Confidence-Weighted Elastic Gaussian Networks To Predict Protein Flexibility

Anonymous Authors[1]

Abstract

Rapid protein flexibility prediction is useful for evaluating novel folds generated by deep learning models. Gaussian Network Models (GNMs) provide a physics-inspired framework for this task but assume uniform spring constants, ignoring the per-residue confidence information available from modern structure predictors. We introduce a zero-parameter modification that weights the GNM spring constants with AlphaFold2 predicted Local Distance Difference Test (pLDDT) scores and inverse squared distance: $\gamma_{ij} = p_i p_j / d_{ij}^2$. On the ATLAS molecular dynamics benchmark (1932 proteins), the mean Pearson correlation increases from $r = 0.765$ to $r = 0.841$. Partial correlation analysis confirms that pLDDT contributes information beyond local geometry. A lightweight Graph Neural Network trained to correct analytical residuals reaches $r = 0.871$, approaching the inter-replica ceiling of $r = 0.88$. These results suggest that pLDDT encodes mechanical information that is not captured by contact geometry alone.

1. Introduction

Proteins are dynamic molecular machines whose biological activity depends on conformational flexibility, which therefore plays a key role in protein-based drug design. Key features, such as catalytic loops, allosteric sites, and binding interfaces, are governed by conformational dynamics that cannot be inferred from a single static structure (Bahar et al., 2010). Although molecular dynamics (MD) simulations provide a detailed view of such behavior based on principles of statistical mechanics, they remain computationally expensive. As generative models for protein design become more prevalent, there is demand for methods that can evaluate the dynamic properties of novel structures without relying on MD simulations. This is particularly evident when integrating drug design tools in agentic environments, where long MD simulations are not desirable.

Gaussian Network Models (GNMs) have served this role for decades (Bahar et al., 1997; Tirion, 1996). By modeling the protein as an elastic network of Cα atoms connected by harmonic springs, GNMs derive per-residue fluctuations analytically from the network topology. Their main limitation is the uniformity assumption, in which every contact within the cutoff is assigned the same spring constant regardless of local chemical or evolutionary context. Distance-weighted variants (wGNM, Hinsen 1999) partially address this by setting $\gamma_{ij} \propto 1/d_{ij}^2$, but contact geometry alone cannot distinguish a buried residue that evolution has kept structurally fixed from one that is locally packed yet conformationally variable. Two contacts at the same distance receive identical spring constants regardless of whether the underlying residues occupy conserved or flexible positions across homologs. Sequence-derived information is therefore necessary to capture this distinction.

AlphaFold2 (Jumper et al., 2021) produces a per-residue confidence score (pLDDT) alongside each predicted structure. High pLDDT at a given position indicates that the model predicted that residue with high confidence, typically because it occupies a well-defined position across evolutionary homologs. Low pLDDT indicates structural variability and hence is an indicator of disorder. A position that evolution has kept structurally fixed is likely to have stronger harmonic constraints in the folded state. We tested whether encoding this signal into GNM spring constants improves flexibility prediction (Figure 1).

The method requires as inputs a protein structure and per-residue pLDDT scores from AlphaFold2. For structures predicted by AlphaFold2 both are available in the output file. For experimental PDB structures, pLDDT must be obtained separately by running AlphaFold2 on the sequence (Mirdita et al., 2022). In the ATLAS benchmark used for evaluation, pLDDT was computed with AlphaFold2 Collab and provided alongside each entry (Vander Meersche et al., 2024).

We show that: (1) pLDDT-weighted spring constants improve GNM predictions across protein sizes with zero trainable parameters. (2) Partial correlation analysis on 1932 proteins confirms that pLDDT contributes flexibility infor-

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Submitted to the 2026 Workshop on Generative and Agentic AI for Biology (ICML 2026). Do not distribute.

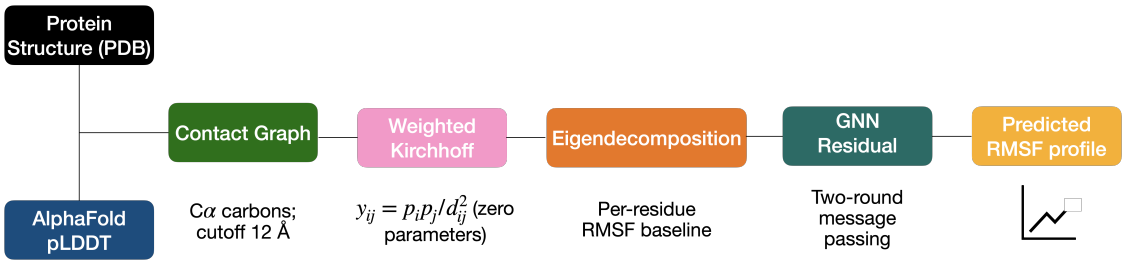

Figure 1. Method overview schematic.

mation not redundant with local packing geometry. (3) A GNN trained on this physics prior reached near the inter-replica performance ceiling, and (4) we characterize the conditions under which the method degrades.

2. Related Work

Elastic network models. The GNM (Bahar et al., 1997) and anisotropic network model (ANM, Atilgan et al. 2001) are standard tools for flexibility prediction. Distance-dependent spring constants were introduced by Hinsen (1999). Extensions incorporating residue types and empirical potentials have been explored. Using pLDDT as a multiplicative factor in GNM spring constants has not, to our knowledge, been proposed previously.

pLDDT in flexibility modeling. A closely related prior work is CABS-flex 3.0 (Wróblewski et al., 2025), which uses AlphaFold2 pLDDT scores to modulate distance restraints in coarse-grained simulations. In their Rigid-pLDDT mode, restraint strength is assigned based on per-residue pLDDT and secondary structure classification, validated against ATLAS. Our approach differs mechanistically as we incorporate pLDDT into the analytical Kirchhoff matrix as a multiplicative spring constant, without running simulations. The prediction is instantaneous once the structure and pLDDT are available, whereas CABS-flex runs a coarse-grained trajectory.

Machine learning for flexibility. Kouba et al. (2024) train Flexpert-Seq, a sequence-based predictor using a fine-tuned protein language model ($r = 0.76$ on ATLAS), and Flexpert-3D, which combines sequence embeddings with ANM predictions over MD-relaxed structures ($r = 0.82$ on ATLAS). Flexpert-3D uses MD-relaxed structures as input; our method uses PDB structures directly. These results are not directly comparable to our GNN evaluation as they use a different train/test split. Kouba et al. (2024) also establish the inter-replica correlation ceiling of $r = 0.88$ on ATLAS, which we use as a reference throughout.

3. Methods

3.1. pLDDT-Weighted GNM

Let the protein contact graph have nodes at Cα positions and edges connecting pairs with $d_{ij} < r_c = 12$ Å. We define the spring constant for each edge as:

$$\gamma_{ij} = \frac{p_i\, p_j}{d_{ij}^2}, \quad p_i = \frac{\mathrm{pLDDT}(i)}{100} \in [0, 1] \tag{1}$$

The Kirchhoff matrix $\mathbf{K}$ has off-diagonal entries $K_{ij} = -\gamma_{ij}$ and diagonal entries $K_{ii} = \sum_{j \neq i} \gamma_{ij}$. Per-residue reduced mean-square fluctuation (RMSF) is obtained from the pseudo-inverse via eigendecomposition, excluding the zero eigenvalue:

$$\mathrm{RMSF}(i) = \sqrt{\sum_{k:\, \lambda_k > 0} \frac{[\mathbf{u}_k]_i^2}{\lambda_k}} \tag{2}$$

The product $p_i p_j$ means the spring is strong only when both endpoints have high pLDDT. A disordered residue weakens all of its incident springs regardless of its neighbors, consistent with the expectation that a residue without a well-defined equilibrium position cannot transmit strong restoring forces through the contact network. AlphaFold2 pLDDT values are in practice always above zero; the minimum observed in the ATLAS benchmark is above 20, so $p_i > 0$ for all residues and all spring constants are strictly positive. If a residue had pLDDT $= 0$, its incident springs would be zero and that residue would be effectively disconnected from the contact network; this edge case does not occur in our data.

3.2. Partial Correlation Analysis

It is worth noting that pLDDT and local packing density may be correlated, since buried residues often have both many neighbors and high pLDDT, as both reflect structural conservation. Therefore, to test whether pLDDT contributes to flexibility information beyond geometry, we computed the partial correlation $r(\mathrm{pLDDT}, \mathrm{RMSF} \mid \mathrm{wdegree})$ for each protein, where $\mathrm{wdegree}(i) = \sum_j 1/d_{ij}^2$ captures weighted local density. The procedure regresses both pLDDT and

RMSF on wdegree separately and correlates the residuals. A zero partial correlation would indicate that pLDDT adds no information beyond geometry. A non-zero partial correlation means that the component of pLDDT not explained by geometry still predicts flexibility. The ΔR^2 metric quantifies how much additional variance is explained when pLDDT is added to a regression already containing wdegree.

3.3. Graph Neural Network

The analytical formula applies the same function for all residues regardless of broader structural context. We trained a GNN to learn a per-residue additive correction:

$$\widehat{\text{RMSF}}(i) = \text{RMSF}_{\text{base}}(i) + f_\theta(i) \tag{3}$$

where $\text{RMSF}_{\text{base}}$ is the pLDDT-wGNM output and f_θ is a two-round (i.e. two layers) message passing network. Node features are pLDDT (normalized), degree, wdegree, mean neighbor distance, pLDDT $\times$ wdegree, and $\text{RMSF}_{\text{base}}$. The output layer is initialized to zero weights so training starts at the physics baseline. The loss is $1 - r(\widehat{\text{RMSF}}, \text{RMSF}_{\text{MD}})$.

3.4. Benchmark and Evaluation

We used the ATLAS database (Vander Meersche et al., 2024), which provides PDB structures, AlphaFold2 pLDDT scores, and per-residue RMSF from 500 ns MD simulations in triplicate for 1932 proteins after filtering. Performance is measured as mean per-protein Pearson r between predicted and MD RMSF. Analytical methods (GNM, wGNM, pLDDT-GNM, pLDDT-wGNM) have no trainable parameters and are evaluated on all 1932 proteins. For the GNN, we used 800 proteins for training and 200 held-out proteins for validation. All GNN results are reported on these 200 proteins. The model checkpoint was selected based on best validation Pearson r, so reported GNN performance may be slightly optimistic; cross-validation is left for future work. For direct comparison, pLDDT-wGNM is also evaluated on the same 200 proteins ($r = 0.828$), separate from its full 1932-protein evaluation.

4. Results

4.1. Zero-Shot Performance

Table 1 reports results on all 1932 proteins. All adjacent comparisons are significant by paired t-test ($p < 0.0001$). pLDDT-GNM uses pLDDT as the sole spring weight without distance weighting and achieves $r = 0.814$, outperforming wGNM ($r = 0.800$) by $\Delta = +0.015$ ($p < 0.0001$). This indicates that pLDDT encodes flexibility-relevant information beyond geometric proximity. Combining both signals in pLDDT-wGNM gives the best zero-shot result.

Table 1. Mean Pearson r on ATLAS. Analytical methods are evaluated on all 1932 proteins. pLDDT-wGNM (val) and pLDDT-GNN are both evaluated on the same 200 held-out validation proteins.

Method	Mean r	Input
GNM (uniform)	0.765	PDB
wGNM ($1/d^2$)	0.800	PDB
pLDDT-GNM	0.814	PDB + pLDDT
pLDDT-wGNM	0.841	PDB + pLDDT
pLDDT-wGNM (val)	0.828	PDB + pLDDT
pLDDT-GNN	0.871	PDB + pLDDT

Gains are consistent across protein sizes: $+0.091$ for small proteins ($L < 100$, $n = 394$), $+0.075$ for medium ($100 \leq L < 300$, $n = 1058$), and $+0.067$ for large ($L \geq 300$, $n = 480$).

4.2. pLDDT Contributes Beyond Geometry

The partial correlation $r(\text{pLDDT}, \text{RMSF} \mid \text{wdegree})$ across all 1932 proteins is -0.562 ± 0.254 ($t = -97.06$, $p < 0.0001$), significant in 1809 of 1932 proteins. The negative sign reflects that, after accounting for local packing, residues with high pLDDT tend to be less flexible than wdegree alone predicts. A buried residue may have many contacts (high wdegree) yet still be evolutionarily variable (low pLDDT), in which case geometry overestimates its rigidity; pLDDT corrects for this. Adding pLDDT to wdegree in a linear regression increases mean R^2 from 0.416 to 0.661, a gain of $\Delta R^2 = 0.244 \pm 0.169$ ($t = 63.43$, $p < 0.0001$).

4.3. Failure Analysis

Stratifying by the within-protein standard deviation of pLDDT reveals a clear pattern (Table 2). When pLDDT varies substantially within a protein (Q4, SD > 9.7), mean $r = 0.901$. When pLDDT is nearly uniform (Q1, SD < 4.0), performance drops to $r = 0.765$, matching the uniform GNM baseline. If all residues have similar pLDDT, the modification to spring constants is nearly uniform and the model reduces to wGNM.

Table 2. Performance stratified by within-protein pLDDT standard deviation ($n = 1932$ proteins, quartile split).

Quartile	SD range	n	Mean r
Q1 (low var.)	0.0 to 4.0	483	0.765
Q2	4.0 to 6.5	483	0.828
Q3	6.5 to 9.7	483	0.870
Q4 (high var.)	> 9.7	483	0.901

Inspection of the 20 worst-performing proteins reveals two patterns. The first is proteins with uniformly high pLDDT and very low variance (SD < 2.0), such as 2ad6_A ($r = 0.349$, mean pLDDT $= 98.7$, SD $= 0.7$) and 1gnt_A

($r = 0.344$, SD $= 0.8$). These are well-ordered proteins where AlphaFold2 is confident everywhere, leaving pLDDT with no discriminative power. The second pattern is proteins with low mean pLDDT and high variance (SD > 15), such as 1dd3_A ($r = 0.224$, mean pLDDT $= 78.7$, SD $= 19.3$), which likely contain intrinsically disordered regions where the GNM framework is less appropriate. Overall, 82 proteins (4.2%) have $r < 0.6$ and 4 proteins (0.2%) have $r < 0.3$.

4.4. Graph Neural Network

The pLDDT-wGNM zero-shot model is the primary contribution of this work. As an extension, we trained a GNN to learn residual corrections on top of the physics baseline, asking whether additional performance can be recovered through learning. On the 200 held-out validation proteins, pLDDT-wGNM achieves $r = 0.828$. The GNN trained on the remaining 800 proteins reaches $r = 0.871$ on the same set ($\Delta = +0.043$, $t = 11.0$, $p < 0.0001$, win rate 85%). An ablated model trained without $\text{RMSF}_{\text{base}}$ as a node feature achieves $r = 0.869$, a difference of 0.002, confirming the gain reflects learned graph-structural patterns rather than rescaling of the physics baseline.

5. Conclusion

As generative models for protein design continue to expand the known fold space, there is a critical need for rapid evaluators of conformational stability. Our zero-shot learning pLDDT-weighted GNM provides a computationally efficient feedback mechanism that can be integrated into generative pipelines or agentic workflows to ensure that designed sequences possess the desired mechanical properties. Despite these advantages, the method's performance is intrinsically linked to the quality and variance of the input pLDDT scores.

Limitations

The method requires both a protein structure and AlphaFold2 pLDDT scores. For experimental PDB structures these must be obtained separately, typically by running ColabFold on the sequence. Performance degrades when pLDDT shows low within-protein variance. The GNN was evaluated on a single train/validation split; cross-validation would give more robust estimates. The ATLAS benchmark was constructed from high-quality X-ray structures of well-folded proteins; performance on intrinsically disordered proteins or computationally designed structures has not been assessed. Comparisons with Flexpert use published numbers that may not correspond to the same train/test split as our GNN evaluation.

Impact Statement

This paper presents work in machine learning applied to structural biology. The methods could be used to evaluate flexibility of computationally designed proteins. We do not foresee specific ethical concerns beyond those generally associated with advances in protein design.

References

Atilgan, A. R. et al. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophysical journal*, 80(1):505–515, 2001.

Bahar, I., Atilgan, A. R., and Erman, B. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Folding and Design*, 2(3):173–181, 1997.

Bahar, I., Lezon, T. R., Yang, L.-W., and Eyal, E. Global dynamics of proteins: bridging between structure and function. *Annual review of biophysics*, 39:23–42, 2010.

Hinsen, K. Analysis of domain motions by approximate normal mode calculations. *Proteins: Structure, Function, and Bioinformatics*, 33(3):417–429, 1999.

Jumper, J. et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Kouba, P., Planas-Iglesias, J., Damborsky, J., Sedlar, J., Mazurenko, S., and Sivic, J. Learning to engineer protein flexibility. *arXiv preprint arXiv:2412.18275*, 2024.

Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., and Steinegger, M. ColabFold: making protein folding accessible to all. *Nature Methods*, 19:679–682, 2022.

Tirion, M. M. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Physical review letters*, 77(9):1905, 1996.

Vander Meersche, Y. et al. Atlas: protein flexibility description from atomistic molecular dynamics simulations. *Nucleic Acids Research*, 52(D1):D384–D392, 2024.

Wróblewski, K., Zalewski, M., Kuriata, A., and Kmiecik, S. CABS-flex 3.0: an online tool for simulating protein structural flexibility and peptide modeling. *Nucleic Acids Research*, 53:W95–W101, 2025.

A. Kirchhoff Matrix Validation

The modified Kirchhoff matrix satisfies all required properties for all 1932 proteins in the benchmark. Symmetry error was below 10^{-15} in all cases. Row sums were below 10^{-15} in all cases. All proteins had exactly one zero eigenvalue, corresponding to the rigid-body translation mode. All remaining eigenvalues were strictly positive, confirming that the pseudo-inverse is well-defined. These properties hold by construction whenever $\gamma_{ij} > 0$ for all edges, which is guaranteed since $p_i > 0$ and $d_{ij} > 0$ for all connected pairs.

B. Residue Alignment Verification

For each protein we verified that the number of Cα atoms in the PDB file matches the number of rows in both the RMSF TSV and the pLDDT TSV provided by ATLAS. Spot checks confirmed exact alignment for 1ab1_A (46 residues), 1a62_A (130 residues), and 16pk_A (415 residues). Proteins where any file had fewer than 15 residues after alignment were excluded from the benchmark.

C. Comparison with CABS-flex pLDDT Integration

CABS-flex 3.0 (Wróblewski et al., 2025) also integrates AlphaFold2 pLDDT into a flexibility prediction pipeline through its Rigid-pLDDT mode. In CABS-flex, pLDDT is used to assign distance restraint categories in a coarse-grained simulation where residues with high pLDDT receive stronger restraints, limiting their displacement during the trajectory. In our method, pLDDT modulates the spring constants of the analytical Kirchhoff matrix, so the prediction is obtained through a single eigendecomposition without any simulation.

The two approaches are therefore complementary. CABS-flex produces a trajectory and structural ensemble, useful for visualizing conformational changes and identifying allosteric pathways. Our method produces a per-residue RMSF profile in under one second per protein, which suits fast screening or use as a fitness function within iterative design workflows. Both approaches rely on the premise that pLDDT encodes information about mechanical rigidity, through different mechanistic implementations.

D. Failure Mode Analysis

Table 3 reports the complete failure stratification results across all 1932 proteins.

Table 3. Full failure analysis stratification.

Stratification	Group	n	Mean r	$r < 0.5$
by pLDDT SD	Q1 SD $\in [0, 4)$	483	0.765	—
	Q2 SD $\in [4, 6.5)$	483	0.828	—
	Q3 SD $\in [6.5, 9.7)$	483	0.870	—
	Q4 SD ≥ 9.7	483	0.901	—
by length	Small $L < 100$	394	0.863	4
	Medium 100 to 299	1058	0.847	15
	Large $L \geq 300$	480	0.811	12
by mean pLDDT	$[70, 80)$	12	0.785	2
	$[80, 90)$	134	0.888	1
	$[90, 100]$	1785	0.838	28

The pLDDT SD stratification is the most informative predictor of method performance. Mean pLDDT alone is a poor predictor, as the high-pLDDT group (≥ 90) contains the largest number of $r < 0.5$ cases in absolute terms (28 proteins), but this group also contains 93% of all proteins, so the relative failure rate is low. In practice, computing the within-protein SD of pLDDT takes one line of code and gives a reliable estimate of expected performance. If SD is below approximately 4.0, pLDDT-wGNM is unlikely to improve substantially over wGNM.

E. GNN Implementation Details

Architecture. Two message-passing rounds, hidden dimension 64, LayerNorm, SiLU activations, dropout 0.1. For each directed edge $(i \to j)$, the message is computed as MLP$([h_i; h_j])$ where h_i and h_j are the current hidden states. Messages are averaged at each destination node and added as a residual to the current hidden state. Output: Linear(65, 32), SiLU, Dropout(0.1), Linear(32, 1), zero-initialized output weights.

Training. AdamW optimizer, learning rate 10^{-3}, weight decay 10^{-4}, cosine annealing to 10^{-5} over 100 epochs, gradient clipping at norm 0.5. Best checkpoint selected by validation Pearson r. 800 training proteins, 200 validation proteins, random seed 42.

Ablation. Without $\text{RMSF}_{\text{base}}$ as a node feature: $r = 0.869$ vs. $r = 0.871$ with it included, a difference of 0.002. This confirms the performance gain comes from the other node features rather than from rescaling the physics baseline.