Sequence-context-aware decoding enables robust reconstruction of protein dynamics from crystallographic B-factors

Yiquan Wang*[1,2,3], Minnuo Cai[1,3], Yahui Ma[1], and Kai Wei *[1]

[1]Xinjiang Key Laboratory of Biological Resources and Genetic Engineering, College of Life Science and Technology, Xinjiang University, Urumqi, Xinjiang, China
[2]College of Mathematics and System Science, Xinjiang University, Urumqi, Xinjiang, China
[3]Shenzhen X-Institute, Shenzhen, China

Abstract

X-ray crystallography provides the majority of protein structures, yet the B-factors associated with these coordinates are often influenced by crystal packing and refinement protocols, limiting their utility for quantifying solution-state dynamics. Consequently, a gap remains between the abundance of static PDB structures and the availability of quantitative dynamic information. Here, we investigate the extent to which sequence context can help recover intrinsic motion from crystallographic data. We present the B-Factor Corrector (BFC), a fine-tuned protein language model that treats B-factor analysis as a sequence-to-dynamics translation task. By leveraging deep contextual embeddings, BFC mitigates crystal lattice effects to recover intrinsic flexibility, yielding a Pearson correlation of 0.80 with ground-truth conformational fluctuations derived from structural ensembles. Furthermore, the model significantly improves the reconstruction of fluctuation profiles in physical space (in Å, $r = 0.49$ vs baseline $r = 0.11$ upon rescaling), suggesting that large-scale static structural data can be repurposed to decode accurate relative dynamic patterns latent within the Protein Data Bank.

Introduction

Protein function is fundamentally dynamic, relying on conformational ensembles that extend beyond static coordinates [1–4]. While X-ray crystallography has contributed over 190,000 structures to the Protein Data Bank (PDB) [5], these models are traditionally viewed as static snapshots [6]. The atomic displacement parameter, or B-factor, provides the primary experimental metric for motion; however, its interpretation remains challenging. Influenced by non-biological factors such as crystal packing contacts, solvent conditions, and refinement heterogeneity, B-factors are frequently regarded as qualitative indicators of local disorder rather than quantitative measures of solution-state dynamics [7–10].

To bridge the gap between static structures and dynamic function, researchers typically rely on Molecular Dynamics (MD) simulations [11] or NMR spectroscopy [12]. While accurate, these methods are computationally expensive or experimentally demanding, leaving the vast majority of PDB entries without quantitative dynamic profiling. Analytical approaches like Gaussian Network Models (GNM) [13] and Anisotropic Network Models (ANM) [14] offer rapid approximations but, as they rely on coarse-grained contact topology while treating amino acids as uniform mechanical nodes, often overlook the subtle chemical specificity governed by the local sequence environment.

Here, we hypothesized that the discrepancy between crystallographic B-factors and solution-state dynamics is not random noise, but rather a systematic artifact encoded in the local physicochemical context. We reasoned that protein sequences store an evolutionary memory of structural plasticity that persists even when the physical coordinate is constrained by the crystal lattice [15]. To test this, we developed the B-Factor Corrector (BFC), a framework that leverages the deep contextual embeddings of the ESM-2 protein language model [16]. Unlike traditional physical models, BFC treats the recovery of dynamics as a sequence-context-aware translation task that filters out crystal-induced rigidity to reveal the underlying solution-state fluctuations. Our results indicate that static crystallographic data, when interpreted through the lens of a language model, contains latent signals governing protein motion. This allows for the high-fidelity reconstruction of relative flexibility patterns and a substantial improvement in estimating absolute fluctuation magnitudes, potentially enabling the retrospective dynamic analysis of the PDB at scale.

Results

Sequence context enables robust reconstruction of RMSF profiles

We first investigated whether a language model could learn the mapping between noisy crystallographic data and solution-state behavior. We benchmarked BFC on a comprehensive dataset of proteins containing both crystal structures and corresponding structural ensemble profiles derived from PDBFlex [17]. Comparison with these ensemble-based

*Corresponding authors: Yiquan Wang (ethan@stu.xju.edu.cn), Kai Wei (kaiwei@xju.edu.cn)

ground truths highlighted discrepancies in the raw crystallographic data. Our analysis showed that normalized experimental B-factors exhibited a low correlation with the ensemble RMSF (Pearson Correlation Coefficient, PCC = 0.15) in the test set (Figure 1b). This suggests that raw B-factors, often influenced by lattice contacts, may diverge significantly from intrinsic flexibility.

In contrast, BFC recovered a signal consistent with the consensus flexibility of the structural ensembles, yielding a PCC of 0.80 on the test set (Figure 1b). To rigorously assess the recovery of absolute physical magnitudes (in Å) independent of global scaling artifacts, we employed an oracle rescaling strategy where both baseline and model predictions were standardized to match the first and second moments of the ground-truth ensemble. Even with this correction, the traditional physical formula failed to capture the fluctuation landscape (PCC = 0.11, SCC = 0.25), indicating that the relative distribution of B-factors is fundamentally distorted. Conversely, BFC achieved a robust reconstruction of the residue-wise flexibility profile (PCC = 0.49, SCC = 0.80) and a significantly lower Mean Absolute Error (0.12 Å vs. 0.56 Å, Figure 1c). These results suggest that BFC effectively decodes the intrinsic dynamic sequence pattern latent within the static crystal structure, correcting local distortions that analytical formulas cannot address.

End-to-end fine-tuning captures biophysical nuances

To understand the biophysical basis of this performance, we isolated the contribution of sequence context by comparing BFC against baseline models (Figure 2a,b). A gradient-boosted decision tree (XGBoost) trained solely on B-factor values did not improve correlation significantly. This observation implies that the relationship between B-factors and RMSF is likely not a simple non-linear mapping and suggests that experimental data alone, without sequence context, may be insufficient for denoising. Furthermore, we compared BFC (fully fine-tuned) with a frozen ESM-2 model, where only the output head was trained. The frozen model yielded limited improvements, suggesting that general-purpose evolutionary representations may not be inherently aligned with specific physical dynamics.

The significant performance gap achieved by end-to-end fine-tuning suggests that the model adapts its attention mechanisms to encode sequence-dependent physical properties. Detailed stratification reveals that BFC effectively learns distinct dynamic signatures. It maintains high accuracy in buried hydrophobic cores (Figure 2d), likely by recognizing the stability conferred by high packing density. Conversely, in solvent-exposed loops, typically enriched with flexible residues like Glycine and Proline, the model achieves its most dramatic improvement over analytical baselines (Figure 2c). This indicates that BFC weighs sequence-encoded intrinsic propensities against lattice constraints to recover solution-state dynamics.

Denoising crystal packing artifacts in antibody CDR loops

The biological imperative for this decoding approach is best illustrated by its ability to identify and correct misleading structural signals within our test set. In the crystal structure of an antibody fragment (PDB: 3efd [18]), crystal packing interactions artificially stabilize the Complementarity-Determining Region (CDR) loops, leading to low B-factors that falsely suggest rigidity [19, 20]. A naive interpretation of the B-factors would fail to capture the functional plasticity required for antigen binding [21–23].

Guided by the sequence context of the CDR loops, BFC identified the experimental rigidity as a potential artifact. The model predicted a high-flexibility profile that aligned with the ground-truth ensemble variability, effectively correcting the discrepancy in the input signal (Figure 3). The increase in correlation for this specific case (+1.494 PCC) suggests that BFC does not merely smooth input data but captures sequence-dependent physical properties. The model appears to weigh sequence-encoded intrinsic propensities against experimentally observed lattice constraints, helping to recover solution-state dynamics masked by the crystallization process.

Deciphering disease-critical dynamics latent in the PDB

To validate the utility of our model for biological discovery, we applied BFC to the ATLAS database [24], an external repository of high-quality molecular dynamics simulations. Crucially, this dataset is distinct from our ensemble-based training distribution, allowing us to test whether the model generalizes from static structural clusters to time-resolved dynamic trajectories. We analyzed the SH2 domain of SLAM-associated protein (PDB: 1D4T [25]), a critical signaling module. Consistent with the functional nature of this region, the experimental B-factors for the loop spanning residues 68–73 exhibited a distinct peak of flexibility (Figure 4c).

BFC accurately reproduced this localized peak of flexibility. Structural analysis indicates that this loop acts as a gatekeeper for the binding of the SLAMF1 ligand (Figure 4e). Notably, the Arg68 residue, located at the apex of this predicted flexible region, is the site of a missense mutation associated with X-linked lymphoproliferative syndrome type 1 (XLP1) [26]. Statistical analysis confirmed that the BFC-predicted flexibility for this loop was significantly higher than the scaffold background ($p < 0.001$), identifying a functional hotspot and validating the model's capacity to capture disease-relevant dynamics from sequence context alone (Figure 4b). This capacity to pinpoint function-critical dynamics suggests that BFC can serve as a valuable tool for re-analyzing existing PDB entries to generate hypotheses regarding protein function and disease mechanisms.

Discussion

Our study offers a perspective that contrasts with the view that crystallographic B-factors are too noisy for quantitative dynamic analysis. By treating B-factors as a signal influenced by environmental factors, we show that deep learn-

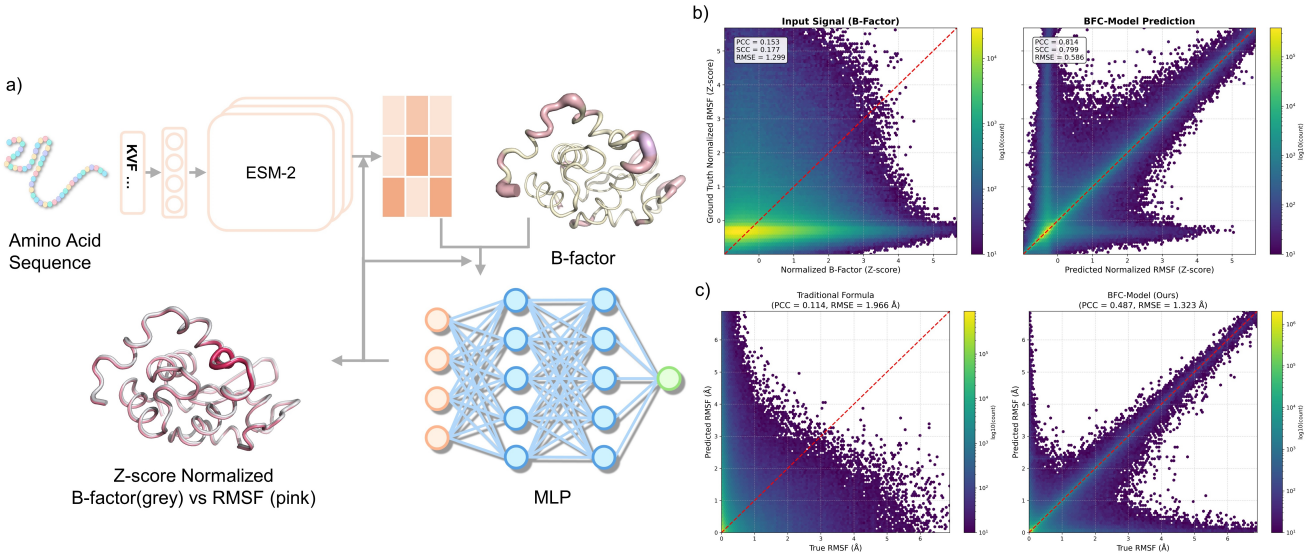

Figure 1: **Sequence-context-aware decoding enables quantitative recovery of RMSF from crystallographic B-factors. a,** Schematic of the B-Factor Corrector (BFC) workflow. The model extracts deep sequence embeddings via a fine-tuned ESM-2 and combines them with experimental B-factors to predict intrinsic dynamics. The structural visualizations illustrate the data transformation: the top structure displays the raw input B-factor distribution, while the bottom superposition reveals the discrepancy between the normalized input B-factors (grey tube) and the ground-truth ensemble RMSF (pink gradient). **b, c,** Global performance evaluation on the test set. Density scatter plots compare the correlation of predictions against ground-truth ensemble RMSF values. **b,** In normalized space (Z-score), BFC (right) effectively denoises the input B-factor signal (left), recovering a high correlation (PCC=0.80). **c,** Evaluation of profile reconstruction in physical units (Å). To isolate profile fidelity from experimental amplitude variance, both the Traditional Formula (left) and BFC (right) were rescaled to match the ground-truth moments (Oracle Rescaling). Even with correct global scaling, the traditional formula fails to capture the relative fluctuation ranking (SCC=0.25), whereas BFC accurately reconstructs the physical landscape (SCC=0.80), establishing a linear correlation with solution-state reality absent in raw data.

ing models can help infer underlying solution-state dynamics using sequence context. This mirrors recent advances in protein design [27–30], where deep learning bridges the gap between sequence and physical realizability. Our findings suggest that the ”noise” in B-factors is not random, but rather a convolution of intrinsic motion and environmental constraints. By leveraging the protein language model as an evolutionary prior, BFC bypasses the limitations of topology-based physical models to recover the intrinsic dynamic landscape directly from sequence. Crucially, unlike topological models that require complete complex coordinates to define mechanical constraints, BFC operates on isolated chains by inferring environmental effects—such as interface rigidification—that are implicitly imprinted onto the input B-factors. While the precise prediction of absolute fluctuation amplitudes remains constrained by the lack of prior knowledge regarding global scaling factors (PCC=0.49), BFC achieves a high-precision reconstruction of the relative dynamic distribution (PCC=0.80), significantly outperforming traditional physical models.

This finding opens the door to dynamic mining of the PDB, allowing researchers to revisit over 190,000 static structures and extract quantitative dynamic information without the computational cost of molecular dynamics simulations [31–35]. While BFC is not a replacement for MD in studying complex transitions, it provides a computationally efficient approximation of intrinsic conformational flexibility, serving as a bridge between static structural biology and dynamic biochemistry. We anticipate that BFC will facilitate the large-scale retrospective analysis of PDB entries to investigate mechanisms of protein function and disease.

Methods

Dataset Construction and Preprocessing

To rigorously test the hypothesis that solution-state dynamics can be quantitatively recovered from crystallographic data, we constructed a large-scale, high-fidelity benchmark connecting static PDB structures with dynamic MD ensembles.

Data Acquisition and Filtering. We retrieved high-resolution X-ray structures from the RCSB Protein Data Bank (PDB) and matched them with ground-truth Root Mean Square Fluctuation (RMSF) profiles derived from structural ensembles in the PDBFlex database [17]. Specifically, ground-truth RMSF values were calculated based on the coordinate variance of Cα atoms across the aligned ensemble members to capture backbone flexibility. To ensure that the model learns from reliable physical signals rather than noise,

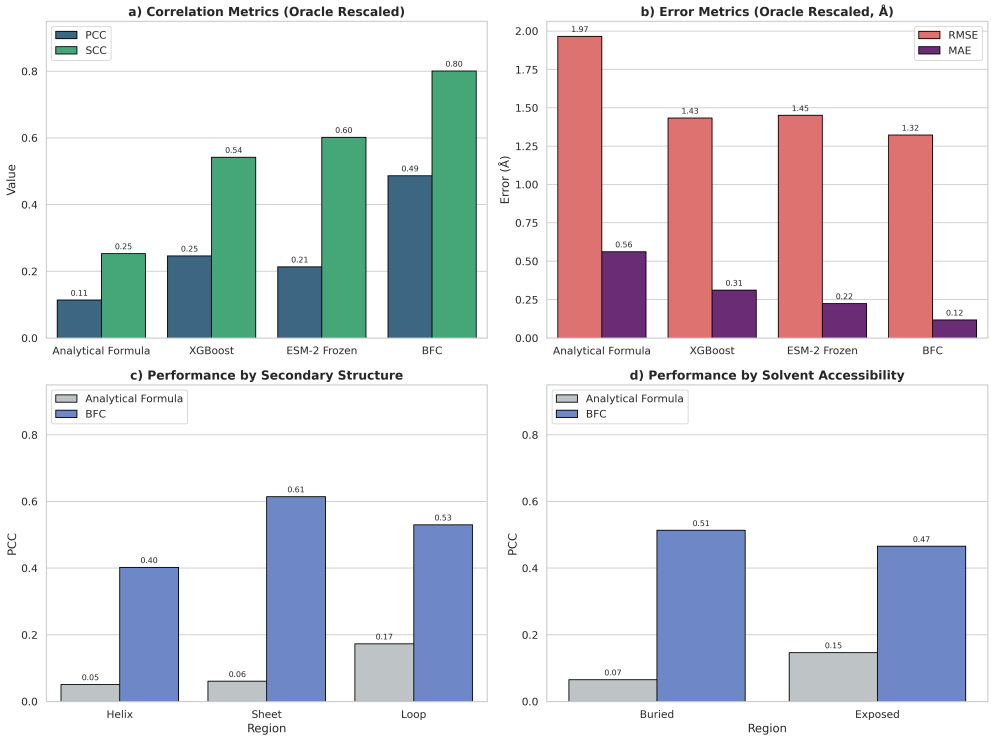

Figure 2: **BFC outperforms baselines and demonstrates robust generalization. a, b,** Systematic benchmarking of dynamic prediction accuracy. BFC (fine-tuned) significantly outperforms the analytical Gaussian Network Model (GNM), XGBoost, and pre-trained ESM-2 (frozen) in both correlation metrics (PCC/SCC) and error metrics (RMSE/MAE). The performance gap highlights the necessity of end-to-end task adaptation. **c, d,** Performance stratified by structural context. BFC maintains robust performance across secondary structures (c) and solvent accessibility levels (d), achieving high accuracy even in flexible loops and solvent-exposed regions where traditional physical formulas typically fail.

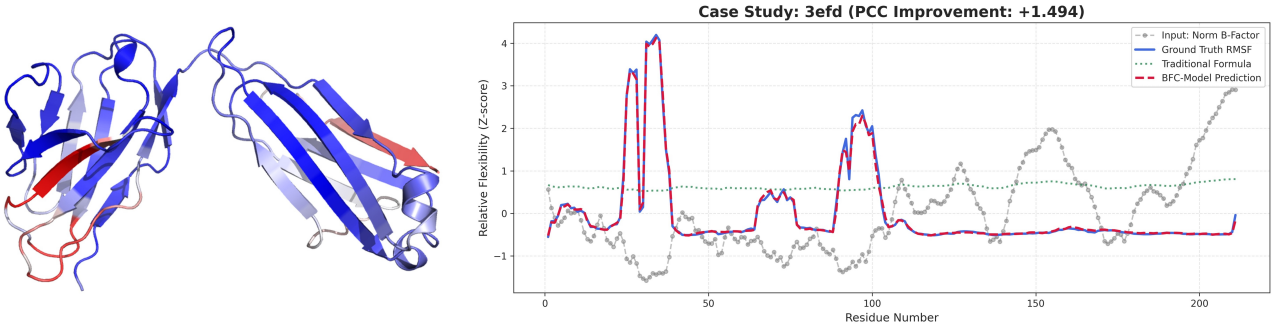

Figure 3: **Correction of crystal packing artifacts in antibody CDR loops (PDB: 3efd).** Comparison of normalized dynamic profiles reveals a major discrepancy in the CDR loop regions. Experimental B-factors (gray dotted line) falsely suggest rigidity due to crystal contacts. BFC (red dashed line) utilizes sequence context to identify these artifacts, recovering the high flexibility profile that matches the ground-truth ensemble RMSF (blue solid line). The structural map (left) visualizes the predicted flexible regions (red) versus rigid scaffolds (blue).

we applied a strict quality control pipeline: (1) *Experimental Constraints*: Beyond relying on database metadata, we explicitly parsed PDB headers (specifically `EXPDTA` records) to confirm X-ray diffraction as the method. Only structures with a resolution ≤ 3.0 Å were retained. (2) *Sequence Processing and Alignment*: A strict residue-level mapping was enforced. Non-standard Selenomethionine (MSE) residues were mapped to Methionine (MET) to handle common crystallographic phasing derivatives, while other non-standard residues were treated as gaps. Chains with significant missing electron density (gaps $> 10\%$) or mismatches between Cα coordinates and the sequence registry were discarded. (3) *Statistical Quality Control*: To filter out corrupted data entries or flat-lined profiles, we calculated the standard deviation of in-

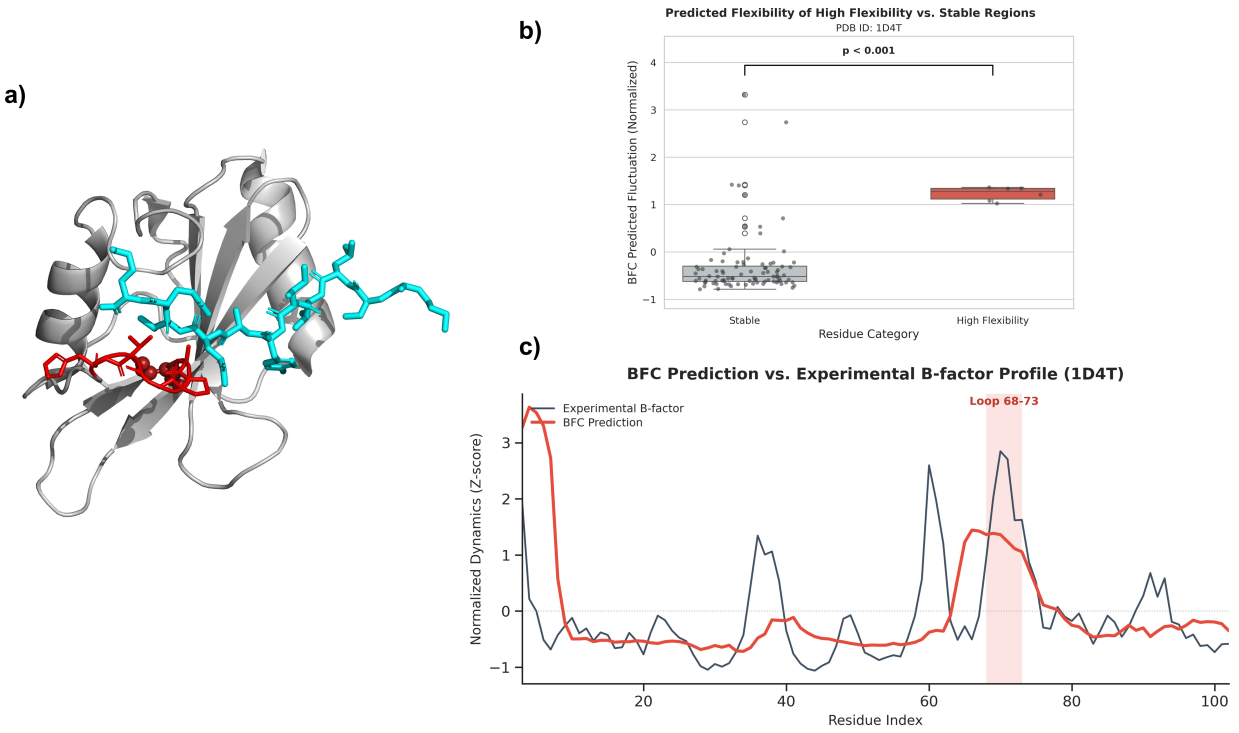

Figure 4: **Identification of disease-associated flexible motifs (PDB: 1D4T). a,** Structural visualization of the SH2 domain. BFC identifies a highly flexible loop (residues 68-73, red sticks) accommodating the SLAMF1 ligand (cyan). The Arg68 residue (red spheres), associated with XLP1 disease, is located within this flexible region. **b,** Statistical validation. The BFC-predicted flexibility for the binding loop (red box) is significantly higher than the protein scaffold background ($p < 0.001$). **c,** Dynamic profile comparison. BFC prediction (red line) accurately reproduces the flexibility peak at the binding loop (68-73) seen in the experimental profile (dark gray), confirming the model ability to pinpoint pathogenicity-relevant dynamics.

put B-factors and ground-truth RMSF for each chain. Entries exhibiting effectively zero variance ($\sigma < 10^{-6}$) were considered artifactual and removed. (4) *Context Sufficiency*: Peptide chains shorter than 50 residues were excluded to ensure sufficient sequence context for the language model. The final curated dataset comprises 280,487 protein chains derived from 132,386 unique PDB entries.

Input Representation and Normalization. Consistent with our goal of decoding intrinsic signals solely from crystallographic data, we minimized the use of hand-crafted features. The model inputs are restricted to two modalities: the raw amino acid sequence (providing physicochemical context) and the experimental B-factors (providing the latent dynamic signal). To address the variability in B-factor magnitudes caused by differing refinement protocols and crystal packing densities across the PDB, we standardized the input B-factors (B_i) and target RMSF values (R_i) using chain-specific Z-score normalization:

$$x'_i = \frac{x_i - \mu_{chain}}{\sigma_{chain}} \tag{1}$$

where μ_{chain} and σ_{chain} denote the mean and standard deviation for the chain. This normalization compels the model to focus on the relative distribution of fluctuations along the sequence rather than absolute global magnitudes, which are often artifactual.

Data Partitioning. To prevent data leakage and ensure that the model generalizes to novel protein folds, the dataset was split at the PDB-entry level. The 132,386 unique entries were randomly partitioned into training (207,465 chains), validation (52,268 chains), and testing (20,754 chains) sets. This strategy ensures that homologous chains derived from the same crystal structure do not appear simultaneously in training and evaluation partitions.

Feature Engineering and Input Representation

Minimalist Sequence-Only Input Strategy. To rigorously evaluate the hypothesis that the solution-state dynamics can be decoded directly from the interplay between evolutionary constraints and experimental noise, we adopted a minimalist input strategy. We deliberately excluded explicit structural descriptors derived from coordinate geometry (such as secondary structure assignments, solvent accessibility, or contact maps) and theoretical computations (such as Normal Mode Analysis). The model relies strictly on two input modalities:

- **Sequence Embeddings:** The raw amino acid sequence of each individual chain serves as the sole source of physicochemical and evolutionary context. Sequences were tokenized and processed by the pre-trained ESM-

2 language model (`esm2_t6_8M_UR50D`) to generate token-level representations $\mathbf{h}_i \in \mathbb{R}^{320}$.

- **Normalized B-Factors:** The crystallographic B-factor is provided as the only experimental observation. To normalize the vast differences in refinement scales and crystal qualities across the PDB, raw B-factors were standardized using chain-level Z-score normalization prior to input:

$$B'_i = \frac{B_i - \mu_{chain}}{\sigma_{chain}} \tag{2}$$

This design forces the model to treat the B-factor as a noisy proxy for dynamics and use the sequence context to "denoise" or "translate" it into solution-state RMSF, without relying on engineered physical features.

B-Factor Corrector (BFC) Architecture

BFC treats dynamics prediction as a token-level regression task. The architecture consists of the ESM-2 encoder followed by a specialized decoding head.

Encoder. We employed the ESM-2 transformer as a feature extractor. Unlike approaches that freeze the language model, we performed full parameter fine-tuning on the ESM-2 backbone. This allows the attention mechanisms to adapt specifically to the task of recognizing crystallographic artifacts and correlating them with sequence motifs.

Prediction Head. The embeddings $\mathbf{h}_i$ are concatenated with the normalized B-factor scalar B'_i. This composite vector is passed through a regression head consisting of a linear projection (dim=256), Layer Normalization, and a ReLU activation function. A Dropout layer ($p = 0.1$) is applied for regularization before the final linear projection to a scalar value representing the predicted RMSF.

Baseline Models

We benchmarked BFC against three distinct classes of models to isolate the source of performance gains:

Analytical Formula: A traditional physical conversion where RMSF is derived directly from B-factors assuming an isotropic harmonic oscillator:

$$\text{RMSF} = \sqrt{\frac{3B}{8\pi^2}} \tag{3}$$

XGBoost (B-Factor Only): To determine if the relationship between B-factors and RMSF is simply a non-linear mapping independent of sequence context, we trained a gradient-boosted decision tree regressor (XGBoost [36]) using *only* the normalized B-factors as input features. Hyperparameters (max_depth=5, n_estimators=1000) were optimized on the validation set.

Frozen ESM-2: To assess the necessity of fine-tuning, we trained the same prediction head described above on top of fixed embeddings from the pre-trained ESM-2 model, without updating the transformer weights.

Training and Optimization

Loss Function. We minimized the Mean Absolute Error (MAE) between the predicted and ground-truth RMSF values. Crucially, to handle variable sequence lengths within batches, we implemented a **masking mechanism** where padding tokens (assigned a placeholder label of -1.0) are strictly excluded from the computation. The loss is calculated as:

$$\mathcal{L} = \frac{1}{N_{valid}} \sum_{i=1}^{N_{total}} m_i \cdot |y_i - \hat{y}_i| \tag{4}$$

where $m_i \in \{0, 1\}$ is the binary mask indicating valid residues, and N_{valid} is the total number of non-padding tokens in the batch.

Optimization. Training was performed using the AdamW optimizer [37] with a learning rate of 5×10^{-5} and a batch size of 64. We employed a linear learning rate scheduler with warmup, where the learning rate increases linearly for the first 10% of training steps and decays linearly thereafter.

Implementation. All models were implemented in PyTorch 2.6.0 [38] and trained on a single NVIDIA GeForce RTX 5090 (CUDA 13.0). To balance computational efficiency with data integrity, input sequences were truncated to a maximum length of 1,024 residues. Statistical analysis confirms that this threshold covers 99.37% of the dataset (Supplementary Figure S5), ensuring negligible information loss. Training proceeded for a fixed duration of 100 epochs, and the checkpoint achieving the highest Pearson Correlation Coefficient (PCC) on the validation set was selected for testing.

Evaluation Metrics

To comprehensively assess model performance across both relative fluctuation patterns and absolute physical magnitudes, we employed four metrics. All metrics were calculated individually for each protein chain k with length L_k, and then averaged over the total number of chains in the test set (N). Let $\mathbf{y} = (y_1, \ldots, y_{L_k})$ denote the predicted RMSF profile and $\hat{\mathbf{y}} = (\hat{y}_1, \ldots, \hat{y}_{L_k})$ denote the ground-truth ensemble RMSF.

Pearson Correlation Coefficient (PCC): Measures the linear correlation between the predicted and experimental profiles, assessing the model's ability to capture the shape of the fluctuation landscape.

$$\text{PCC} = \frac{1}{N} \sum_{k=1}^{N} \frac{\sum_{i=1}^{L_k} (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{L_k} (y_i - \bar{y})^2} \sqrt{\sum_{i=1}^{L_k} (\hat{y}_i - \bar{\hat{y}})^2}} \tag{5}$$

Spearman Correlation Coefficient (SCC): Evaluates the monotonic rank-order correlation. This metric is robust to outliers and assesses whether the model correctly ranks residues from most rigid to most flexible.

$$\text{SCC} = \frac{1}{N} \sum_{k=1}^{N} \left(1 - \frac{6 \sum_{i=1}^{L_k} d_i^2}{L_k (L_k^2 - 1)} \right) \tag{6}$$

where d_i is the difference between the ranks of y_i and $\hat{y}_i$.

Root Mean Square Error (RMSE): Quantifies the average magnitude of error in physical units (Å). This is the primary metric for assessing quantitative accuracy, penalizing large deviations more heavily than small ones.

$$\text{RMSE} = \frac{1}{N}\sum_{k=1}^{N}\sqrt{\frac{1}{L_k}\sum_{i=1}^{L_k}(y_i - \hat{y}_i)^2} \tag{7}$$

Mean Absolute Error (MAE): Provides a linear score of error magnitude, offering a direct interpretation of the average deviation per residue.

$$\text{MAE} = \frac{1}{N}\sum_{k=1}^{N}\left(\frac{1}{L_k}\sum_{i=1}^{L_k}|y_i - \hat{y}_i|\right) \tag{8}$$

Physical Space Reconstruction Assessment (Oracle Rescaling). Since solution-state fluctuation amplitudes vary independently of sequence due to experimental conditions (e.g., temperature, crystal quality), direct prediction of absolute RMSF magnitudes from static coordinates is ill-posed. To decouple intrinsic profile accuracy from global amplitude scaling, we applied an oracle rescaling strategy for the physical space evaluation presented in Figure 1c. For each protein chain, both the BFC output (Z-score) and the baseline physical estimates derived from B-factors were linearly transformed to match the mean (μ_{GT}) and standard deviation (σ_{GT}) of the ground-truth ensemble RMSF. This transformation ensures that the evaluation metrics (RMSE, PCC, SCC) strictly measure the fidelity of the fluctuation landscape shape (profile reconstruction) rather than systematic scaling errors.

External Validation and Functional Annotation

To benchmark the biological fidelity of BFC predictions, we utilized the ATLAS database of standardized molecular dynamics simulations. Ground-truth RMSF profiles were established by computing the ensemble average of the three independent trajectory replicates provided for each entry, ensuring robustness against stochastic sampling errors.

For the functional case studies presented (e.g., PDB 1D4T and 1CUO), critical functional regions such as ligand-binding loops and active site adjacencies were identified and mapped according to the UniProt Knowledgebase [39] annotations. To quantitatively verify the correspondence between predicted dynamics and biological function, we implemented a statistical scoring workflow. The BFC-predicted normalized fluctuations of these UniProt-annotated regions were compared against the remaining structural scaffold. A Welch's t-test for unequal variances was employed to assess significance, with a threshold of $p < 0.001$ confirming that the model selectively assigns significantly higher flexibility to function-critical motifs compared to the stable protein core.

Declaration of Interests

The authors declare no competing interests.

Funding

This work was supported by the Natural Science Foundation of Xinjiang Uygur Autonomous Region (Grant Number: 2024D01C216) and the "Tianchi Talents" introduction plan.

Code availability

Source code for BFC and the pre-trained model weights are available at https://github.com/wyqmath/BFactorCorrector.

References

[1] H. Frauenfelder, S. G. Sligar, and P. G. Wolynes. The energy landscapes and motions of proteins. *Science*, 254(5038):1598–1603, 1991.

[2] P. Csermely, R. Palotai, and R. Nussinov. Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. *Trends in biochemical sciences*, 35(10):539–546, 2010.

[3] R. Nussinov, Y. Liu, W. Zhang, and H. Jang. Protein conformational ensembles in function: roles and mechanisms. *RSC chemical biology*, 4(11):850–864, 2023.

[4] R. B. Fenwick, H. van den Bedem, J. S. Fraser, and P. E. Wright. Integrated description of protein dynamics from room-temperature X-ray crystallography and NMR. *Proceedings of the National Academy of Sciences*, 111(4):E445–E454, 2014.

[5] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, ..., and P. E. Bourne. The protein data bank. *Nucleic acids research*, 28(1):235–242, 2000.

[6] M. A. DePristo, P. I. De Bakker, and T. L. Blundell. Heterogeneity and inaccuracy in protein structures solved by X-ray crystallography. *Structure*, 12(5):831–838, 2004.

[7] D. W. Li and R. Brüschweiler. All-atom contact model for understanding protein dynamics from crystallographic b-factors. *Biophysical journal*, 96(8):3074–3081, 2009.

[8] O. Carugo. B-factor accuracy in protein crystal structures. *Acta Crystallographica Section D: Structural Biology*, 78(1):69–74, 2022.

[9] E. Klyshko, J. S. H. Kim, L. McGough, V. Valeeva, E. Lee, R. Ranganathan, and S. Rauscher. Functional protein dynamics in a crystal. *Nature Communications*, 15(1):3244, 2024.

[10] E. Eyal, S. Gerzon, V. Potapov, M. Edelman, and V. Sobolev. The limit of accuracy of protein modeling: influence of crystal packing on protein structure. *Journal of molecular biology*, 351(2):431–442, 2005.

[11] Martin Karplus and J Andrew McCammon. Molecular dynamics simulations of biomolecules. *Nature structural biology*, 9(9):646–652, 2002.

[12] A. Lasorsa and P. C. van Der Wel. Solid-state nmr protocols for unveiling dynamics and (drug) interactions of membrane-bound proteins. *Protein Science*, 34(4):e70102, 2025.

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

[14] A. R. Atilgan, S. R. Durell, R. L. Jernigan, M. C. Demirel, O. Keskin, and I. Bahar. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophysical journal*, 80(1):505–515, 2001.

[15] Y. Wang, M. Cai, Y. Dong, Y. Ma, and K. Wei. From signal to symphony: Exploring 2d sequence representations for protein function prediction. *Journal of Chemical Information and Modeling*, 2025.

[16] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, ..., and A. Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

[17] T. Hrabe, Z. Li, M. Sedova, P. Rotkiewicz, L. Jaroszewski, and A. Godzik. Pdbflex: exploring flexibility in protein structures. *Nucleic acids research*, 44(D1):D423–D428, 2016.

[18] S. Uysal, V. Vásquez, V. Tereshko, K. Esaki, F. A. Fellouse, S. S. Sidhu, and A. Kossiakoff. Crystal structure of full-length KcsA in its closed conformation. *Proceedings of the National Academy of Sciences*, 106(16):6644–6649, 2009.

[19] F. C. Spoendlin, M. L. Fernández-Quintero, S. S. Raghavan, H. L. Turner, A. Gharpure, J. R. Loeffler, and C. M. Deane. Predicting the conformational flexibility of antibody and T cell receptor complementarity-determining regions. *Nature Machine Intelligence*, pages 1–13, 2025.

[20] M. L. Fernández-Quintero, G. Georges, J. M. Varga, and K. R. Liedl. Ensembles in solution as a new paradigm for antibody structure prediction and design. *MAbs*, 13(1):1923122, jan 2021.

[21] R. J. Blackler, S. Müller-Loennies, B. Pokorny-Lehrer, M. S. Legg, L. Brade, H. Brade, ..., and S. V. Evans. Antigen binding by conformational selection in near-germline antibodies. *Journal of Biological Chemistry*, 298(5):101901, 2022.

[22] M. L. Fernández-Quintero, J. Kraml, G. Georges, and K. R. Liedl. CDR-H3 loop ensemble in solution–conformational selection upon antibody binding. *MAbs*, 11(6):1077–1088, August 2019.

[23] C. N. Liu, L. M. Denzler, O. E. Hood, and A. C. Martin. Do antibody CDR loops change conformation upon binding? *MAbs*, 16(1):2322533, December 2024.

[24] Y. Vander Meersche, G. Cretin, A. Gheeraert, J. C. Gelly, and T. Galochkina. ATLAS: protein flexibility description from atomistic molecular dynamics simulations. *Nucleic acids research*, 52(D1):D384–D392, 2024.

[25] F. Poy, M. B. Yaffe, J. Sayos, K. Saxena, M. Morra, J. Sumegi, and M. J. Eck. Crystal structures of the XLP protein SAP reveal a class of SH2 domains with extended, phosphotyrosine-independent sequence recognition. *Molecular cell*, 4(4):555–561, 1999.

[26] C. Booth, K. C. Gilmour, P. Veys, A. R. Gennery, M. A. Slatter, H. Chapel, and H. B. Gaspar. X-linked lymphoproliferative disease due to SAP/SH2D1A deficiency: a multicenter study on the manifestations, management and outcome of the disease. *Blood, The Journal of the American Society of Hematology*, 117(1):53–62, 2011.

[27] J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, ..., and D. Baker. Robust deep learning–based protein sequence design using ProteinMPNN. *Science*, 378(6615):49–56, 2022.

[28] J. L. Watson, D. Juergens, N. R. Bennett, B. L. Trippe, J. Yim, H. E. Eisenach, ..., and D. Baker. De novo design of protein structure and function with RFdiffusion. *Nature*, 620(7976):1089–1100, 2023.

[29] J. Dauparas, G. R. Lee, R. Pecoraro, L. An, I. Anishchenko, C. Glasscock, and D. Baker. Atomic context-conditioned protein sequence design using LigandMPNN. *Nature Methods*, pages 1–7, 2025.

[30] T. Hayes, R. Rao, H. Akin, N. J. Sofroniew, D. Oktay, Z. Lin, ..., and A. Rives. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, 2025.

[31] J. Jung, W. Nishima, M. Daniels, G. Bascom, C. Kobayashi, A. Adedoyin, and K. Y. Sanbonmatsu. Scaling molecular dynamics beyond 100,000 processor cores for large-scale biophysical simulations. *Journal of computational chemistry*, 40(21):1919–1930, 2019.

[32] M. Sahil, S. Sarkar, and J. Mondal. Long-time-step molecular dynamics can retard simulation of protein-ligand recognition process. *Biophysical Journal*, 122(5):802–816, 2023.

[33] E. Prašnikar, M. Ljubič, A. Perdih, and J. Borišek. Machine learning heralding a new development phase in molecular dynamics simulations. *Artificial intelligence review*, 57(4):102, 2024.

[34] A. Lappala. The next revolution in computational simulations: Harnessing AI and quantum computing in molecular dynamics. *Current Opinion in Structural Biology*, 89:102919, 2024.

[35] T. Cui, Y. Zhou, and T. Wang. Recent advances in artificial intelligence–driven biomolecular dynamics simulations based on machine learning force fields. *Current Opinion in Structural Biology*, 95:103191, 2025.

[36] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. Association for Computing Machinery.

[37] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[38] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... & Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, volume 32, 2019.

[39] T. UniProt Consortium. Uniprot: the universal protein knowledgebase. *Nucleic acids research*, 46(5):2699–2699, 2018.

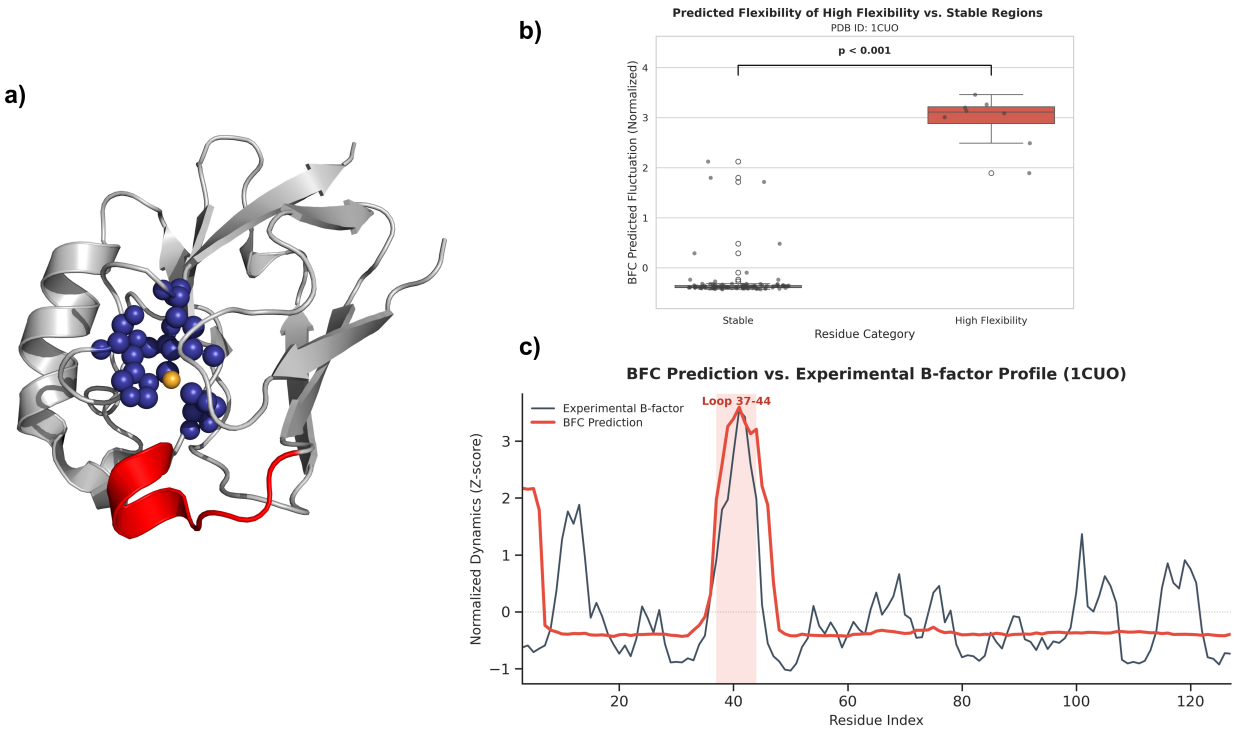

Figure S1: **BFC decodes the functional duality of rigidity and flexibility in Azurin (PDB: 1CUO). a,** Structural visualization of the Azurin active site. The landscape reveals a sophisticated dynamic duality: BFC correctly assigns structural rigidity to the core residues (blue spheres) that form a precise scaffold to coordinate the copper ion (orange). Simultaneously, the model identifies a distinct, highly flexible loop (residues 37-44, highlighted in red) adjacent to the core. This loop is known to mediate the dynamic protein-protein interactions required for electron transfer. **b,** Statistical validation. A quantitative comparison shows that the BFC-predicted flexibility for the functional loop (red box) is significantly higher than that of the protein's stable regions ($p < 0.001$, Welch's t-test). **c,** Dynamic profile comparison. The BFC prediction (red line) demonstrates strong agreement with the experimental B-factor profile (dark gray), accurately reproducing the localized flexibility peak at the 37-44 loop region while maintaining low values for the rigid core.

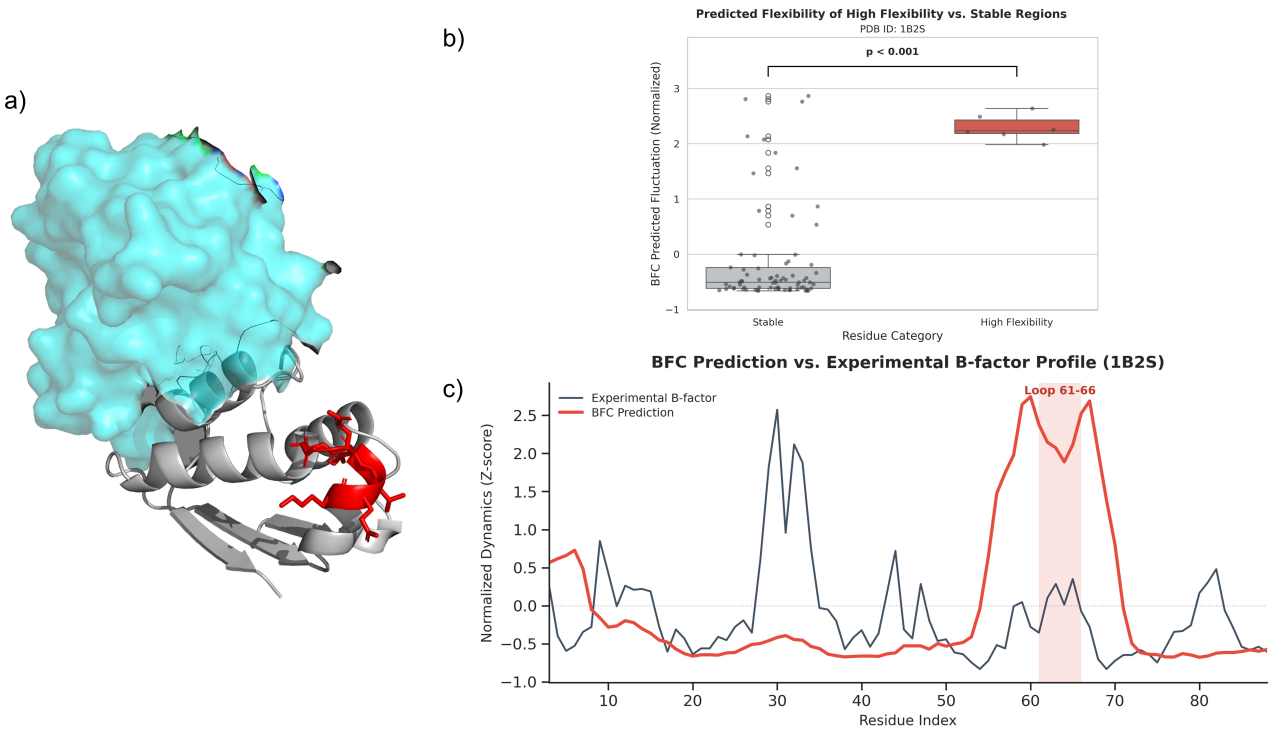

Figure S2: **BFC corrects crystal artifacts to reveal intrinsic dynamics in Barstar (PDB: 1B2S). a,** Structural context of the Barstar (Chain E, gray ribbon) and Barnase (cyan surface) complex. The BFC-predicted high-flexibility region (residues 61-66) is highlighted in red. The visualization confirms that this loop is solvent-exposed and located away from the rigid binding interface, suggesting a high propensity for intrinsic motion. **b,** Statistical validation of the prediction. The BFC-predicted fluctuation scores for the identified loop (red box) are significantly higher than those of the rest of the protein ($p < 0.001$, Welch's t-test), indicating a distinct dynamic signature. **c,** Comparison of normalized dynamic profiles. A major discrepancy is observed in the 61-66 loop region (shaded red): experimental B-factors (dark gray) suggest rigidity likely due to crystal packing artifacts or model fitting constraints. In contrast, BFC (red line) correctly recovers the high intrinsic flexibility physically expected for a solvent-exposed loop. This demonstrates the model's ability to transcend the limitations of raw crystallographic data and decode true biophysical dynamics solely from sequence context.

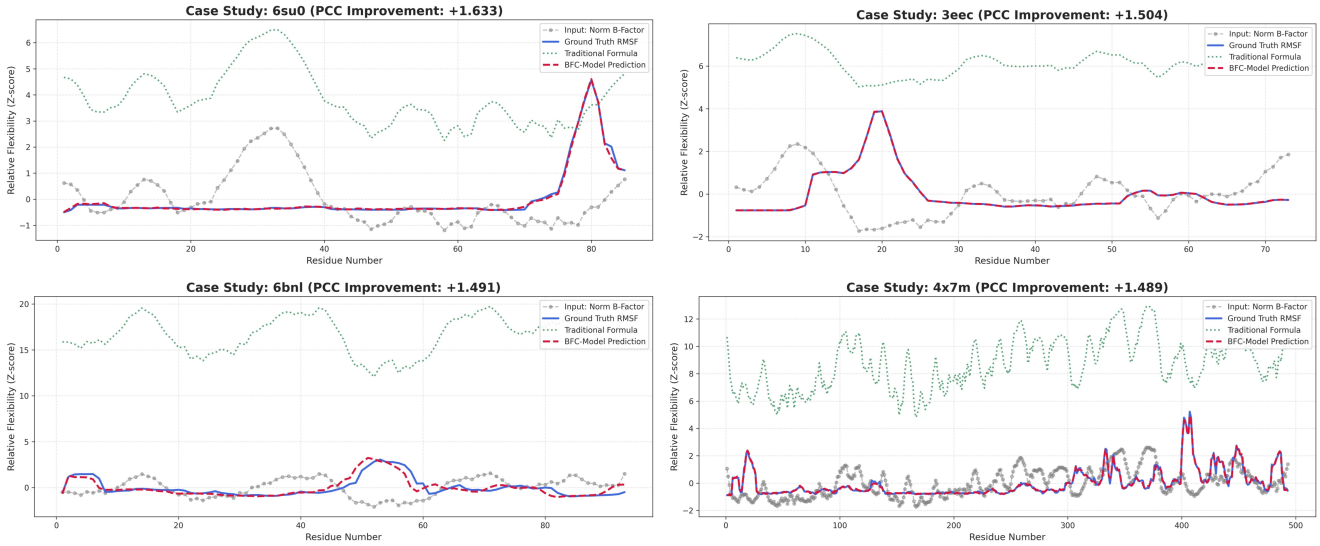

Figure S3: **Extended Case Studies of dynamic profile reconstruction.** This figure illustrates detailed flexibility predictions for four proteins with varying sequence lengths, including 6su0, 3eec, 6bnl, and 4x7m. The plots display the BFC-Model Prediction (red dashed line) in comparison to the Ground Truth RMSF (blue solid line), the Input Norm B-Factor (gray connected dots), and the Traditional Formula (green dotted line). The BFC model exhibits consistent alignment with the ground truth trajectory across all cases, effectively identifying distinct flexibility peaks in both short peptides and larger protein structures such as 4x7m. The PCC improvement scores noted in the panel headers quantify the increase in Pearson correlation coefficient achieved by the model relative to the normalized B-factor input. These metrics validate the capacity of BFC to mitigate experimental artifacts and accurately recover intrinsic biophysical dynamics.

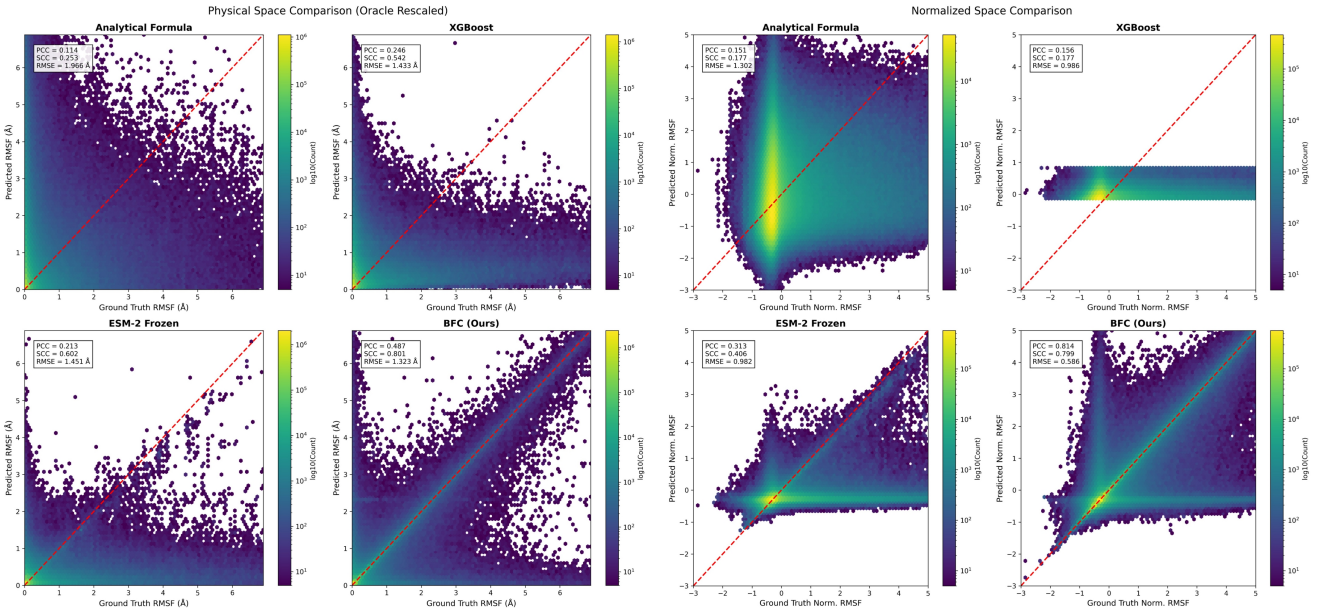

Figure S4: **Comprehensive scatter plot analysis of model performance in physical and normalized spaces.** This figure presents a density-based comparison of predicted flexibility versus ground truth RMSF across four different methods. The panels are organized into two groups: the Physical Space Comparison on the left measures fluctuations in Angstroms, while the Normalized Space Comparison on the right evaluates performance using Z-scores. The color gradient represents the logarithmic density of data points, where yellow indicates high-concentration regions and purple indicates low density. The red dashed line marks the ideal identity relationship where the prediction equals the ground truth. Baseline methods such as the Analytical Formula and XGBoost exhibit significant dispersion and poor correlation. Notably, in the Physical Space Comparison (left columns), all predictions were subjected to Oracle Rescaling (matched to ground-truth mean/std) to ensure a fair comparison of profile shapes. Despite this, the analytical formula shows broad scatter (SCC=0.25), confirming that B-factors are locally distorted. While the Frozen ESM-2 model shows improved correlation, it still retains considerable variance around the diagonal. In contrast, the BFC model demonstrates superior convergence along the identity line in both evaluation metrics. The statistical indicators provided in each panel confirm that BFC achieves the highest Pearson Correlation Coefficient (PCC) and Spearman Correlation Coefficient (SCC) combined with the lowest Root Mean Square Error (RMSE), validating its robustness across different feature spaces.

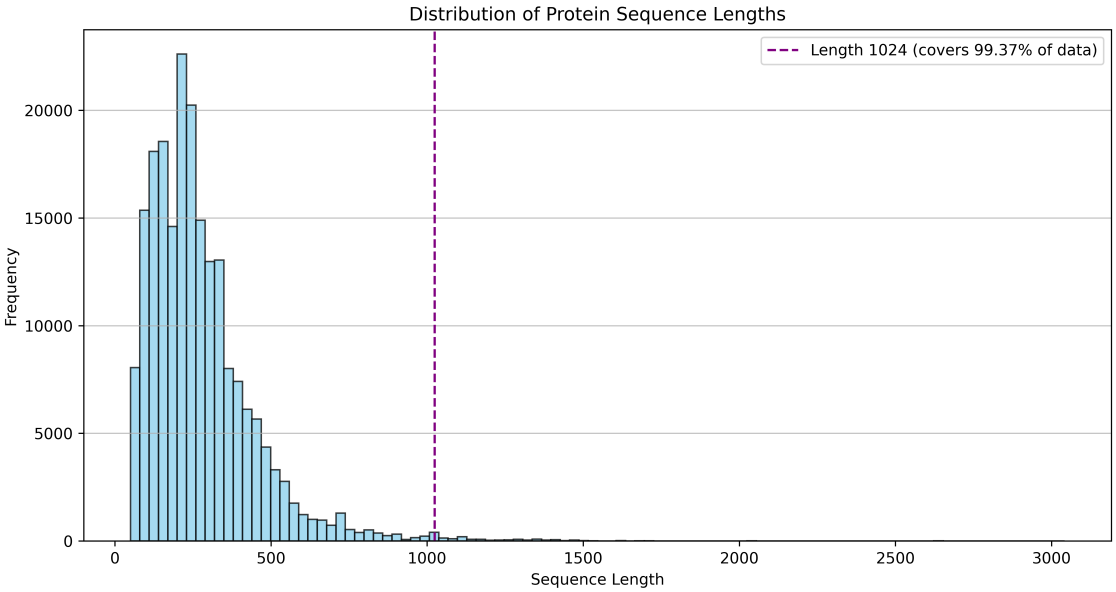

Figure S5: **Distribution of protein sequence lengths.** The vertical dashed line marks the truncation threshold of 1,024 residues set for training and testing. This cutoff covers 99.37% of the total dataset (206,168 out of 207,465 sequences have a length $\leq$ 1,024), effectively reducing computational overhead while preserving the vast majority of biological data.