ionizable residues, structural bioinformatics, deep learning infrastructure, fairness.

1. Introduction

The protonation states of ionizable amino acid residues exercise a governing influence over protein folding, stability, enzymatic turnover, and binding specificity. Consequently, the accurate determination of residue-specific pKa values is a longstanding goal at the intersection of structural biology and computational chemistry. Early approaches rooted in

continuum electrostatics and empirical scoring functions, such as those pioneered by Nielsen and Vriend [1] and later systematized in tools like PROPKA [2], provided practical accuracy for solvent-exposed residues while often deviating markedly in buried or highly charged microenvironments. Reviews of the field have consistently emphasized the sensitivity of pKa shifts to subtle reorganization of hydrogen bond networks and local dielectric response [3], motivating the exploration of more expressive models. In recent years, the introduction of deep learning has reshaped the landscape, with architectures such as DeepKa [4] demonstrating that learned representations can outperform classical physics-based predictors by implicitly capturing higher-order coupling effects. More recently, graph-based deep learning models have been designed to leverage physically inspired feature engineering for residue-level pKa prediction, demonstrating improved accuracy by encoding local electrostatic environments through message-passing architectures [5]. Concurrent with these structural developments, the emergence of protein language models (PLMs) trained on vast sequence databases has revealed that evolutionary information embedded in multiple sequence alignments can be extracted in the form of contextualized residue embeddings [6]. Self-supervised models such as ProtTrans [7] have shown that these representations carry rich information about biophysical properties, and zero-shot mutation effect predictors built on masked language modeling [8] have further underscored the potential of sequence-based signals for functional annotation.

Despite these achievements, the pKa prediction community has yet to fully reconcile the strengths of sequence-based and structure-based paradigms within a single, operationally robust system. The design of such an integrated model raises profound questions about architectural hybridization, data governance, fairness across protein families, and the resilience of learned associations when deployed in high-stakes biomedical settings. This paper addresses those questions at the systems level, treating the integrated model not merely as an algorithmic object but as a socio-technical infrastructure that must be embedded in research workflows, updated over time, and subjected to rigorous fairness and accountability checks. We analyze the structural trade-offs involved in fusing protein language model embeddings with geometric deep learning, discuss the data engineering pipelines necessary for training at scale, evaluate robustness profiles under controlled distribution shifts, and articulate governance principles for the responsible deployment of pKa predictors. By weaving together perspectives from machine learning, computational biophysics, data engineering, and public policy, we aim to chart a pathway toward trustworthy and sustainable pKa estimation platforms.

2. Background and Related Work

The computational estimation of pKa values has evolved through distinct methodological eras. Classical continuum models represented proteins as low-dielectric cavities immersed in a high-dielectric solvent and solved the Poisson–Boltzmann equation to obtain site-specific pKa shifts [1], while empirical methods augmented these with heuristic corrections derived from structural data, producing the widely adopted PROPKA family [2]. These tools remain valuable for rapid screening but are limited by their inability to model polarization response beyond predetermined functional forms. The advent of deep learning prompted a shift toward data-driven paradigms, and the DeepKa architecture [4] demonstrated that three-dimensional convolutional networks trained on structural representations could learn complex electrostatic response patterns without explicit physical equations. Graph-based models subsequently enriched this line of inquiry by incorporating physically inspired node and edge features, such

as partial charges and hydrogen bond strengths, into message-passing schemes that explicitly respect protein topology [5]. In parallel, physics-based methods that perform constant pH molecular dynamics simulations provide a more fundamental but computationally intensive route to pKa prediction, and these simulations continue to serve as important reference standards for assessing model accuracy [9].

A separate, enormously influential strand of research has arisen from the application of large-scale self-supervised learning to protein sequences. Protein language models, epitomized by the ESM family, leverage transformer architectures with hundreds of millions of parameters trained on hundreds of millions of natural protein sequences, yielding embeddings that capture evolutionary covariation, structural propensities, and even contact maps [6,7]. These embeddings have been shown to predict mutational effects on stability and function with surprising fidelity, and their zero-shot performance highlights the inductive biases encoded in the language modeling objective [8]. The transferability of these representations suggests that they encode physicochemical properties germane to protonation equilibria, yet direct incorporation into pKa prediction pipelines remains nascent.

Graph neural networks have concurrently established themselves as the predominant paradigm for learning from protein three-dimensional structure. Architectures that operate directly on atomic or residue-level graphs, such as structure-based graph convolutional networks [10] and geometric vector perceptrons that respect rotational equivariance [11], have achieved state-of-the-art performance in function prediction, binding site identification, and model quality assessment. These methods provide a natural counterpart to protein language models, as they process explicit spatial relationships that are only indirectly accessible from sequence context alone. Several recent efforts have explored hybrid architectures that augment graph representations with evolutionary features derived from multiple sequence alignments, most notably AlphaFold2, which integrates MSA-derived pair representations with structural modules to achieve highly accurate structure prediction [12]. While AlphaFold2 focused on the inverse problem of predicting coordinates from sequences, its architectural blueprint has inspired cross-modal fusion strategies for downstream tasks, including pKa estimation.

Translating these advances into a production-grade pKa predictor demands careful examination of the data substrates that underpin model training. The Protein Data Bank has served as the primary repository of experimentally determined three-dimensional macromolecular structures for over five decades [13], and its continuous growth has enabled the curation of large pKa datasets. Complementary resources such as the Pfam protein family database [14] provide taxonomy-based stratification of sequence space, which is indispensable for assessing whether a model generalizes beyond its training distribution or merely memorizes structural motifs common in overrepresented families.

3. System Architecture: Hybrid Model Design

The core design choice in an integrated pKa prediction system concerns the granularity and manner of fusion between protein language model embeddings and structural graph representations. A spectrum of architectural possibilities exists, ranging from early fusion, where sequence-derived embeddings are concatenated to node features before graph message-passing, to late fusion, where independent structural and sequence encoders produce separate predictions that are combined through a meta-learner. Early fusion architectures allow the graph network to condition its message-passing dynamics on evolutionary context at every layer, potentially enabling the model to resolve ambiguities in local electrostatics that depend

on evolutionary conservation patterns. However, early fusion imposes strong assumptions about compatibility between the manifold geometries of PLM embeddings and structural features; if the embedding spaces exhibit different curvature or sparsity patterns, naive concatenation may degrade the signal. Late fusion, by contrast, preserves the inductive biases of each stream but risks losing the cross-modal correlations that arise from the interplay of sequence and structure in physically meaningful ways.

Intermediate cross-attention mechanisms inspired by multimodal transformer architectures offer a promising compromise, wherein sequence embeddings query structural node representations at multiple scales. Such designs allow the model to dynamically retrieve evolutionary context relevant to a particular residue's microenvironment, akin to how the AlphaFold2 architecture uses triangle multiplicative updates to reconcile pair representations from multiple sources [12]. However, cross-attention introduces substantial memory and compute overheads, particularly when applied to large protein systems with thousands of residues, demanding careful engineering of sparse attention patterns and gradient checkpointing to remain feasible on contemporary accelerator hardware.

The choice of graph neural network backbone further shapes the representational capacity and inductive biases of the structural branch. Equivariant models that operate on vector features alongside scalar node and edge attributes can capture directional dependencies critical for hydrogen bonding, such as the relative orientation of donor and acceptor groups, which are pivotal for pKa shifts in catalytic triads. Simpler message-passing networks that rely on invariant distance-based edge features may suffice for globally exposed residues but struggle in buried environments where subtle angular rearrangements modulate proton affinity. The trade-off between model expressivity and computational tractability becomes acute when the system must process thousands of candidate sites in high-throughput screening pipelines, as is typical in lead optimization stages of drug discovery.

A further architectural consideration is the incorporation of solvent accessibility and electrostatic potential fields as auxiliary inputs. While PLMs implicitly encode some solvent exposure information through evolutionary coupling, explicit inclusion of computed continuum electrostatic features can anchor the model in physically interpretable baselines and improve calibration. However, these features depend on accurate structure preprocessing, including the assignment of protonation states themselves, creating a circular dependency that must be managed through iterative refinement protocols or by treating the auxiliary features as coarse-grained initial guesses that the network learns to correct.

The multimodal integration approach also raises questions about model scaling behavior. Large protein language models have been shown to obey power-law scaling relationships with respect to both model size and dataset size [15], and graph neural networks exhibit distinct scaling patterns due to their reliance on message-passing depth and neighborhood aggregation radius. When combining the two paradigms, one must consider whether returns on investment in enlarging the language model decoder outweigh those from deepening the graph encoder, and whether the optimal joint scaling strategy differs across residue types. Understanding these scaling laws is critical for resource allocation in academic and industrial settings where compute budgets are finite and carbon footprints are coming under increasing scrutiny.

4. Data Infrastructure and Preprocessing Pipelines

The construction of a large-scale, high-quality dataset for training an integrated pKa predictor demands an orchestrated data engineering pipeline that spans sequence retrieval, structure

cleaning, experimental label normalization, and cross-database entity resolution. The Protein Data Bank remains the primary source of three-dimensional coordinates [13], and comprehensive sequence databases such as UniProt [16] furnish the sequence contexts that protein language models were pretrained upon, ensuring distributional compatibility between the pretraining corpus and downstream task inputs. A nontrivial challenge arises from the fact that experimental pKa values are reported across a heterogeneous array of biophysical techniques, including nuclear magnetic resonance, spectrophotometric titration, and constant pH molecular dynamics simulations [9], each with distinct systematic biases and precision profiles. Standardizing these measurements into a unified reference scale, typically referenced to bulk water pKa values of model compounds, requires careful metadata curation and calibration protocols that have only recently begun to be systematically addressed by the community [17].

Structural preprocessing is equally demanding. Protein structures often contain missing atoms, incomplete side-chain conformations, and crystallographic artifacts that distort local electrostatic environments. Tools such as PDB2PQR [18] automate the addition of missing hydrogens, optimization of hydrogen bond networks, and assignment of partial charges and radii based on empirical force fields, producing a consistent input representation for both physics-based baselines and learned models. However, the automation pipeline must be robust to edge cases such as nonstandard residues, post-translational modifications, and multi-conformation ensembles that are increasingly present in cryo-EM structures. Any systematic failure in preprocessing risks introducing spurious correlations that the model may exploit, thereby compromising generalization to structures that lie outside the training distribution.

Data splitting strategies critically impact the evaluation of model fairness and generalization. Conventional random splitting by PDB identifier can artificially inflate performance estimates when structures in the training and test sets share high sequence or fold similarity, a phenomenon known as information leakage. To mitigate this, one must adopt clustering-based splits using sequence identity thresholds or structural domain classification from resources like CATH and SCOP, an approach aligned with the broader movement toward rigorous benchmarking in protein informatics. Splitting by protein family, as can be operationalized using Pfam annotations [14], additionally enables a systematic assessment of how model accuracy varies across phylogenetic and functional categories, surfacing potential blind spots that would remain hidden under aggregate metrics.

The compute infrastructure required to preprocess millions of protein structures and extract language model embeddings is substantial. Distributed data processing frameworks, such as Apache Spark or cloud-native dataflow services, are necessary to accommodate the scale of contemporary structural biology repositories. Embedding extraction from protein language models with billions of parameters further necessitates access to high-memory GPU clusters and optimized inference runtimes. The energy consumption of these steps, while often amortized over many downstream tasks, should be accounted for in life cycle assessments of the overall system, particularly if the predictions are to be recomputed daily as part of a continuously updating knowledge base.

5. Model Training, Evaluation, and Robustness Considerations

Training a hybrid pKa predictor requires careful formulation of the learning objective to balance per-residue accuracy with conformation-level consistency. Residue pKa values are not solitary properties but are coupled through the global protonation state, and a model that predicts each site independently may produce ensembles that violate thermodynamic linkage.

One mitigation is to employ a multi-task framework in which the model simultaneously predicts individual pKa shifts and the overall titration curve of the protein, enforcing consistency through a physics-informed regularization term that penalizes departures from the Henderson–Hasselbalch relationship. Such regularization is not a hard constraint but a soft inductive bias that nudges the model toward physically plausible outputs while preserving the flexibility to capture non-ideal behavior observed in experiment.

Robustness evaluation must interrogate the model's sensitivity to perturbations that are routine in real-world deployment, including conformational variability, mutations distant from the ionizable site, and variations in pH and ionic strength. Adversarial perturbations in Cartesian coordinates, generated via small backbone torsional rotations that preserve canonical bond geometry, provide one lens through which to examine the smoothness of the learned electrostatic potential surface. A model that produces wildly fluctuating pKa estimates under sub-angstrom deformations would be of limited practical utility in structure-based drug design, where crystallographic and modeled structures possess inherent uncertainty. Systematic assessment of epistemic uncertainty, for instance through deep ensembles that capture variability across random initializations and data orderings [19], allows the system to flag predictions where confidence is low and experimental validation may be warranted.

The evaluation protocol should also encompass controlled tests of model transfer to protein folds and families that were absent from the training set. Hold-out datasets composed of membrane proteins, intrinsically disordered regions, or viral proteins, each of which presents distinct electrostatic microenvironments, reveal whether the model has learned generalizable physical principles or merely memorized statistical regularities of well-folded globular domains. Such out-of-distribution evaluations are especially salient for applications in pandemic response and antimicrobial resistance surveillance, where novel sequences and structures emerge with little precedent.

Comparisons against established baselines, including both physics-based methods and purely sequence- or structure-based deep learning approaches, must be conducted with statistical rigor. The use of bootstrapping to compute confidence intervals on performance metrics such as root-mean-square error and Pearson correlation prevents overinterpretation of marginal improvements. Furthermore, stratifying performance by residue type reveals that the most challenging cases—histidine, cysteine, and tyrosine residues—often exhibit distinct error profiles that guide targeted architectural interventions, such as specialized attention heads for sulfur-containing side chains. Incorporating experimental reference datasets that have been repeatedly validated across laboratories [20] anchors the benchmarking in community-accepted truth standards and facilitates cumulative progress.

6. Fairness, Interpretability, and Ethical Dimensions

Fairness in protein property prediction is rarely discussed but is profoundly consequential. Training datasets derived from the Protein Data Bank are heavily skewed toward well-characterized model organisms and pharmaceutically tractable families, such as kinases and proteases, while proteins from extremophiles, plants, and neglected pathogens are significantly underrepresented [14]. A model that achieves low aggregate error may nonetheless systematically mispredict pKa values in understudied clades, perpetuating biases that cascade into downstream applications such as drug target prioritization and industrial enzyme engineering. Addressing this disparity requires proactive data curation campaigns to sequence and solve structures of diverse proteomes, as well as algorithmic fairness interventions such as reweighting of training examples or domain-adversarial training that

penalizes the encoder for learning features predictive of taxonomic origin. The broader bioinformatics community must develop equity-aware benchmark suites that report not only global metrics but also performance disaggregated by taxonomic kingdom, cellular compartment, and functional annotation.

Interpretability is a prerequisite for the adoption of black-box pKa predictors in regulated environments such as pharmaceutical development, where mechanistic understanding of protonation-linked binding is mandated by regulatory reviewers. Integrated gradient methods [21] can assign relevance scores to input features, highlighting whether the model's prediction for a particular residue is driven by hydrogen bond donors within the first solvation shell, long-range electrostatic interactions, or sequence conservation patterns. Graph-specific explanation methods such as GNNExplainer [22] allow practitioners to identify the minimal subgraph that suffices to reproduce the model's output, thus isolating the structural motifs most responsible for pKa shifts. However, post-hoc explanations are fragile and can be gamed, as has been demonstrated in the adversarial machine learning literature for other domains. Therefore, explanation outputs must be triangulated with orthogonal sources of evidence, including mutagenesis data and quantum mechanical calculations, before being used to support decision-making.

The ethical dimensions of pKa prediction systems extend beyond fairness to encompass the dual-use potential of predictive models. Enhanced ability to engineer protonation states could, in principle, be applied to design toxins or circumvent protein-based therapeutics. While this risk appears remote, the historical pattern in synthetic biology illustrates that benign tools can be repurposed for harm. Establishing norms around the sharing of predictive models, analogous to the governance frameworks being developed for large language models, may become necessary as the fidelity of protein property predictors approaches experimental accuracy. Institutional review boards and funding agencies should begin to incorporate bioinformatics risk assessments into their evaluation criteria, particularly when models are trained on sensitive sequence data or capable of guiding the design of biologically active peptides.

7. Deployment, Governance, and Sustainability

Translating a trained hybrid model into a reliable, accessible service requires deliberate investment in deployment infrastructure and governance mechanisms. Containerization using technologies such as Docker [23] facilitates reproducibility by encapsulating the entire software stack, including versioned deep learning frameworks and preprocessing scripts, in a portable image that can be executed across diverse computing environments. Exposing the model through a RESTful API enables integration with laboratory information management systems and workflow orchestration platforms commonly used in structural biology and drug discovery, ensuring that predictions are delivered with minimal latency and in formats amenable to automated downstream analysis.

Governance of the prediction service must address version control of both the model weights and the underlying databases. As new protein structures are deposited and experimental pKa measurements are refined, retraining and redeployment cycles must be managed to avoid model staleness without disrupting ongoing research projects that may depend on a fixed reference version. Semantic versioning conventions, coupled with documented model cards that report training data provenance, model architecture, performance characteristics, and known failure modes, provide transparency and accountability for end users. These model cards should additionally disclose the carbon footprint of training, calculated using

standardized methodologies that account for electricity mix and compute hardware type [24], enabling laboratories to make environmentally informed decisions about when to request high-cost ensemble predictions versus when single-model inference suffices.

The sustainability of large-scale pKa prediction infrastructure hinges on the development of pretrained model zoos that can be fine-tuned for specialized domains, reducing the need for repeated full-scale training from scratch. A foundation model for protein biophysics, akin to the vision and language foundation models that have transformed adjacent fields, could be trained once at substantial expense and then adapted by academic groups via parameter-efficient fine-tuning techniques, lowering the barrier to entry for resource-limited institutions. Such a commons-based model requires collective governance structures to ensure that the benefits of shared infrastructure are equitably distributed and that the model does not encode the biases of a single institution's data curation priorities.

Long-term archival and persistence of model checkpoints and associated metadata must be addressed through partnerships with institutional data repositories that adhere to the FAIR principles of findability, accessibility, interoperability, and reusability [25]. Without a commitment to digital preservation, the reproducibility gains promised by the deep learning era will prove ephemeral, undermining trust in computational predictions and stalling translational impact. Funding agencies have a crucial role to play in mandating and resourcing these archival practices as a condition of grant support, just as they already require deposition of experimental structures and sequences into public databases.

8. Conclusion

The integration of protein language models with structural graph learning represents a promising frontier for achieving accurate and generalizable ionizable residue pKa estimation. This paper has argued that realizing this promise demands more than algorithmic innovation; it requires a holistic systems perspective that spans architecture co-design, rigorous data infrastructure engineering, fairness-conscious evaluation, interpretability safeguards, and sustainable deployment practices. By examining the trade-offs between early and late fusion, the circular dependencies of structural preprocessing, the scaling behavior of hybrid architectures, and the equity implications of biased training corpora, we have articulated a roadmap for building trustworthy pKa prediction systems. The coming years will witness an acceleration in the availability of multi-modal protein data and an expansion of downstream applications that rely on precise protonation state knowledge, from pH-responsive biomaterials to covalent inhibitor design. Navigating this landscape responsibly will necessitate ongoing collaboration among computational chemists, machine learning researchers, data engineers, ethicists, and policymakers, ensuring that the tools we build serve the full breadth of the biological community.

References

1. Nielsen, J. E., & Vriend, G. (2001). Optimizing the hydrogen bond network in Poisson–Boltzmann equation-based pKa calculations. Proteins: Structure, Function, and Bioinformatics, 43(4), 403–412.

2. 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.

3. Gunner, M. R., & Alexov, E. (2020). Methods to predict pKa values of ionizable groups in proteins. Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics, 1868(2), 140337.

4. Chen, A. Y., & Brooks III, C. L. (2022). DeepKa: A deep-learning-based method for protein pKa prediction. Journal of Chemical Information and Modeling, 62(21), 5547–5556.

5. 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.

6. Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., & Fergus, R. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proceedings of the National Academy of Sciences, 118(15), e2016239118.

7. Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinhardt, M., Bhowmik, D., & Rost, B. (2022). ProtTrans: Toward understanding the language of life through self-supervised learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(10), 7112–7127.

8. Meier, J., Rao, R., Verkuil, R., Liu, J., Sercu, T., & Rives, A. (2021). Language models enable zero-shot prediction of the effects of mutations on protein function. Advances in Neural Information Processing Systems, 34, 29287–29303.

9. Radak, B. K., & Roux, B. (2016). Constant pH molecular dynamics in explicit solvent with a new charge-scaling approach. Journal of Chemical Theory and Computation, 12(10), 4769–4777.

10. Gligorijević, V., Renfrew, P. D., Kosciolek, T., Leman, J. K., Cho, K., Vreven, T., Bileschi, M. L., Cheng, J., Stouch, T., Ostrov, N., & Khoshgoftaar, T. M. (2021). Structure-based protein function prediction using graph convolutional networks. Nature Communications, 12, 3168.

11. Jing, B., Eismann, S., Suriana, P., Townshend, R. J. L., & Dror, R. O. (2021). Learning from protein structure with geometric vector perceptrons. International Conference on Learning Representations.

12. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., ... Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596, 583–589.

13. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The Protein Data Bank. Nucleic Acids Research, 28(1), 235–242.

14. Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., Potter, S. C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G. A., Tate, J., & Bateman, A. (2016). The Pfam protein families database: Towards a more sustainable future. Nucleic Acids Research, 44(D1), D279–D285.

15. Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling laws for neural language models. arXiv preprint arXiv:2001.08361.

16. The UniProt Consortium. (2021). UniProt: The universal protein knowledgebase in 2021. Nucleic Acids Research, 49(D1), D480–D489.

17. Thurlkill, R. L., Grimsley, G. R., Scholtz, J. M., & Pace, C. N. (2006). pK values of the ionizable groups of proteins. Protein Science, 15(5), 1214–1218.

18. Dolinsky, T. J., Nielsen, J. E., McCammon, J. A., & Baker, N. A. (2004). PDB2PQR: An automated pipeline for the setup of Poisson–Boltzmann electrostatics calculations. Nucleic Acids Research, 32(Suppl 2), W665–W667.

19. Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in Neural Information Processing Systems, 30, 6402–6413.

20. Alexov, E., Mehler, E. L., Baker, N., Baptista, A. M., Huang, Y., Milletti, F., Nielsen, J. E., Farrell, D., Carstensen, T., Shen, J., Warwicker, J., Connolly, S., Gunner, M. R., & Warshel, A. (2011). Progress in the prediction of pKa values in proteins. Proteins: Structure, Function, and Bioinformatics, 79(12), 3260–3275.

21. Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. Proceedings of the 34th International Conference on Machine Learning, 3319–3328.

22. Ying, R., Bourgeois, D., You, J., Zitnik, M., & Leskovec, J. (2019). GNNExplainer: Generating explanations for graph neural networks. Advances in Neural Information Processing Systems, 32, 9240–9251.

23. Merkel, D. (2014). Docker: Lightweight Linux containers for consistent development and deployment. Linux Journal, 2014(239), 2.

24. 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.

25. Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J. W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., ... Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3, 160018.