
Seek and You Shall Fold

Nadav Bojan Sellam^{*†}, Meital Bojan^{*†}, Paul Schanda[†], Alex Bronstein^{†‡}

[†]IST Austria [‡]Technion, Israel

^{*} Equal contribution

Abstract

Accurate protein structures are essential for understanding biological function, yet incorporating experimental data into protein generative models remains a major challenge. Most predictors of experimental observables are non-differentiable, making them incompatible with gradient-based conditional sampling. This is especially limiting in nuclear magnetic resonance, where rich data such as chemical shifts are hard to directly integrate into generative modeling. We introduce a framework for *non-differentiable guidance* of protein generative models, coupling a continuous diffusion-based generator with *any* black-box objective via a tailored genetic algorithm. We demonstrate its effectiveness across three modalities: pairwise distance constraints, nuclear Overhauser effect restraints, and for the first time chemical shifts. These results establish chemical shift guided structure generation as feasible, expose key weaknesses in current predictors, and showcase a general strategy for incorporating diverse experimental signals. Our work points toward automated, data-conditioned protein modeling beyond the limits of differentiability.

1 Introduction

Proteins are the "backbone" of all cellular processes, serving as the structural framework, molecular machines, and regulatory signals that sustain life. Knowing the conformations that proteins adapt – an ensemble of three-dimensional structures – is crucial to understanding how they function and how they interact with other proteins, as well as for designing drugs that modulate their actions. Experimental workflows typically begin with measurements and then seek a structure that best explains the data, but this mapping remains a major bottleneck [3, 25]. Machine learning has transformed protein modeling: AlphaFold [15] delivers highly accurate structure predictions, while generative models such as RFdiffusion [35] enable controllable design, and Boltz [30] learns thermodynamically plausible distributions to sample alternative states. BioEmu [19] further captures near-equilibrium variability from molecular dynamics (MD) simulations. Yet none of these models are designed to directly *condition* structure generation on experimental measurements, forcing practitioners to resort to expensive MD refinement to generate states consistent with data.

A common strategy for protein-data conditioned modeling is to guide generative diffusion models using differentiable forward models [23, 24, 7]. Although mature, well-validated predictors exist for many experimental observables, most are non-differentiable and therefore cannot be used directly for gradient-based guidance. Consequently, current approaches must either approximate these predictors with differentiable surrogates or restrict themselves to modalities with analytic gradients. While effective when such models exist, this approach has several drawbacks: (i) it requires differentiable models for every observable, (ii) gradients are often noisy or ill-conditioned, leading to unstable convergence, (iii) guidance strength must be tuned under a fixed diffusion schedule, and (iv) the process is prone to overfitting to noise and producing off-manifold or broken structures that must be discarded or relaxed. These limitations call for alternative guidance methods that remain robust in the presence of non-differentiable predictors and experimental noise.

Our contributions. We introduce a general framework for data-conditioned protein generative modeling that overcomes the limitations of gradient-based approaches by performing *non-differentiable guidance*. Specifically:

1. **General non-differentiable guidance:** We introduce a method to couple a continuous protein structure generator with *any* black-box scoring function, guiding the generator’s outputs using a tailored genetic algorithm. This allows seamless integration of mature but non-differentiable experimental predictors.
2. **Application to nuclear magnetic resonance (NMR):** We demonstrate the framework on several experimental and synthetic constraints. For pairwise distances and nuclear Overhauser effects (NOEs), our approach yields strong improvements. For chemical shifts, we provide the first proof-of-concept demonstration of direct generative guidance, a long-standing goal in NMR structural biology.

Focus on NMR. In this work, we focus on NMR spectroscopy, which reports on local structure and dynamics and remains a key tool for protein characterization [16]. Chemical shifts are the most widely measured and information-rich NMR observables, yet predictors such as UCBShift [20, 31] and SHIFTX2 [9] are non-differentiable, making them incompatible with gradient-based guidance.

Integrating chemical shifts into structure generation has long been sought in NMR structural biology but has been infeasible due to this differentiability barrier. As chemical shift prediction improves, our framework could dramatically accelerate or even automate peak assignment—one of the most time-consuming bottlenecks in NMR, often requiring days or weeks on high-performance computing resources. Additional background is provided in Appendix B.

2 Methodology

We optimize a generator’s latent variables to maximize agreement with experimental data under a black-box score function (Fig. 1). Let $x \in \mathcal{X}$ be a protein structure decoded from a latent $z \in \mathcal{Z}$, and let $S : \mathcal{X} \rightarrow \mathbb{R}$ denote a score derived from experimental observables (e.g., NOEs, chemical shifts). Our goal is to find $x^* = \arg \max_{x \in \mathcal{X}} S(x)$, where $x = G(z_T, 0)$ is obtained by sampling $z_T \sim \mathcal{N}(0, I)$ at diffusion time T and denoising to $t = 0$ with the generator G . Since S is typically non-differentiable, we employ a *protein-specific* genetic algorithm (GA) to search candidate structures, presented in Alg. 1.

Generator. We use BioEmu [19] as a structural prior. BioEmu is a lightweight diffusion model, trained to produce thermodynamically consistent protein structures. Each structure is represented by a residue-wise latent $z \in \mathbb{R}^{N \times d}$, with N residues and per-residue dimension d , enabling localized perturbations. We further adapt its stochastic SDE sampler to a deterministic probability flow ODE, making $G(z_T, t)$ reproducible and enabling controlled, repeatable perturbations.

Genetic algorithm for structure optimization. Our GA [13, 8, 6] maintains a population of protein structures and is designed to keep changes localized while departing from vanilla GAs by: (1) *disabling crossover*, which would induce destabilizing global jumps; (2) applying *residue-wise perturbations* in the generator’s latent space; and (3) introducing a *perturbation time scheduler* that anneals from global to local moves.

Residue-wise perturbation: Our perturbation strategy is designed to keep changes localized, preserving stability while still enabling controlled exploration. Given a diffusion time t , each parent $x^{(i)}$ is mapped to its intermediate latent $z_t^{(i)} = G^{-1}(x^{(i)}, t)$ by running the generator backward, so that forwarding $z_t^{(i)}$ to $t=0$ exactly reconstructs $x^{(i)}$. At the same t , a stochastic resample $\hat{z}_t^{(i)}$ is drawn from the reverse SDE, which forwards to a different but distributionally consistent structure. Next, a residue index $a \sim \text{Uniform}(\{1, \dots, N\})$ is selected, and the corresponding row of $z_t^{(i)}$ is

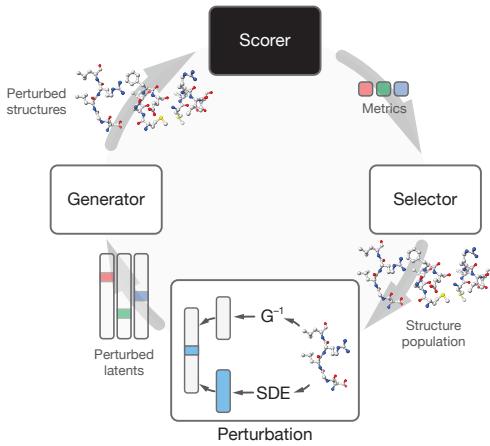


Figure 1: Overview of method. Latents are perturbed and decoded with BioEmu, scored by experimental data, and evolved with a genetic algorithm.

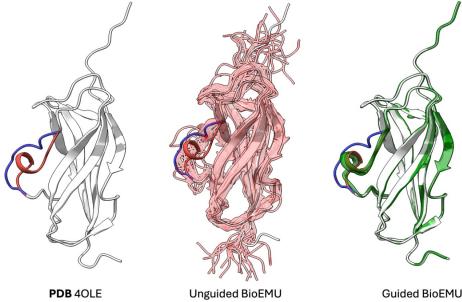


Figure 2: **Pairwise distance guidance of 4OLE.** Alternative conformations of residues 60 – 68 were guided from conformation B (blue) toward conformation A (red) using C_α pairwise distance restraints. Without guidance, BioEmu produces structures resembling conformation B, whereas guidance enables recovery of the helical conformation A.

replaced with $\hat{z}_t^{(i)}[a]$, yielding a perturbed latent $z_t^{(i,\text{pert})}$. Forwarding $z_t^{(i,\text{pert})}$ to $t=0$ then produces the perturbed child structure $x^{(i,\text{child})} = G(z_t^{(i,\text{pert})}, 0)$.

Perturbation time scheduler: The perturbation time t_k decreases linearly across generations, $t_k = t_{\max} - \frac{k}{K}(t_{\max} - t_{\min})$, enabling broad exploration at early steps and fine refinement later. This annealing schedule prevents the plateauing observed with a fixed t , and is directly analogous to annealing molecular dynamics, where the temperature is gradually lowered to transition from broad exploration to local refinement. We provide a more detailed discussion of the scheduler and its impact in Appendix E.

This population-based search maintains diversity, avoids greedy local updates that can lead to poor minima, and is not tied to a fixed diffusion schedule, allowing iterative refinement until convergence. Additional implementation details are provided in Appendix A and F.

3 Results

Experimental setup. Our experiments test whether genetic guidance can recover accurate folds in proteins where state-of-the-art generators, including BioEmu and AlphaFold3 (AF3), yield erroneous or partially misfolded structures. We focus on challenging cases with available NMR data, where both methods struggle to reproduce the correct fold. In all experiments, the search was initialized from an AF3 prediction, unless stated otherwise. This setting provides a stringent test: success requires not only generating physically plausible structures, but also satisfying experimental constraints that reveal limitations of current predictors. We begin with a toy example using pairwise distance restraints, then proceed to experiments with NOE-derived restraints, and finally evaluate chemical shift scoring. Complete implementation details are provided in Appendix F, with additional results in Appendix D.

Pairwise distance guidance As a toy experiment, in preparation for NOE-based restraints, we derived pairwise distance restraints from conformation A of 4OLE, using $C_\alpha-C_\alpha$ distances of at most 5 Å to mimic NOE-derived contacts, with an L_1 loss on these distances. This protein exhibits significant alternative conformations in residues 60–68: conformation A contains a helical segment, while conformation B forms a strand. Starting from the conformation B PDB, we guided the system

Algorithm 1: Genetic guidance for protein structures

Input: scorer S ; generator G ; total generations K ;
 t_{\min}, t_{\max}
 $\{z_T^{(i)}\}_{i=1}^P \sim \mathcal{N}(0, I)$
for $k = 0 \dots K$ **do**
 // (1) Scoring
 $s^{(i)} \leftarrow S(G(z_T^{(i)}, 0))$
 // (2) Selection
 Retain elites; sample parents via tournament
 // (3) Controlled perturbation
 $t_k \leftarrow t_{\max} - \frac{k}{K}(t_{\max} - t_{\min})$
 for each selected parent $x^{(i)}$ **do**
 $z_{t_k}^{(i)} \leftarrow G^{-1}(x^{(i)}, t_k)$
 $\hat{z}_{t_k}^{(i)} \leftarrow \text{SDE}^{-1}(x^{(i)}, t_k)$
 $a \sim \text{Uniform}(\{1, \dots, N\})$
 $z_{t_k}^{(i)}[a] \leftarrow \hat{z}_{t_k}^{(i)}[a]$
 $x^{(i,\text{child})} \leftarrow G(z_{t_k}^{(i)}, 0)$
 // (4) Population update
 Next generation \leftarrow elites + children

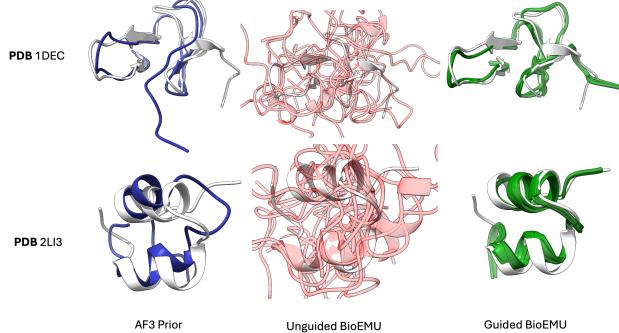


Figure 3: NOE-guided structures. Guided BioEmu with NOE-derived restraints on peptides 1DEC and 2LI3. Reference PDBs are in gray. AF3 priors contained misfolded regions inconsistent with NOEs, unguided BioEmu did not match the PDB, while guided BioEmu reduced violations and recovered the experimental folds.

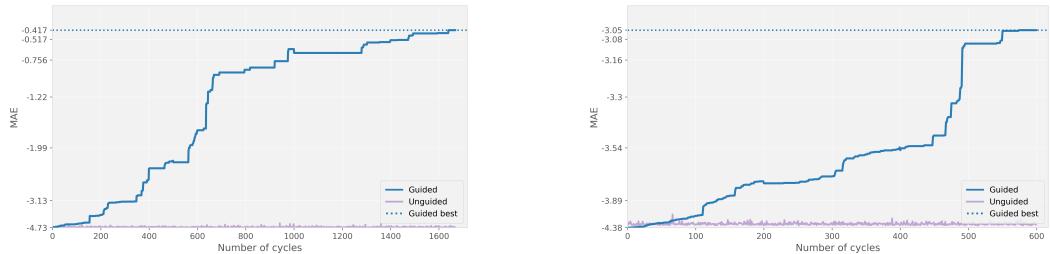


Figure 4: Chemical shift guidance metric. BioEmu guided using UCBShift-predicted chemical shifts. Plots show mean absolute error (MAE) of the best structure as a function of optimization cycles. **Left:** peptide 1DEC with synthetic PDB-derived shifts. **Right:** protein 1DFU with experimental shifts (BMRB 4395). In both cases, guidance improved the chemical-shift metric, though convergence to the experimental ensemble was not achieved.

toward conformation A. Guidance recovered the helical state, which is otherwise inaccessible to BioEmu alone (Fig. 2). This shows our method can capture structural differences spanning several residues. The corresponding metric graph is provided in Appendix F.1.1.

NOE restraints guidance We next evaluated guidance using nuclear Overhauser effect (NOE)-derived restraints. NOE cross-peaks arise from through-space dipolar couplings between protons within approximately 5–6 Å and are conventionally converted into distance restraints, making them a key source of spatial proximity information in NMR. As such, they provide a useful test case for assessing the effectiveness of our guidance method in capturing local geometry.

To assess performance, we used a non-differentiable violation score defined as the fraction of violated restraints multiplied by the average magnitude of their violations (Appendix F). This metric penalizes both widespread inconsistencies and severe outliers, making it well-suited for guiding structural refinement. Figure 3 illustrates results for two peptides (1DEC and 2LI3), with additional cases and quantitative results shown in Appendix D.3. In both cases, AlphaFold predictions served as priors but contained misfolded regions inconsistent with the NOE data. Unguided BioEmu produced incorrect structures, whereas guided BioEmu reduced violations and recovered the experimental folds, showing that NOEs are a strong signal for refinement.

Chemical shift guidance The most challenging guidance experiments used chemical shifts, with predicted values for generated structures computed using the widely adopted UCBShift model [31]. We tested two settings: synthetic shifts for peptide 1DEC and experimental shifts for protein 1DFU. For 1DEC (also used in the NOE experiments), synthetic shifts were generated by running the deposited PDB structure through UCBShift and then attempting to recover them during optimization, providing a controlled baseline fully consistent with the scoring model. For 1DFU (BMRB ID 4395), we used calibrated experimental shifts from RefDB [39], selecting this protein because both an NMR ensemble and matched assignments are available.

In both experiments, guidance was restricted to C_α chemical shifts to emphasize backbone geometry. As shown in Fig. 4, the chemical shift objective proved more challenging than pairwise-distance or

NOE-based objectives, requiring substantially more optimization cycles to achieve improvement. The score consistently improved over cycles, although plateauing behavior was observed at later stages (Appendix F.1.3). Despite these improvements, the resulting structures did not converge toward the experimental PDB ensemble, indicating that chemical shift agreement does not necessarily translate into structural accuracy. This matches previous reports [2] that UCBShift predictions provide limited structural signal. Guided structures are provided in Appendix D.4.

4 Related Work

Gradient-free and evolutionary approaches are increasingly explored for molecular and biomolecular design. Early work such as MUDM [10] introduced a multi-objective diffusion framework for 3D molecule generation, using posterior approximation and Monte Carlo sampling to compute low-variance conditional gradients without retraining. In the protein domain, evolutionary diffusion has been applied to sequence generation in discrete space [1], but this does not address continuous 3D structure generation or conditioning structural samples on experimental data. More recently, there has been a surge of concurrent work extending these ideas to new molecular settings, including evolutionary multi-objective diffusion for molecular design [33] and reward-guided fine-tuning via iterative distillation to improve sample efficiency for biomolecular objectives [32]. Additional discussion of concurrent work in vision and general diffusion guidance is provided in Appendix C.

While these advances highlight the promise of gradient-free methods, none address the unique challenges of protein structure generation. Proteins inhabit a highly constrained, high-dimensional conformational space where even small perturbations can drastically change the resulting structure, and naive search often fails to converge. Unlike the discrete sequence domain, where every amino acid sequence corresponds to a valid backbone, the continuous structural domain is far more delicate: uninformed perturbations in latent space easily push candidates off-manifold, yielding non-physical conformations. To our knowledge, this work is the first to develop a tailored evolutionary search method for 3D protein structure generation with respect to *arbitrary* black-box objectives.

5 Discussion

We introduced a novel framework for experiment-guided protein structure generation that integrates non-differentiable objectives with generative diffusion models. By coupling BioEmu with a tailored genetic algorithm, we showed that black-box predictors ranging from pairwise distance restraints to NOE violation metrics and chemical shift scores can guide the search toward experimental consistency. Signal strength varied: pairwise restraints served as a proof of concept, NOEs enabled robust recovery of experimental folds, and chemical shifts improved agreement with the scoring model but did not yield structurally accurate ensembles. Limitations arise from both the generator and the guidance signals: BioEmu constrains the accessible conformational space, and current chemical shift predictors such as UCBShift remain limited by insufficient physical grounding, underscoring the need for more accurate and biophysically informed models. Future work includes stronger structural priors, physically informed and differentiable chemical shift models, ensemble-level objectives, and extending the framework to other modalities such as cryo-EM, crosslinking, or spectroscopy. Together, these directions would enable a more faithful integration of experimental data and reinforce genetic guidance as a general strategy for conditioning protein generative models on black-box objectives that fall outside the scope of differentiability.

Acknowledgments and Disclosure of Funding

Supported by ISTA and Technion. Thanks to the Bronstein and Schanda labs.

References

- [1] Sarah Alamdari, Nitya Thakkar, Rianne Van Den Berg, Neil Tenenholz, Bob Strome, Alan Moses, Alex Xijie Lu, Nicolo Fusi, Ava Pardis Amini, and Kevin K Yang. Protein generation with evolutionary diffusion: sequence is all you need. *BioRxiv*, pages 2023–09, 2023.

- [2] Meital Bojan, Sanketh Vedula, Advaith Maddipatla, Nadav Bojan Sellam, Federico Napoli, Paul Schanda, and Alex M Bronstein. Representing local protein environments with atomistic foundation models. *arXiv preprint arXiv:2505.23354*, 2025.
- [3] Stephen K Burley, Charmi Bhikadiya, Chunxiao Bi, Sebastian Bitrich, Li Chen, Gregg V Crichlow, Cole H Christie, Kenneth Dalenberg, Luigi Di Costanzo, Jose M Duarte, et al. Rcsb protein data bank: powerful new tools for exploring 3d structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic acids research*, 49(D1):D437–D451, 2021.
- [4] G Marius Clore and Angela M Gronenborn. Nmr structures of proteins and protein complexes. *FASEB Journal*, 8(8):682–690, 1994.
- [5] Valentin De Bortoli, Emile Mathieu, Michael Hutchinson, James Thornton, Yee Whye Teh, and Arnaud Doucet. Riemannian score-based generative modelling. *Advances in neural information processing systems*, 35:2406–2422, 2022.
- [6] A. E. Eiben and J. E. Smith. *Introduction to Evolutionary Computing*. Springer, 2003.
- [7] Alessandro Fadini, Mingjian Li, Airlie J McCoy, Thomas C Terwilliger, Randy J Read, Doeke Hekstra, and Mohammed AlQuraishi. AlphaFold as a prior: Experimental structure determination conditioned on a pretrained neural network. *bioRxiv*, 2025.
- [8] David E. Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley, 1989.
- [9] Beomsoo Han, Yifeng Liu, Simon W. Ginzinger, and David S. Wishart. Shiftx2: significantly improved protein chemical shift prediction. *J. Biomol. NMR*, 2011.
- [10] Xu Han, Caihua Shan, Yiffei Shen, Can Xu, Han Yang, Xiang Li, and Dongsheng Li. Training-free multi-objective diffusion model for 3d molecule generation. In *The Twelfth International Conference on Learning Representations*.
- [11] Haoran He, Jiajun Liang, Xintao Wang, Pengfei Wan, Di Zhang, Kun Gai, and Ling Pan. Scaling image and video generation via test-time evolutionary search. *arXiv preprint arXiv:2505.17618*, 2025.
- [12] Jeffrey C Hoch, Kumaran Baskaran, Harrison Burr, John Chin, Hamid R Eghbalnia, Toshimichi Fujiwara, Michael R Gryk, Takeshi Iwata, Chojiro Kojima, Genji Kurisu, et al. Biological magnetic resonance data bank. *Nucleic acids research*, 51(D1):D368–D376, 2023.
- [13] John H. Holland. *Adaptation in Natural and Artificial Systems*. University of Michigan Press, 1975.
- [14] Purvish Jajal, Nick John Eliopoulos, Benjamin Shiue-Hal Chou, George K Thiruvathukal, James C Davis, and Yung-Hsiang Lu. Inference-time alignment of diffusion models with evolutionary algorithms. *arXiv preprint arXiv:2506.00299*, 2025.
- [15] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.
- [16] Ian R. Kleckner and Mark P. Foster. An introduction to nmr-based approaches for measuring protein dynamics. *Methods Enzymol.*, 2010.
- [17] Piotr Klukowski, Fred F Damberger, Frédéric H-T Allain, Hideo Iwai, Harindranath Kadavath, Theresa A Ramelot, Gaetano T Montelione, Roland Riek, and Peter Güntert. The 100-protein nmr spectra dataset: A resource for biomolecular nmr data analysis. *Scientific data*, 11(1):30, 2024.
- [18] Jin Sub Lee and Philip M Kim. Flowpacker: protein side-chain packing with torsional flow matching. *Bioinformatics*, 41(3):btaf010, 2025.

- [19] Sarah Lewis, Tim Hempel, José Jiménez-Luna, Michael Gastegger, Yu Xie, Andrew YK Foong, Victor García Satorras, Osama Abdin, Bastiaan S Veeling, Iryna Zaporozhets, Yaoyi Chen, Soojung Yang, Adam E. Foster, Arne Schneuing, Jigyasa Nigam, Federico Barbero, Stumper Vincent, Andrew Campbell, Jason Yim, Marten Lienen, Yu Shi, Shuxin Zheng, Hannes Schulz, Usman Munir, Roberto Sordillo, Ryota Tomioka, Cecilia Clementi, and Frank Noé. Scalable emulation of protein equilibrium ensembles with generative deep learning. *Science*, page eadv9817, 2025.
- [20] Jie Li, Kochise C. Bennett, Yuchen Liu, Michael V. Martin, and Teresa Head-Gordon. Accurate prediction of chemical shifts for aqueous protein structure for “real world” cases using machine learning. *Chem. Sci.*, 2020. preprint arXiv:1912.02735.
- [21] Xiner Li, Masatoshi Uehara, Xingyu Su, Gabriele Scalia, Tommaso Biancalani, Aviv Regev, Sergey Levine, and Shuiwang Ji. Dynamic search for inference-time alignment in diffusion models. *arXiv preprint arXiv:2503.02039*, 2025.
- [22] Cheng Lu, Zhijie Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. In *Advances in Neural Information Processing Systems*, volume 35, pages 5775–5787, 2022.
- [23] Advaith Maddipatla, Nadav Bojan Sellam, Meital Bojan, Sanketh Vedula, Paul Schanda, Ailie Marx, and Alex M Bronstein. Inverse problems with experiment-guided alphafold. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025. PMLR 267; arXiv:2502.09372.
- [24] Sai Advaith Maddipatla, Nadav Bojan Sellam, Sanketh Vedula, Ailie Marx, and Alex M Bronstein. Generative modeling of protein ensembles guided by crystallographic electron densities. *arXiv preprint arXiv:2412.13223*, 2024.
- [25] John L Markley, Robert Powers, Eldon L Ulrich, and William M Westler. The future of nmr-based structural biology. *Journal of Magnetic Resonance*, 281:98–109, 2017.
- [26] Eli Fritz McDonald, Taylor Jones, Lars Plate, Jens Meiler, and Alican Gulsevin. Benchmarking alphafold2 on peptide structure prediction. *Structure*, 31(1):111–119, 2023.
- [27] Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. *arXiv preprint arXiv:2102.09672*, 2021.
- [28] Michael Nilges. Automated noe assignment and structure calculation with cyana. *Methods in Molecular Biology*, 60:51–88, 1997.
- [29] Arthur G Palmer. Nmr characterization of the dynamics of biomacromolecules. *Chemical Reviews*, 101(9):3113–3130, 2001.
- [30] Saro Passaro, Gabriele Corso, Jeremy Wohlwend, Mateo Reveiz, Stephan Thaler, Vignesh Ram Somnath, Noah Getz, Tally Portnoi, Julien Roy, Hannes Stark, et al. Boltz-2: Towards accurate and efficient binding affinity prediction. *BioRxiv*, pages 2025–06, 2025.
- [31] A. L. Ptaszek et al. Ucbshift 2.0: Bridging the gap from backbone to side chain protein chemical shift prediction for protein structures. *J. Am. Chem. Soc.*, 2024.
- [32] Xingyu Su, Xiner Li, Masatoshi Uehara, Sunwoo Kim, Yulai Zhao, Gabriele Scalia, Ehsan Hajiramezanali, Tommaso Biancalani, Degui Zhi, and Shuiwang Ji. Iterative distillation for reward-guided fine-tuning of diffusion models in biomolecular design. *arXiv preprint arXiv:2507.00445*, 2025.
- [33] Ruiqing Sun, Dawei Feng, Sen Yang, Yijie Wang, and Huaimin Wang. Evolutionary training-free guidance in diffusion model for 3d multi-objective molecular generation. *arXiv preprint arXiv:2505.11037*, 2025.
- [34] Gian Marco Visani, William Galvin, Michael Pun, and Armita Nourmohammad. H-packer: Holographic rotationally equivariant convolutional neural network for protein side-chain packing. In *Machine Learning in Computational Biology*, pages 230–249. PMLR, 2024.

- [35] Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Henry E Eisenach, Whitney Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. Broadly applicable and accurate protein design by integrating structure prediction networks and diffusion generative models. *Nature*, 620(7976):1089–1100, 2023.
- [36] Zhao Wei, Chin Chun Ooi, Abhishek Gupta, Jian Cheng Wong, Pao-Hsiung Chiu, Sheares Xue Wen Toh, and Yew-Soon Ong. Evolvable conditional diffusion. *arXiv preprint arXiv:2506.13834*, 2025.
- [37] David S Wishart and David A Case. Nmr chemical shifts as a tool for understanding protein structure and dynamics. *Current Opinion in Structural Biology*, 15(5):572–578, 2005.
- [38] Kurt Wüthrich. *NMR of Proteins and Nucleic Acids*. Wiley, 1986.
- [39] Haiyan Zhang, Stephen Neal, and David S Wishart. Refdb: a database of uniformly referenced protein chemical shifts. *Journal of biomolecular NMR*, 25:173–195, 2003.

A Extended Background

A.1 BioEmu Training and Sampling

BioEmu [19] is an E(3)-equivariant conditional diffusion model trained to sample residue-level protein conformations from a learned equilibrium distribution. Each protein is represented in a residue-wise backbone frame representation, where residue i is parameterized by a tuple (r_i, Q_i) consisting of the C_α position $r_i \in \mathbb{R}^3$ and a local orientation frame $Q_i \in SO(3)$ constructed via Gram–Schmidt orthogonalization of the N–C α and C–C α bonds.

The forward noising process is modeled as a pair of independent stochastic differential equations (SDEs) acting on positions and orientations:

$$dr_t = -\frac{1}{2}\beta(t)r_t dt + \sqrt{\beta(t)} dW_t, \quad (1)$$

$$dQ_t = \mathcal{L}_{SO(3)}(Q_t) dt + \sqrt{2} d\mathcal{W}_t^{SO(3)}, \quad (2)$$

where $\beta(t)$ follows a cosine noise schedule [27], W_t is standard Brownian motion in \mathbb{R}^3 , and $d\mathcal{W}_t^{SO(3)}$ denotes Brownian motion on the rotation group $SO(3)$ with generator $\mathcal{L}_{SO(3)}$ corresponding to the isotropic Gaussian distribution on rotations (IGSO(3)) [5]. This formulation ensures that positional noise is variance-preserving and that rotational noise is uniformly distributed over the manifold, enabling the model to learn orientation-equivariant denoising updates.

BioEmu learns a time-dependent score network $s_\theta(r_t, Q_t, t)$ that estimates the gradient of the log-density of the perturbed data:

$$s_\theta(r_t, Q_t, t) \approx \nabla_{r_t, Q_t} \log p_t(r_t, Q_t), \quad (3)$$

and generates samples by integrating the reverse-time SDE with learned score estimates:

$$d(r_t, Q_t) = \left[-\frac{1}{2}\beta(t)(r_t, Q_t) - \beta(t)s_\theta(r_t, Q_t, t) \right] dt + \sqrt{\beta(t)} d\bar{W}_t, \quad (4)$$

where $d\bar{W}_t$ denotes time-reversed Brownian motion.

Training data. BioEmu is first pre-trained on clustered AlphaFoldDB structures to encourage conformational diversity. It is then fine-tuned on more than 200 ms of equilibrium MD trajectories, reweighted by Markov state models to better match equilibrium populations. Finally, it is refined using Property Prediction Fine-Tuning (PPFT) with MEGAscale protein stability measurements, aligning the learned distribution with experimental free energies.

Side-chain modeling. In its original form, BioEmu employs Hpacer [34] for reconstructing side-chain conformations from the generated backbones. In this work, we instead integrate FlowPacer [18], a flow-based packing model that provides faster inference and improved physical plausibility of side-chain placements. This substitution improves side-chain accuracy while reducing runtime, leading to more reliable structures for downstream genetic guidance.

A.1.1 DPM-Solver for Fast Sampling

To generate samples efficiently, BioEmu employs DPM-Solver [22], a high-order numerical ODE solver specifically designed for diffusion probabilistic models. Given a learned score function $s_\theta(x_t, t)$, the corresponding probability flow ODE is

$$\frac{dx_t}{dt} = f_\theta(x_t, t) = -\frac{1}{2}\beta(t)[x_t + s_\theta(x_t, t)], \quad (5)$$

where $\beta(t)$ is the noise schedule. DPM-Solver integrates this ODE using a second- or third-order solver, yielding high-fidelity samples with significantly fewer function evaluations compared to standard ancestral samplers:

$$x_{t_{k-1}} = x_{t_k} + h_k \Phi^{(m)}(f_\theta, x_{t_k}, t_k),$$

where $h_k = t_{k-1} - t_k$ is the integration step and $\Phi^{(m)}$ is an m -th order multi-step update (typically $m = 2$ or $m = 3$). This formulation enables BioEmu to generate thousands of near-equilibrium

conformations in only 30–50 denoising steps, compared to hundreds of steps required by Euler–Maruyama or predictor–corrector methods.

DPM-Solver supports both a deterministic ODE solver and a stochastic variant based on the reverse SDE:

$$dx_t = f_\theta(x_t, t) dt + \sqrt{\beta(t)} d\bar{W}_t, \quad (6)$$

where $d\bar{W}_t$ is the time-reversed Brownian motion term, reintroducing noise at each step to maintain diversity. BioEmu’s default procedure uses the stochastic variant (6), whereas in this work we employ the deterministic ODE formulation (5) to improve prevent drastic changes to the protein and ensure stable convergence under genetic algorithm-based optimization.

B Biological Background

NMR spectroscopy for proteins. Nuclear magnetic resonance (NMR) spectroscopy is a widely used technique for probing protein structure and dynamics in solution [38]. NMR exploits the magnetic properties of atomic nuclei (most commonly ^1H , ^{13}C , and ^{15}N) and measures their resonance frequencies in a strong magnetic field. Each experiment produces a spectrum in which peaks correspond to specific nuclear spins, providing atom-level information that reflects the local chemical environment and, indirectly, the 3D structure of the protein. Unlike X-ray crystallography, NMR is performed in solution, making it particularly valuable for studying flexible regions, conformational exchange, and dynamics [29].

Chemical shifts and peak assignment. The chemical shift is the fundamental NMR observable: the resonance frequency of a nucleus relative to a reference compound. It is highly sensitive to the local electronic environment and therefore encodes detailed information about backbone torsion angles, hydrogen bonding, and secondary structure [37]. In practice, chemical shifts appear as peaks in 1D or 2D spectra, but before they can be used for structure modeling, these peaks must be *assigned* to specific atoms in the protein sequence. Peak assignment is notoriously challenging: large proteins produce thousands of overlapping peaks, noise and artifacts can obscure weaker signals, and experimental conditions can shift peaks, making manual assignment a laborious, error-prone process that often requires days or weeks of expert effort [25]. Automating or accelerating chemical shift based assignment remains one of the major bottlenecks in NMR-driven structure determination.

NOE distance restraints. Another major source of structural information in NMR is the nuclear Overhauser effect (NOE), which reports on through-space dipolar couplings between nearby protons. NOEs are typically measured using nuclear Overhauser effect spectroscopy (NOESY), a 2D or 3D NMR experiment in which cross-peaks indicate spatial proximity between proton pairs. These cross-peaks can be converted into approximate upper bounds on inter-proton distances, typically up to 5 Å [38, 4]. Such distance restraints are invaluable for constraining tertiary structure, as they directly encode spatial proximity between residues. However, NOESY spectra are often highly crowded, and many peaks can correspond to multiple possible proton pairs, making NOE assignment a combinatorial problem that usually requires iterative assignment and structure refinement cycles [28].

C Extended Related Work

Beyond biology, there has been a recent surge of interest in combining diffusion models with evolutionary or black-box search methods in other domains. Concurrent work in the vision domain has explored several complementary approaches for gradient-free guidance: large-scale test-time evolution search to optimize image and video fidelity [11], evolutionary algorithms for inference-time alignment of generated samples with downstream objectives [14], and black-box conditional generation via evolvable diffusion, which frames conditioning as an optimization problem over the diffusion process [36]. Dynamic search methods have also been proposed to improve inference-time alignment of diffusion models, offering a gradient-free alternative to direct score-based guidance [21].

D Extended experiments

D.1 Mean vs. best metric

Throughout the paper, unless otherwise noted, we report the best-structure metric per cycle, defined as the minimum loss (or maximum score) observed within the population at that cycle. This choice provides a consistent summary across modalities and emphasizes the emergence of the highest-quality candidates during the search.

For completeness, Fig. 5 shows the mean-structure metric per cycle plotted alongside the best-structure metric using the same definitions as in the main text. Across targets and modalities, the mean traces closely mirror the best traces and exhibit the same trends (improvements and occasional plateaus), supporting our choice to emphasize the best-structure metric in the main figures.

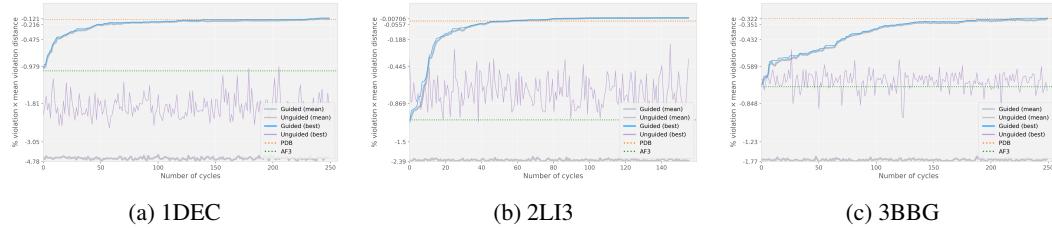


Figure 5: Best vs. mean metric per cycle. For each target, we plot the best-structure metric (solid) and the mean-structure metric (faint) using the same definitions as in the main text for NOE guidance (Y-axis: fraction of violated restraints \times mean violation distance; X-axis: number of cycles). Dashed baselines indicate PDB and AF3.

D.2 Pairwise distances experiments

This section provides extended results for pairwise distance guidance. As described in Section 3, pairwise $C_\alpha-C_\alpha$ distances within 5 Å were used as surrogate NOE restraints to guide the search. Figure 6 reports results for 4OLE, which exhibits two alternate conformations (A: helical, B: strand). Unguided BioEmu produces structures resembling altloc B but never recovers conformation A, whereas guidance consistently improves the best-structure metric and enables recovery of the helical conformation. These results illustrate how pairwise distance information can provide a strong corrective signal, even in challenging cases where the unguided prior fails.

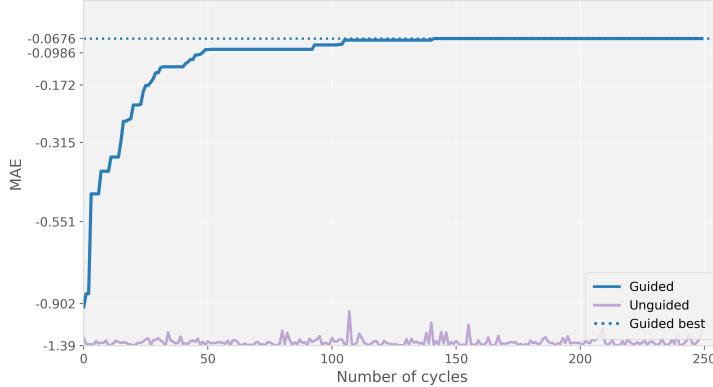


Figure 6: 4OLE pairwise distance guidance. BioEmu guided with C_α pairwise distance restraints. Y-axis: mean absolute error (MAE). The search was initialized from altloc B, which BioEmu alone resembles but never recovers as conformation A. Guidance enables consistent improvement in the best-structure metric and recovery of the helical conformation A.

D.3 NOE experiments

This section provides extended results for NOE-guided experiments. In all cases, candidate structures were scored using the NOE violation metric (fraction of violated restraints multiplied by mean violation distance; see Appendix F). Figures 7–11 report optimization traces for multiple proteins. Table 1 reports metric comparisons in violation percentage, mean violation and median violation for the reference PDB, the AF3 prior and the best guided sample. Consistently across all cases, BioEmu guided with NOE-derived restraints improves the best-structure metric over cycles, always outperforming the AlphaFold prior and the unguided BioEmu baseline, and often matching or surpassing the reference PDB performance. These consistent results demonstrate the effectiveness of the tailored genetic algorithm in leveraging experimental restraints to refine generative models toward experimentally consistent structures.

PDB	Method	Loss	Percent violation	Mean violation	Median violation
1dec	PDB	0.1279	0.2211	0.5787	0.4981
	AF3	1.0658	0.3347	3.1846	1.5388
	Guided-best	0.1184	0.2110	0.5613	0.3344
2kkz	PDB	0.1445	0.2629	0.5496	0.4081
	AF3	0.2836	0.3042	0.9324	0.5786
	Guided-best	0.1492	0.2579	0.5784	0.4093
1pqx	PDB	0.0429	0.1215	0.3533	0.3389
	AF3	0.1409	0.1497	0.9417	0.3457
	Guided-best	0.0594	0.1215	0.4894	0.3405
2li3	PDB	0.0259	0.0644	0.4016	0.3664
	AF3	1.1002	0.2441	4.5076	3.9301
	Guided-best	0.0001	0.0068	0.0080	0.0076
3bbg	PDB	0.3200	0.5303	0.6035	0.4873
	AF3	0.7224	0.6086	1.1870	0.5474
	Guided-best	0.3180	0.5126	0.6203	0.4672

Table 1: Comparison of NOE metrics across the AlphaFold (AF3) prior, the PDB reference, and the best structure in the guided population (min-loss). Bold indicates the best (lowest) value for each metric within a PDB. Details of the evaluation procedure are provided in Appendix F.2.

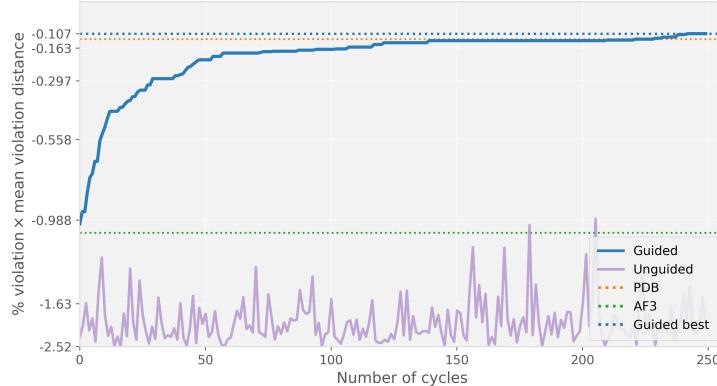


Figure 7: **1DEC NOE guidance.** BioEmu guided with NOE-derived restraints. Y-axis: fraction of violated restraints multiplied by mean violation distance. Guidance yields consistent improvement in the best-structure metric, surpassing the PDB according to this measure.

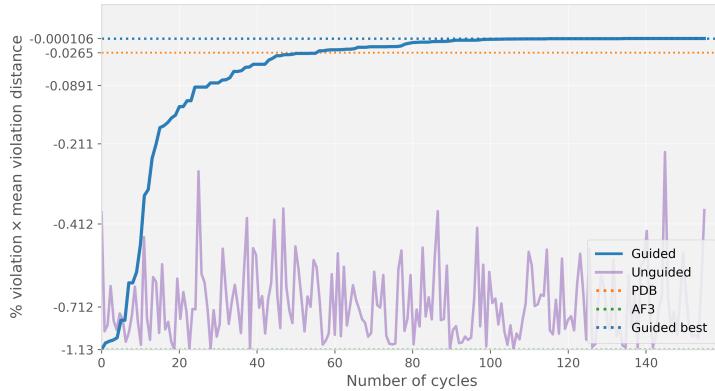


Figure 8: **2LI3 NOE guidance.** BioEmu guided with NOE-derived restraints. Y-axis: fraction of violated restraints multiplied by mean violation distance. Guidance yields consistent improvement in the best-structure metric, surpassing the PDB according to this measure.

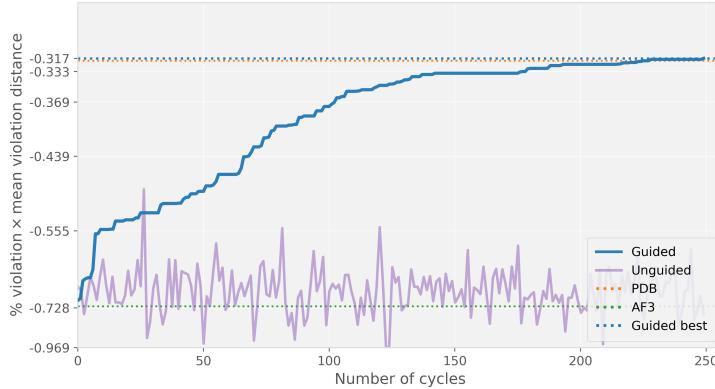


Figure 9: **3BBG NOE guidance.** BioEmu guided with NOE-derived restraints. Y-axis: fraction of violated restraints multiplied by mean violation distance. Guidance yields consistent improvement in the best-structure metric, reaching performance comparable to the PDB according to this measure.

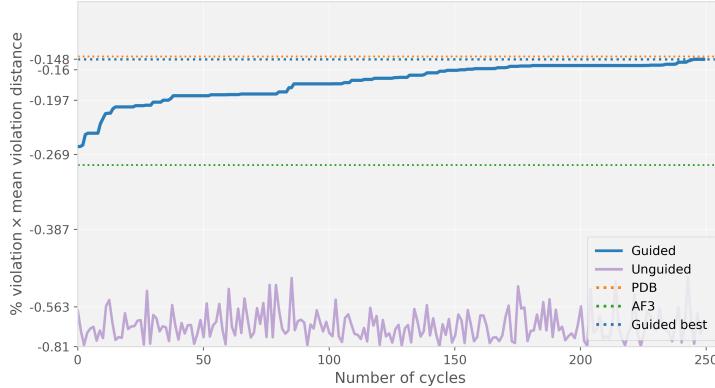


Figure 10: **2KKZ NOE guidance.** BioEmu guided with NOE-derived restraints. Y-axis: fraction of violated restraints multiplied by mean violation distance. Guidance yields consistent improvement in the best-structure metric, reaching performance comparable to the PDB according to this measure.

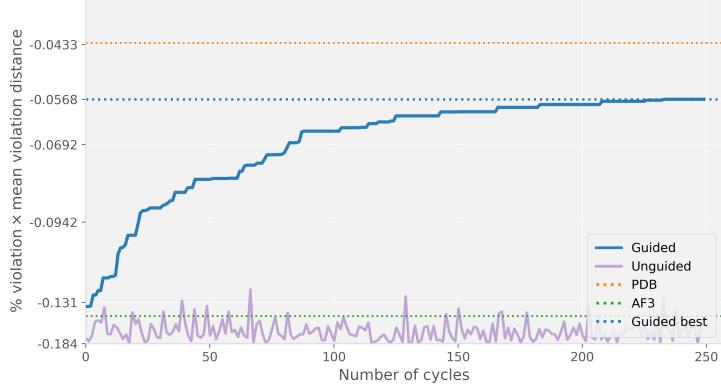


Figure 11: **1PQX NOE guidance.** BioEmu guided with NOE-derived restraints. Y-axis: fraction of violated restraints multiplied by mean violation distance. Guidance yields consistent improvement in the best-structure metric.

D.3.1 NOE cumulative violation distributions

Setup. For each target we plot the cumulative distribution of NOE violation magnitudes (including zeros) in Fig. 12, i.e., the fraction of restraints with violation $\leq x$ as a function of the threshold x (in Å). Curves that rise faster/earlier indicate fewer and/or smaller violations. Each panel compares the reference PDB structure, the AlphaFold3 (AF3) prior, and the best guided structure (minimum NOE loss) from our search.

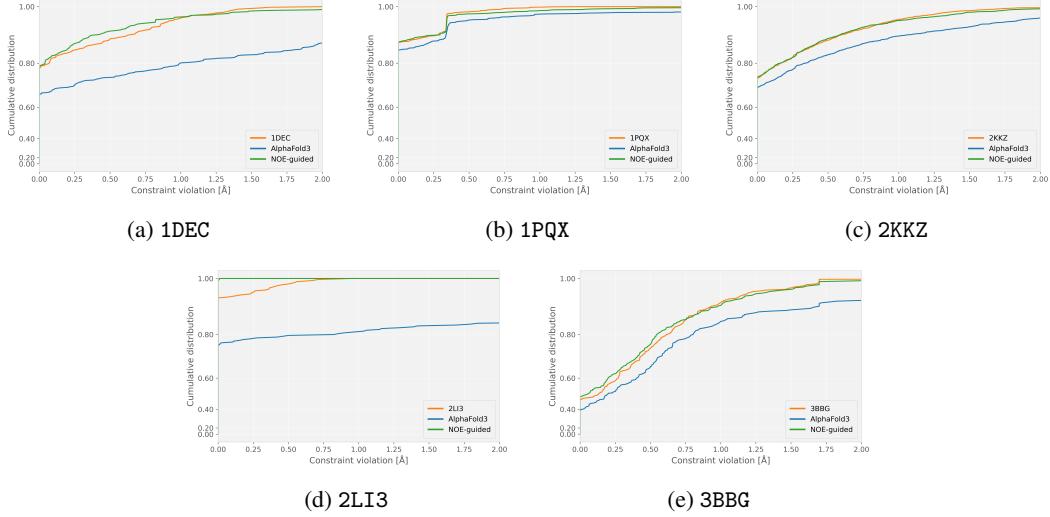


Figure 12: **Cumulative violation distributions for NOE restraints.** Faster-rising curves indicate fewer/smaller violations. Guided consistently outperforms AF3 and is comparable to or better than the PDB across targets.

Summary. Across all targets, the guided curve consistently dominates the AF3 curve (higher at nearly all thresholds), showing uniformly fewer/smaller violations than the AF3 prior. Relative to the PDB reference, guided is either comparable or better over most of the range (notably strong on 2LI3), and on the remaining cases it closely tracks the PDB curve. These results mirror the scalar NOE metrics in the main text/table and visualize how guidance reduces both violation frequency and severity.

D.4 Chemical shift experiments

This section provides extended results for chemical shift–guided experiments. Structures guided with UCBShift-predicted shifts are shown in Fig. 13. The chemical shift metric, defined in Appendix F.1.3, consistently improved over optimization cycles, although plateauing behavior was observed at certain stages. Despite this metric improvement, the resulting guided structures did not converge toward the experimental PDB ensembles, underscoring the discrepancy between chemical shift agreement and structural similarity. These results highlight both the utility of chemical shift data for guidance and the current limitations of predictors such as UCBShift, which lack explicit physical grounding.

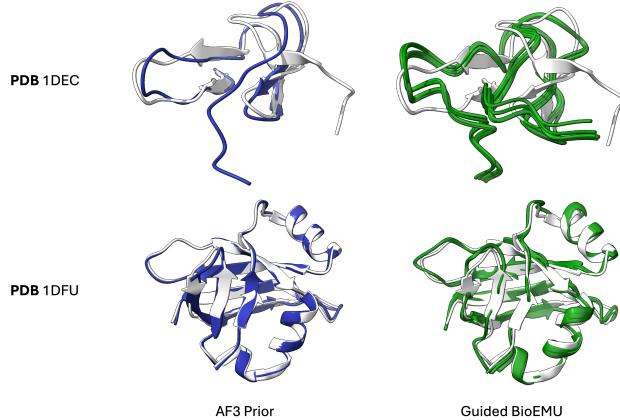


Figure 13: **Chemical shift guided structures.** Guiding BioEmu with UCBShift shifts predictions. Reference PDB structures shown in white. 1DEC was guided with synthetic PDB-derived shifts. 1DFU was guided with experimental shifts (BMRB 4395).

E Plateau in Metric Graphs

When running our genetic-guided algorithm with a constant perturbation time, we observed plateaus in the evaluation metric. To address this, we modified the perturbation diffusion time to follow a linear schedule. This adjustment helps because, as the search progresses, the required mutations need to become more fine-grained. A fixed diffusion time cannot adequately support convergence, since the resulting mutations introduce excessive structural changes. In contrast, gradually reducing the diffusion time across cycles enables faster and more accurate convergence.

However, plateaus still occur. We hypothesize that this happens because the genetic search requires spending varying numbers of cycles at different perturbation times, and a simple linear schedule cannot capture this behavior. As future work, we plan to design a more adaptive scheduling strategy in which the perturbation time is adjusted dynamically—only when a plateau is detected.

This trend is illustrated in Figures 7–11.

F Technical details

The following section provides full technical details of our method. We first describe the **score functions** used to evaluate candidate structures during genetic optimization (Section F.1), including pairwise distance matching, NOE-derived distance restraints, and chemical shift prediction with UCBShift. Next, we outline the **evaluation metrics** (Section F.2) used to quantify structural agreement with experimental data. We then provide details of **data curation** (Section F.3), specifying how reference structures and experimental constraints were selected and processed for each modality. Finally, we summarize **runtime analysis** (Section F.4) and **hyperparameter settings** (Section F.5) to support replication of our experiments.

F.1 Score models

To evaluate candidate protein structures in our genetic algorithm, we used three complementary scoring models: NOE-based distance restraints, generic pairwise distance matching, and chemical shift prediction (UCBShift [31]).

F.1.1 Pairwise distance model

The pairwise model provides a general framework for comparing inter-atomic distances against reference values. Inputs are a CSV file of atom pairs with target distances. Given a structure, distances are computed as

$$d_{ij} = \|\mathbf{r}_i - \mathbf{r}_j\|_2.$$

Three loss modes are supported:

- 12: mean squared error between predicted and target distances,
- 11: mean absolute error,
- topk: mean absolute error of the k largest deviations, highlighting worst-case mismatches.

In practice, we found that the 11 formulation provided the most stable and informative signal for guiding the genetic algorithm, and it was therefore used as the default.

F.1.2 NOE restraint model

The NOE scoring function enforces satisfaction of nuclear Overhauser effect (NOE)-derived distance restraints. Input restraints are provided as a CSV file with atom–atom pairs, residue indices, and upper/lower distance bounds.

Since BioEmu does not generate explicit hydrogen atoms, we approximate NOE restraints by enforcing them on the heavy atoms to which the hydrogens are covalently bound. For example, H–H restraints are mapped to the corresponding C–C or C–N distances. This approximation preserves the geometric character of the NOE but avoids inconsistencies due to missing hydrogens.

For each experimental restraint between atoms i and j , we compute the interatomic distance

$$d_{ij} = \|\mathbf{r}_i - \mathbf{r}_j\|_2,$$

and measure its deviation from the allowed bounds ℓ_{ij} and u_{ij} as

$$L_{ij} = \max(0, \ell_{ij} - d_{ij}) + \max(0, d_{ij} - u_{ij}).$$

Thus $L_{ij} = 0$ whenever the restraint is satisfied, and grows linearly with the degree of violation otherwise.

Because NMR assignments are often uncertain, restraints are grouped into *ambiguous sets* $G \subseteq \mathcal{R}$ with an effective OR-logic: it is sufficient that one member of the set is satisfied. Accordingly, the violation of a group is defined as the smallest of its constituent penalties,

$$L_G = \min_{(i,j) \in G} L_{ij}.$$

Let \mathcal{G} be the collection of all such groups, and $\mathcal{V} = \{G \in \mathcal{G} : L_G > 0\}$ the subset of violated groups. The overall NOE score combines both frequency and severity of violations:

$$L_{\text{NOE}} = \frac{|\mathcal{V}|}{|\mathcal{G}|} \cdot \frac{1}{|\mathcal{V}|} \sum_{G \in \mathcal{V}} L_G.$$

In other words, the final loss equals the fraction of violated groups multiplied by the average magnitude of those violations.

This design penalizes both widespread mild inconsistencies and a few severe outliers. Since it involves max and min operations together with discrete violation indicators, the function is *non-differentiable*. We therefore use L_{NOE} strictly as a black-box objective in our genetic search.

Distances vs. peak intensities. NOE signals in NMR originate from dipolar couplings between protons within roughly 6 Å. The intensity of a cross-peak in a NOESY spectrum follows an inverse-sixth power dependence on the internuclear distance, $I \propto r^{-6}$. After peak assignment, these intensities are typically converted into effective distance restraints according to

$$r = r_{\text{ref}} \left(\frac{I_{\text{ref}}}{I} \right)^{1/6},$$

where r_{ref} and I_{ref} denote the distance and intensity of a reference pair.

For structural ensembles, physical consistency requires computing the mean of the calibrated intensities across conformations and then transforming this average intensity back into a distance. Directly averaging distances neglects the r^{-6} relationship and can bias the restraints. In the present work we use the simpler distance-based formulation, while a fully intensity-averaged treatment will be incorporated in future extensions.

Ensemble effects. Because NOE cross-peaks reflect population-weighted averages over conformational ensembles, restraints are fundamentally ensemble quantities. In this work, however, we applied the constraints independently to each generated structure. A more rigorous treatment would evaluate NOE violations over an ensemble of conformations, thereby capturing conformational heterogeneity; we leave this extension for future work.

F.1.3 Chemical shift model (UCBShift)

To incorporate chemical shift data, we used UCBShift [31], a widely used chemical shift predictor trained on large experimental databases. Our implementation calls both the X and Y modules of UCB-Shift to obtain predictions for the supplied structure via the `calc_sing_pdb` routine. Experimental reference shifts are read from a CSV file and aligned by residue index to the predicted values.

The loss is computed as either

$$L_{l1} = \frac{1}{N} \sum_i |y_i - \hat{y}_i|, \quad L_{l2} = \frac{1}{N} \sum_i (y_i - \hat{y}_i)^2,$$

depending on the chosen mode.

Although the L_1 and L_2 loss functions are differentiable, the overall score is *not differentiable*, because the UCBShift algorithm itself is non-differentiable. Thus, gradients with respect to structural coordinates are not available, and the model can only be used as a black-box scoring function within our genetic algorithm.

F.2 Evaluation metric

F.2.1 Pairwise distances

Evaluation for the pairwise distance experiments is based directly on the loss functions defined in Appendix F.1.1. Given reference distances and predicted distances, the score is computed using the chosen formulation (11, 12, or topk), with the L_1 version serving as our default. Thus, the pairwise evaluation metric coincides exactly with the pairwise loss used during optimization.

F.2.2 NOE restraints

We evaluate structural agreement with NOE restraints using the violation terms defined in Appendix F.1.2. For each restraint group $G \in \mathcal{G}$ we obtain a group-level violation L_G , where $L_G = 0$ indicates satisfaction and $L_G > 0$ indicates a violation. From these values we report three descriptive metrics:

- **Violated NOE restraints percentage:** the fraction of groups with nonzero violation,

$$p_{\text{viol}} = \frac{|\{G \in \mathcal{G} : L_G > 0\}|}{|\mathcal{G}|}.$$

- **Median violation:** the median magnitude of violations across the violated groups (reported in Å),

$$m_{\text{viol}} = \text{median}(\{L_G : L_G > 0\}).$$

- **Mean violation:** the mean magnitude of violations across the violated groups (reported in Å),

$$\mu_{\text{viol}} = \frac{1}{|\{G \in \mathcal{G} : L_G > 0\}|} \sum_{G:L_G>0} L_G.$$

As in the NOE model, we also summarize performance with a scalar loss

$$L_{\text{NOE}} = p_{\text{viol}} \cdot \mu_{\text{viol}},$$

which penalizes both the prevalence and severity of violations.

F.2.3 Chemical shift

Evaluation against chemical shift data follows the loss definitions given in Appendix F.1.3. Predicted shifts \hat{y}_i are compared to experimental reference values y_i using either the L_1 or L_2 loss, with L_1 as the default. Accordingly, the chemical shift evaluation metric coincides with the scoring function used during optimization.

F.3 Data curation

F.3.1 Pairwise distances

For the pairwise distance experiments we selected the protein 4OLE, which corresponds to a domain from the NBR1 gene, a neighbor of human BRCA1 that has been implicated in breast cancer. This entry is particularly challenging because it contains two alternative conformations (altloc A and altloc B). Conformation A adopts a helical arrangement, whereas conformation B forms a strand. Since BioEmu by default generates structures closer to altloc B, we instead used altloc A as the reference. The main structural differences are concentrated in residues 60–68, alongside smaller variations that are less consequential. From the reference structure we computed all C_α – C_α pairwise distances and retained only those pairs within 5 Å. This cutoff was chosen to approximate the short-range nature of NOE-derived contacts, providing a simplified proxy for NMR-like restraints. The resulting set of pairwise contacts was used as the experimental input for the genetic guidance procedure.

F.3.2 NOE restraints

Structure selection We evaluated our method on proteins where the BioEmu generator alone produced inaccurate folds. From the benchmark of [26], we first considered the 20 peptides reported as most challenging for AlphaFold. This yielded three peptides—1DEC, 2LI3, and 3BBG—on which BioEmu also generated erroneous or partially misfolded conformations.

We further selected proteins from the 100 NMR spectra database [17], choosing 1PQX and 2KKZ because BioEmu struggled to reproduce their experimental folds. Together, these five cases form a targeted testbed: experimentally validated NMR ensembles that expose weaknesses of both AlphaFold and BioEmu, allowing us to test whether our genetic guidance procedure can repair BioEmu’s predictions.

NOE restraint extraction NOE-derived distance restraints were obtained directly from the NMR STAR files associated with each PDB entry. We parsed the restraint tables using the `pynmrstar` library, selecting only those entries explicitly labeled as NOE constraints (`_Gen_dist_constraint_list.Constraint_type`). Other types of structural information present in the files—such as hydrogen-bond, dihedral, or RDC-derived restraints—were excluded so that the scoring relied solely on NOE distance geometry.

Ambiguous assignments involving multiple possible proton pairs were retained without disambiguation, in order to reflect the inherent uncertainty of NMR peak assignment. Lower and upper bounds were taken directly from the STAR fields `_Gen_dist_constraint_list.Distance_lower_bound_val` and

`_Gen_dist_constraint_list.Distance_upper_bound_val`, respectively. When a lower bound was missing, it was explicitly set to 0 Å to enforce a physically valid minimum distance. All bounds were kept in their original Ångström units, with no additional normalization or thresholding, to preserve fidelity to the experimental input data.

F.3.3 Chemical shifts

The evaluation of a realistic case was done using experimental chemical shifts from RefDB [39], a curated subset of the BMRB [12] containing calibrated assignments.

Additional details are described in Section 3

F.4 Runtime analysis

Unless otherwise stated, all runs were executed on a single NVIDIA H100 GPU with 30 CPU cores used for parallel score evaluation. Excluding the UCBShift experiments, wall-clock time varied between 1.0 and 3.5 hours per target, primarily as a function of sequence length and atom count. Small proteins such as 1DEC completed in roughly 1 hour, whereas larger targets like 4OLE required about 3 hours. The UCBShift runs exhibit higher variance due to predictor latency and the larger number of generations (Section F.5); we therefore report only cycle counts for those experiments.

Hardware: 1×H100 GPU, 30 CPU cores; parallelized black-box scoring.

F.5 Hyperparameters

We tuned four hyperparameter families: (i) the number of generations (cycles), (ii) the diffusion-time scheduler that maps generation index k to a perturbation time t_k , (iii) the tournament size for parent selection, and (iv) the elitism keep for survivor selection.

Population and selection. Unless noted otherwise, we used a population size of 100, tournament size = 5, and elitism keep = 5. These values provided a stable exploration–exploitation balance across targets.

Cycles (generations). Most experiments ran for 250 cycles. Exceptions are summarized below:

Target	Modality	Cycles
1DEC	UCBShift (chemical shifts)	1700
1DFU	UCBShift (chemical shifts)	600
2LI3	NOE	153
<i>All others</i>	Pairwise / NOE	250

Diffusion-time scheduler. The scheduler specifies which perturbation diffusion time t_k is used at generation k . We tuned this mapping to control the global→local search progression. Concretely, we select (t_{\min}, t_{\max}) and a schedule shape, then set t_k accordingly. In practice, early generations use larger t_k to promote broader moves; later generations reduce t_k to focus on local refinements. For difficult objectives (e.g., UCBShift) we biased the schedule toward spending more generations at intermediate t_k to mitigate early plateaus. For all other experiments, we started from diffusion index 0 and increased it by 1 every 4 cycles, capping the maximum index at 40. The generative process spans indices 0 (completely noisy) through 50 (completely clean).