ENSEMBLE-IF: TOWARDS PROTEIN DESIGN FOR TARGET CONFORMATIONAL POPULATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Designing amino acid sequences that adopt one or more target conformations is an important problem in protein engineering, with applications in antibody and GPCR design. However, many use cases require controlling not only which conformational states are accessible, but also their equilibrium populations, a property that existing inverse folding-based design methods leave unconstrained. In this work, we introduce ENSEMBLE-IF, an inverse folding-based pipeline for designing protein sequences that exhibit target equilibrium populations. ENSEMBLE-IF combines pretrained inverse folding models with machine-learned conformational ensemble generation, scoring sequence-structure compatibility across sampled conformations to identify mutations that directionally reshape the conformational landscape. We validate designs with molecular dynamics simulations on events defined from reference trajectories, achieving an over 90% reduction in mean target-band distance error relative to baselines while maintaining structural integrity.

1 INTRODUCTION

Proteins are macromolecules that facilitate nearly all biological processes. The design of proteins that exhibit specific desired properties is therefore invaluable for the development of novel drugs and therapeutics. To accomplish this, methods for predicting a protein's sequence given a particular backbone structure, known as *inverse folding* (Yue & Dill, 1992), have become prevalent in the field of protein design (Watson et al., 2023). In recent years, inverse folding models like ProteinMPNN (Dauparas et al., 2022) have shown promise in predicting amino acid sequences compatible with given backbone crystal structures, opening doors for protein design across various domains within computational biology (Goverde et al., 2024; Gasser et al., 2025; de Haas et al., 2024).

However, recent innovations in inverse folding have placed far less emphasis on the prediction of protein sequences that satisfy particular dynamical constraints, not just a set of 3D crystal structures. Emerging works such as DynamicMPNN (Abrudan et al., 2026), SwitchCraft (Jing et al., 2026), and ProteinMPNN-MSD (Liu et al., 2025) have demonstrated progress towards *multi-state design*, which predicts protein sequences that fold into multiple conformational states. Compatibility with multiple states, however, does not determine the *relative equilibrium populations* of those states. At equilibrium, the occupancy of a conformational state depends on its free energy such that even modest sequence-induced changes in the relative free energies of competing states can substantially redistribute their populations (Norn et al., 2021; Vendruscolo & Dobson, 2005). Consequently, even though a sequence may be compatible with certain active and inactive conformations, it may occupy them in proportions that are unsuitable for its desired biological function. Mechanisms such as allosteric regulation, signaling, molecular recognition, and enzyme activity are particularly relevant when the design objective crucially relies on shifts in the populations of certain conformational states, not just the existence of those states (Motlagh et al., 2014; Nussinov et al., 2023).

In this work, we therefore consider an alternative problem setting that we will call *ensemble-level design* that tackles questions such as the following: if a protein naturally interconverts between open and closed states, can we design sequence mutations that shift the equilibrium population of the open state from 20% to a desired range of 60–70%? Such problems in protein design are much more difficult than sequence prediction for static backbone structures. To solve this, we introduce ENSEMBLE-IF, an inverse folding-based framework for editing sequences that shift conformational

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.

populations towards user-specified target bands. By combining pretrained inverse folding models with machine-learned conformational ensemble generation, our approach uses sequence-structure compatibility across sampled conformations to identify mutations predicted to directionally reshape the underlying conformational landscape. Unlike previous methods for protein design, ENSEMBLE-IF specifically targets the reshaping of equilibrium populations instead of static sets of conformations, moving beyond the single- and multi-state design paradigms and paving the way towards more powerful control over protein function. Our contributions are as follows:

- We introduce ENSEMBLE-IF, a novel machine learning-based pipeline for designing proteins that satisfy target equilibrium populations.
- We extend protein sequence design to the *target-population* setting, formulating one of the first methods that explicitly designs mutations to achieve user-specified equilibrium occupancies over conformational states.
- On geometric events mined from mdCATH, ENSEMBLE-IF achieves over 90% reduction in mean target-band distance error over relevant baselines under rigorous molecular dynamics (MD) experiments while maintaining superior native contact recovery compared to baselines in 4/6 tested target bands.

2 PRELIMINARIES AND RELATED WORK

2.1 SINGLE- AND MULTI-STATE INVERSE FOLDING

Proteins are chains of amino acids, also known as *residues*. Each chain is represented with a residue sequence, where each token represents an amino acid. In this work, we restrict our scope to the standard 20 naturally-occurring amino acids. For a given protein with n residues, we denote its sequence and 3D backbone structure by S and $\boldsymbol{X} \in \mathbb{R}^{n\times 4\times 3}$, respectively. Predicting $\boldsymbol{X}$ given S is known as protein folding (Dill et al., 2008). Consequently, predicting S given $\boldsymbol{X}$ is known as inverse folding (Yue & Dill, 1992). We call the problem of learning $p(S \mid \boldsymbol{X})$ *single-state* inverse folding and the problem of learning $p(S \mid \{\boldsymbol{X}_i\}_{i=1}^k)$ *multi-state* inverse folding, where each $\boldsymbol{X}_i$ is a possible conformation of S. Knowing what sequences will fold into a known structure or set of structures is highly valuable for designing novel proteins with desired effects (Bowie et al., 1991; Dahiyat & Mayo, 1997; Sumida et al., 2024). ProteinMPNN (Dauparas et al., 2022) and ESM-IF (Hsu et al., 2022) are among the first single-state inverse folding models trained on crystal structures from the Protein Data Bank (Berman et al., 2000). Several following single-state design models have demonstrated similar capabilities (Gao et al., 2023; Akpinaroglu et al., 2023; Song et al., 2024). Among multi-state inverse folding models, DynamicMPNN (Abrudan et al., 2026) leverages graph representations of multiple backbone structures to predict sequences that exhibit strong re-foldability with AlphaFold 3 (Abramson et al., 2024). ABACUS-T (Liu et al., 2025) tackles multi-state design using a similar approach, adding a protein language model (Lin et al., 2023) and multiple sequence alignments (MSAs) to enhance performance.

2.2 ENSEMBLE-LEVEL INVERSE FOLDING

Single- and multi-state design target sequences that fold into static backbone structures; however, in reality, proteins are hardly static macromolecules, capable of moving and vibrating in ways intimately related to their function (Nussinov et al., 2023; Corbella et al., 2023). For example, G Protein Coupled Receptors (GPCRs) are cell-surface proteins that transduce extracellular signals by sampling an ensemble of inactive, intermediate, and active conformations (Wootten et al., 2018). Metamorphic proteins interconvert between distinct folded conformations whose relative populations determine key functions (Alexander et al., 2009). Complete protein design therefore requires specifying not only accessible structures but also their equilibrium populations and dynamical properties. ENSEMBLE-IF aims to remedy this gap, designing sequences to control the populations of equilibrium states, thereby providing fine-grained ways of controlling protein function and taking steps towards more complete protein design.

Compared to single- and multi-state design, ensemble-level design is less explored. Caliby (Shuai et al., 2025) uses a graph neural network-based Potts model (Wu, 1982) to predict sequences compatible with a given protein's conformational ensemble. However, Caliby conditions designs on

existing ensembles rather than treating ensemble behavior itself as the design objective. Next, VibeGen (Ni & Buehler, 2026) designs sequences to match prescribed low-frequency vibrational modes using an iterative designer/predictor loop trained with normal-mode analysis (NMA) supervision (Case, 1994). While this is conceptually close to ENSEMBLE-IF, NMA profiles characterize local collective motions around structures rather than equilibrium populations of conformational states. Similarly, Flexpert-Design (Kouba et al., 2025) designs sequences that are compatible with a per-residue flexibility profile, which is again related but not well-posed for this particular setting. Finally, Conformational Biasing (Cavanagh et al., 2026) is a framework that uses relative inverse folding scores to identify mutations that favor specific conformations. However, score differences do not directly specify equilibrium populations because favoring one conformation does not specify how strongly it should be populated. ENSEMBLE-IF on the other hand moves towards more targeted population programming. Figure 1 illustrates the difference between the various sequence design paradigms as well as the niche that our work aims to fill.

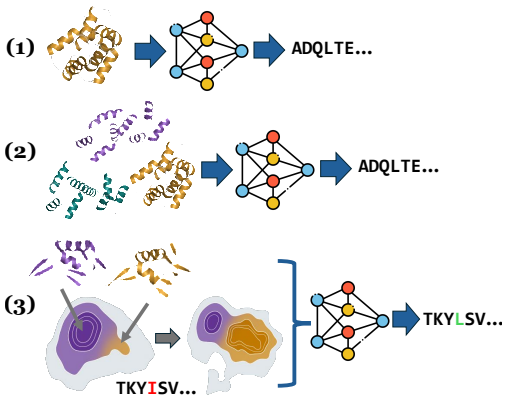


Figure 1: Comparison of protein sequence design paradigms. **(1)** illustrates single-state design where sequences are predicted given a single static structure. **(2)** depicts multi-state design where multiple structures constrain the design space. Our work tackles **(3)**, representing ensemble-level mutational design targeting the control of ensemble state populations.

2.3 PROBLEM DEFINITION

ENSEMBLE-IF seeks to solve the problem of controlling equilibrium conformation populations. Let S_0 stand for the native sequence of a protein, and let $\mathcal{T}_0 = \{\boldsymbol{X}_t\}_{t=1}^{T}$ stand for a reference equilibrium trajectory of length T with time-indexed conformations $\boldsymbol{X}_t \in \mathcal{X}$ in the equilibrium conformational space $\mathcal{X}$. Each $\boldsymbol{X}_t$ may also be called a *frame*, similar to videos. In this work, we define an *event* as a criterion on conformations $E\colon \mathcal{X} \to \{0,1\}$, such as whether two residues or region centroids are within a certain distance of each other. Next, we define a *state* $\Omega_E = \{\boldsymbol{X} \in \mathcal{X} \mid E(\boldsymbol{X}) = 1\}$ as a set of conformations satisfying E. Finally, we define the empirical *population* of this state as

$$\hat{\mathcal{P}}_E(\mathcal{T}) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}\left[\boldsymbol{X}_t \in \Omega_E\right] = \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}\left[E(\boldsymbol{X}_t)\right],$$

or the fraction of frames where the event occurs. Let ρ_S denote the equilibrium conformational distribution of sequence S under fixed physical conditions. Our design objective is to identify mutations to S_0 such that

$$\mathcal{P}_E(S) = \mathbb{E}_{\boldsymbol{X}\sim\rho_S}[E(\boldsymbol{X})] \in B, \qquad B = [b-\epsilon, b+\epsilon] \subseteq [0,1], \tag{1}$$

where b is the desired population and ϵ is its tolerance. Formally, we wish to predict $p(S \mid S_0, \mathcal{T}_0, E, B, \epsilon)$, or the distribution of sequences that are compatible with a given native sequence, a reference equilibrium trajectory, an event definition, and a state population target band. Compared with single- and multi-state inverse folding, this task is significantly more difficult because designed mutations need to not only satisfy structural compatibility but also manipulate populations to fall within a prescribed range.

3 METHOD

3.1 DESIGNING SEQUENCES TOWARDS A TARGET STATE POPULATION

The central idea behind ENSEMBLE-IF is to let the requested state population determine how strongly each sampled conformation influences sequence proposals, then iteratively evaluate which

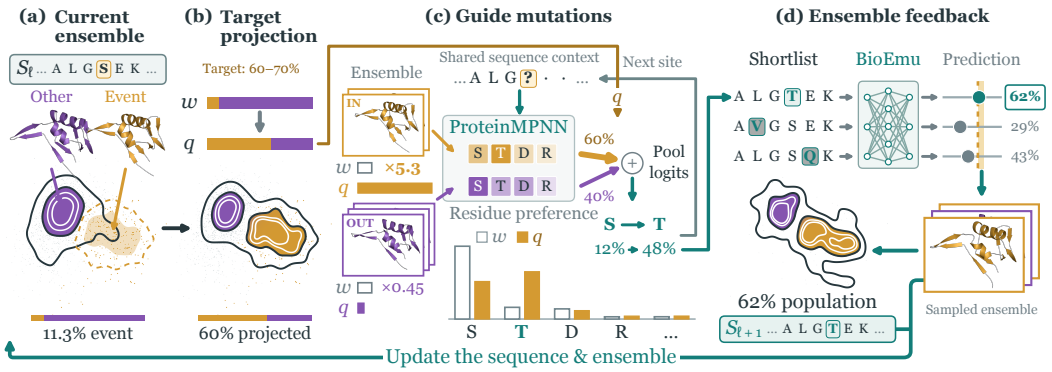


Figure 2: Outline of the ENSEMBLE-IF framework. In **(a)**, for a given round ℓ, we take the incumbent sequence S_ℓ and incumbent ensemble and the desired state population and project the empirical frame weights onto the desired population state target band in **(b)**. During guidance in **(c)**, those ensemble in-state frames are weighted higher while out-of-state frames are discounted, biasing the logits towards sequences more compatible with the ensemble frames. Finally, in **(d)**, we take candidate mutated sequences and screen them with BioEmu, maintaining the best improvements in state populations. The best candidate sequence and its predicted ensemble become the new incumbents.

proposals produce the desired population shift. Similar to Hsu et al. (2022) and Cavanagh et al. (2026), we note that single-state inverse folding likelihoods have been demonstrated to act as compatibility scores across alternative conformations even if the underlying model has only been trained on single structures. With this insight, rather than selecting a particular favored structure and designing for it, we determine how much probability mass must move between a state and its complement, and use this redistribution to determine how strongly each conformation should influence the designed sequence.

Concretely, ENSEMBLE-IF maintains an incumbent sequence S_ℓ and a proposal ensemble $\mathcal{E}_\ell$ at each round ℓ of mutation proposal, starting from the native sequence S_0 and reference trajectory $\mathcal{T}_0$. Each round consists of three steps: (1) projection of the ensemble onto the target population band, (2) proposing compatible sequences through weighted inverse folding, and (3) evaluating the proposed sequences using a fast equilibrium conformer generator to decide which sequences to retain and use for the next round of proposals.

3.2 PROJECTING ENSEMBLES ONTO TARGET POPULATIONS

A population target constrains the total probability of the event state Ω_E, but does not specify how conformations should be distributed. We therefore seek an ensemble that satisfies the target band while introducing as little variation as possible. Let the proposal ensemble for a given round be $\mathcal{E} = \{\boldsymbol{X}_i\}_{i=1}^N$, omitting the round index for clarity. Assign empirical uniform frame weights $w_i = 1/N$, and let $e_i = E(\boldsymbol{X}_i)$ and $u = \sum_{i=1}^N w_i e_i$ denote frame memberships and the current ensemble population, respectively. To satisfy the target band while preserving as much of the empirical ensemble as possible, we obtain target weights q by solving

$$q = \arg\min_{v \in \Delta_N} D_{\mathrm{KL}}(v \,||\, w) \quad \text{subject to} \quad \sum_{i=1}^N v_i e_i \in B \tag{2}$$

where Δ_N is the probability simplex over N samples, and $D_{\mathrm{KL}}(v \,||\, w)$ is the KL-divergence between a candidate reweighted distribution v and the empirical distribution w over those same conformations. Minimizing this divergence identifies the distribution closest to the empirical ensemble, in the KL sense, whose event population satisfies the target band B.

For a binary event with $0 < u < 1$, these projection weights q have a closed-form solution. Let $u_B = \min\{b + \epsilon, \max\{b - \epsilon, u\}\}$ be the population in B closest to u. Then

$$q_i = \begin{cases} w_i \, u_B / u, & e_i = 1, \\ w_i \, (1 - u_B)/(1 - u), & e_i = 0. \end{cases} \tag{3}$$

This projection ensures that the event state receives total mass u_B and its complement receives mass $1 - u_B$ while keeping the relative weights of conformations within each state unchanged. To illustrate, if the event is underpopulated relative to B, then $u_B > u$, and so frames where the event occurs receive more weight ($q_i > w_i$ for $e_i = 1$), while the remaining frames receive less ($q_i < w_i$ for $e_i = 0$). If u already lies in B, then $q = w$, and no redistribution occurs. These weights are subsequently used to generate sequence proposals with a single-state inverse folding model.

3.3 Sequence Design via Weighted Inverse Folding

To convert projected ensemble weights into sequence proposals, we combine ProteinMPNN (Dauparas et al., 2022) residue logits across conformations during decoding. We opt for ProteinMPNN for its simplicity and strong performance in inverse folding in general. For a given decoding order π and decoding step j, let h_j denote the sequence context available when sampling position π_j. ProteinMPNN then produces amino-acid logits $z_{ij} \in \mathbb{R}^{20}$ conditioned on this sequence context and each conformation $\boldsymbol{X}_i$. We then conduct multi-state design by combining these logits with the projected weights from Equation (3) and sampling from:

$$\bar{z}_j^q(a) = \sum_{i=1}^{N} q_i z_{ij}(a), \qquad r(a \mid h_j) = \frac{\exp\big(\bar{z}_j^q(a)/\tau\big)}{\sum_{a' \in \mathcal{A}} \exp\big(\bar{z}_j^q(a')/\tau\big)}, \tag{4}$$

where $\tau > 0$ is the sampling temperature, and $\mathcal{A}$ is the permitted amino-acid alphabet. Each sampled residue is then shared across all conformations before proceeding to the next position. The weights q remain fixed throughout a proposal, while the sequence context evolves with each sampled residue.

Intuitively, Equation (4) can be seen as reconciling the potentially different amino-acid preferences of the decoder for each conformation in the ensemble. That is, we can express the pooled decoder in terms of the inverse folding model's conditional probabilities like so:

$$r(a \mid h_j) \propto \exp\left(\frac{1}{\tau} \sum_{i=1}^{N} q_i \log p_\theta(a \mid \boldsymbol{X}_i, h_j)\right) = \prod_{i=1}^{N} p_\theta(a \mid \boldsymbol{X}_i, h_j)^{q_i/\tau}. \tag{5}$$

In this way, the decoder favors amino acids with high average log-compatibility under the projected ensemble with temperature τ as a concentration parameter. Via Equation (5), population reweighting can be seen as influencing sequence design by relating shifts in mass towards the event state as increases in the relative influence of its conformer-specific residue preferences. The target population therefore determines how the decoder balances compatibility with the event state against compatibility with its complement at each residue decision.

To isolate the effect of population reweighting, let $\bar{z}_j^w(a)$ denote the same logit average computed with w instead of q. Let $\bar{z}_j^+(a)$ and $\bar{z}_j^-(a)$ denote the mean logits within the event state and its complement, respectively, with the empirical weights renormalized within each class. Substituting Equation (3) gives

$$\bar{z}_j^q(a) = \bar{z}_j^w(a) + (u_B - u)\big[\bar{z}_j^+(a) - \bar{z}_j^-(a)\big]. \tag{6}$$

The second term is therefore a state-specific contrast whose sign and magnitude follow the required population correction and vanish when $u \in B$. Because its scale need not match the population change induced by a mutation, we introduce guidance strength η:

$$\bar{z}_j^{(\eta)}(a) = \bar{z}_j^w(a) + \eta\left[\bar{z}_j^q(a) - \bar{z}_j^w(a)\right], \tag{7}$$

which replaces $\bar{z}_j^q$ in Equation (4). Here, $\eta = 1$ recovers the projected decoder, $\eta = 0$ designs against the empirical ensemble, and $\eta > 1$ amplifies the target-specific preference. Unlike changing τ, which rescales all logits, η changes only the contrast induced by the population objective. In our experiments, we treat η as a hyperparameter to tune.

3.4 Iterative Refinement with Ensemble Feedback

While projection satisfies the target band by construction, a sequence designed against the projected ensemble is not guaranteed to satisfy it. A proposed sequence's equilibrium population may drift significantly, motivating the need to evaluate the sequence's own ensemble within the design loop.

Since running MD for every sequence proposal would be extremely costly, we use a fast machine-learned conformer generator as an oracle to sample each mutant's equilibrium ensemble, opting for BioEmu (Lewis et al., 2025) for its strong sampling ability and faithfulness to equilibrium structures.

At each round, we generate K sequence proposals and record their decoding log-probabilities. Let $\mathcal{D}$ denote the set of designable residue positions. We score each proposal using

$$\mathrm{score}(S;\pi) = \sum_{j:\pi_j \in \mathcal{D}} \log r(S_{\pi_j} \mid h_j). \tag{8}$$

We then deduplicate the top $M \le K$ candidates and sample N_{eval} BioEmu conformers for each. For $\mathcal{Y}(S) = \{\widetilde{\boldsymbol{X}}_i(S)\}_{i=1}^{N_{\text{eval}}}$, define

$$\hat{u}(S) = \frac{1}{N_{\text{eval}}} \sum_{i=1}^{N_{\text{eval}}} E\big(\widetilde{\boldsymbol{X}}_i(S)\big), \qquad d_B(S) = \max\{b - \epsilon - \hat{u}(S),\ 0,\ \hat{u}(S) - (b+\epsilon)\}. \tag{9}$$

The distance d_B measures the population error, vanishing to 0 for an estimated population inside the band. We select the best candidate S^* with the smallest such distance, breaking ties by proposal score, and accept it only if $d_B(S^*) < d_B(S_\ell)$, i.e. the event distance to the target band induced by the proposed sequence is shorter than that of the incumbent sequence. Upon acceptance, we update

$$S_{\ell+1} = S^*, \quad \mathcal{E}_{\ell+1} = \mathcal{Y}(S^*) \tag{10}$$

The next projection step responds to the population sampled for accepted sequence, continually guiding subsequent mutations. This iterative refinement procedure runs until the predicted ensemble reaches B or after at most L rounds. If no candidate is accepted, we report a failure to improve within the proposal budget. See Section A for hyperparameters used in the following experiments and Figure 2 for an illustration of the full ENSEMBLE-IF pipeline.

4 EXPERIMENTS

To evaluate ENSEMBLE-IF, we seek to influence the state populations of geometrically-defined events given their native sequences and trajectories. For each event, we define a desired target band occupancy and validate each of our designs using MD under settings as close as possible to those used to generate the native trajectories.

Datasets. To discover events, we obtain native sequences, structures, and trajectories from mdCATH (Mirarchi et al., 2024), a dataset of all-atom MD simulations of 5,398 protein domains from the CATH protein structural classification database (Sillitoe et al., 2021). mdCATH provides five production replicas per domain across five temperatures with frames saved every 1 ns. We used the provided 320K trajectories pre-equilibrated at 300K for 20 ns to target more stationary, native-state-like dynamics compared to the higher-temperature trajectories (Sugita & Okamoto, 1999). Each replica contains 400-450 ns of simulation time, allowing generally enough time for global structural observables to stabilize (Mirarchi et al., 2024). For the sake of more efficient downstream MD validation, we restrict our search to single-chain domains containing at most 250 residues.

Event mining. We denote mdCATH replicas as R0 through R4 and use R0 and R1 for event discovery. Before identifying events, we first generate candidate observables that define each event. In this work, we define candidate observables as the virtual Cβ distances between two residues or the distances between the virtual Cβ centroids of two residue regions. To construct a region, we take each eligible residue in each frame and include residues whose virtual Cβ distance to that anchor is at most $r \in \{6\text{Å}, 8\text{Å}, 10\text{Å}\}$, averaged over the R0 and R1 frames. We retain regions containing at least 3-5 residues and hold their membership fixed across subsequent frames. To avoid targeting state populations dominated by disproportionately flexible regions, we disregard the five residues at each terminus of each system (Dziadek et al., 2024) and require a sequence separation between singleton residues or cross-region residue pairs of at least 8 positions. This lets us focus on events centered around smoother internal rearrangements rather than local fluctuations arising from weakly constrained termini or nearby residues.

This results in 61.6 million residue-pair or region-pair combinations for event selection. For each candidate, we divide its distances into two groups by minimizing the within-group squared error and

set the event state threshold as the midpoint between their means. Candidates are then ranked by separation of those centers and then by distance magnitude. We maintain only the highest-ranking candidate per-configuration and domain, reducing the number of candidate observables to 19.6K. We then apply a series of quality filters, such as ensuring that 90% of R0–R2 frames must have a TM-score (Zhang & Skolnick, 2004) of at least 0.7 w.r.t. the reference PDB structure. See Section B for full details. After sequence deduplication, our event candidates shrink to 345 across 199 domains. We finally sort the candidates according to a weighted score of these criteria and evaluate the top-3 domains: `1divA02`, `1kwgA03`, and `2v0cA03`. See Section B for a full description of the weighted score.

Target band definition. For these domains and their mined events, we define the target population bands as shifts from each event state's native occupancy in R3. For `1divA02` and `1kwgA03`, the target is shifted -30 percentage points (pp) from native occupancies of 64.89% and 65.56%, respectively. For `2v0cA03`, native occupancy is 47.4%, and we use this domain to test a target near-full occupancy of 98% as an edge case experiment. For comprehensiveness, we also test bands in the opposite direction, namely $+30$ pp for `1divA02`, $+15$ pp for `1kwgA03`, and -30 pp for `2v0cA03`. We relax all targets with a 3 pp tolerance window, clipping bands to the valid occupancy range ([0,100]).

Molecular dynamics validation. To validate our designs, we replicate the original mdCATH MD setup as closely as possible. For each design, we predict an equilibrium all-atom structure using ESMFold2 (Candido et al., 2026) with MSAs enabled and use its coordinates to initialize 3 MD replicas using OpenMM 8.3.1 (Eastman et al., 2024) with the CHARMM22* force field (Vanommeslaeghe et al., 2010). Similar to the settings used in mdCATH, we pre-equilibrate the structure for 10 ns with restrained NPT at 300K followed by 10 ns of unrestrained NPT at the same temperature before launching 3 production replicas for 400 ns at 320K. Like mdCATH, we solvate each protein in a periodic cubic box of TIP3P water with at least 9Å of padding, add Na^+ and Cl^- ions to neutralize the system, and target an NaCl concentration of 0.15M. The full MD validation setup details are documented in Section C.1.

Baselines. We include standard ProteinMPNN (Dauparas et al., 2022) as a static-structure control and use Conformational Biasing (CB) (Cavanagh et al., 2026) as our primary population-shifting baseline. CB is not tailored to the specific problem of ensemble population specification, and so we adapt it to our benchmark. CB scores mutations by the difference between standardized changes in ProteinMPNN sequence log-likelihood for two representative conformations, yielding a directional conformational bias. To adapt CB to our setting, we partition each native ensemble according to the given binary event and use the Cα-RMSD medoids from the on- and off-states as the representative structures. For a protein of length L, we then exhaustively score all $19L$ single mutations and rank them by the difference in standardized ProteinMPNN scores favoring the on-state when the native occupancy is below the target band, or the off-state when it is above. Because this captures only the direction towards population changes, we then screen the top-5 highest-ranked mutations using 768 samples from BioEmu (3 seeds $\times$ 256 conformers) and keep the mutation whose predicted ensemble is closest to the target band. For experiments with multiple rounds of mutations, see Section E.

4.1 MDCATH RESULTS

Figure 3 depicts each method's running event occupancy over time, with mean-pooled occupancies represented by dark lines and their individual replicas depicted by faint lines. As expected, ProteinMPNN fails to reach the target band in nearly all cases, serving as a reference point. Across the board, ENSEMBLE-IF achieves a smaller running distance to the target band compared to CB, particularly in both directions for `1kwgA03`. Its improvements over CB in `1divA02` are more limited, exhibiting wide replica variability when targeting the [31.9, 37.9] band. ENSEMBLE-IF falls within three of the six target bands where its two misses on 1divA02 are within 1.03 pp of the band, whereas the low-occupancy 2v0cA03 target remains substantially undershot. However, across the board, ENSEMBLE-IF is closer to the target band, as evidenced by Table 1, where it achieves over 90% lower mean target-band distance error than CB.

We also evaluate methods on native contact recovery, measured as the fraction of preserved residue-residue pairs from the deposited PDB structure that are within 8Å of each other. Higher native contact recovery indicates greater faithfulness to the original fold of the protein, which is impor-

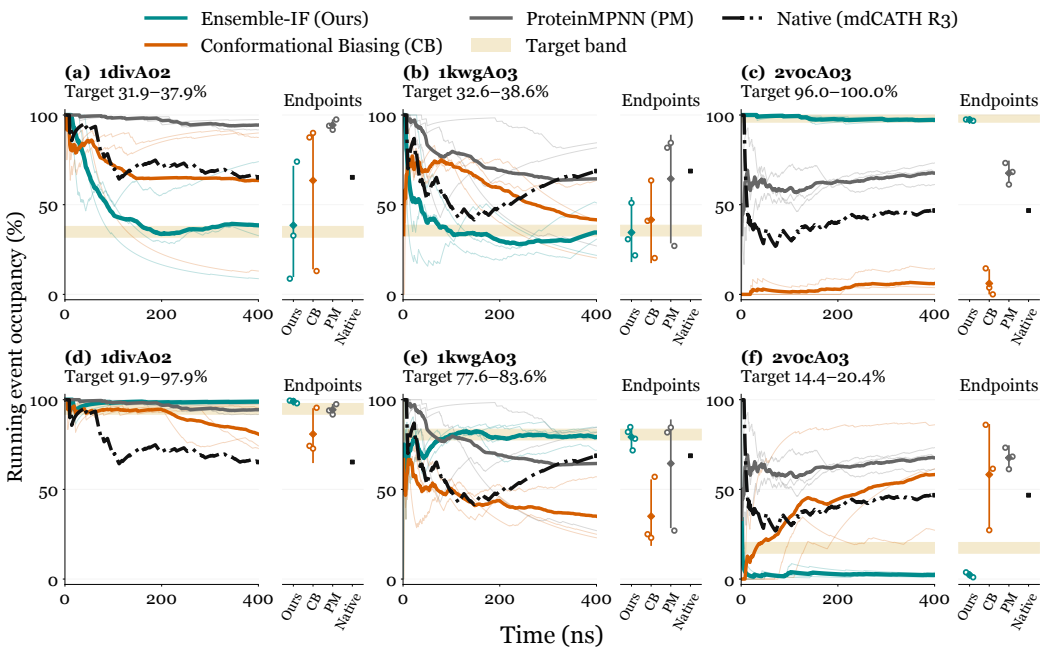


Figure 3: MD validation of the running occupancy of the mined event over time across targets. ENSEMBLE-IF exhibits the strongest adherence to the target band with CB following behind by comparison. Methods generally see high variability across their replicas for `1divA02` and `1kwgA03`, although their means exhibit distinct shifts in performance.

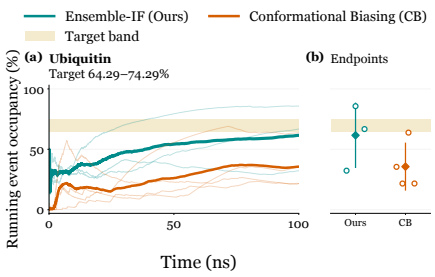
Table 1: Event occupancy and native contact recovery across target bands. For occupancy, **bold** indicates the value closer to the target band; for contact recovery, **bold** indicates the higher mean.

Domain	Target band (%)	Event occupancy (%)		Native contact recovery (%)	
		Ensemble-IF	CB	Ensemble-IF	CB
1divA02	[31.89, 37.89]	**38.50**	63.50	**80.49 ± 1.30**	80.43 ± 2.55
	[91.89, 97.89]	**98.92**	80.83	**88.85 ± 0.20**	85.72 ± 2.13
1kwgA03	[32.56, 38.56]	**34.50**	41.58	83.07 ± 4.65	**88.15 ± 0.73**
	[77.56, 83.56]	**79.19**	35.00	85.60 ± 1.40	**88.53 ± 2.13**
2v0cA03	[96.00, 100.00]	**97.25**	6.08	**89.99 ± 0.39**	42.11 ± 0.75
	[14.40, 20.40]	**2.33**	58.25	**83.67 ± 2.08**	36.34 ± 4.48

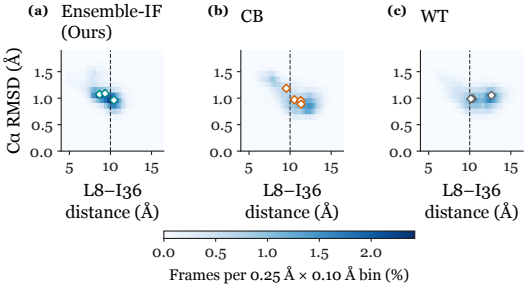
tant to show that demonstrated targeted population control does not simply result in total structural disruption. On this metric in Table 1, ENSEMBLE-IF fares better than CB on four of the six target bands, especially those belonging to `2v0cA03`. CB retains more native contacts for both `1kwgA03` targets, although these are also cases in which ENSEMBLE-IF exhibits its clearest improvements in event occupancy, suggesting that there is a point after which staying too close to the native fold prevents population control. Overall, these results suggest that ENSEMBLE-IF's control over ensemble populations is generally achieved without requiring large disruptions to the native structure.

4.2 UBIQUITIN CASE STUDY

The previous experiments on mdCATH domains evaluate model performance on state populations derived from a purely geometry-based trajectory mining algorithm. Conformational shifts mined in a way similar to these experiments are therefore not guaranteed to be biologically relevant. To further validate our method, we add a case study examining ubiquitin, a small, 76-residue regulatory protein found in nearly all eukaryotic cells (Goldstein et al., 1975). Ubiquitin in its apo state is known to exhibit a "pincer" motion that modulates the geometry of its primary hydrophobic recognition surface,



(a) Running occupancy for the ubiquitin L8–I36 event. Across replicas, ENSEMBLE-IF improves on moving the pincer movement population closer to the desired target band.



(b) Relationship between the L8–I36 distance and structural RMSD during MD. Unlike CB, ENSEMBLE-IF successfully shifts most of its replicas' state populations to the left.

affecting a substantial fraction of known ubiquitin-protein interactions (Lange et al., 2008; Peters & de Groot, 2012). This makes it an ideal candidate for ensemble-level design that is biologically relevant. In this case study, we examine a 100 ns trajectory provided by Yovanno et al. (2023) initialized using the canonical unmodified ubiquitin "Major" conformation. Within this trajectory, we apply the filters from Section 4 and identify the virtual Cβ distance between L8 and I36 falling below 10Åas an event, corroborating prior work in computational biology (Agrata & Komander, 2025). See Figure 5 for a visualization of the event's positive and negative states. These residues lie on the β_1–β_2 and α_1–β_3 loops, which directly characterize the documented pincer motion.

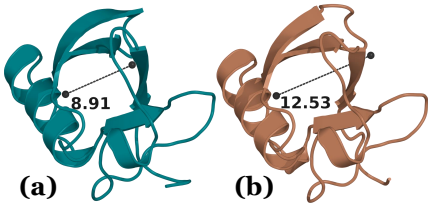


Figure 5: Ubiquitin conformations from the 100 ns trajectory. **(a)** depicts the event-positive state, where the L8-I36 virtual Cβ distance $d = 8.91$Å $<$ 10Å. **(b)** depicts the event-negative state where $d = 12.53$Å > 10Å.

For this case study, we evaluate ENSEMBLE-IF alongside CB. We target a population occupancy shift of $+20$ pp with a tolerance of 5 pp, aiming to move the native occupancy from 49.29 pp into the band $[64.29 \text{ pp}, 74.29 \text{ pp}]$. The MD setup by Yovanno et al. (2023) differs slightly from mdCATH, simulating with the CHARMM36 force field for only 100 ns, taking time steps of 2 fs instead of 4, saving frames every 10 ps instead of every 1 ns, and equilibrating for 1.2 ns total at 310K rather than 20 ns at 300K. We replicate as much of their setup as possible, which we document in Section C.2.

In Figure 4a, we see that ENSEMBLE-IF generally achieves trajectories whose running event occupancy is closer to the target band over CB in a biologically relevant system. We also report the relationship between the L8–I36 distance and the all-residue Cα root mean squared deviation (RMSD) w.r.t. the native deposited structure for the ubiquitin trajectory. We see that ENSEMBLE-IF's design is able to push replicas towards the left of the event threshold while maintaining limited coordinate deviation. Contrast this with CB, which has more vertical variability and fails to push most of its replicas within the target band.

5 CONCLUSION

To conclude, ENSEMBLE-IF is a novel framework for tackling ensemble-level protein design for targeting conformational populations. Among protein design methods, it is one of the few designed for ensemble-level design at this level of granularity, achieving state occupancies closer to the specified target bands than competing baselines while maintaining structural integrity of the native fold. Future work includes combining this work with multi-state design models to design proteins that exhibit unseen conformational states as well as specify their occupancies. Additionally, alleviating the need for an oracle for evaluating mutant candidate sequences would greatly increase the scalability of the method. In sum, ENSEMBLE-IF tackles a crucial obstacle in the achievement of fully programmable biology, a central goal of machine-learned protein engineering.

AI USE STATEMENT

GPT-6 was used to review and refine the phrasing and presentation of this manuscript. GPT-5.6- and GPT-6-based coding agents were also used to audit and implement basic features of the method and all figure generation code, all of which was verified for correctness by the primary author. All other components of the research process, e.g. writing, ideation, formulation, arguments, experimental setups, design choices, etc. are human-made. We take responsibility for the final content of this work, including all text and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The ENSEMBLE-IF algorithm is described in detail in Section 3, including ensemble projection, weighted inverse folding, and iterative refinement with ensemble feedback; the hyperparameters used in our experiments are provided in Section A. The mdCATH event mining procedure is summarized in Section 4, with the complete filtering and candidate-ranking procedure given in Section B. Molecular dynamics validation is described in Section 4, with the full simulation and system-preparation protocols for the mdCATH and ubiquitin experiments provided in Section C, Section C.1, and Section C.2.

REFERENCES

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J. Ballard, Joshua Bambrick, Sebastian W. Bodenstein, David A. Evans, Chia-Chun Hung, Michael O'Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvilė Žemgulytė, Eirini Arvaniti, Charles Beattie, Ottavia Bertolli, Alex Bridgland, Alexey Cherepanov, Miles Congreve, Alexander I. Cowen-Rivers, Andrew Cowie, Michael Figurnov, Fabian B. Fuchs, Hannah Gladman, Rishub Jain, Yousuf A. Khan, Caroline M. R. Low, Kuba Perlin, Anna Potapenko, Pascal Savy, Sukhdeep Singh, Adrian Stecula, Ashok Thillaisundaram, Catherine Tong, Sergei Yakneen, Ellen D. Zhong, Michal Zielinski, Augustin Žídek, Victor Bapst, Pushmeet Kohli, Max Jaderberg, Demis Hassabis, and John M. Jumper. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493—-500, 2024.

Alex Abrudan, Sebastian Pujalte Ojeda, Chaitanya K. Joshi, Matthew Greenig, Felipe Engelberger, Alena Khmelinskaia, Jens Meiler, Michele Vendruscolo, and Tuomas Knowles. Multi-state protein sequence design with dynamicMPNN. In *The Fourteenth International Conference on Learning Representations*, 2026.

Rashmi Agrata and David Komander. Ubiquitin-A structural perspective. *Mol. Cell*, 85(2):323–346, January 2025.

Deniz Akpinaroglu, Kosuke Seki, Amy Guo, Eleanor Zhu, Mark JS Kelly, and Tanja Kortemme. Structure-conditioned masked language models for protein sequence design generalize beyond the native sequence space. *bioRxiv*, pp. 2023–12, 2023.

Patrick A Alexander, Yanan He, Yihong Chen, John Orban, and Philip N Bryan. A minimal sequence code for switching protein structure and function. *Proc. Natl. Acad. Sci. U. S. A.*, 106(50):21149–21154, December 2009.

H M Berman, J Westbrook, Z Feng, G Gilliland, T N Bhat, H Weissig, I N Shindyalov, and P E Bourne. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, January 2000.

J U Bowie, R Lüthy, and D Eisenberg. A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, 253(5016):164–170, July 1991.

Salvatore Candido, Thomas Hayes, Alexander Derry, Roshan Rao, Zeming Lin, Robert Verkuil, Bryan Z Wu, Jin Sub Lee, Elise S Bruguera, Jehan A Keval, Mykhailo Kopylov, John E Pak, Wesley Wu, Neil Thomas, Samson Mataraso, Alvin Hsu, Ashton C Trotman-Grant, Kilian Fatras, Allan dos Santos Costa, Rohil Badkundri, Halil Akın, Deniz Oktay, Jonathan Deaton, Elizabeth Montabana, Hrishita Sitwala, Yue Yu, Marius Wiggert, Dylan Alexander Carlin, Anthony W Goering, Tomasz Blazejewski, Mccullen Sandora, Michael Hla, Tina Z Jia, Leon H Kloker,

Nicholas J Sofroniew, Masatoshi Uehara, Jassi Pannu, Sharrol Bachas, Daniel S Liu, Tom Sercu, and Alexander Rives. Language modeling materializes a world model of protein biology. *bioRxiv*, June 2026.

David A. Case. Normal mode analysis of protein dynamics. *Current Opinion in Structural Biology*, 4(2):285–290, 1994. ISSN 0959-440X. doi: 10.1016/S0959-440X(94)90321-2. Funding Information: This work was supported by NIH grant GM-45811. I thank Rafael BrOschweiler for helpful discussions.

Peter E Cavanagh, Andrew G Xue, Shizhong A Dai, Albert Qiang, Tsutomu Matsui, and Alice Y Ting. Computational design of conformation-biasing mutations to alter protein functions. *Science*, 391(6790):eadv7953, March 2026.

Marina Corbella, Gaspar P Pinto, and Shina C L Kamerlin. Loop dynamics and the evolution of enzyme activity. *Nat. Rev. Chem.*, 7(8):536–547, August 2023.

B I Dahiyat and S L Mayo. De novo protein design: fully automated sequence selection. *Science*, 278(5335):82–87, October 1997.

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.

Robbert J de Haas, Natalie Brunette, Alex Goodson, Justas Dauparas, Sue Y Yi, Erin C Yang, Quinton Dowling, Hannah Nguyen, Alex Kang, Asim K Bera, Banumathi Sankaran, Renko de Vries, David Baker, and Neil P King. Rapid and automated design of two-component protein nanomaterials using ProteinMPNN. *Proc. Natl. Acad. Sci. U. S. A.*, 121(13):e2314646121, March 2024.

Ken A Dill, S Banu Ozkan, M Scott Shell, and Thomas R Weikl. The protein folding problem. *Annu. Rev. Biophys.*, 37(1):289–316, 2008.

Ł J Dziadek, A K Sieradzan, C Czaplewski, M Zalewski, F Banaś, M Toczek, W Nisterenko, S Grudinin, A Liwo, and A Giełdoń. Assessment of four theoretical approaches to predict protein flexibility in the crystal phase and solution. 20(17):7667–7681, September 2024.

Peter Eastman, Raimondas Galvelis, Raúl P Peláez, Charlles R A Abreu, Stephen E Farr, Emilio Gallicchio, Anton Gorenko, Michael M Henry, Frank Hu, Jing Huang, Andreas Krämer, Julien Michel, Joshua A Mitchell, Vijay S Pande, João Pglm Rodrigues, Jaime Rodriguez-Guerra, Andrew C Simmonett, Sukrit Singh, Jason Swails, Philip Turner, Yuanqing Wang, Ivy Zhang, John D Chodera, Gianni De Fabritiis, and Thomas E Markland. OpenMM 8: Molecular dynamics simulation with machine learning potentials. *J. Phys. Chem. B*, 128(1):109–116, January 2024.

Zhangyang Gao, Cheng Tan, and Stan Z. Li. Pifold: Toward effective and efficient protein inverse folding. In *The Eleventh International Conference on Learning Representations*, 2023.

Hans-Christof Gasser, Diego A Oyarzún, Javier Antonio Alfaro, and Ajitha Rajan. Tuning ProteinMPNN to reduce protein visibility via MHC class I through direct preference optimization. *Protein Eng. Des. Sel.*, 38(gzaf003):gzaf003, January 2025.

G Goldstein, M Scheid, U Hammerling, D H Schlesinger, H D Niall, and E A Boyse. Isolation of a polypeptide that has lymphocyte-differentiating properties and is probably represented universally in living cells. *Proc. Natl. Acad. Sci. U. S. A.*, 72(1):11–15, January 1975.

Casper A Goverde, Martin Pacesa, Nicolas Goldbach, Lars J Dornfeld, Petra E M Balbi, Sandrine Georgeon, Stéphane Rosset, Srajan Kapoor, Jagrity Choudhury, Justas Dauparas, Christian Schellhaas, Simon Kozlov, David Baker, Sergey Ovchinnikov, Alex J Vecchio, and Bruno E Correia. Computational design of soluble and functional membrane protein analogues. *Nature*, 631 (8020):449–458, July 2024.

Chloe Hsu, Robert Verkuil, Jason Liu, Zeming Lin, Brian Hie, Tom Sercu, Adam Lerer, and Alexander Rives. Learning inverse folding from millions of predicted structures. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 8946–8970. PMLR, 17–23 Jul 2022.

Bowen Jing, Mihir Bafna, Anisha Parsan, Heyuan Michael Ni, David Kwabi-Addo, Bryan D. Bryson, Adam Klivans, and Bonnie Berger. Switchcraft: A programmatic framework for designing state-switching proteins. In *Forty-third International Conference on Machine Learning*, 2026.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Andrew J Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, August 2021.

Petr Kouba, Joan Planas-Iglesias, Jiri Damborsky, Jiri Sedlar, Stanislav Mazurenko, and Josef Sivic. Learning to engineer protein flexibility. In *The Thirteenth International Conference on Learning Representations*, 2025.

Oliver F Lange, Nils-Alexander Lakomek, Christophe Farès, Gunnar F Schröder, Korvin F A Walter, Stefan Becker, Jens Meiler, Helmut Grubmüller, Christian Griesinger, and Bert L de Groot. Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution. *Science*, 320(5882):1471–1475, June 2008.

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, Stimper 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*, pp. eadv9817, 2025. doi: 10.1126/science.adv9817.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574.

Yufeng Liu, Rui Wu, Xinyu Wang, Sheng Wang, Linghui Chen, Fudong Li, Quan Chen, and Haiyan Liu. Enhancing functional proteins through multimodal inverse folding with ABACUS-T. *Nat. Commun.*, 16(1):10177, November 2025.

Antonio Mirarchi, Toni Giorgino, and Gianni De Fabritiis. MdCATH: A large-scale MD dataset for data-driven computational biophysics. *Sci. Data*, 11(1):1299, November 2024.

Hesam N Motlagh, James O Wrabl, Jing Li, and Vincent J Hilser. The ensemble nature of allostery. *Nature*, 508(7496):331–339, April 2014.

Bo Ni and Markus J Buehler. VibeGen: Agentic end-to-end de novo protein design for tailored dynamics using a language diffusion model. *Matter*, 9(5):102706, May 2026.

Christoffer Norn, Basile I M Wicky, David Juergens, Sirui Liu, David Kim, Doug Tischer, Brian Koepnick, Ivan Anishchenko, Foldit Players, David Baker, and Sergey Ovchinnikov. Protein sequence design by conformational landscape optimization. *Proc. Natl. Acad. Sci. U. S. A.*, 118 (11):e2017228118, March 2021.

Ruth Nussinov, Yonglan Liu, Wengang Zhang, and Hyunbum Jang. Protein conformational ensembles in function: roles and mechanisms. *RSC Chem. Biol.*, 4(11):850–864, November 2023.

Jan H Peters and Bert L de Groot. Ubiquitin dynamics in complexes reveal molecular recognition mechanisms beyond induced fit and conformational selection. *PLoS Comput. Biol.*, 8(10): e1002704, October 2012.

Richard W. Shuai, Tianyu Lu, Subhang Bhatti, Petr Kouba, and Po-Ssu Huang. Ensemble-conditioned protein sequence design with caliby. *bioRxiv*, 2025. doi: 10.1101/2025.09.30.679633.

Ian Sillitoe, Nicola Bordin, Natalie Dawson, Vaishali P Waman, Paul Ashford, Harry M Scholes, Camilla S M Pang, Laurel Woodridge, Clemens Rauer, Neeladri Sen, Mahnaz Abbasian, Sean Le Cornu, Su Datt Lam, Karel Berka, Ivana Hutařová Varekova, Radka Svobodova, Jon Lees, and Christine A Orengo. CATH: increased structural coverage of functional space. *Nucleic Acids Res.*, 49(D1):D266–D273, January 2021.

Zhenqiao Song, Tinglin Huang, Lei Li, and Wengong Jin. SurfPro: Functional protein design based on continuous surface. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 46074–46088. PMLR, 21–27 Jul 2024.

Yuji Sugita and Yuko Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.*, 314(1-2):141–151, November 1999.

Kiera H Sumida, Reyes Núñez-Franco, Indrek Kalvet, Samuel J Pellock, Basile I M Wicky, Lukas F Milles, Justas Dauparas, Jue Wang, Yakov Kipnis, Noel Jameson, Alex Kang, Joshmyn De La Cruz, Banumathi Sankaran, Asim K Bera, Gonzalo Jiménez-Osés, and David Baker. Improving protein expression, stability, and function with ProteinMPNN. *J. Am. Chem. Soc.*, 146 (3):2054–2061, January 2024.

K Vanommeslaeghe, E Hatcher, C Acharya, S Kundu, S Zhong, J Shim, E Darian, O Guvench, P Lopes, I Vorobyov, and A D Mackerell, Jr. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. Comput. Chem.*, 31(4):671–690, March 2010.

Michele Vendruscolo and Christopher M Dobson. Towards complete descriptions of the free-energy landscapes of proteins. *Philos. Trans. A Math. Phys. Eng. Sci.*, 363(1827):433–50; discussion 450–2, February 2005.

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, Basile I M Wicky, Nikita Hanikel, Samuel J Pellock, Alexis Courbet, William Sheffler, Jue Wang, Preetham Venkatesh, Isaac Sappington, Susana Vázquez Torres, Anna Lauko, Valentin De Bortoli, Emile Mathieu, Sergey Ovchinnikov, Regina Barzilay, Tommi S Jaakkola, Frank DiMaio, Minkyung Baek, and David Baker. De novo design of protein structure and function with RFdiffusion. *Nature*, 620 (7976):1089–1100, August 2023.

Denise Wootten, Arthur Christopoulos, Maria Marti-Solano, M Madan Babu, and Patrick M Sexton. Mechanisms of signalling and biased agonism in G protein-coupled receptors. *Nat. Rev. Mol. Cell Biol.*, 19(10):638–653, October 2018.

F Y Wu. The potts model. *Rev. Mod. Phys.*, 54(1):235–268, January 1982.

Remy A Yovanno, Alvin Yu, Tyler J Wied, and Albert Y Lau. Phosphorylation at ser65 modulates ubiquitin conformational dynamics. *Structure*, 31(7):884–890.e2, July 2023.

K Yue and K A Dill. Inverse protein folding problem: designing polymer sequences. *Proc. Natl. Acad. Sci. U. S. A.*, 89(9):4163–4167, May 1992.

Yang Zhang and Jeffrey Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins*, 57(4):702–710, December 2004.

A ENSEMBLE-IF HYPERPARAMETER SETTINGS

Throughout all of our experiments, we use $K = 64$ sequence proposals per round of sequence design, evaluating the $M = 4$ top sequence candidates. We use $N_{\text{eval}} = 128$ BioEmu conformers for ensemble feedback, noting that this is the biggest driver of runtime costs given the time needed to retrieve an MSA and subsequently sample 128 structures from a diffusion model. We saw the best performance for nearly all target bands at $\eta = 1$ besides `1kwgA03`'s $[77.6\%, 83.6\%]$ target band, in which case we use $\eta = 0.5$. We do not impose a cap on the number of mutations per proposal, although one could be enforced during proposal generation, for example by restricting the designable positions or constraining each proposed sequence's Hamming distance from the incumbent sequence.

B MDCATH FILTERING AND FINAL CANDIDATE SCORING

This section details how the events are filtered and how final events are scored and ranked for final candidate selection.

Event filtering. We apply the following filters to enhance the quality of the mined events.

- **Global motion confounding.** On R0 and R1, the absolute Pearson correlations of the event distance with both deposited-reference Cα RMSD and radius of gyration must be below 0.6. This reduces the chance that an event simply tracks structural distortion/expansion.
- **Fold retention.** At least 90% of the R0–R2 frames must have a TM-score (Zhang & Skolnick, 2004) of at least 0.7 w.r.t. the deposited crystal structure. This favors events that respect the reference fold and are not dominated by a small set of highly distorted conformations.
- **Population support.** Event occupancy in R2 must be between 0.1 and 0.9 with a binary effective sample size of at least 50. This excludes trivially constant events and limits reliance on using highly correlated frames as independent events.
- **Temporal drift.** We divide R2 into two sections and require each section's occupancy to differ from the full-sized trajectory occupancy by at most 5 pp to ensure that the event is not short-lived.
- **Threshold robustness.** To avoid events that are dependent on a small change to the fitted cutoff, we ensure that moving the distance cutoff by 0.25Å changes R2 occupancy by no more than 1 pp.

After filtering, we apply the following metrics for a final event score to rank the candidates by.

Weighted event ranking. We rank retained events using a weighted combination of replica consistency, temporal persistence, distance-density separation, and threshold stability:

$$s_{\text{rank}}(E) = 0.30\,P(-\sigma_{\text{rep}}) + 0.35\,P(\tau_{\text{dwell}}) + 0.20\,P(-v_{\text{KDE}}) + 0.15\,P(-s_c), \tag{11}$$

where P denotes the percentile rank among the retained events, normalized to $[0, 1]$ and computed separately for each statistic. This transformation makes diagnostics with different units and numerical scales comparable. Larger scores are preferred, with negative signs indicating criteria for which smaller values are desirable. The coefficients are heuristic weights that place the greatest emphasis on temporal persistence and consistency across replicas.

Replica consistency. Let $\widehat{p}_r = \widehat{\mathcal{P}}_E(\mathcal{T}_r)$ denote the event occupancy in replica r. We measure between-replica variation as

$$\sigma_{\text{rep}} = \text{SD}(\widehat{p}_0, \ldots, \widehat{p}_4)\,. \tag{12}$$

Smaller values favor events with consistent population estimates across the five trajectories, reducing preference for events whose observed occupancy depends strongly on the sampled replica.

Temporal persistence. The dwell statistic τ_{dwell} is the median duration of uninterrupted visits to either binary state. A visit is a maximal consecutive sequence of saved frames with unchanged event membership, ending when the event distance crosses the fixed cutoff. Larger values favor persistent state assignments rather than rapid switching near the boundary. This statistic characterizes persistence in the discretely sampled trajectories; it does not by itself establish metastability.

Distance-density separation. Let $\widehat{f}$ be a kernel density estimate of the event-distance distribution, c the event cutoff, and μ_{low} and μ_{high} the two discovery-group means. We define the cutoff-valley ratio as

$$v_{\mathrm{KDE}} = \frac{\widehat{f}(c)}{\sqrt{\widehat{f}(\mu_{\mathrm{low}})\widehat{f}(\mu_{\mathrm{high}})}}. \tag{13}$$

A smaller ratio favors a cutoff lying in a relatively low-density region between the discovery centers, rather than partitioning a densely populated portion of the distance distribution. The discovery centers serve as reference locations and need not coincide with density maxima.

Threshold stability. Let $\widehat{p}_2(a)$ denote the R2 occupancy obtained by applying the distance cutoff a. We measure sensitivity to the event boundary as

$$s_c = \max_{\delta\in\{-h,h\}} \left|\widehat{p}_2(c+\delta) - \widehat{p}_2(c)\right|, \qquad h = 0.25\,\text{Å}. \tag{14}$$

Smaller values favor population estimates that remain stable under small perturbations of the fitted cutoff. Unlike the binary robustness filter, this term distinguishes the relative threshold sensitivity of events that have already passed the admission criteria.

To further pare down the list of candidates to a reasonable set for MD, a subsequent post-filter ranking accepted the top-3 events using

$$S = 0.30\,P(-\sigma_{\mathrm{rep}}) + 0.35\,P(t_{\mathrm{dwell}}) + 0.20\,P(-v) + 0.15\,P(-s_c), \tag{15}$$

where P denotes the percentile rank among accepted events. Here, σ_{rep} is the standard deviation of the five replica occupancies, t_{dwell} is the median binary-state dwell duration, and s_c is the largest R2 occupancy change under the cutoff perturbations above. The valley ratio v is the kernel-density estimate at c divided by the geometric mean of the estimated densities at the two discovery-group means. Higher scores therefore favor consistent populations across replicas, longer state visits, a clearer density trough at the cutoff, and lower threshold sensitivity.

C Molecular Dynamics Experimental Setup

We evaluate sequence designs using explicit-solvent, all-atom molecular dynamics (MD), with separate physical protocols for the mdCATH domains and ubiquitin in an effort to reproduce the setups of their origin.

C.1 mdCATH Setup

Systems and structure preparation. We simulate selected designs and controls for the three mdCATH domains `1divA02`, `1kwgA03`, and `2v0cA03`. Each structure is initialized from its own MSA-conditioned ESMFold2 structure prediction (Candido et al., 2026). PDBFixer repairs missing atoms in existing residues without inserting missing residue segments. We use MoleculeKit's `systemPrepare` to assign protonation states at pH 7 and optimize the hydrogen-bond network. The N- and C-termini are capped with ACE and CT3 (N-methylamide), respectively. Native disulfide bonds are retained only when both residues in the mapped native pair remain cysteines.

Each protein is solvated in a periodic cubic TIP3P water box with at least 9 Å of protein padding and an edge length below 100 Å. Na^+ and Cl^- ions are added to neutralize the system and target an NaCl concentration of 0.150 M.

Force field and integration settings. Simulations use ACEMD 4.1.4 with an OpenMM 8.3.1 backend (Eastman et al., 2024), running on CUDA in mixed precision for efficiency. We use the CHARMM22* protein parameter set with TIP3P water and ions.

We use Langevin dynamics with an integration time step of 4 fs and a friction coefficient of $0.1\,\mathrm{ps}^{-1}$. Bonds involving hydrogen are constrained, water molecules are rigid, and the integrator constraint tolerance is 10^{-5}. Hydrogen mass repartitioning is applied with hydrogen masses of 4 amu, including water hydrogens. Electrostatics are treated using particle-mesh Ewald (PME) with a real-space cutoff of 9 Å, and van der Waals interactions are switched between 7.5 and 9 Å.

Equilibration and production. Following system construction, we perform up to 5,000 energy minimization steps. Each structure then undergoes 10 ns of restrained equilibration in the isothermal–isobaric (NPT) ensemble at 300 K and 1 bar. During this stage, overlapping positional restraints are applied to protein Cα atoms and all protein heavy atoms, with restraint strengths of $1.0\,\mathrm{kcal\,mol^{-1}\,\AA^{-2}}$ and $0.1\,\mathrm{kcal\,mol^{-1}\,\AA^{-2}}$, respectively; C$\alpha$ atoms therefore belong to both restraint groups. A further 10 ns of unrestrained NPT equilibration uses the same temperature and pressure. An isotropic Monte Carlo barostat is active during both equilibration stages.

We equilibrate each structure once. Its production replicas share the resulting coordinates, periodic box, and saved velocities, but use distinct Langevin random seeds, similar to mdCATH. Production is performed at 320 K in the constant-volume (NVT) ensemble, with the barostat disabled and no positional restraints. Our experiments comprise three 400 ns production replicas, saving coordinates and forces every 250,000 integration steps, i.e. 1 ns intervals.

C.2 Ubiquitin Setup

Reference trajectory and target event. We use the archived 100 ns trajectory of unmodified ubiquitin provided by Yovanno et al. (2023), initialized from the canonical "Major" conformation, as the native reference. The event is defined by the virtual Cβ distance between residue positions 8 and 36 falling below 10 Å, denoted L8–I36 according to the native sequence. These residue positions and the distance threshold remain fixed when evaluating mutated sequences. We use the target band $B = [0.6429,\ 0.7429]$.

The archived native trajectory was generated using NAMD. Because the available archive does not include complete simulation inputs, full-system restarts, water/box/ion records, or exact parameter files, our validation protocol is a local reconstruction rather than an exact reproduction of the source simulation. The integration and equilibration settings below describe our executed validation runs. The source Major trajectory supplies contextual native occupancy for defining the state event.

Systems and structure preparation. We validate the selected Ensemble-IF design with guidance strength $\eta = 4.5$ and the R74P Conformational Biasing (CB) control. As in the mdCATH case, each sequence has its own MSA-conditioned ESMFold2/ESMC predicted starting structure and is prepared and equilibrated separately. PDBFixer adds missing atoms in existing residues, including the terminal OXT atom, without inserting missing residue segments. OpenMM adds hydrogens at pH 7. His68, when retained, is assigned the neutral δ-protonated tautomer, while introduced histidines receive neutral tautomers according to OpenMM's pH-7 geometry-based procedure. Introduced cysteines are treated as reduced, and both termini are charged and uncapped.

Both systems are solvated in initially cubic 7.0 nm boxes of CHARMM-modified TIP3P water, with Na^+ and Cl^- ions added for neutralization and 0.15 M added salt. The Ensemble-IF system contains 10,822 waters, 29 Na^+ ions, and 29 Cl^- ions while the CB system contains 10,845 waters, 31 Na^+ ions, and 30 Cl^- ions. The final box vectors after NPT equilibration are recorded separately for each system.

Force field and integration settings. Simulations use OpenMM `8.3.1.dev-6e13f13` (Eastman et al., 2024) on CUDA in mixed precision. We use `LangevinMiddleIntegrator` with a 2 fs time step, a friction coefficient of $1\,\mathrm{ps}^{-1}$, and a temperature of 310 K throughout equilibration and production.

Table 2: Confidence and structural agreement of selected starting folds before MD.

Shift	Method	pLDDT ↑			scRMSD (Å) ↓		
		1divA02	1kwgA03	2v0cA03	1divA02	1kwgA03	2v0cA03
-	ProteinMPNN	77.87	76.29	84.31	0.61	0.94	0.68
Positive shift	CB	73.10	76.40	46.74	0.64	0.80	9.77
	ENSEMBLE-IF	76.45	79.31	74.14	0.63	1.00	1.39
Negative shift	CB	74.96	77.94	47.40	0.62	0.77	10.77
	ENSEMBLE-IF	72.47	75.76	57.17	2.47	1.46	1.61

Bonds involving hydrogen are constrained, water molecules are rigid, and the integrator constraint tolerance is 10^{-5}. Hydrogen mass repartitioning is not used. PME electrostatics use a 1.2 nm real-space cutoff and an Ewald error tolerance of 5×10^{-4}. Lennard–Jones interactions are switched between 1.0 and 1.2 nm. A Monte Carlo barostat maintains 1 atm during NPT equilibration and production, attempting volume moves every 25 integration steps. However, it is disabled during NVT equilibration.

Equilibration and production. Each structure undergoes energy minimization with a maximum of 10,000 iterations, followed by 0.2 ns of restrained NVT equilibration and 1.0 ns of restrained NPT equilibration. Both dynamics stages are performed at 310 K, with the NPT pressure set to 1 atm. During both stages, source-protein heavy atoms are restrained to their starting coordinates using a harmonic potential $\frac{1}{2}kr^2$ per restrained atom, where r is its displacement and $k = 1{,}000\,\mathrm{kJ\,mol^{-1}\,nm^{-2}}$. Velocities are drawn after minimization and projected onto the constraints before NVT equilibration. The subsequent NPT stage continues from the NVT positions and velocities.

Each structure is equilibrated once for this total of 1.2 ns. All production replicas for that sequence start from its common equilibrated full-system positions and periodic box, with positional restraints removed. Unlike the mdCATH protocol, each ubiquitin production replica receives freshly drawn 310 K velocities. Distinct seeds are used for velocity initialization, Langevin dynamics, and barostat moves.

We run three 100 ns production replicas for the selected ENSEMBLE-IF design and four 100 ns replicas for CB R74P. The fourth CB replica uses the same fixed starting fold and equilibration endpoint as its first three replicas. Production is performed in the NPT ensemble at 310 K and 1 atm. Protein coordinates and thermodynamic quantities are saved every 5,000 integration steps, corresponding to 10 ps. Each replica therefore contributes 10,000 production frames, from 0.01 through 100.00 ns.

D FOLDING QUALITY

Here, we provide details regarding the folding quality of the designs generated by CB and ENSEMBLE-IF as well as ProteinMPNN, acting as a control. ProteinMPNN is known for its ability to predict sequences with high foldability (Watson et al., 2023), establishing itself as a mainstay of *de novo* design. This is typically measured in terms of predicted local distance difference test (pLDDT) scores (Jumper et al., 2021) and self-consistency root mean squared deviation (scRMSD), which is measured in angstroms. pLDDT generally measures folding confidence by a structure prediction oracle while scRMSD is a re-foldability metric that measures how well an oracle's predicted structure aligns with the original backbone. In our experiments, we use ESMFold2 (Candido et al., 2026) as our oracle.

In Table 2, we see that ProteinMPNN given the deposited static backbone structure naturally achieves the highest pLDDT and lowest scRMSD metrics on average, highlighting its strength as an inverse folding model. On the other hand, the multi-state and ensemble-level design models exhibit generally lower metrics given that they are intended to influence ensemble populations rather than foldability into a single structure. That said, both CB and ENSEMBLE-IF generally achieve good folding metrics across the board besides `2v0cA03`, implying that their outputs are able to not

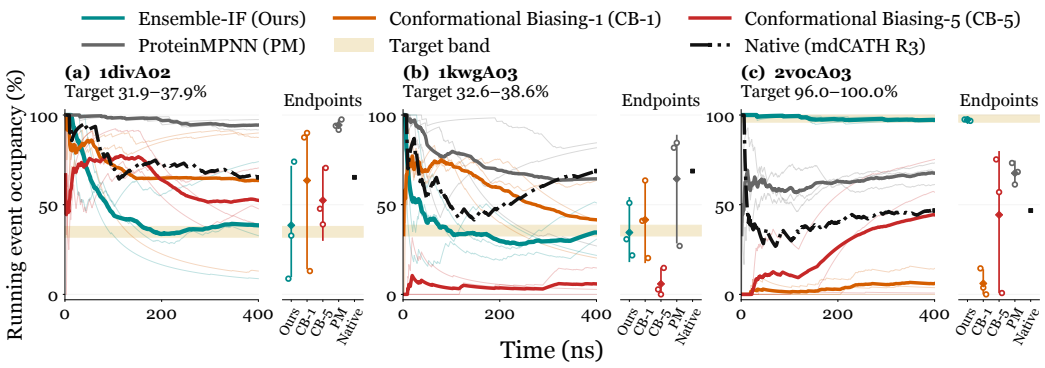


Figure 6: Running event occupancy comparing CB with 5 rounds of mutations (CB-5) versus 1 (CB-1). Results are mixed, being directionally correct similar to CB-1, although still falling short of the target band.

only shape ensemble populations but also produce usable designs in most *de novo* workflows, an important application of protein design.

E Adding Mutations for Conformational Biasing

The main results in this work use CB using a single round of scoring and BioEmu validation. However, this need not be the case. We also experimented with CB using multiple rounds of mutation proposal to accumulate mutational effects for better population control. Figure 6 depicts results on the mdCATH systems compared alongside with CB using a single mutation. We denote CB with 5 mutations as CB-5 and CB with a single mutation as CB-1. We see that CB-5 is able to make improvements towards the bands in `1divA02` and `2v0cA03`, although the benefits are limited, with CB-5 never reaching the target band at all and losing to CB-1 for `1kwgA03`. We hypothesize that this uneven performance reflects the indirect relationship between CB's relative ProteinMPNN scores on representative structures and mutation-induced changes in ensemble occupancy. While BioEmu screening explicitly considers the target interval, repeated structural biasing may not reliably control the magnitude of the population shift realized in MD, resulting in insufficient population shifts.