RETHINKING SELF-CONSISTENCY IN PROTEIN GENERATIVE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-consistency is the standard metric for evaluating structural plausibility in protein generative models: a generated sequence–structure pair is accepted if an independent folding model refolds the sequence to the generated structure within a fixed RMSD threshold. However, self-consistency rejects 44% of native sequence–structure pairs, a gap previously overlooked as a folding model limitation. In this paper, we connect self-consistency to the biophysical ideal of structural plausibility–sufficiently low free energy in that protein's equilibrium ensemble–which makes two failure modes visible: *flexibility failures*, where a broad ensemble cannot be summarized by a single refold, and *folder failures*. We use ATLAS molecular dynamics trajectories to empirically analyze native structure rejections through this lens, and find that both failure modes are enriched among flexible proteins. We propose ensemble self-consistency using conformational ensemble predictors, improving native protein recall across the flexibility spectrum and rescuing co-design generations previously rejected, without compromising specificity on a Rosetta decoy dataset. Since flexibility is often central to protein function, a metric that penalizes it misdirects generative model development; ensemble self-consistency offers a more faithful framework, and our formalization and diagnostics let it evolve as structure prediction models improve.

1 INTRODUCTION

Protein structure generative models (Watson et al., 2023; Geffner et al., 2025a; Lin et al., 2024; Zambaldi et al., 2024) are transforming protein design, generating nanomolar binders to viral proteins (Didi et al., 2026b) and small molecules (Stark et al., 2025). Quality evaluation of generated protein structures is crucial to guide the development of generative models and to select the structures for costly experimental verification. Yet assessing the plausibility of protein tertiary structure is nontrivial. Standard physical validity metrics (Chen et al., 2010) are largely local; a generative model can easily satisfy the bond geometry and avoid steric clashes while still producing unrealistic folds.

To assess global plausibility, the field has converged on the self-consistency pipeline (Yim et al., 2023). The premise is that folding models have learned the manifold of plausible folds, so an implausible tertiary structure should not be reproduced from its sequence. In this pipeline, given an all-atom structure (which implies a sequence), an independent folding model–commonly ESMFold (Lin et al., 2023)–predicts a structure from the sequence, and the structure is accepted if the two agree within a fixed RMSD threshold, typically 2Å (Yim et al., 2023). Self-consistency is now a headline metric in protein co-design (Geffner et al., 2025a; Rector-Brooks et al., 2026) and is often used as an early filter before experimental validation (Didi et al., 2026a).

Despite its central role, the self-consistency pipeline has received little systematic scrutiny. In our PDB subset of 1,520 native protein structures, 44% fail the standard self-consistency test, even though such structures should be accepted by any reasonable plausibility metric. Similar native failure rates have also been noted in Lu et al. (2025), but the underlying failure modes still remain unclear. In this work, we ask a more nuanced, fundamental question: *when self-consistency rejects a real protein, what exactly has failed?*

Answering this requires first specifying what self-consistency is meant to approximate. We first formalize structural plausibility as membership in the high probability region of a protein's equilibrium

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

ensemble. Canonical self-consistency, then, can be viewed as a proxy that replaces ensemble membership with proximity to a single predicted structure. This formalization makes two failure modes of native rejection visible.

A *flexibility failure* arises when the equilibrium ensemble is broad, so a plausible conformation need not match a single refolded conformation within 2Å. A *folder failure* arises when the folding model's prediction is itself a poor surrogate for the sequence-conditioned ensemble. Empirically, using the ATLAS database (Vander Meersche et al., 2024), we decompose native protein failures into these two modes and find that **both** are enriched among flexible proteins. The self-consistency false negatives are not random: they concentrate on flexibility, the core property of protein biology that underlies allostery, catalysis, and binding.

To address this bias, we propose *ensemble self-consistency*, which replaces comparison to a single deterministic refold with comparison to a sampled conformation ensemble. This better approximates the ideal acceptance criterion of whether a structure lies within the sequence-conditioned plausible ensemble. Empirically, flexibility failures are nearly eliminated by moderate sample sizes (≥ 16), and replacing deterministic ESMFold self-consistency with SimpleFold ensemble self-consistency significantly improves recall on native proteins and experimentally characterized de novo designs (Chronowska et al., 2025), while preserving high specificity on a hard Rosetta decoy benchmark (Park et al., 2016). Finally, we re-evaluate published co-design models and show that ensemble self-consistency rescues flexible generations that canonical self-consistency rejects.

Contribution. Our contributions are threefold: **(i)** A formal account of self-consistency as an approximation to an ideal structural plausibility, and use this to decompose native failures into flexibility and folder modes, **(ii)** Empirically characterize these two failure modes on native proteins and show that both increase with conformational flexibility, **(iii)** *Ensemble self-consistency*, a drop-in replacement that nearly eliminates flexibility failures at moderate sample sizes (≥ 16).

2 Related works

Local physical validity metrics. Protein structure quality has long been assessed with local physical and stereochemical criteria, including bond geometry, Ramachandran outliers, steric clashes (Chen et al., 2010; Adams et al., 2010). These metrics are essential for refining and ranking experimental structure characterizations, where the global fold is established but local details may be uncertain; however, they do not assess whether the global fold is plausible.

Self-consistency and (co-)designability. The standard per-sample evaluation metric for protein generative models is self-consistency. Paired with folding-model confidence, it is termed **co-designability**. ESMFold (Lin et al., 2023) is the most commonly used folding model, while DISCO (Rector-Brooks et al., 2026) uses Chai-1 (team et al., 2024) to incorporate ligand input for enzyme design. Crucially, folding models must not require multiple sequence alignments (MSAs) as input, because de novo sequences generally lack meaningful MSAs. This rules out direct use of structure or ensemble predictors that require MSAs for strong performance, such as AlphaFold2/3 (Jumper et al., 2021; Abramson et al., 2024) and BioEmu (Lewis et al., 2025). When evaluating the models that only generate protein backbones, sequences are first designed with an inverse folding model such as ProteinMPNN (Dauparas et al., 2022), and the metric is termed **designability**. We discuss the implications of ProteinMPNN redesign in Section G, but our evaluation scope is the self-consistency component of co-designability in sequence-structure co-design models.

Distributional metrics for generated protein sets. Evaluation of protein generative models has largely focused on distributional metrics (Geffner et al., 2025b; Faltings et al., 2026; Lu et al., 2025), which compare generated set to reference set of native proteins. These metrics are motivated by the observation that even highly (co-)designable models can exhibit limited fold-level or secondary-structure diversity (Lin et al., 2024), such as overproducing helical bundles. Distributional metrics are important but orthogonal to our work. Because distributional metrics are often computed after self-consistency filtering, biases in the per-sample filter will propagate to distributional metrics (Sec. 5.4).

Self-consistency failures in PDB and AFDB. Prior work has noted that self-consistency is imperfect and that optimizing for it can distort generated protein distributions, especially secondary structure and tertiary motifs composition (Lu et al., 2025). We show that the problem is more fundamental;

Two sets of proteins can closely match the distribution of secondary structure composition while having completely different flexibility (Sec. 4.2). Distortion in secondary structure is one possible artifact of optimizing for self-consistency, but it does not explain the flexibility bias we focus on in this work. (Reidenbach et al., 2025) report that only 26% of the AlphaFold protein structure database (AFDB) is co-designable, and attributes this to a mismatch between native sequences and synthetic structures, motivating consistency-distilled datasets built through ProteinMPNN sequence design and refolding. We show self-consistency failures persist even for native structures paired with their native sequences. Thus, the bias is not a property of the synthetic dataset; it is built into the metric itself. Constructing metric-distilled datasets can only reinforce it.

3 Formalizing self-consistency

In this section, we first specify what an ideal *ensemble-level* model and *metric* would mean, and how this relates to the self-consistency test. Building on this perspective, we then formally characterize the self-consistency test's two failure modes.

Ideal ensemble-level criterion. For a sequence $\mathbf{x}$, let $\rho_{\mathbf{x}}$ denote the equilibrium density over structural coordinates $\mathbf{z}$ that the sequence can adopt. After marginalizing over unmodeled degrees of freedom, such as solvent, ligands, and binding partners, this structural ensemble can be written in Boltzmann form: $\rho_{\mathbf{x}}(\mathbf{z}) \propto \exp(-\mathbf{u}_{\mathbf{x}}(\mathbf{z}))$, where $\mathbf{u}_{\mathbf{x}}(\mathbf{z})$ is a (dimensionless) free energy.

This ensemble is the biophysical object that an ideal design model would desire to match. Thus, an ideal design model would not merely output a single structure for a given sequence $\mathbf{x}$, but instead output a distribution over structures, which we denote by $\hat{\rho}_{\mathbf{x}}$. The ideal ensemble-level criterion is that this designed ensemble be close to the true equilibrium ensemble. For example, using the forward KL divergence, we desire $\mathcal{D}_{\mathrm{KL}}(\hat{\rho}_{\mathbf{x}}, \rho_{\mathbf{x}}) < \varepsilon$, or equivalently,

$$\mathbb{E}_{z' \sim \hat{\rho}_{\mathbf{x}}}[-\log \rho_{\mathbf{x}}(\mathbf{z}')] = \mathcal{D}_{\mathrm{KL}}(\hat{\rho}_{\mathbf{x}}, \rho_{\mathbf{x}}) + H(\hat{\rho}_{\mathbf{x}}) < \varepsilon + H(\hat{\rho}_{\mathbf{x}}), \tag{1}$$

where $H(\hat{\rho}_{\mathbf{x}})$ is the entropy of the designed ensemble. Small $\mathcal{D}_{\mathrm{KL}}$ implies that structures sampled from the designed ensemble have high probability under the true sequence-conditioned equilibrium ensemble.

Single-structure generations. For models that generate a single structure, we can view the generated ensemble $\hat{\rho}_{\mathbf{x}}$ as *coarse-grained* point mass distribution at $\mathbf{z}$. Under this limit, equation 1 reduces to

$$-\log \rho_{\mathbf{x}}(\mathbf{z}) = \mathbb{E}_{z' \sim \hat{\rho}_{\mathbf{x}}}[-\log \rho_{\mathbf{x}}(\mathbf{z}')] < \varepsilon + H(\hat{\rho}_{\mathbf{x}}).$$

Therefore, after absorbing the constant and the tolerance, an ideal plausibility criterion for a single generated structure becomes $-\log \rho_{\mathbf{x}}(\mathbf{z}) < \gamma$ for some $\gamma > 0$.

Canonical self-consistency as an approximation. In practice, as $\rho_{\mathbf{x}}$ is unknown, the self-consistency test replaces this equilibrium ensemble with a deterministic folding-model prediction, such as ESMFold, namely $\hat{f}_\theta(\cdot)$. The standard self-consistency test accepts a generated pair $(\mathbf{x}, \mathbf{z})$ if $\mathbf{d}(\mathbf{z}, \hat{f}_\theta(\mathbf{x})) < \tau$, where $\mathbf{d}$ denotes root mean square deviation (RMSD). In other words, self-consistency uses a geometric comparison as a proxy for the ideal probability-based criterion:

$$-\log \rho_{\mathbf{x}}(\mathbf{z}) < \gamma \quad \Rightarrow \quad \mathbf{d}(\mathbf{z}, \hat{f}_\theta(\mathbf{x})) < \tau, \quad \tau = 2\,\text{Å}.$$

Two failure modes. This approximation suffers from an apparent *false rejection*: a given pair $(\mathbf{x}, \mathbf{z})$ is acceptable under the point-to-ensemble criterion (blue), but rejected by empirical self-consistency (green) i.e, $-\log \rho_{\mathbf{x}}(\mathbf{z}) < \gamma$ but $\mathbf{d}(\mathbf{z}, \hat{f}_\theta(\mathbf{x})) \geq \tau$. We identify the source of such *false rejections* through two scenarios.

- **Flexibility failure**: Both the evaluated structure and the folder prediction are plausible conformations, yet they are separated by more than the RMSD threshold:

$$-\log \rho_{\mathbf{x}}(\mathbf{z}) < \gamma, \ -\log \rho_{\mathbf{x}}(\hat{f}_\theta(\mathbf{x})) < \gamma, \quad \mathbf{d}(\mathbf{z}, \hat{f}_\theta(\mathbf{x})) \geq \tau.$$

  The rejection is caused by the collapse of a broad or multimodal ensemble to a single structural representative.

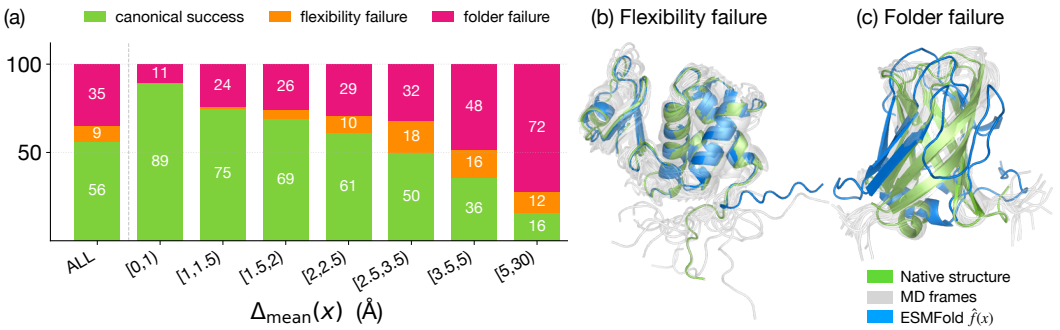

Figure 1: **(a) Flexibility-binned native protein recall** of ESMFold on ATLAS database. **(b) Example of flexibility failure**, where ESMFold prediction is close to a plausible conformation produced by MD but is far from the native structure. **(c) Example of folder failure**, where the MD ensemble is tight but ESMFold prediction (optimally aligned to the native structure) is incorrect.

- **Folder failure**: The evaluated structure is plausible under the sequence-conditioned ensemble, but the deterministic folder prediction is not:

$$-\log \rho_{\mathbf{x}}(\mathbf{z}) < \gamma, \ -\log \rho_{\mathbf{x}}(\hat{f}_\theta(\mathbf{x})) \geq \gamma, \quad \mathbf{d}(\mathbf{z}, \hat{f}_\theta(\mathbf{x})) \geq \tau.$$

  In this case, the refolded structure lies in a low-probability region of the sequence-conditioned ensemble. The rejection, therefore, reflects an error in the folding surrogate.

4 Two Failure Modes of Canonical Self-Consistency

In Sec. 3, we defined two failure modes of canonical self-consistency in terms of the unobserved sequence-conditioned equilibrium ensemble $\rho_{\mathbf{x}}$. In this section, we examine how these failures appear in native proteins. Native proteins provide a positive-control setting, for which the deposited experimental structure should be accepted by a plausibility metric. Therefore, a rejection of this pair is a false negative of the evaluation pipeline.

Our goal in this section is to turn the formal definitions from Sec. 3 into empirical labels. We use short-time molecular dynamics (MD) frames as a finite empirical approximation to the sufficiently low-energy basin of each sequence, allowing us to distinguish *flexibility failures* from *folder failures*.

4.1 Empirically Labeling Failure Modes in ATLAS Database

We use the ATLAS database (Vander Meersche et al., 2024), a large collection of all-atom MD trajectories initialized from experimentally characterized protein structures. Each ATLAS protein is simulated for three independent 100ns replicates. For each native sequence $\mathbf{x}$, let $\mathcal{Z}_{\mathbf{x}} = \{\mathbf{z}_{\mathbf{x}}^j\}$ denote the set of MD frames, and let $\mathbf{z}_{\mathbf{x}}^0$ denote the deposited experimental structure used to initialize the simulations, that is, $-\log \rho_{\mathbf{x}}(\mathbf{z}_{\mathbf{x}}^0) < \gamma$ presumably holds.

At the *dataset* level, ATLAS is *not* enriched for flexible proteins. Its filtering is based on structure quality, completeness, and fold-level diversity, yielding a broad and relatively non-redundant sample of characterized protein space. Therefore, we treat ATLAS as a representative set of the structurally characterized protein space.

At the *protein* level, the finite MD frames can serve as an *empirical proxy* of the high-probability region of the ensemble, which we instantiate as follows.

$$\underline{-\log \rho_{\mathbf{x}}(\hat{f}_\theta(\mathbf{x})) < \gamma} \quad \Rightarrow \quad \min_{\mathbf{z} \in \mathcal{Z}_{\mathbf{x}}} \mathbf{d}(\mathbf{z}, \hat{f}_\theta(\mathbf{x})) < \tau_{\text{MD}}.$$

This empirical approximation is plausible: ATLAS trajectories are short, intended to estimate the flexibility profile of proteins in solution, not to exhaustively sample rare events, unfolding, or large conformational rearrangements. Thus, ATLAS provides a useful basis to study plausible conformations. (We revisit this in Section 5.2.)

Table 1: Distributional distances and structural diversity for ATLAS subsets median-split along each axis. Rg denotes radius of gyration.

Axis	FPD	MMD^2	Diversity (%)
α-helix %	115.56	0.3206	72.6% / 77.8%
β-strand %	105.93	0.2984	75.2% / 78.4%
coil %	64.94	0.1620	74.5% / 81.3%
length	79.93	0.1916	73.0% / 71.1%
Rg	25.19	0.1074	72.1% / 72.7%
Δ_{mean}	10.30	0.0047	79.4% / 79.4%
random	1.79	0.00	86.4% / 85.2%

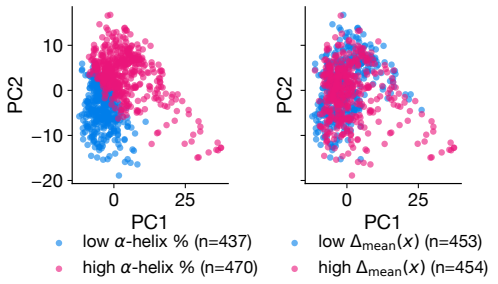

Figure 2: PCA of ESM3 embeddings for median splits by helix content and by Δ_{mean}.

In this regard, two failure modes reduce to:

$$\begin{aligned}
&\textbf{Canonical success:} && \mathbf{d}(\mathbf{z}_{\mathbf{x}}^0, \hat{f}_\theta(\mathbf{x})) < \tau, \\
&\textbf{Flexibility failure:} && \mathbf{d}(\mathbf{z}_{\mathbf{x}}^0, \hat{f}_\theta(\mathbf{x})) \geq \tau \quad \text{and} \quad \min_{\mathbf{z}\in\mathcal{Z}_{\mathbf{x}}} \mathbf{d}(\mathbf{z}, \hat{f}_\theta(\mathbf{x})) < \tau_{\text{MD}} \\
&\textbf{Folder failure:} && \mathbf{d}(\mathbf{z}_{\mathbf{x}}^0, \hat{f}_\theta(\mathbf{x})) \geq \tau \quad \text{and} \quad \min_{\mathbf{z}\in\mathcal{Z}_{\mathbf{x}}} \mathbf{d}(\mathbf{z}, \hat{f}_\theta(\mathbf{x})) \geq \tau_{\text{MD}}.
\end{aligned}$$

where $\mathbf{d}$ denotes RMSD and $\tau_{\text{MD}} = \tau = 2$Å.

We quantify the flexibility of each protein by the mean RMSD from the initial structure $\mathbf{z}_{\mathbf{x}}^0$ to the MD frames, denoted $\Delta_{\text{mean}}(\mathbf{x})$. This quantity often exceeds 2Å (App. B), already showing the caveat of using a fixed distance cutoff regardless of the varying flexibility of proteins. We then bin ATLAS proteins by $\Delta_{\text{mean}}(\mathbf{x})$ and report the fraction of each success/failure class within each bin. As shown in Fig. 1a, canonical self-consistency success decreases sharply with increasing flexibility. Flexibility failures begin to appear once $\Delta_{\text{mean}}(\mathbf{x}) > 1.5$Å and increase with protein flexibility, as expected. More importantly, folder failures also increase with flexibility, indicating that flexible proteins are harder for the folding surrogate to predict.

We refer to this binned decomposition as *flexibility-binned native protein recall*, and use it as the diagnostic throughout the rest of the paper.

4.2 Distributional metrics can miss flexibility collapse

The failure-mode analysis above shows that self-consistency penalizes proteins along the flexibility axis. A generative model optimized for self-consistency can therefore improve its score by avoiding flexible proteins. A natural question is whether existing distributional metrics would detect this collapse.

We partition the self-consistent ATLAS subset into two equally sized groups using the median Δ_{mean}, and then compute distributional metrics between the two groups to test whether these metrics detect a large difference in flexibility. Because the scale of distributional metrics is often difficult to interpret, we contextualize this split by comparing it to analogous median splits along other axes: secondary structure composition (α-helix, β-strand, and coil fraction), sequence length, radius of gyration, and a random control split. For each split, we compare the resulting protein sets using standard distributional statistics from the protein design literature, including structural diversity and Fréchet Protein Distance (FPD) (Lu et al., 2025).

Structural diversity is computed as the percentage of distinct structural clusters under a TM-score cutoff of 0.5, using FoldSeek (van Kempen et al., 2022); we report the diversity of each side of the split. FPD measures the distance between two protein sets as the Fréchet distance between Gaussian fits in ESM3 (Hayes et al., 2025) embedding space, following Lu et al. (2025). Because our sample size is modest and Fréchet distance can be unstable in this regime, we also report maximum mean discrepancy (MMD) (Gretton et al., 2012), a kernel two-sample statistic that compares distributions without assuming Gaussianity.

Table 1 shows that the split by Δ_{mean} produces much smaller distances than the split by other properties. Diversity is also always similar in two sets; Protein sets can exhibit fold-level diversity while suppressing one of the property axes. Figure 2 is consistent with this result: ESM3 embeddings

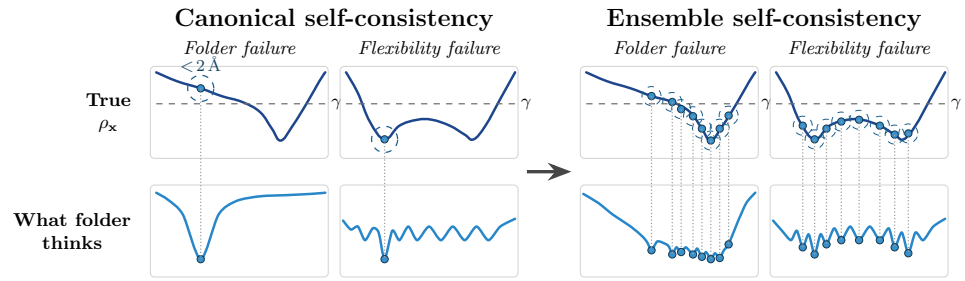

Figure 3: Canonical vs. ensemble self-consistency under the two failure modes. Above denotes the true sequence-conditioned ensemble $\rho_{\mathbf{x}}$; circles of radius ε around each folder prediction show the 2 Å acceptance neighborhood. **Canonical (left):** a single deterministic prediction either lands above the energy cutoff γ (folder failure) or covers only part of a broad basin below γ (flexibility failure). **Ensemble (right):** the folder produces a distribution; an acceptance neighborhood is taken around each sample, providing coverage of multiple modes and rescuing both failure types.

separate proteins split by helix content much more clearly than proteins split by flexibility. Protein generative models already exhibit high FPD to PDB, often in the range of 20-200 (Lu et al., 2025), largely due to distortions in secondary structure composition. In the presence of such dominant distortions, flexibility collapse may be difficult to detect with current distributional metrics. We further compare secondary structure composition and radius of gyration between the low- and high-flexibility sets in Fig. 8. The two sets are nearly indistinguishable under helix fraction, strand fraction, and coil fraction, and the radius of gyration differs only marginally.

If a generator systematically loses flexible proteins by optimizing for self-consistency, standard distributional metrics may not flag the failure. These results motivate the need for a better self-consistency criterion that does not penalize flexibility, as well as a method for estimating the flexibility of generated proteins.

5 Ensemble self-consistency

The flexibility failure analysis in Sec. 4 suggests that the deterministic structure estimate of canonical self-consistency is restrictive by construction: a single point cannot represent a basin whose plausible conformations span more than $\tau = 2$ Å. In light of this, we consider an ensemble-aware version of self-consistency (Fig. 3). Concretely, an ensemble conformation predictor θ produces a distribution $\mathbf{z} \sim \hat{\rho}_{\mathbf{x},\theta}$. Given N predicted samples $\hat{\mathbf{z}}_1 \ldots, \hat{\mathbf{z}}_N \sim \hat{\rho}_{\mathbf{x},\theta}$, we define ensemble self-consistency as $\min_i \mathbf{d}(\mathbf{z}, \hat{\mathbf{z}}_i) < \tau$, i.e., ensemble self-consistency examines whether $\mathbf{z}$ is close to at least one sample from the predicted conformation ensemble.

In the limit as $N \to \infty$, ensemble self-consistency *mitigates* the flexibility failure mode. Intuitively, if $\hat{\rho}_{\mathbf{x},\theta}$ assigns sufficient probability mass to regions that are highly probable under $\rho_{\mathbf{x}}$, then eventually at least one sampled structure will fall in such a region, regardless of how broad or multi-modal the plausible set is.

At finite N, ensemble self-consistency can still fail even when the predicted samples are individually plausible: the predicted ensemble may sample from the correct conformation, yet the finite sample may not include a conformation close to the particular evaluated structure. Therefore, we extend the diagnosis from Sec. 4 to distinguish *finite-sampling* flexibility failures and folder failures.

Recasting failure mode definitions. We recall the notation from Sec. 4, including $\mathcal{Z}_{\mathbf{x}} = \{\mathbf{z}_{\mathbf{x}}^j\}$ and $\mathbf{z}_{\mathbf{x}}^0$. For an ensemble predictor with samples $\hat{\mathbf{z}}_1, \ldots, \hat{\mathbf{z}_N}$, we recast two failure mode as:

$$\begin{aligned}
&\textbf{Ensemble success:} && \min_i \mathbf{d}(\mathbf{z}_{\mathbf{x}}^0, \hat{\mathbf{z}}_i) < \tau \\
&\textbf{Finite-sample flexibility failure:} && \min_i \mathbf{d}(\mathbf{z}_{\mathbf{x}}^0, \hat{\mathbf{z}}_i) \geq \tau \quad \text{and} \quad \forall i,\ \min_{\mathbf{z}' \in \mathcal{Z}_{\mathbf{x}}} \mathbf{d}(\mathbf{z}', \hat{\mathbf{z}}_i) < \tau_{\text{MD}}, \\
&\textbf{Folder failure:} && \min_i \mathbf{d}(\mathbf{z}_{\mathbf{x}}^0, \hat{\mathbf{z}}_i) \geq \tau \quad \text{and} \quad \exists i,\ \min_{\mathbf{z}' \in \mathcal{Z}_{\mathbf{x}}} \mathbf{d}(\mathbf{z}', \hat{\mathbf{z}}_i) \geq \tau_{\text{MD}}.
\end{aligned}$$

The above definition follows our intuition. In finite-sample flexibility failure, the predictor is sampling within the MD-sampled plausible basin, but the finite sample set does not cover the particular reference

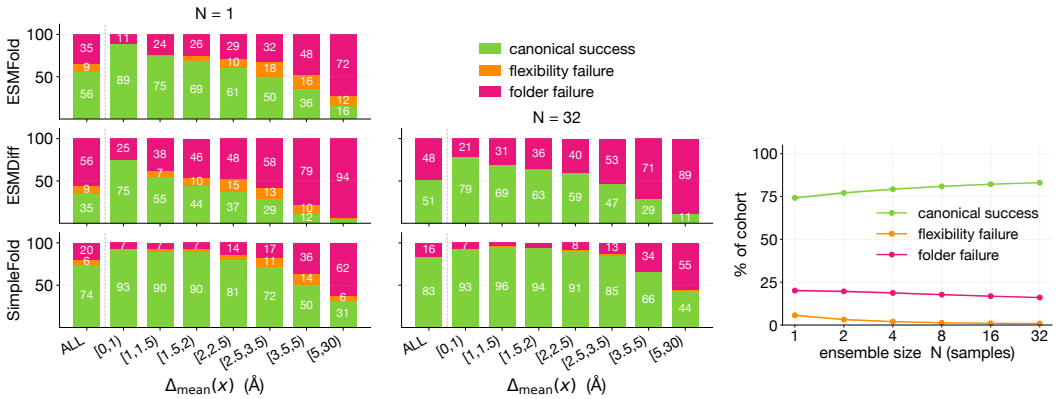

Figure 4: **(a)** Failure-mode decomposition for canonical ESMFold and ensemble self-consistency using ESMDiff and SimpleFold. **(b)** Effect of SimpleFold ensemble size on canonical success and failure modes in the full ATLAS set.

conformation. In folder failure, the predicted distribution assigns mass to conformations that are not supported by the MD ensemble. Note that this decomposition falls into that in Sec. 4.1 when $N = 1$.

5.1 Empirical gains in flexibility-binned native protein recall

In this section, using our definition of ensemble self-consistency, we examine where current MSA-free ensemble models stand in native protein recall. We evaluate two representative MSA-free conformation ensemble predictors, ESMDiff (Lu et al., 2024) and SimpleFold (Wang et al., 2025). For each ATLAS sequence $\mathbf{x}$, we draw $N = 32$ samples from each ensemble predictor and evaluate ensemble self-consistency against the deposited native structure $\mathbf{z}^0_{\mathbf{x}}$. We then apply the failure mode decomposition defined above within flexibility bins of $\Delta_{\text{mean}}(\mathbf{x})$ used in Sec. 4.

As shown in Figure 4a, increasing the ensemble size to $N = 32$ with either ESMDiff or SimpleFold nearly eliminates the flexibility failures observed under the $N = 1$ criterion. This is the failure mode ensemble self-consistency is designed to address, and the decomposition makes clear that an ensemble is what resolves them. In Fig. 4b, SimpleFold's flexibility failures are nearly eliminated by $N = 16$, implying that a moderate number of samples is usually sufficient to cover the plausible basin of a protein.

A substantial fraction of targets still exhibit folder failure. Folder failure also decreases with N, since more samples increase the chance that at least one prediction lands near the native structure (see Fig. 3 for illustration). However, this only helps when the folding model assigns non-negligible probability mass to the correct plausible basin. When the model's predicted landscape is biased, additional samples provide limited benefit, so folder failure decays slowly with N.

SimpleFold is the strongest **current** choice, outperforming both ESMFold and ESMDiff even at $N = 1$. Increasing the SimpleFold ensemble to $N = 16$ further raises overall recall from 74% to 82%, and recall in the most flexible bin from 50% to 65%. We therefore use SimpleFold $N = 16$ for the remaining analyses, while emphasizing that the choice should be revisited as ensemble predictors improve–*flexibility-binned native protein recall* provides the diagnostic for that comparison.

5.2 Robustness of the measure

In this section, we examine the robustness and generalizability of our failure mode analysis and ensemble self-consistency.

Robustness of labels to the simulation time. To assess the sensitivity of our failure labels to the simulation time, we repeat our analysis on MEGASim (Lewis et al., 2025), comparing the first 1 μs with the full 12–22 μs trajectories. Across SimpleFold ensemble sizes $N \in \{1, 4, 8, 16, 32, 64\}$, extending the trajectories reduces the folder failure fraction by only 1.4–2.7 percentage points, with

Table 2: Self-consistency and structural diversity of co-design models under canonical (N=1) versus ensemble (N=16) using SimpleFold. **SC-N**: percentage of generations accepted by SF-N. **Div-N**: structural diversity. LP=La-Proteina.

Model	SC-1	SC-16	Div-1	Div-16
DISCO	41.4%	66.4%	13.8%	18.0%
LP ($\eta = 0.1$)	30.6%	47.8%	17.2%	27.6%
LP ($\eta = 0.2$)	26.8%	45.0%	8.0%	10.4%
LP ($\eta = 0.3$)	22.4%	35.6%	8.4%	9.8%
LP ($\eta = 0.4$)	15.6%	25.4%	7.0%	8.8%
RFdiffusion3	14.6%	23.4%	5.4%	6.2%

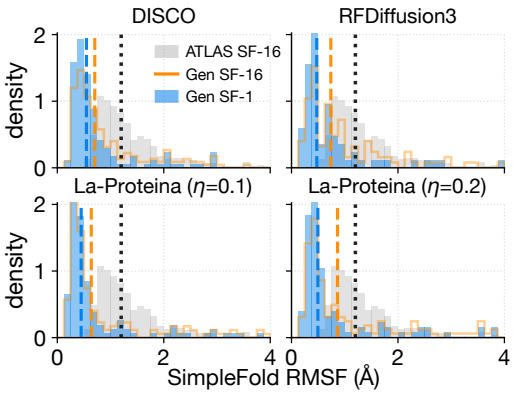

Figure 5: Distribution of SimpleFold RMSF of ATLAS and generated (Gen) self-consistent subset under SF-16 and -1. Dotted lines denote the median.

corresponding increases in flexibility failures. This supports the stability of our analysis. We provide the detailed results and setup in Appendix I.

Validation on de novo proteins. We next test whether the recall gains of ensemble self-consistency extend to experimentally characterized de novo proteins from the Protein Design Archive (Chronowska et al., 2025). Increasing the SimpleFold ensemble size from $N = 1$ to $N = 32$ improves recall from 80.7% to 90.3% on the X-ray subset (155 proteins), and from 69.9% to 90.9% on the NMR subset (44 proteins). These results show that the gains extend beyond native proteins to experimentally characterized de novo designs.

Sensitivity to τ. Finally, we repeat the ATLAS failure-mode analysis and co-design evaluation across RMSD thresholds $\tau \in \{1.5, 2.0, 2.5, 3.0\}$ Å and find that our main conclusions remain largely unchanged. We provide the full results in Appendix H.

5.3 Predicted ensembles as a flexibility proxy

The ideal distributional metric for detecting flexibility collapse would compare the flexibility distribution of generated proteins to that of native proteins. Since even short-time MD simulation is costly, we ask whether SimpleFold ensembles can predict flexibility computed via MD. For each protein, we compute the mean per-residue root mean square fluctuation (RMSF), which is the standard deviation of C_α position around its mean across an ensemble, from the $N = 16$ SimpleFold samples, and compare it to the RMSF computed over the ATLAS trajectory. We find a moderate correlation (Pearson $r = 0.431$, Spearman $\rho = 0.550$, though Fig. 9 shows the absolute scale is compressed), letting us use SimpleFold-derived RMSF as a proxy for detecting protein flexibility collapse in the analyses that follow.

5.4 Re-evaluating co-design models using ensemble self-consistency

We re-evaluate three co-design models: DISCO (Rector-Brooks et al., 2026), La-Proteina (Geffner et al., 2025a) at four noise scales $\eta \in \{0.1, 0.2, 0.3, 0.4\}$, and RFdiffusion3 (Watson et al., 2023) under ensemble self-consistency with SimpleFold. For each model, we generate 100 samples at each of five lengths $\{100, 200, 300, 400, 500\}$, for a total of 500 generations per model, and report percentage of self-consistent samples and structural diversity under two settings: **SF-1**, canonical self-consistency using a single SimpleFold prediction, and **SF-16**, ensemble self-consistency over $N = 16$ SimpleFold samples. Table 2 shows all six model variants gain both self-consistency and structural diversity under SF-16 relative to SF-1.

Self-consistent generations under SF-16 span a broader RMSF range than those under SF-1, and their median RMSF is closer to that of the self-consistent ATLAS subset (Fig. 5). However, a gap remains even under SF-16. This is partly because both DISCO and La-Proteina were evaluated using sampling settings chosen to maximize co-designability, which already enriches the sampled distribution for

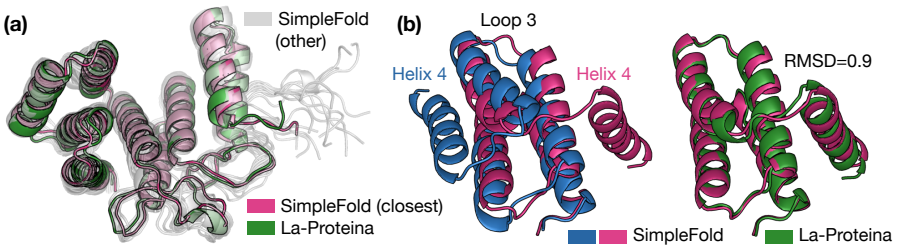

Figure 6: **Example La-Proteina generations rescued by ensemble self-consistency. (a)** Structural ensemble of 16 SimpleFold refold predictions (gray) for a La-Proteina–generated structure (green), highlighting variability in the C-terminal loop. The closest SimpleFold member (pink) aligns with the La-Proteina at RMSD = 1.7Å. **(b)** Two SimpleFold refold predictions, one from each of two clusters, distinguished by a hinge motion in loop 3 that reorients helix 4. The smaller cluster representative (pink) matches the La-Proteina conformation (green) at RMSD = 0.9Å.

less flexible proteins. Raising La-Proteina's noise scale does increase median RMSF, but at a steep cost to self-consistency and structural diversity (Table 2). The same limitation we observed for native proteins appears here: flexible proteins are harder to predict, and harder to generate.

Figure 6a shows SimpleFold predicting diverse C-terminal loop conformations for a La-Proteina-generated sequence, with one refold matching the generated loop at RMSD = 1.7Å. In Figure 6b, the 16 SimpleFold refold predictions form two clusters (13 and 3 members), separated by a hinge motion in loop 3 that ultimately reorients helix 4; one of these clusters closely matches the La-Proteina generation (RMSD=0.9Å). This example illustrates that modeling conformational heterogeneity can have effects beyond local fluctuations: flexibility in loops connecting secondary structure elements can drive larger-scale motions that easily exceed a global 2Å RMSD cutoff.

5.5 Ensemble self-consistency preserves high specificity

A natural concern with ensemble self-consistency is that taking a minimum of N predictions could make the criterion too permissive, leading to false positives. We test this using hard negative samples derived from the Rosetta decoy set (Park et al., 2016). These decoys are non-native or near-native alternative conformations generated for proteins with known native structures via local fragment perturbations. Thus, they are much more challenging negatives than randomly mismatched sequence-structure pairs. To avoid labeling plausible alternative conformations as negatives, we intersect this set with ATLAS MD trajectories and retain only decoys that are at least 2Å away from every MD frame of the corresponding protein. This yields 935 decoys across 10 targets (App. F).

On this specificity test, ensemble self-consistency produces only a single additional false positive out of 935: SimpleFold-32 accepts 2 decoys, SimpleFold-1 accepts 1. The min-of-32 RMSD is also only 0.1-0.3Å lower than the single sample RMSD (Fig. 10), suggesting that the SimpleFold ensembles remain tight for these rigid targets. These results show that ensemble self-consistency retains high specificity on this Rosetta decoy set.

6 Conclusion

In this work, we disentangle two primary failure modes of the self-consistency pipeline and provide both conceptual and empirical analyses to quantify their respective impacts. Building on this understanding, we extend the standard self-consistency framework to explicitly account for conformational heterogeneity through ensemble predictions. This approach improves recall on native proteins and experimentally characterized de novo designs while maintaining high specificity on the evaluated Rosetta decoy set. Ensemble self-consistency may also be useful for downstream tasks such as motif scaffolding and binder design, where conformational flexibility can be functionally important.

AI USE STATEMENT

We used generative AI tools to assist with language editing and rephrasing. We did not use generative AI tools to generate experimental data, develop methods, interpret experimental results, or formulate the scientific conclusions of this work. All AI-assisted text and presentation suggestions were reviewed and revised by the authors, and all references identified with AI assistance were independently verified against the original sources. The authors take full responsibility for the final content of this work.

REFERENCES

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, 2024.

Paul D Adams, Pavel V Afonine, Gábor Bunkóczi, Vincent B Chen, Ian W Davis, Nathaniel Echols, Jeffrey J Headd, L-W Hung, Gary J Kapral, Ralf W Grosse-Kunstleve, et al. Phenix: a comprehensive python-based system for macromolecular structure solution. *Biological crystallography*, 66 (2):213–221, 2010.

Vincent B Chen, W Bryan Arendall, Jeffrey J Headd, Daniel A Keedy, Robert M Immormino, Gary J Kapral, Laura W Murray, Jane S Richardson, and David C Richardson. Molprobity: all-atom structure validation for macromolecular crystallography. *Biological crystallography*, 66(1):12–21, 2010.

Marta Chronowska, Michael J Stam, Derek N Woolfson, Luigi F Di Costanzo, and Christopher W Wood. The protein design archive (pda): insights from 40 years of protein design. *nature biotechnology*, 43(5):669–671, 2025.

Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J Ragotte, Lukas F Milles, Basile IM Wicky, Alexis Courbet, Rob J de Haas, Neville Bethel, et al. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378(6615):49–56, 2022.

Kieran Didi, Danny Reidenbach, Matthew Penner, Supriya Ravi, Marshall Case, Mike Nichols, Erik Swanson, Alex Reis, Maggie Prescott, Yue Qian, Dongming Qian, Jingjing Yang, Weiji Li, Le Li, Daichi Shonai, Sean Gay, Bhoomika Basu Mallik, Ho Yeung Chim, Liurong Chen, Miguel Atienza Juantay, Hubert Klein, Anna Macintyre, Maxim Secor, Daniele Granata, Zhonglin Cao, Guoqing Zhou, Tomas Geffner, Xi Chen, Micha Livne, Zuobai Zhang, Tianjing Zhang, Michael M. Bronstein, Martin Steinegger, Kristine Deibler, Scott Soderling, Alena Khmelinskaia, Florian Hollfelder, Christian Dallago, Emine Kucukbenli, Arash Vahdat, Pierce Ogden, and Karsten Kreis. Latent generative search unlocks de novo design of untapped biomolecular interactions at scale. `https://research.nvidia.com/labs/genair/proteina-complexa/assets/proteina_complexa_validation.pdf`, 2026a.

Kieran Didi, Zuobai Zhang, Guoqing Zhou, Danny Reidenbach, Zhonglin Cao, Sooyoung Cha, Tomas Geffner, Christian Dallago, Jian Tang, Michael M. Bronstein, Martin Steinegger, Emine Kucukbenli, Arash Vahdat, and Karsten Kreis. Scaling atomistic protein binder design with generative pretraining and test-time compute. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026b.

Felix Faltings, Hannes Stark, Tommi Jaakkola, and Regina Barzilay. Protein fid: Improved evaluation of protein structure generative models. *Bioinformatics*, 42(4), 2026.

Tomas Geffner, Kieran Didi, Zhonglin Cao, Danny Reidenbach, Zuobai Zhang, Christian Dallago, Emine Kucukbenli, Karsten Kreis, and Arash Vahdat. La-proteina: Atomistic protein generation via partially latent flow matching. *arXiv preprint arXiv:2507.09466*, 2025a.

Tomas Geffner, Kieran Didi, Zuobai Zhang, Danny Reidenbach, Zhonglin Cao, Jason Yim, Mario Geiger, Christian Dallago, Emine Kucukbenli, Arash Vahdat, et al. Proteina: Scaling flow-based protein structure generative models. *arXiv preprint arXiv:2503.00710*, 2025b.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The journal of machine learning research*, 13(1):723–773, 2012.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, et al. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, 2025.

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.

Sarah Lewis, Tim Hempel, José Jiménez-Luna, Michael Gastegger, Yu Xie, Andrew YK Foong, Victor García Satorras, Osama Abdin, Bastiaan S Veeling, Iryna Zaporozhets, et al. Scalable emulation of protein equilibrium ensembles with generative deep learning. *Science*, 389(6761): eadv9817, 2025.

Yeqing Lin, Minji Lee, Zhao Zhang, and Mohammed AlQuraishi. Out of many, one: Designing and scaffolding proteins at the scale of the structural universe with genie 2. *arXiv preprint arXiv:2405.15489*, 2024.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

Jiarui Lu, Xiaoyin Chen, Stephen Zhewen Lu, Chence Shi, Hongyu Guo, Yoshua Bengio, and Jian Tang. Structure language models for protein conformation generation. *arXiv preprint arXiv:2410.18403*, 2024.

Tianyu Lu, Melissa Liu, Yilin Chen, Jinho Kim, and Po-Ssu Huang. Assessing generative model coverage of protein structures with shapes. *Cell Systems*, 16(8), 2025.

Hahnbeom Park, Philip Bradley, Per Greisen Jr, Yuan Liu, Vikram Khipple Mulligan, David E Kim, David Baker, and Frank DiMaio. Simultaneous optimization of biomolecular energy functions on features from small molecules and macromolecules. *Journal of chemical theory and computation*, 12(12):6201–6212, 2016.

Jarrid Rector-Brooks, Théophile Lambert, Marta Skreta, Daniel Roth, Yueming Long, Zi-Qi Li, Xi Zhang, Miruna Cretu, Francesca-Zhoufan Li, Tanvi Ganapathy, Emily Jin, Avishek Joey Bose, Jason Yang, Kirill Neklyudov, Yoshua Bengio, Alexander Tong, Frances H. Arnold, and Cheng-Hao Liu. General multimodal protein design enables dna-encoding of chemistry. 2026. URL https://arxiv.org/abs/2604.05181.

Danny Reidenbach, Zhonglin Cao, Zuobai Zhang, Kieran Didi, Tomas Geffner, Guoqing Zhou, Jian Tang, Christian Dallago, Arash Vahdat, Emine Kucukbenli, et al. Consistent synthetic sequences unlock structural diversity in fully atomistic de novo protein design. *arXiv preprint arXiv:2512.01976*, 2025.

Hannes Stark, Felix Faltings, MinGyu Choi, Yuxin Xie, Eunsu Hur, Timothy O'Donnell, Anton Bushuiev, Talip Uçar, Saro Passaro, Weian Mao, et al. Boltzgen: Toward universal binder design. *bioRxiv*, pp. 2025–11, 2025.

Chai Discovery team, Jacques Boitreaud, Jack Dent, Matthew McPartlon, Joshua Meier, Vinicius Reis, Alex Rogozhonikov, and Kevin Wu. Chai-1: Decoding the molecular interactions of life. *BioRxiv*, pp. 2024–10, 2024.

Michel van Kempen, Stephanie S Kim, Charlotte Tumescheit, Milot Mirdita, Cameron LM Gilchrist, Johannes Söding, and Martin Steinegger. Foldseek: fast and accurate protein structure search. *Biorxiv*, pp. 2022–02, 2022.

Yann Vander Meersche, Gabriel Cretin, Aria Gheeraert, Jean-Christophe Gelly, and Tatiana Galochkina. Atlas: protein flexibility description from atomistic molecular dynamics simulations. *Nucleic acids research*, 52(D1):D384–D392, 2024.

Yuyang Wang, Jiarui Lu, Navdeep Jaitly, Josh Susskind, and Miguel Angel Bautista. Simplefold: Folding proteins is simpler than you think. *arXiv preprint arXiv:2509.18480*, 2025.

Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.

Jason Yim, Andrew Campbell, Andrew YK Foong, Michael Gastegger, José Jiménez-Luna, Sarah Lewis, Victor Garcia Satorras, Bastiaan S Veeling, Regina Barzilay, Tommi Jaakkola, et al. Fast protein backbone generation with se (3) flow matching. *arXiv preprint arXiv:2310.05297*, 2023.

Vinicius Zambaldi, David La, Alexander E Chu, Harshnira Patani, Amy E Danson, Tristan OC Kwan, Thomas Frerix, Rosalia G Schneider, David Saxton, Ashok Thillaisundaram, et al. De novo design of high-affinity protein binders with alphaproteo. *arXiv preprint arXiv:2409.08022*, 2024.

A SETTINGS

A.1 INFERENCE

We used the official repositories, checkpoints, and default inference hyperparameters for ESMFold, SimpleFold, ProteinMPNN, DISCO, La-Proteina, and RFDiffusion3 unless otherwise specified. For ProteinMPNN, we used the Cα-only model with the default noise scale of 0.1. For ESMFold, we used the 3B checkpoint. For SimpleFold, we used the 3B checkpoint and sampled with temperature of 0.8, following the setting used in the SimpleFold paper for multi conformation sampling. For DISCO, we used the full-effort setting, corresponding to 4 recycles and 200 diffusion steps. For La-Proteina, we used the unconditional checkpoint intended for proteins up to 500 residues and sampled with noise scales $\eta = \eta_x = \eta_z$, using the default value of 0.1 and additional values of 0.2, 0.3, and 0.4. All other sampling parameters were left at their defaults.

A.2 EVALUATION

We compute RMSD after Kabsch alignment using Cα atoms only. Our official codebase includes all scores computed on the ATLAS database and generated structures, together with scripts to reproduce every figure and table reported in this paper.

A.3 COMPUTE

All ProteinMPNN, ESMDiff, and ESMFold inference was performed on NVIDIA A100 GPUs (80GB), using under 300 GPU-hours in total. All SimpleFold inference and co-design model generation was performed on AMD MI300A GPUs with the cuBLAS backend, using approximately 1,200 GPU-hours: roughly 1,000 GPU-hours for SimpleFold ($N = 32$) predictions on $1{,}520 + 500 \times 6 = 4{,}520$ samples, and around 200 GPU-hours for co-design model sampling.

B STATISTICS OF ATLAS DATABASE

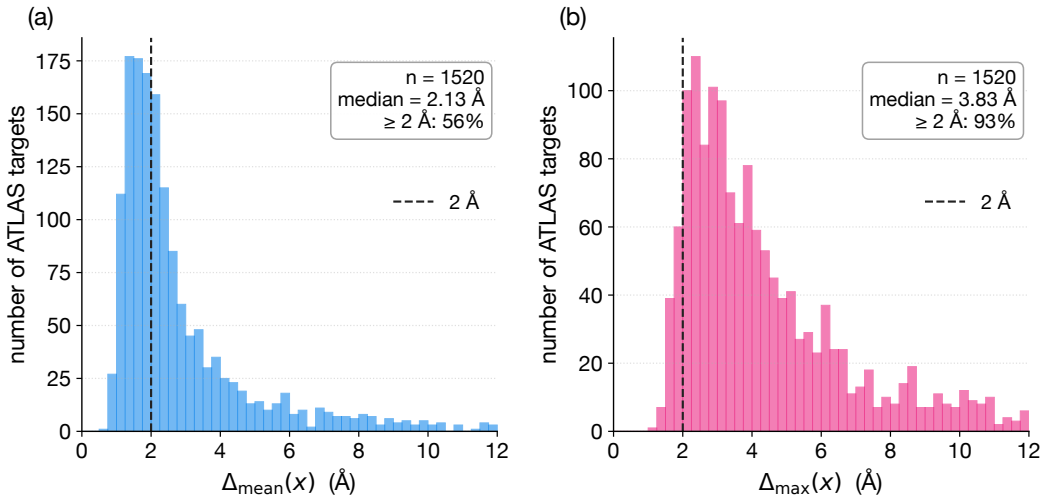

Figure 7: **Distribution of MD displacement in ATLAS. (a)** Mean RMSD to deposited structure across MD frames, Δ_{mean}. **(b)** Maximum RMSD to deposited structure across MD frames, Δ_{max}.

Figure 7 shows that ATLAS proteins often exhibit conformations beyond the fixed 2 Å threshold used in canonical self-consistency. The median Δ_{mean} is 2.13 Å, and 56% of proteins have mean MD displacement above 2 Å. The maximum displacement is even larger, with median $\Delta_{\text{max}} = 3.83$ Å and 93% of proteins exceeding 2 Å. This supports the use of proposed ensemble self-consistency: a fixed RMSD cutoff can reject native proteins simply because their plausible conformational ensemble is broad.

C PROPERTIES OF LOW- AND HIGH-Δ_{mean} SPLITS

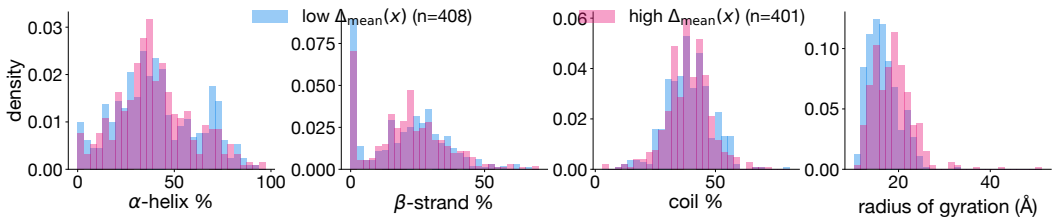

Figure 8: Distribution of secondary structure composition and radius of gyration for the low- and high-Δ_{mean} splits.

D SIMPLEFOLD FLEXIBILITY VS. MD FLEXIBILITY

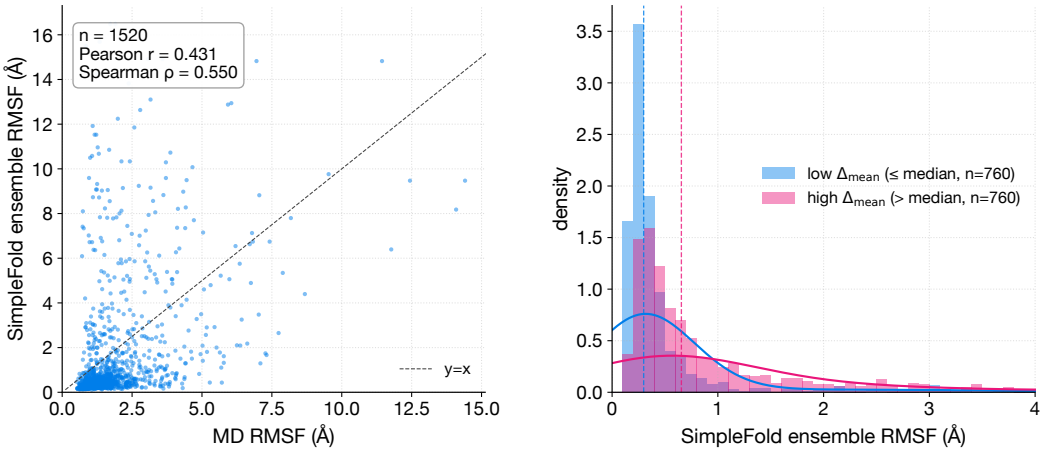

Figure 9: **Comparison between SimpleFold-predicted flexibility and ATLAS MD flexibility.** Per-residue Cα RMSF is computed from SimpleFold ensembles and from ATLAS MD trajectories. SimpleFold captures the relative flexibility of proteins with moderate correlation to MD, but underestimates the magnitude of fluctuations.

Figure 9 compares flexibility estimated from SimpleFold ensembles with flexibility measured from ATLAS MD. SimpleFold RMSF is moderately correlated with MD RMSF, suggesting that ensemble predictors can provide a useful proxy for flexibility. However, SimpleFold systematically underestimates the magnitude of MD fluctuations, so we use it as a diagnostic rather than as a calibrated replacement for MD.

E SELF-CONSISTENCY OF CO-DESIGN MODELS BY GENERATION LENGTH

Table 3–7 report the length-wise breakdown of the results in Table 2. LP denotes La-Proteina.

Table 3: Results for $L = 100$.

Model	SC-1	SC-16	Div-1	Div-16
DISCO	66.0%	83.0%	34.0%	37.0%
LP ($\eta = 0.1$)	71.0%	83.0%	40.0%	47.0%
LP ($\eta = 0.2$)	62.0%	78.0%	33.0%	41.0%
LP ($\eta = 0.3$)	60.0%	78.0%	38.0%	45.0%
LP ($\eta = 0.4$)	46.0%	65.0%	29.0%	38.0%
RFdiffusion3	44.0%	58.0%	16.0%	18.0%

Table 4: Results for $L = 200$.

Model	SC-1	SC-16	Div-1	Div-16
DISCO	56.0%	75.0%	16.0%	18.0%
LP ($\eta = 0.1$)	52.0%	76.0%	28.0%	46.0%
LP ($\eta = 0.2$)	40.0%	72.0%	15.0%	27.0%
LP ($\eta = 0.3$)	29.0%	47.0%	10.0%	15.0%
LP ($\eta = 0.4$)	23.0%	38.0%	9.0%	14.0%
RFdiffusion3	21.0%	42.0%	5.0%	6.0%

Table 5: Results for $L = 300$.

Model	SC-1	SC-16	Div-1	Div-16
DISCO	38.0%	67.0%	9.0%	14.0%
LP ($\eta = 0.1$)	22.0%	45.0%	12.0%	26.0%
LP ($\eta = 0.2$)	20.0%	45.0%	7.0%	13.0%
LP ($\eta = 0.3$)	13.0%	28.0%	6.0%	8.0%
LP ($\eta = 0.4$)	4.0%	15.0%	3.0%	6.0%
RFdiffusion3	5.0%	12.0%	4.0%	4.0%

Table 6: Results for $L = 400$.

Model	SC-1	SC-16	Div-1	Div-16
DISCO	31.0%	66.0%	9.0%	20.0%
LP ($\eta = 0.1$)	6.0%	26.0%	6.0%	15.0%
LP ($\eta = 0.2$)	7.0%	23.0%	5.0%	8.0%
LP ($\eta = 0.3$)	6.0%	14.0%	5.0%	8.0%
LP ($\eta = 0.4$)	3.0%	4.0%	2.0%	2.0%
RFdiffusion3	3.0%	4.0%	2.0%	2.0%

Table 7: Results for $L = 500$.

Model	SC-1	SC-16	Div-1	Div-16
DISCO	16.0%	41.0%	6.0%	7.0%
LP ($\eta = 0.1$)	2.0%	9.0%	1.0%	7.0%
LP ($\eta = 0.2$)	5.0%	7.0%	3.0%	3.0%
LP ($\eta = 0.3$)	4.0%	11.0%	3.0%	4.0%
LP ($\eta = 0.4$)	2.0%	5.0%	1.0%	3.0%
RFdiffusion3	0.0%	1.0%	–	1.0%

F ROSETTA DECOY SPECIFICITY TEST

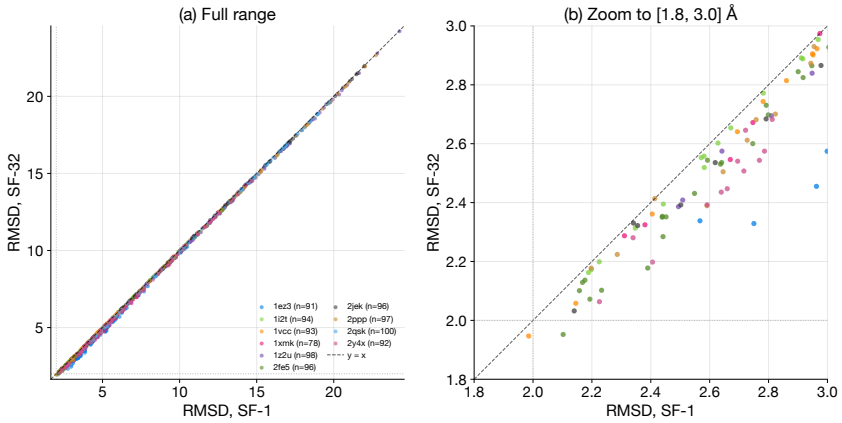

Figure 10: **Ensemble self-consistency preserves specificity on hard negative decoys.** Each point is a decoy structure paired with its native sequence, colored by target. SF-1 denotes the RMSD to the first SimpleFold prediction, corresponding to deterministic self-consistency; SF-32 denotes the minimum RMSD over 32 SimpleFold predictions, corresponding to ensemble self-consistency. The dotted lines mark the 2 Å acceptance threshold.

We identify 10 Rosetta decoy targets that overlap with ATLAS: `1ez3_B`,`1i2t_A`,`1vcc_A`,`1xmk_A`, `1z2u_A`,`2fe5_A`,`2jek_A`,`2ppp_A`, `2qsk_A`,`2y4x_A`. Each target contains 800-1,944 decoys, totaling 13,526 decoys.

Because the decoy set contains low-energy, plausible conformations, we apply an additional filtering procedure to construct negative set. For each decoy, we first compute its Cα RMSD to the experimental reference structure and discard decoys within 2Å, since these structures are effectively native-like. To obtain a balanced set across decoy difficulty, we bin the remaining decoys for each target into 10 quantiles by RMSD to the native structure and uniformly sample 10 decoys per bin, yielding approximately 100 decoys per target. For each sampled decoy, we then compare it against the ATLAS MD ensemble for the corresponding native sequence. We load the three ATLAS trajectories, restrict all structures to Cα atoms, concatenate the frames, and align the ATLAS Cα sequence to the decoy sequence by exact substring match. We retain a decoy for the specificity test only if it is more than 2Å from every ATLAS MD frame. This yields 935 decoys across 10 targets.

For each retained negative pair, consisting of a native sequence and a decoy structure, we evaluate SimpleFold self-consistency using 32 SimpleFold samples from the native sequence. We report two scores: SF-1, the RMSD from the decoy to the first SimpleFold sample, and SF-32, the minimum RMSD to any of the 32 samples. A decoy is accepted if its RMSD is below 2 Å.

This experiment therefore measures specificity: whether single sample or ensemble self consistency incorrectly accepts near-native but implausible decoy conformations. As shown in Fig. 10, SF-32 slightly reduces the RMSD for some decoys, as expected when taking the best of 32 samples. However, it rarely moves these hard negatives below the acceptance threshold: SF-32 introduces only one additional false positive out of 935 decoys, accepting 2 decoys compared to 1 accepted by SF-1. Thus, the ensemble self-consistency improves recall for flexible native conformations without sacrificing specificity on this decoy set.

G PROTEINMPNN REDESIGN IS NOT A REMEDY

The folder failure of Sec. 3 is sequence-specific: for the same plausible structure, some sequences are easy for the folder and some are not. We test this directly using ProteinMPNN redesign. For each ATLAS protein $\mathbf{x}$ with deposited structure $\mathbf{z}^0_{\mathbf{x}}$, we generate $N = 8$ MPNN-redesigned sequences $\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_8 \sim \text{ProteinMPNN}(\cdot \mid \mathbf{z}^0_{\mathbf{x}})$ and refold each with ESMFold.

Fig. 11a shows that MPNN redesigns yield only a marginal increase in canonical self-consistency overall, and the improvement is concentrated on rigid proteins; in the most flexible bin canonical success rate falls from 16% to 8%. Fig. 11b shows that even where aggregate acceptance rates are similar, the per-protein RMSD differences are large, with a spread that widens with flexibility.

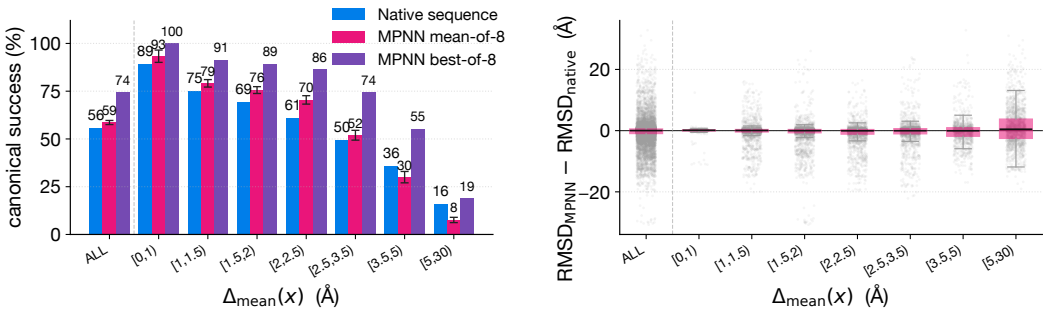

Figure 11: Canonical self-consistency on ATLAS proteins, comparing native sequences against ProteinMPNN redesigns. **(a)** Acceptance rate per flexibility bin. For MPNN mean-of-8, bars show the mean and error bars the bootstrapped success rate over 100 trials, each sampling a different MPNN redesign per protein. **(b)** Per-protein RMSD difference $\mathbf{d}(\mathbf{z}_0, \hat{f}(\tilde{\mathbf{x}}_i)) - \mathbf{d}(\mathbf{z}_0, \hat{f}(\mathbf{x}))$.

A natural response is that one could let generative models output $(\mathbf{x}, \mathbf{z})$ pairs but replace $\mathbf{x}$ with an MPNN redesign $\tilde{\mathbf{x}}$ which agrees most with $\mathbf{z}$, for both evaluation and for downstream design. We argue against this for two reasons. First, sequence selection is a modeling choice about the energy landscape, not merely a way to express a structure. A model that designs a flexible enzyme picks sequences that produce that flexibility; replacing them with MPNN redesigns flattens the landscape toward a single-state interpretation, destroying exactly the property the model was designing for. Second, MPNN inherits the same flexibility-correlated failure. The gain from MPNN redesign diminishes sharply with flexibility: in the most flexible bin, mean acceptance is similar to native sequence. Therefore, redesign provides no rescue precisely where canonical self-consistency fails the most.

H ROBUSTNESS TO THE RMSD THRESHOLD

We repeated the ATLAS failure-mode decomposition and the co-design SC-1/SC-16 analysis across all six generative model variants at $\tau \in \{1.5, 2.0, 2.5, 3.0\}$ Å. Our main conclusions hold throughout this range: both failure modes increase with flexibility, and increasing the SimpleFold ensemble size from $N = 1$ to $N = 16$ improves native recall. Generative model rankings are also unchanged, except for La-Proteina ($\eta = 0.4$) and RFdiffusion3, the two lowest-performing models, whose differences are not statistically significant at any tested threshold. These threshold-sensitivity analyses can be reproduced using the submitted code by setting `PASS_THRESHOLD_A = 2.0` in `_common.py` (line 58) to the desired threshold and rerunning the analysis.

Relaxing the threshold does not substitute for ensemble self-consistency. SimpleFold-16 achieves 82.6% native recall at $\tau = 2$ Å, whereas canonical ESMFold self-consistency attains only 78.4% even at $\tau = 6$ Å. SimpleFold-1 requires $\tau = 3.46$ Å to match the recall of SimpleFold-16 at $\tau = 2$ Å. This threshold relaxation, however, substantially increases false positives on the Rosetta decoy dataset. Increasing the ensemble size from $N = 1$ to $N = 32$ at fixed $\tau = 2$ Å increases the false-positive count from 1 to 2, whereas relaxing the threshold from $\tau = 2$ to 3.5 Å increases it from 1 to 141. Thus, ensemble self-consistency improves native recall while retaining the selectivity of a stringent RMSD threshold.

Table 8: **Failure mode decomposition on MEGASim**. Each paired entry reports the percentage of systems in that category using the first 1 μs / full 12–22 μs trajectory. Success rates are unchanged by the reference trajectory duration.

	$N = 1$	$N = 4$	$N = 8$	$N = 16$	$N = 32$	$N = 64$
Folder failure	16.2 / 13.5	14.4 / 12.3	14.0 / 12.2	13.3 / 11.7	13.3 / 11.7	12.9 / 11.5
Flexibility failure	10.0 / 12.7	7.0 / 9.1	4.8 / 6.6	4.1 / 5.6	3.0 / 4.5	2.2 / 3.7
Success	73.8	78.6	81.2	82.7	83.8	84.9

I Robustness to simulation time and remaining limitations

ATLAS is short trajectories initialized from the experimentally deposited structures; it is not sampling from entire sequence-conditioned equilibrium ensemble $\rho_{\mathbf{x}}$. While it justifies us using ATLAS MD frames as plausible (sufficiently low free energy) conformations that protein can adopt in a solution, it also limits the scope of our analysis. Consequently, our failure decomposition may overestimate folder failures: a refolded structure that lies outside the ATLAS MD ensemble is not necessarily implausible, but may instead correspond to a conformation that was not sampled within the short trajectory.

To quantify the extent to which the limited duration of the ATLAS trajectories causes flexibility failures to be labeled as folder failures, we ran our failure mode analysis on the MEGASim dataset, a dataset released with BioEmu containing 217 systems simulated for more than 10 μs each. These include 90 de novo designed proteins, making the dataset particularly relevant to our study. Across SimpleFold ensemble sizes $N = 1$–64, we compared the decomposition obtained using the first 1 μs of each trajectory with that obtained using the full 12–22 μs trajectory.

Table 8 supports two conclusions. First, 12–22$\times$ extended simulation produces a systematic but modest shift from folder failures to flexibility failures, **only** by 1.4–2.7 percentage points. The failure mode decomposition is stable with respect to simulation time. Second, our main findings are independently reproduced on MEGASim, a long MD dataset that includes de novo designed proteins: as the SimpleFold ensemble size increases, recall improves and both folder and flexibility failures decrease, and flexibility failures are nearly eliminated at moderate ensemble sizes.

I.1 Limitations

Nevertheless, stability over the tested time window does not establish comprehensive coverage of the sequence-conditioned equilibrium ensemble $\rho_{\mathbf{x}}$. Even the longer MEGASim trajectories may miss conformations separated by large free-energy barriers, including rare transitions, large-scale domain rearrangements, and fold-switching behavior. Finally, our analysis does not directly address intrinsically disordered proteins.