TOWARDS *de novo* ENZYME DESIGN WITH MULTISTATE GIBBS SAMPLING

Anonymous authors
Paper under double-blind review

ABSTRACT

Enzyme catalysis is fundamentally multistate, yet most computational design methods optimize proteins around a single catalytic geometry. We introduce *Multistate Gibbs Sampling* (MUGS), a framework for designing one sequence to support multiple catalytic states by alternating between motif conditioned structure prediction and shared sequence updates. We find that Gibbs sampling improves atomic motif scaffolding in the single state setting, then extend to multiple states by curating an initial benchmark of experimentally characterized enzymes (MAME) and validate on a five state serine hydrolase catalytic cycle. Across all, MUGS refines and rescues existing designs and can also generate multistate compatible sequences *de novo*. These results suggest a path toward enzyme design that treats catalytic trajectories, rather than isolated states, as the primary object of design.

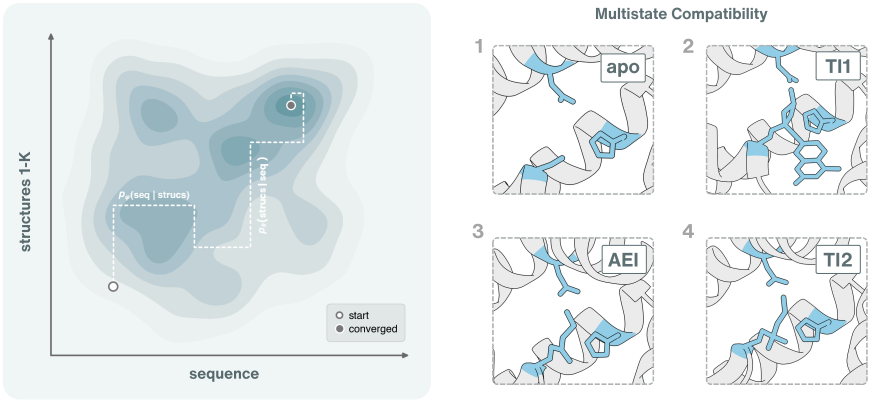


Figure 1: Gibbs sampling for multistate enzyme design.

1 INTRODUCTION

The ability to catalyze arbitrary desired chemical reactions, even those unseen in nature, remains a longstanding challenge in chemistry. *De novo* enzyme design has long been viewed as a path toward such *programmable catalysis* (Lovelock et al., 2022), leveraging the extraordinary selectivity and efficiency with which enzymes carry out chemical transformations while casting this challenge to that of protein design. Such capability would have broad potential for sustainable chemical manufacturing, pharmaceutical synthesis, and even unlocking novel chemistry (Kissman et al., 2024; Reetz et al., 2024; Garcia-Borras et al., 2017).

Computational approaches to enzyme design traditionally start from a *theozyme*, a minimal arrangement of catalytic functional groups for a single state of the desired reaction (Tantillo et al., 1998). Early approaches searched over existing protein backbones for sites capable of accommodating this arrangement (Röthlisberger et al., 2008; Jiang et al., 2008; Siegel et al., 2010). More recent generative models (e.g., RFdiffusion2) instead design entirely new protein structures around the prescribed catalytic atoms, termed *atomic motif scaffolding* (Ahern et al., 2025). These approaches have since then been extended to sequence-structure co-design and to reactive intermediate conditioning (Didi

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.

et al., 2026; Rector-Brooks et al., 2026), yet enzyme design remains predominantly formulated around a single catalytic state.

Catalysis, however, is inherently multistate. Many enzymes proceed through multistep mechanisms, traversing a reaction coordinate along which the same active site must accommodate distinct substrates, transition states, covalent intermediates, and products (Smith et al., 2008; Riziotis et al., 2022). This makes active site preorganization particularly challenging: compatibility with one catalytic state does not guarantee compatibility with another. For serine hydrolase design, Lauko et al. (2025) introduced a single state scaffolding, multistate filtering strategy, using RFdiffusion to generate proteins around one geometry and PLACER to evaluate compatibility with remaining states along the catalytic cycle. Selecting for multistate compatibility substantially improved experimental success. However, this generate-then-filter paradigm seems to require substantial sampling to identify designs that happen to accommodate all states when generation is specified for only one. We instead ask whether multistate context can be treated as a first class component of generation itself.

We propose *Multistate Gibbs Sampling* (MUGS), a framework for designing or refining a single protein sequence against multiple catalytic states simultaneously. MUGS alternates between sampling motif conditioned structures for each state and updating a shared sequence conditioned on those structures. To enable the structure step, we train a motif conditioned structure predictor that folds proteins around atomic motifs. We find that Gibbs sampling improves motif recovery even in the single state setting, and that MUGS extends this to satisfaction of multiple states. For evaluation, we curate an initial Multistate Atomic Motif Enzyme (MAME) benchmark by identifying homologous experimental structures that capture distinct states along multistep catalytic cycles. Across these examples, MUGS can rescue and refine failed designs from existing enzyme design methods, or design without a starting sequence. To our knowledge, this is the first *de novo* enzyme design framework to incorporate multiple catalytic states directly during generation.

2 METHOD

2.1 GIBBS SAMPLING FOR PROTEIN DESIGN

Let $\mathbf{s}$ denote a protein sequence, $\mathbf{x}$ its structure, potentially in complex with a binding partner or ligand, and $\mathbf{c}$ the condition specifying the design task. Ideally, protein design would sample directly from the joint distribution,

$$p\left(\mathbf{s}, \mathbf{x} \mid \mathbf{c}\right) \tag{1}$$

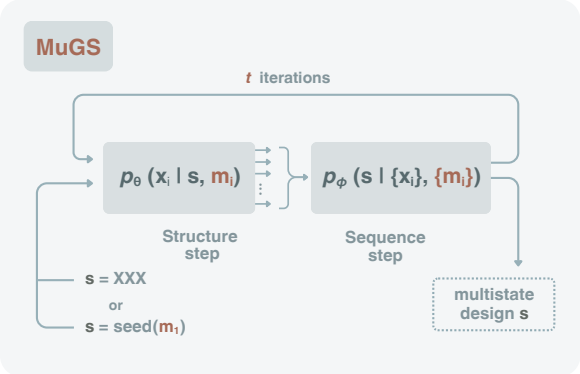
Unfortunately, direct sampling from the true joint remains intractable. Modern protein models, however, provide approximate access to its reciprocal conditionals: structure predictors model $p_\theta(\mathbf{x} \mid \mathbf{s}, \mathbf{c})$, while inverse-folding models model $p_\phi(\mathbf{s} \mid \mathbf{x}, \mathbf{c})$. This suggests a natural Gibbs style sampling procedure (Geman and Geman, 1984; Casella and George, 1992), alternating between draws from the structure and sequence conditionals.

Indeed, several recent protein design methods can be interpreted through this lens. ProteinHunter and HalluDesign cycle between structure prediction and sequence redesign for binder design (Cho et al., 2025; Fang et al., 2025); under this formulation, $\mathbf{c}$ would correspond to the specified binding target. Fry et al. (2026) propose NISE, which similarly alternates protein–ligand co-structure prediction with sequence design using LASErMPNN, with $\mathbf{c}$ corresponding to the target small molecule. Although introduced under different terminology, these works also make use of alternations between reciprocal conditionals.

Multistate Gibbs Sampling. For multistate enzyme design, we instead consider a joint distribution over all structural states and their shared sequence,

$$p\left(\mathbf{s}, \{\mathbf{x}_i\}_{i=1}^{K} \mid \{\mathbf{m}_i\}_{i=1}^{K}\right) \tag{2}$$

where $\mathbf{x}_i$ and $\mathbf{m}_i$ correspond to the structure and catalytic motif at state i, respectively. Here, we define each motif by its relevant catalytic atoms together with the state specific reacting species, allowing them to differ in ligand identity, covalent connectivity, and geometry across the reaction.



Algorithm 1: Multistate Gibbs Sampling

```
# motifs {m_i}_{i=1}^K
# iterations T
if s^(0) is unset:
  s^(0) ← seed(m_1)
for t in range(T):
  for i in range(1, K+1):
    x_i^(t+1) ~ p_θ(· | s^(t), m_i)
  s^(t+1) ~ p_φ(· | {x_i^(t+1)}_{i=1}^K, {m_i}_{i=1}^K)
return s^(T)   # or best sample
```

Figure 2: Overview of MUGS alternation of structure and sequence conditionals.

Structure step. We approximate the multistate structure conditional by sampling each state separately given the shared sequence and its corresponding motif,

$$p_\theta\left(\{\mathbf{x}_i\}_{i=1}^K \mid \mathbf{s}, \{\mathbf{m}_i\}_{i=1}^K\right) = \prod_{i=1}^K p_\theta\left(\mathbf{x}_i \mid \mathbf{s}, \mathbf{m}_i\right) \tag{3}$$

Sequence step. The sequence conditional couples the states through a shared sequence,

$$p_\phi\left(\mathbf{s} \mid \{\mathbf{x}_i\}_{i=1}^K, \{\mathbf{m}_i\}_{i=1}^K\right) \tag{4}$$

We instantiate these conditionals using atomic motif conditioned structure prediction and tied inverse folding, respectively, as described in the following sections (2.2, 2.3). The starting sequence can be initialized in a `seed(·)` step from any design method, including RFdiffusion (Watson et al., 2023), and subsequently refined for multistate compatibility. If no starting design is provided, the seed can instead be hallucinated (Cho et al., 2025). Together, alternating samples from these conditionals defines an approximate blocked Gibbs sampling procedure, which we call MUGS (Fig. 2; Algorithm 1).

2.2 MOTIF-CONDITIONED STRUCTURE PREDICTION

The structure conditional $p_\theta(\mathbf{x} \mid \mathbf{s}, \mathbf{m})$ requires a model that folds a specified sequence while conditioning on an atomic motif. Standard cofolding models approximate $p(\mathbf{x} \mid \mathbf{s})$; some support structural templates, but we show that these do not directly encode catalytic side chain geometry together with reacting species (Appendix A.4). Conversely, atomic motif scaffolders model $p(\mathbf{x} \mid \mathbf{m})$, generating scaffolds around a motif rather than folding a given sequence. We therefore train MCFOLD for both sequence and motif conditioning, $p_\theta(\mathbf{x} \mid \mathbf{s}, \mathbf{m})$.

We initialize MCFOLD from a pretrained AlphaFold3-style cofolding model (Abramson et al., 2024; Wohlwend et al., 2025) and fine tune it on atomic motifs randomly sampled from native training structures (Algorithm 4). Each motif is represented by centered atomic coordinates $\mathbf{x}_{\text{motif}}$ and an atom mask $\mathbf{a}_{\text{motif}}$. These features are injected into the atom representations of the input encoder, diffusion encoder, and decoder,

$$\mathbf{h} \leftarrow \mathbf{h} + f_{\text{coord}}(\mathbf{x}_{\text{motif}}) + f_{\text{mask}}(\mathbf{a}_{\text{motif}})$$

At inference, we enforce the prescribed motif geometry after each denoising step by rigidly aligning and replacing the predicted motif coordinates, analogous to masked replacement strategies used for diffusion inpainting (Lugmayr et al., 2022). Ablations show that motif conditioning improves placement and coordinate replacement is necessary to enforce exact geometry without degrading structure quality (Appendix A.4).

2.3 MULTISTATE INVERSE FOLDING

For the sequence conditional $p_\phi\left(\mathbf{s} \mid \{\mathbf{x}_i\}_{i=1}^K, \{\mathbf{m}_i\}_{i=1}^K\right)$, we use tied LigandMPNN (Dauparas et al., 2022). During autoregressive decoding, corresponding residue positions are tied across states: LigandMPNN computes an amino acid distribution for each state, averages these distributions, and

samples a single residue identity shared by all states. This provides a direct approximation to the shared sequence conditional while retaining each state's ligand and atomic context. Tied MPNN variants have previously been used for multistate sequence design, including two state hinge proteins (Praetorius et al., 2023) and general multistate redesign in SWITCHCRAFT (Jing et al., 2026).

2.4 INFERENCE TIME SCALING

Though MUGS itself can be viewed as scaling at inference time through its repeated Gibbs updates, this compute need not be spent only on longer trajectories (*depth*). The same formulation also allows scaling in *width* by rollouts of multiple trajectories in parallel. We consider two such strategies.

Beam MuGS. Analogous to beam search (Freitag and Al-Onaizan, 2017), we maintain a set of B trajectories, $\mathcal{B}_t = \{(\mathbf{s}_t^{(b)}, \mathbf{x}_t^{(b)})\}_{b=1}^{B}$. Each trajectory is propagated by the MUGS kernel with R_x structure branches and R_s sequence proposals per structure, yielding a candidate set $\mathcal{C}_{t+1}$. Candidates are scored by a multistate objective F and pruned back to beam width B:

$$|\mathcal{C}_{t+1}| = BR_xR_s, \qquad \mathcal{B}_{t+1} = \mathrm{TopB}_F(\mathcal{C}_{t+1})$$

In practice, R_x can correspond to multiple diffusion samples from the structure model, while R_s can correspond to multiple sequence proposals from the inverse folder. The corresponding algorithmic modifications to vanilla MUGS are highlighted in Algorithm 2 and simply amount to adding an additional loop over width at the relevant sampling steps, followed by a reduction step. Notably, with a different update ordering and particular choices of R_x and R_s, Beam MUGS takes a form closely related to NISE (Fry et al., 2026) (see Appendix A.2).

SMC MuGS. If we instead want to preserve stochastic exploration while biasing search toward better multistate solutions, we can use Sequential Monte Carlo (SMC), iterating propagation under the MUGS kernel with importance weighting and resampling according to improvement under the multistate objective. For candidate $c \in \mathcal{C}_{t+1}$ with parent $\pi(c) \in \mathcal{P}_t$,

$$w_{t+1}^{(c)} \propto \exp\left[\beta\left(F_{t+1}^{(c)} - F_t^{(\pi(c))}\right)\right]$$

$$\tilde{c}_1, \ldots, \tilde{c}_B \overset{\text{iid}}{\sim} \mathrm{Cat}\left(\frac{\mathbf{w}_{t+1}}{\sum_c w_{t+1}^{(c)}}\right)$$

where β controls the strength of selection and the categorical resampling draws B trajectories with replacement according to the normalized weights. This is similar in spirit to Feynman–Kac steering (Singhal et al., 2025), where base model trajectories are reweighted and resampled using reward tilting potentials. Refer to Algorithm 3 and additional information provided in Appendix A.2.

In our experiments, since our focus was increasing the number of successful designs, we use F to tilt sampling toward the success criterion. More generally, however, the SMC formulation supports arbitrary steering through any evaluable potential. For example, setting $F = -E(\mathbf{x})$ for a molecular energy function could represent a Boltzmann-style tilt, with weights proportional to $\exp(-\beta E)$ that preferentially retain trajectories moving toward lower energy configurations.

2.5 REACTION COORDINATE FREE ENERGY SKETCHING

Separately, a useful byproduct of training a structure model that conditions jointly on sequence and an atomic motif (Sec. 2.2, MCFold) is the ability to reconstruct protein environments at successive points along a reaction coordinate. This could provide a practical route to approximating the reaction free energy landscape. For a reaction coordinate ξ, the free energy profile reflects how likely the system is to occupy configurations at each point along the reaction,

$$G(\xi) = -k_B T \log p(\xi) + C$$

so frequently sampled states have lower free energy, while rarely sampled regions form barriers. In particular, the activation barrier $\Delta G^{\ddagger} = G(\xi^{\ddagger}) - G(\xi_{\mathrm{R}})$ is a key determinant of catalytic rate, while the relative heights of competing pathways can distinguish mechanisms and explain changes in activity or selectivity. Such profiles have therefore become a central computational tool for identifying

enzyme reaction mechanisms, locating rate-limiting barriers, and understanding how the protein environment stabilizes catalytic states (Gao et al., 2006; Warshel and Florián, 1998).

Computationally, estimating $G(\xi)$ requires two ingredients: an energy model that assigns energies and forces to each molecular configuration, and sufficient sampling of configurations along the reaction coordinate. A standard energy model for enzymatic reactions is QM/MM, which treats the reacting region quantum mechanically while modeling the surrounding protein and solvent classically,

$$U(\mathbf{x}) = E_{\mathrm{QM}}(\mathbf{x}_Q) + U_{\mathrm{MM}}(\mathbf{x}_M) + U_{\mathrm{QM/MM}}(\mathbf{x}_Q, \mathbf{x}_M), \qquad \mathbf{F}(\mathbf{x}) = -\nabla_{\mathbf{x}} U(\mathbf{x})$$

Molecular dynamics then uses these forces to propagate the system and sample configurations along the reaction coordinate, typically with enhanced sampling to access rare reactive states (Hu and Yang, 2008; Senn and Thiel, 2009; Bernardi et al., 2015). Because QM force evaluations are expensive, recent work has used learned potentials to accelerate reactive simulations while retaining explicit trajectory sampling (Pan et al., 2021; Gao et al., 2026).

MCFOLD provides a complementary route. Given catalytic geometries $\{\mathbf{m}(\xi_i)\}_{i=1}^{K}$ along a reaction coordinate ξ, we independently reconstruct a sequence-specific protein environment for each state,

$$\mathbf{x}_i \sim p_\theta(\mathbf{x} \mid \mathbf{s}, \mathbf{m}(\xi_i))$$

Rather than explicitly propagating the protein environment through the full reaction trajectory, these predicted environments can be physically scored at each state to obtain an approximate profile $\{\widehat{G}(\xi_i)\}_{i=1}^{K}$. In practice, a small amount of local physical relaxation may be required before scoring. We refer to the resulting approximate profile as a *reaction free energy sketch*. This construction is orthogonal to the choice of energy model and can therefore be combined with either QM/MM or learned potentials.

3 RESULTS

3.1 SINGLE STATE SCAFFOLDING

We first ask whether Gibbs sampling improves single state atomic motif scaffolding on the Atomic Motif Enzyme (AME) benchmark introduced by RFdiffusion2 (RFD2). For each RFD2 backbone, we use only one of its eight LigandMPNN sequences as the MUGS seed, including both benchmark successes and failures. We also evaluate *de novo* MUGS, initialized from an all-X sequence with only motif residues fixed. Designs ($n = 100$ per motif) are evaluated using the same criterion as Ahern et al. (2025): refold motif RMSD < 1.5 Å and no ligand backbone clash (Fig. 3).

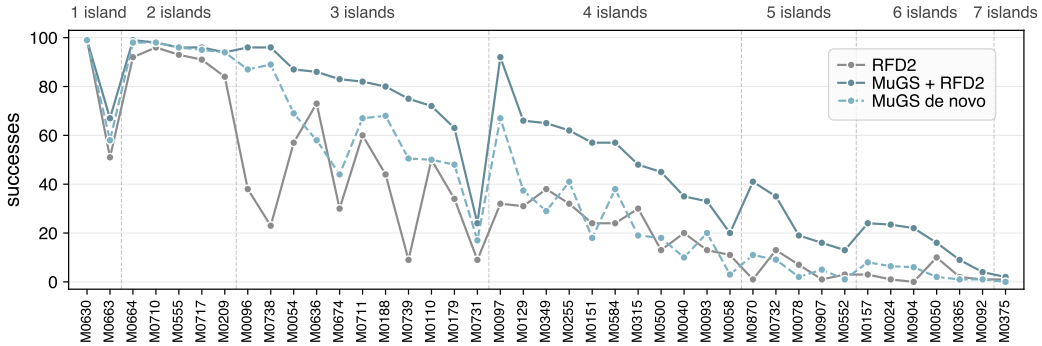


Figure 3: Single state atomic motif scaffolding on the AME benchmark (Ahern et al., 2025).

MUGS rescues 1,029 failing RFD2 designs and improves motif RMSD on 68% of paired designs, matching or exceeding RFD2 on all 41 motifs. Gains are largest for more fragmented active sites. Without a sequence seed, de novo MUGS still outperforms RFD2 on 27 of 41 motifs and remains competitive on simpler, contiguous sites.

Because MUGS also increases the number of structure samples, we ask whether its gains can be explained by additional folding alone. We therefore repeatedly refold each RFD2 sequence using

the same structure-sampling budget as the MuGS trajectory. Best-of-N sampling improves motif RMSD only modestly and never reaches the 1.5 Å threshold, supporting that the gains come from the Gibbs sampling procedure rather than the additional folds.

3.2 Multistate refinement

MAME curation. We construct the Multistate Atomic Motif Enzyme (MAME) set as an extension of the RFdiffusion2 AME benchmark, itself based on M-CSA (Ribeiro et al., 2018) catalytic annotations. We search the PDB and associated literature for additional experimentally determined states of the same reactions, yielding 5 candidate motif sets. For each example, we retain the catalytic residues annotated in the corresponding M-CSA entry and restrict the initial set to three states. These states span distinct points along the catalytic process, including apo enzymes, substrate or cofactor bound complexes, product states, and experimentally trapped covalent intermediates. Full curation details are provided in Appendix B.1.

Multistate evaluation. For each motif, we start from all 100 RFD2 designs, including both successes and failures on the original AME benchmark, and run MuGS with up to $K = 3$ states provided during Gibbs sampling, using a single state subset for each K and prioritizing reactive-intermediate states when available. This tests whether MuGS can make initially single state compatible designs multistate compatible, or even rescue designs that fail the original single-state criterion into solutions that succeed across all states. Evaluation is always performed against all three states: each designed sequence is refolded separately in every state, functional motif tip RMSD is measured after alignment on the catalytic motif, and a design succeeds only if all three refolded states pass the threshold.

Table 1: Multistate AME recovery. Threshold must be met across all states.

Method	< 1.5 Å	< 1.75 Å	< 2.0 Å
RFD2	2	8	20
MuGS $K = 1$	2	12	32
MuGS $K = 2$	10	51	114
MuGS $K = 3$	17	59	127

Conditioning on multiple states consistently improves multistate compatibility (Table 1, Fig. 9). Much of the gain appears from $K = 1$ to $K = 2$, while the value of a third state is motif dependent, suggesting that different catalytic states may have different relative importance for producing a compatible sequence. Across all motifs, however, the central result is consistent: explicitly designing across multiple states yields more sequences that satisfy every experimental state than single state design, supporting MuGS as a direct route toward designing proteins for catalytic trajectories.

Inference time scaling. As discussed in Section 2.4, MuGS allows inference time compute to be allocated not only along a single trajectory, but also across multiple particles that are propagated and pruned in parallel using either Beam or SMC updates. To test how this alternative allocation of compute affects multistate recovery, we focus on M0054 from MAME, where vanilla MuGS produced only one multistate success across 100 starting RFD2 designs. We initialize Beam MuGS with $B = 4$, $R_x = 1$, and $R_s = 2$, and use the same population configuration for SMC MuGS ($\beta = 1$), with the two methods differing only in how candidates are selected for the next iteration. To eliminate variation from sequence initialization and ensure comparison, Beam, SMC, and vanilla MuGS are initialized from the exact same starting backbones and seed sequences in the same order. The two search variants are compute matched and evaluated using the same multistate success criterion as above, requiring motif RMSD < 1.5 Å in all three states.

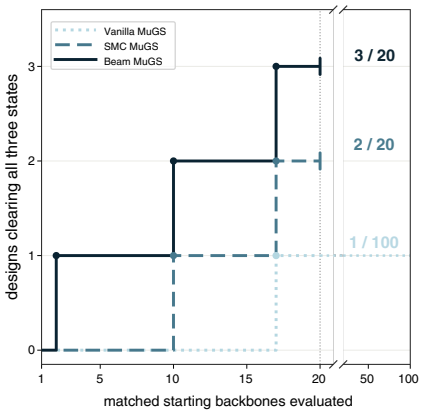


Figure 4: Comparison of MuGS inference time scaling variants on M0054.

Beam MuGS reaches its first successful design by the second starting backbone, a second by the tenth, and a third by the seventeenth, yielding three successes within the first 20 starting designs

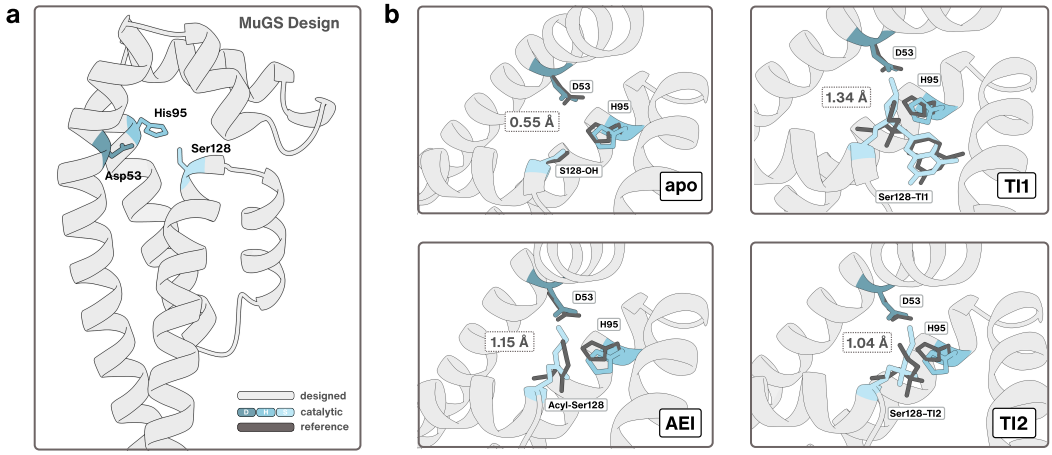


Figure 5: **MuGS serine hydrolase design**. **(a)** MuGS *de novo* design of serine hydrolase using motif states derived from Lauko et al. (2025). **(b)** Fold compatibility of MuGS design for relevant states along reaction coordinate.

(Fig. 4). SMC reaches two successes within the same 20 designs, while vanilla MuGS reaches only one. Moreover, the successful starting designs correspond, with the vanilla success also recovered by SMC and both SMC successes recovered by Beam, suggesting that the scaling progressively rescues additional starting designs rather than simply identifying a different subset of solutions. Notably, both search variants recover more within 20 starting designs than vanilla MuGS does across all 100 RFD2 designs, where only a single design succeeds. This suggests that allocating the inference compute to more targeted exploration within each starting design could potentially be more effective than distributing it across a larger number of independent starts.

3.3 *De novo* DESIGN OF MULTISTATE SERINE HYDROLASE

Given MuGS's success in multistate refinement, we next ask whether it can design *de novo* across a complete catalytic cycle. The experimentally validated serine hydrolase from Lauko et al. (2025) provides a natural test bed for this setting. The reaction coordinate spans apo and substrate-bound states, followed by two tetrahedral intermediates (TI1 and TI2) separated by an acyl-enzyme intermediate (AEI) (Appendix B.2). We use the best representative TI1, AEI, and TI2 PLACER geometries as motifs and run MuGS *de novo* sequences with all $K = 1, 2, 3$ combinations ($n = 100$ each). Each design is refolded in all five states and scored by its worst state motif RMSD. As a benchmark for success, we use the released wet lab validated serine hydrolase from Lauko et al. (2025), which achieves a worst state RMSD of 1.59 Å across the five states; this threshold is indicated by the bronze line in Fig. 5.

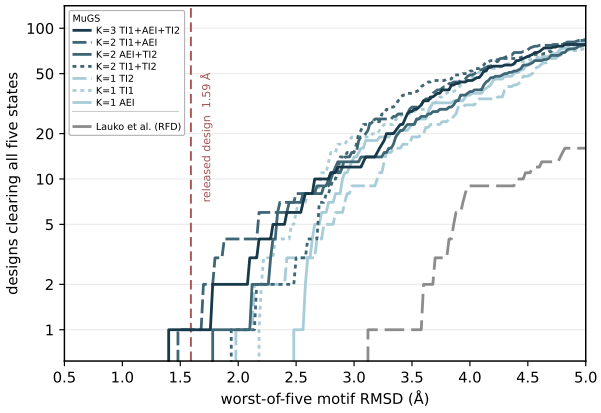


Figure 6: Serine hydrolase design motif RMSD traces.

MuGS markedly shifts the distribution toward designs compatible with the full catalytic cycle, while the Lauko RFD pipeline generates no design better than 3.0 Å within 100 samples (Fig. 6). The fact that Lauko et al. (2025) ultimately identifies a wet lab validated design reaching the 1.59 Å threshold suggests that substantially broader sampling was needed in the single state pipeline. In contrast, both $K = 2$ and $K = 3$ MuGS recover designs reaching this threshold within $n = 100$ *de novo* samples. Figure 5 highlights the top MuGS design and its compatibility across four states.

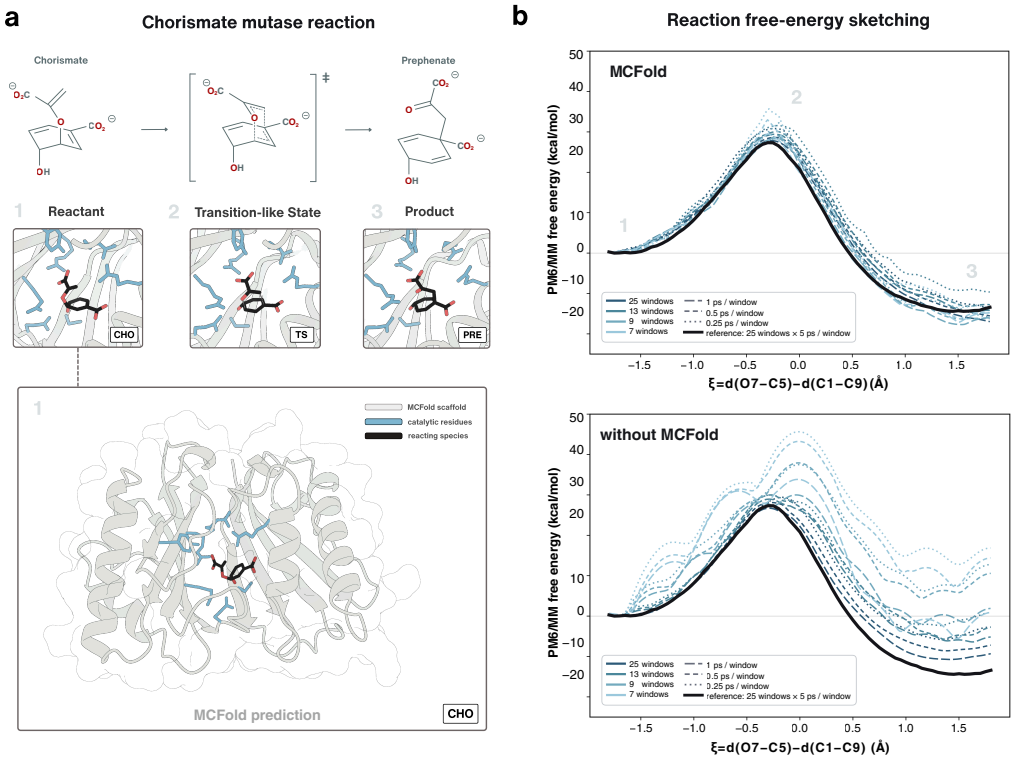


Figure 7: **Reaction free energy sketching with MCFOLD. (a)** Chorismate mutase reaction mechanism (top; adapted from Lee et al., 1995) and MCFOLD reconstructions of reactant, transition-state-like, and product environments, with the full reconstructed reactant state shown below. **(b)** Reaction free energy sketches with MCFOLD (top) or short QM/MM MD alone (bottom), compared with the 125 ps reference profile (black). MCFOLD recapitulates the PM6/MM free energy landscape, whereas naively shortening QM/MM MD does not.

3.4 REACTION FREE ENERGY SKETCHING OF CHORISMATE MUTASE

To evaluate reaction MCFold free energy sketching, we use *Bacillus subtilis* chorismate mutase (BsCM), a well characterized model enzyme for studying catalysis of a concerted pericyclic reaction (Chook et al., 1994; Lee et al., 1995). BsCM converts chorismate to prephenate through a Claisen rearrangement (Fig. 7a; reaction mechanism adapted from Lee et al., 1995), in which the C5–O7 bond is broken as the C1–C9 bond is formed. We therefore describe progress along the reaction by

$$\xi = d(\mathrm{O7}, \mathrm{C5}) - d(\mathrm{C1}, \mathrm{C9}),$$

which varies continuously from reactant to product. The active site comprises seven catalytic residues, Arg7, Arg63, Cys75, Glu78, Arg90, Tyr108, and Arg116, as annotated in M-CSA entry 474 (Ribeiro et al., 2018). These residues form a hydrogen bonding and electrostatic environment that constrains chorismate in a reactive conformation and stabilizes the transition state.

To construct a reaction free energy sketch, we provide MCFOLD with the wild-type sequence and catalytic geometry at each ξ_i, and independently reconstruct the surrounding protein environment, This allows MCFOLD to directly interpolate protein environments accommodating each catalytic geometry across the reaction coordinate, rather than relying on MD to stochastically and sequentially explore these configurations. Each reconstructed state is then subjected to a short restrained QM/MM simulation for local relaxation, and the resulting windows are combined into an approximate free energy profile.

$$\mathbf{m}(\xi_i),\, \mathbf{s} \xrightarrow{\text{MCFOLD}} \mathbf{x}_i \xrightarrow{\text{relax / score}} \widehat{G}(\xi_i)$$

We note that MCFOLD is agnostic to the downstream energy model, so more accurate free energy estimates can be obtained by pairing the reconstructed states with higher-level QM/MM or learned

potentials. For this case study, we use PM6/MM as a computationally tractable testbed, with a 25 window, 5 ps per window reference profile (125 ps total; Fig. 7b, black); while semiempirical QM can introduce error in absolute barrier heights, it allows us to evaluate the reconstruction procedure consistently across many sampling conditions.

Using MCFOLD, we generate reaction free energy sketches across progressively reduced QM/MM relaxation budgets (Fig. 7b, top). Even at the most aggressive setting, with just 0.25 ps of relaxation across seven reconstructed states, the sketch retains a Pearson correlation of $r = 0.99$ with the reference. This amounts to only 1.75 ps of QM/MM sampling in total, $71\times$ less than the reference, almost two orders of magnitude reduction. Representative MCFOLD reconstructions of the reactant, transition, and product states are shown in Fig. 7a.

To test whether reduced sampling alone would be sufficient, we also construct free energy sketches using short QM/MM MD without MCFOLD, progressively reducing the total sampling for each simulation (Fig. 7b, bottom). As the sampling decreases, these short-MD sketches increasingly deviate from the reference, and most profiles incorrectly place the product above the reactant.

4 CONCLUSION

We present MUGS, to our knowledge the first enzyme design framework to model multiple catalytic states directly during generation. Across single state scaffolding, experimentally curated multistate motifs, and a five state serine hydrolase cycle, MUGS improves multistate compatibility and recovers solutions from both existing designs and fully unspecified sequences. Beyond design, we find that motif conditioned structure prediction enables reaction free energy sketching by reconstructing protein environments along the coordinate with substantially reduced physical sampling. Future work can expand our MAME benchmark through additional homologous structure mining or by generating motifs with DFT and molecular potentials, while also developing methods that translate reaction mechanisms directly into state specific catalytic geometries. Ultimately, we believe *programmable catalysis* will require designing proteins not for isolated structures, but for the sequence of chemical states that gives rise to the complex functions found in nature, and even those absent from it. We hope MUGS provides a step in that direction.

AI USE STATEMENT

In this work, we used generative AI tools to assist with code and experiment setup, scripts for dataset curation, plotting code, and language editing and refinement of the manuscript. We did not use generative AI to interpret experimental results, determine the organization of the paper, or create the final figures. All research questions, methodological ideas, and experimental directions were conceived by us; generative AI was used only to discuss and further develop ideas after they had been formulated. We independently conducted the literature review and assessed the relevance and novelty of the research problem. We reviewed all AI assisted work: generated code was inspected and tested for correctness, collected data were checked against their original sources, plotting scripts and outputs were verified against the underlying results, and all manuscript edits were reviewed for factual accuracy. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

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.

Woody Ahern, Jason Yim, Doug Tischer, Saman Salike, Seth M Woodbury, Donghyo Kim, Indrek Kalvet, Yakov Kipnis, Brian Coventry, Han Raut Altae-Tran, et al. Atom level enzyme active site scaffolding using rfdiffusion2. *bioRxiv*, pages 2025–04, 2025.

Brian J Bahnson, Vernon E Anderson, and Gregory A Petsko. Structural mechanism of enoyl-coa hydratase: three atoms from a single water are added in either an e1cb stepwise or concerted fashion. *Biochemistry*, 41(8):2621–2629, 2002.

Christoph Bannwarth, Sebastian Ehlert, and Stefan Grimme. Gfn2-xtb—an accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. *Journal of chemical theory and computation*, 15 (3):1652–1671, 2019.

Jordi Benach, Sılvia Atrian, Roser Gonzàlez-Duarte, and Rudolf Ladenstein. The refined crystal structure of drosophila lebanonensis alcohol dehydrogenase at 1.9 å resolution. *Journal of molecular biology*, 282(2):383–399, 1998.

Jordi Benach, Sılvia Atrian, Roser Gonzàlez-Duarte, and Rudolf Ladenstein. The catalytic reaction and inhibition mechanism of drosophila alcohol dehydrogenase: observation of an enzyme-bound nad-ketone adduct at 1.4 å resolution by x-ray crystallography. *Journal of molecular biology*, 289 (2):335–355, 1999.

Jordi Benach, Jan-Olof Winberg, John-Sigurd Svendsen, Sílvia Atrian, Roser Gonzàlez-Duarte, and Rudolf Ladenstein. Drosophila alcohol dehydrogenase: acetate–enzyme interactions and novel insights into the effects of electrostatics on catalysis. *Journal of molecular biology*, 345(3): 579–598, 2005.

Rafael C Bernardi, Marcelo CR Melo, and Klaus Schulten. Enhanced sampling techniques in molecular dynamics simulations of biological systems. *Biochimica et Biophysica Acta (BBA)-General Subjects*, 1850(5):872–877, 2015.

George Casella and Edward I George. Explaining the gibbs sampler. *The American Statistician*, 46 (3):167–174, 1992.

Yehlin Cho, Griffin Rangel, Gaurav Bhardwaj, and Sergey Ovchinnikov. Protein hunter: exploiting structure hallucination within diffusion for protein design. *bioRxiv*, pages 2025–10, 2025.

Yuh Min Chook, Joseph V Gray, Hengming Ke, and William N Lipscomb. The monofunctional chorismate mutase from bacillus subtilis: structure determination of chorismate mutase and its complexes with a transition state analog and prephenate, and implications for the mechanism of the enzymatic reaction. *Journal of molecular biology*, 240(5):476–500, 1994.

David Cobessi, Frédérique Tête-Favier, Stéphane Marchal, Saıd Azza, Guy Branlant, and André Aubry. Apo and holo crystal structures of an nadp-dependent aldehyde dehydrogenase from streptococcus mutans. *Journal of molecular biology*, 290(1):161–173, 1999.

David Cobessi, Frédérique Tête-Favier, Stéphane Marchal, Guy Branlant, and André Aubry. Structural and biochemical investigations of the catalytic mechanism of an nadp-dependent aldehyde dehydrogenase from streptococcus mutans. *Journal of molecular biology*, 300(1):141–152, 2000.

Subhadra Dalwani, Pradip Kumar Mondal, Werner Schmitz, Rik K Wierenga, and Petri M Pihko. Enantioselective hydration of non-coa enoyl-thioesters by enoyl-coa hydratase (ech): Activation of the active site oxyanion hole with 3, 5-adenosine-diphosphate enables competent catalysis. *JACS Au*, 6(4):2464, 2026.

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, Zuobai Zhang, Guoqing Zhou, Danny Reidenbach, Zhonglin Cao, Sooyoung Cha, Tomas Geffner, Christian Dallago, Jian Tang, Michael Bronstein, et al. Scaling atomistic protein binder design with generative pretraining and test-time compute. In *International Conference on Learning Representations*, volume 2026, pages 61056–61109, 2026.

Minchao Fang, Chentong Wang, Jungang Shi, Fengbai Lian, Qihan Jin, Zhe Wang, Yanzhe Zhang, Peipei Chen, Zhanyuan Cui, Yanjun Wang, et al. Halludesign: Protein optimization and de novo design via iterative structure hallucination and sequence design. *bioRxiv*, pages 2025–11, 2025.

Markus Freitag and Yaser Al-Onaizan. Beam search strategies for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 56–60, 2017.

Benjamin Fry, Kaia Slaw, and Nicholas F Polizzi. Zero-shot design of drug-binding proteins via neural iterative selection- expansion. *Nature*, pages 1–13, 2026.

Jiali Gao, Shuhua Ma, Dan T Major, Kwangho Nam, Jingzhi Pu, and Donald G Truhlar. Mechanisms and free energies of enzymatic reactions. *Chemical reviews*, 106(8):3188–3209, 2006.

Meng Gao, Armin Shayesteh Zadeh, Aniruddha Seal, Siva Dasetty, Siddarth K Achar, Misko Dzamba, Benjamin K Miller, Leif D Jacobson, C Lawrence Zitnick, Brandon M Wood, et al. Quantum-accurate atomistic modeling of enzyme catalysis using a machine learned potential. *arXiv preprint arXiv:2609.09293*, 2026.

Marc Garcia-Borras, Kendall N Houk, and Gonzalo Jiménez-Osés. Computational design of protein function. 2017.

Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6): 721–741, 1984.

David G Gourley, Annette K Shrive, Igor Polikarpov, Tino Krell, John R Coggins, Alastair R Hawkins, Neil W Isaacs, and Lindsay Sawyer. The two types of 3-dehydroquinase have distinct structures but catalyze the same overall reaction. *Nature structural biology*, 6(6):521–525, 1999.

Hao Hu and Weitao Yang. Free energies of chemical reactions in solution and in enzymes with ab initio quantum mechanics/molecular mechanics methods. *Annu. Rev. Phys. Chem.*, 59(1):573–601, 2008.

Lin Jiang, Eric A Althoff, Fernando R Clemente, Lindsey Doyle, Daniela Rothlisberger, Alexandre Zanghellini, Jasmine L Gallaher, Jamie L Betker, Fujie Tanaka, Carlos F Barbas III, et al. De novo computational design of retro-aldol enzymes. *science*, 319(5868):1387–1391, 2008.

Bowen Jing, Mihir Bafna, Anisha Parsan, Heyuan Michael Ni, David Kwabi-Addo, Bryan Bryson, Adam Klivans, and Bonnie Berger. Switchcraft: A programmatic framework for designing state-switching proteins. *arXiv preprint arXiv:2605.31236*, 2026.

Suhng Wook Kim, Sun-Shin Cha, Hyun-Soo Cho, Jeong-Sun Kim, Nam-Chul Ha, Moon-Ju Cho, Soyoung Joo, Kyeong Kyu Kim, Kwan Yong Choi, and Byung-Ha Oh. High-resolution crystal structures of δ5-3-ketosteroid isomerase with and without a reaction intermediate analogue. *Biochemistry*, 36(46):14030–14036, 1997.

Elijah N Kissman, Max B Sosa, Douglas C Millar, Edward J Koleski, Kershanthen Thevasundaram, and Michelle CY Chang. Expanding chemistry through in vitro and in vivo biocatalysis. *Nature*, 631(8019):37–48, 2024.

Anna Lauko, Samuel J Pellock, Kiera H Sumida, Ivan Anishchenko, David Juergens, Woody Ahern, Jihun Jeung, Alexander F Shida, Andrew Hunt, Indrek Kalvet, et al. Computational design of serine hydrolases. *Science*, 388(6744):eadu2454, 2025.

AY Lee, John D Stewart, Jon Clardy, and Bruce Ganem. New insight into the catalytic mechanism of chorismate mutases from structural studies. *Chemistry & biology*, 2(4):195–203, 1995.

W-H Lee, Luis Augusto Perles, Ronaldo AP Nagem, Annette K Shrive, Alastair Hawkins, Lindsay Sawyer, and Igor Polikarpov. Comparison of different crystal forms of 3-dehydroquinase from salmonella typhi and its implication for the enzyme activity. *Biological Crystallography*, 58(5): 798–804, 2002.

Sarah L Lovelock, Rebecca Crawshaw, Sophie Basler, Colin Levy, David Baker, Donald Hilvert, and Anthony P Green. The road to fully programmable protein catalysis. *Nature*, 606(7912):49–58, 2022.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 11451–11461. IEEE, 2022.

Maria Maneiro, Antonio Peon, Emilio Lence, Jose M Otero, Mark J Van Raaij, Paul Thompson, Alastair R Hawkins, and Concepcion Gonzalez-Bello. Insights into substrate binding and catalysis in bacterial type i dehydroquinase. *Biochemical Journal*, 462(3):415–424, 2014.

Xiaoliang Pan, Junjie Yang, Richard Van, Evgeny Epifanovsky, Junming Ho, Jing Huang, Jingzhi Pu, Ye Mei, Kwangho Nam, and Yihan Shao. Machine learning assisted free energy simulation of solution–phase and enzyme reactions. *Journal of chemical theory and computation*, 17(9):5745, 2021.

Saro Passaro, Gabriele Corso, Jeremy Wohlwend, Mateo Reveiz, Stephan Thaler, Vignesh Ram Somnath, Noah Getz, Tally Portnoi, Julien Roy, Hannes Stark, et al. Boltz-2: Towards accurate and efficient binding affinity prediction. *BioRxiv*, 2025.

Florian Praetorius, Philip J. Y. Leung, Maxx H. Tessmer, Adam Broerman, Cullen Demakis, Acacia F. Dishman, Arvind Pillai, Abbas Idris, David Juergens, Justas Dauparas, Xinting Li, Paul M. Levine, Mila Lamb, Ryanne K. Ballard, Stacey R. Gerben, Hannah Nguyen, Alex Kang, Banumathi Sankaran, Asim K. Bera, Brian F. Volkman, Jeff Nivala, Stefan Stoll, and David Baker. Design of stimulus-responsive two-state hinge proteins. *Science*, 381(6659):754–760, 2023.

Jarrid Rector-Brooks, Théophile Lambert, Marta Skreta, Daniel Roth, Yueming Long, Zi-Qi Li, Xi Zhang, Miruna Cretu, Francesca-Zhoufan Li, Tanvi Ganapathy, et al. General multimodal protein design enables dna-encoding of chemistry. *arXiv preprint arXiv:2604.05181*, 2026.

Manfred T Reetz, Ge Qu, and Zhoutong Sun. Engineered enzymes for the synthesis of pharmaceuticals and other high-value products. *Nature Synthesis*, 3(1):19–32, 2024.

António J M Ribeiro, Gemma L Holliday, Nicholas Furnham, Jonathan D Tyzack, Katherine Ferris, and Janet M Thornton. Mechanism and catalytic site atlas (m-csa): a database of enzyme reaction mechanisms and active sites. *Nucleic acids research*, 46(D1):D618–D623, 2018.

Ioannis G Riziotis, António JM Ribeiro, Neera Borkakoti, and Janet M Thornton. Conformational variation in enzyme catalysis: a structural study on catalytic residues. *Journal of Molecular Biology*, 434(7):167517, 2022.

Daniela Röthlisberger, Olga Khersonsky, Andrew M Wollacott, Lin Jiang, Jason DeChancie, Jamie Betker, Jasmine L Gallaher, Eric A Althoff, Alexandre Zanghellini, Orly Dym, et al. Kemp elimination catalysts by computational enzyme design. *Nature*, 453(7192):190–195, 2008.

Hans Martin Senn and Walter Thiel. Qm/mm methods for biomolecular systems. *Angewandte Chemie International Edition*, 48(7):1198–1229, 2009.

Justin B Siegel, Alexandre Zanghellini, Helena M Lovick, Gert Kiss, Abigail R Lambert, Jennifer L St. Clair, Jasmine L Gallaher, Donald Hilvert, Michael H Gelb, Barry L Stoddard, et al. Computational design of an enzyme catalyst for a stereoselective bimolecular diels-alder reaction. *Science*, 329(5989):309–313, 2010.

Raghav Singhal, Zachary Horvitz, Ryan Teehan, Mengye Ren, Zhou Yu, Kathleen McKeown, and Rajesh Ranganath. A general framework for inference-time scaling and steering of diffusion models. *arXiv preprint arXiv:2501.06848*, 2025.

Adam JT Smith, Roger Muller, Miguel D Toscano, Peter Kast, Homme W Hellinga, Donald Hilvert, and KN Houk. Structural reorganization and preorganization in enzyme active sites: comparisons of experimental and theoretically ideal active site geometries in the multistep serine esterase reaction cycle. *Journal of the American Chemical Society*, 130(46):15361–15373, 2008.

Dean J Tantillo, Chen Jiangang, and Kendall N Houk. Theozymes and compuzymes: theoretical models for biological catalysis. *Current opinion in chemical biology*, 2(6):743–750, 1998.

Arieh Warshel and Jan Florián. Computer simulations of enzyme catalysis: finding out what has been optimized by evolution. *Proceedings of the National Academy of Sciences*, 95(11):5950–5955, 1998.

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.

Jeremy Wohlwend, Gabriele Corso, Saro Passaro, Noah Getz, Mateo Reveiz, Ken Leidal, Wojtek Swiderski, Liam Atkinson, Tally Portnoi, Itamar Chinn, et al. Boltz-1 democratizing biomolecular interaction modeling. *BioRxiv*, pages 2024–11, 2025.

Brandon Wood, Misko Dzamba, Xiang Fu, Meng Gao, Muhammed Shuaibi, Luis Barroso-Luque, Kareem Abdelmaqsoud, Vahe Gharakhanyan, John Kitchin, Daniel Levine, et al. Uma: A family of universal models for atoms. *Advances in Neural Information Processing Systems*, 38:129391–129427, 2026.

Filip Yabukarski, Justin T Biel, Margaux M Pinney, Tzanko Doukov, Alexander S Powers, James S Fraser, and Daniel Herschlag. Assessment of enzyme active site positioning and tests of catalytic mechanisms through x-ray–derived conformational ensembles. *Proceedings of the National Academy of Sciences*, 117(52):33204–33215, 2020.

Appendix

A METHOD DETAILS

A.1 MUGS

Blocked Gibbs sampling constructs a Markov chain by repeatedly sampling each block of variables from its conditional distribution given the others. When these conditionals arise from a common joint distribution, each update preserves that joint distribution, making it invariant under the resulting transition kernel. For example, suppose

$$(\mathbf{s}_t, \mathbf{x}_t) \sim p(\mathbf{s}, \mathbf{x})$$

and perform the structure update

$$\mathbf{x}_{t+1} \sim p(\mathbf{x} \mid \mathbf{s}_t)$$

The resulting joint distribution is

$$\int p(\mathbf{s}_t, \mathbf{x}_t) p(\mathbf{x}_{t+1} \mid \mathbf{s}_t)\, d\mathbf{x}_t = p(\mathbf{s}_t) p(\mathbf{x}_{t+1} \mid \mathbf{s}_t) = p(\mathbf{s}_t, \mathbf{x}_{t+1})$$

The sequence update preserves the same joint by the analogous argument, so a full Gibbs iteration leaves the target distribution invariant.

One MuGS iteration can be viewed as an approximate blocked Gibbs update over the current sequence–structure pair. For compactness, let

$$\mathbf{x}_t \coloneqq \{\mathbf{x}_{i,t}\}_{i=1}^{K}, \qquad \mathbf{m} \coloneqq \{\mathbf{m}_i\}_{i=1}^{K}$$

The corresponding transition kernel is

$$K((\mathbf{s}_t, \mathbf{x}_t), (\mathbf{s}_{t+1}, \mathbf{x}_{t+1})) = p_\theta(\mathbf{x}_{t+1} \mid \mathbf{s}_t, \mathbf{m}) p_\phi(\mathbf{s}_{t+1} \mid \mathbf{x}_{t+1}, \mathbf{m}),$$

where

$$p_\theta(\mathbf{x} \mid \mathbf{s}, \mathbf{m}) = \prod_{i=1}^{K} p_\theta(\mathbf{x}_i \mid \mathbf{s}, \mathbf{m}_i)$$

In MUGS, however, p_θ and p_ϕ are learned independently, so we treat the procedure as approximate blocked Gibbs sampling rather than assuming exact compatibility with a common joint.

A.2 MUGS INFERENCE TIME SCALING

MUGS itself can be viewed as increasing compute at inference time through its repeated Gibbs updates, but its formulation also allows control over where this additional inference time compute is spent: either in depth, by extending a trajectory, or in width, by exploring multiple trajectories in parallel. Here we outline several natural extensions along these axes.

A.2.1 BEAM MUGS

MuGS can be extended in width using a beam search (Freitag and Al-Onaizan, 2017) analog, maintaining a beam of B trajectories. Let,

$$\mathcal{B}_t = \left\{ \left(\mathbf{s}_t^{(b)}, \mathbf{x}_t^{(b)} \right) \right\}_{b=1}^{B}$$

For each trajectory b, we draw R_x structure samples,

$$\mathbf{x}_{t+1}^{(b,r_x)} \sim p_\theta\left(\cdot \mid \mathbf{s}_t^{(b)}, \mathbf{m} \right), \qquad r_x = 1, \ldots, R_x,$$

and for each structure sample draw R_s sequence samples,

$$\mathbf{s}_{t+1}^{(b,r_x,r_s)} \sim p_\phi\left(\cdot \mid \mathbf{x}_{t+1}^{(b,r_x)}, \mathbf{m} \right), \qquad r_s = 1, \ldots, R_s.$$

This produces BR_xR_s candidate trajectories, which are scored by a multistate objective F and pruned back to beam width B:

$$\mathcal{B}_{t+1} = \mathrm{TopB}_F \left\{ \left(\mathbf{s}_{t+1}^{(b,r_x,r_s)}, \mathbf{x}_{t+1}^{(b,r_x)} \right) \right\}_{b,r_x,r_s}$$

Here, B controls the number of parallel trajectories retained across iterations, while R_x and R_s control structure and sequence branching factor within each iteration. In practice, R_x can be implemented by drawing multiple diffusion samples from the structure predictor, while R_s corresponds to drawing multiple sequences from an inverse folding model.

Connection to NISE. Our beam formulation is closely related to neural iterative selection–expansion (NISE) (Fry et al., 2026), which similarly alternates between sequence design, structure prediction, and selection. In NISE, each retained protein–ligand structure is expanded into 1,000 sequence samples using LASErMPNN, the resulting sequences are cofolded, and the top three self consistent structures by ligand confidence are retained for the next round. The main difference is the update ordering: NISE expands sequences before predicting structures, whereas Beam MuGS samples structures before sequences. NISE can therefore be viewed as a sequence first analog with large R_s, $R_x = 1$, and beam width $B = 3$.

Algorithm 2: Beam MuGS

```
# motifs {m_i}_{i=1}^K
# beam width B
# structure branches R_x
# sequence branches R_s
# iterations T
for b in range(B):
    s_0^(b) ← seed(m_1)
for t in range(T):
    for b in range(B):
        for r_x in range(R_x):
            for i in range(1, K + 1):
                x_i^(b,r_x) ~ p_θ(· | s_t^(b), m_i)
            for r_s in range(R_s):
                s^(b,r_x,r_s) ~ p_φ(· | {x_i^(b,r_x)}_{i=1}^K, {m_i}_{i=1}^K)
    C_{t+1} = {(s^(b,r_x,r_s), x^(b,r_x))}_{b,r_x,r_s}
    B_{t+1} ← TopB_F(C_{t+1})
return s_T^(b*),    b* = arg max_b F(s_T^(b), x_T^(b))
```

A.2.2 SEQUENTIAL MONTE CARLO MUGS

Rather than deterministically selecting the highest-scoring trajectories, we may want to preserve stochastic exploration while biasing sampling toward desirable regions of the design space. Sequential Monte Carlo (SMC) provides a natural alternative, maintaining a population of trajectories that are propagated through the MuGS transition and resampled according to their weights. Let

$$\mathcal{P}_t = \left\{ \left(\mathbf{s}_t^{(b)}, \mathbf{x}_t^{(b)} \right) \right\}_{b=1}^{B}$$

Each particle is propagated according to the MuGS transition kernel,

$$K((\mathbf{s}_t, \mathbf{x}_t), (\mathbf{s}_{t+1}, \mathbf{x}_{t+1})) = p_\theta(\mathbf{x}_{t+1} \mid \mathbf{s}_t, \mathbf{m})\, p_\phi(\mathbf{s}_{t+1} \mid \mathbf{x}_{t+1}, \mathbf{m})$$

Let $q_{\text{MuGS}}(\tau_t)$ denote the implicit distribution over MuGS trajectories up to iteration t,

$$\tau_t = ((\mathbf{s}_0, \mathbf{x}_0), \ldots, (\mathbf{s}_t, \mathbf{x}_t))$$

We target an exponentially tilted trajectory distribution,

$$p_t^*(\tau_t) \propto q_{\text{MuGS}}(\tau_t) \exp[\beta F(\mathbf{s}_t, \mathbf{x}_t)]$$

Using q_{MuGS} as the proposal, the importance weights are

$$w_t(\tau_t) \propto \frac{p_t^*(\tau_t)}{q_{\text{MuGS}}(\tau_t)} \propto \exp[\beta F(\mathbf{s}_t, \mathbf{x}_t)]$$

The incremental weights used for resampling are

$$w_{t+1}^{(b)} \propto \exp\left[\beta \left(F(\mathbf{s}_{t+1}^{(b)}, \mathbf{x}_{t+1}^{(b)}) - F(\mathbf{s}_t^{(b)}, \mathbf{x}_t^{(b)}) \right) \right]$$

The next particle population $\mathcal{P}_{t+1}$ is obtained by drawing B particles with replacement from the updated population according to $\text{Cat}(w_{t+1}^{(1)}, \ldots, w_{t+1}^{(B)})$. This is similar in nature to the Feynman–Kac formulation used for inference-time steering of diffusion models in Singhal et al. (2025). In practice, F could favor motif recovery across states (e.g., RMSD), structural confidence, or other task-specific multistate design criteria.

Algorithm 3: SMC MuGS

```
# motifs {m_i}_{i=1}^K
# num_particles B
# iterations T
for b in range(B):
  s_b^(0) ← seed(m_1)
for t in range(T):
  for b in range(B):
    for i in range(1, K + 1):
      x_{i,b}^(t+1) ~ p_θ(· | s_b^(t), m_i)
    s_b^(t+1) ~ p_φ(· | {x_{i,b}^(t+1)}_{i=1}^K, {m_i}_{i=1}^K)
    w_b^(t+1) ∝ exp(β[F(s_b^(t+1), x_b^(t+1)) − F(s_b^(t), x_b^(t))])
  P_{t+1} = {(s_b^(t+1), x_b^(t+1))}_{b=1}^B
  P_{t+1} ← Resample(P_{t+1}, w^(t+1))
return s_{b*}^(T),   b* = arg max_b F(s_b^(T), x_b^(T))
```

A.3 MCFOLD

A.3.1 ATOMIC MOTIF SAMPLING

During training, we sample localized atomic motifs from native structures using a protocol analogous to RFdiffusion2 (Ahern et al., 2025), but bias sampling toward compact ligand-centered residue sites to expose the model to the tip atom interactions relevant for enzymes (Algorithm 4). For structures with ligands, we include either the full ligand or, with lower probability, a connected ligand subgraph. We then select nearby protein residues whose side chains lie within 6 Å of the ligand and greedily retain a local cluster using a 7 Å side chain distance cutoff, up to 10 residues. For protein structures without ligands, we similarly start from a random residue and grow a local cluster using the same 7 Å cutoff. Connected subsets of side chain heavy atoms are retained, and the resulting motif is centered and randomly rotated before conditioning the model.

Algorithm 4: Atomic motif sampling

```
def sample_motif(structure):
  # ligand-centered motif
  if has_ligand(structure):
    ligand ← sample_ligand(structure)
    if random() < 0.7:
      motif ← ligand
    else:
      motif ← connected_subgraph( ligand, min_frac=0.6 )
    residues ← local_residues(ligand, contact=6Å, cutoff=7Å, max=10)
  # residue-only motif
  else:
    seed ← random_residue(structure)
    residues ← local_residues(seed, cutoff=7Å, max=10)
  # sample connected side-chain atoms
  for residue in residues:
    motif ← motif ∪ connected_subgraph( sidechain(residue),
min_frac=0.6 )
  # remove absolute coordinate frame
  motif.coords ← center(motif.coords)
  motif.coords ← random_rotate(motif.coords)
  return motif.coords, motif.mask
```

A.4 ARCHITECTURE & TRAINING

We evaluate on 463 PDB entries released between 2021-10-06 and 2023-01-11, all strictly after the 2021-09-30 training cutoff and verified absent from the training set. We ablate the components of MCFOLD to test whether explicit atomic motif conditioning is necessary beyond standard cofolding. Without motif conditioning, the base model does not reliably recover the prescribed atomic geometry, with a mean motif RMSD of 3.97 Å. Fine tuning with atomic motif conditioning reduces this to 0.93 Å, while inference time coordinate replacement enforces the motif nearly exactly. Importantly, these gains do not come at the expense of overall protein structure quality, which remains comparable across all three settings.

We next compare against Boltz-2 template conditioning (Passaro et al., 2025). Though template conditioning may be sufficient for residue level structural constraints, it provides only partial access to an atomic motif: only 43.2% of motif atoms are representable and distal side chain atoms and ligand atoms are omitted. Consistent with this limited expressiveness, Boltz-2 with templating reaches a mean motif RMSD of 3.36 Å, substantially worse than explicit atomic motif conditioning (Table 2).

Table 2: Atomic motif conditioning placement

Method	Cond.	Replace	Mean (Å)	P90 (Å)	< 1 Å	< 2 Å
Boltz-2 + template	–	–	3.36	8.75	48.2%	59.7%
Ablate conditioning + replacement			3.97	10.03	30.1%	48.7%
Ablate replacement	✓		0.93	1.66	69.5%	94.4%
MCFOLD	✓	✓	0.01	0.01	100.0%	100.0%

Table 3: Structure quality (LDDT) under atomic motif conditioning and replacement.

Method	Cond.	Replace	Complex	Intra-prot.	Prot.-prot.	Nuc.-prot.
Ablate conditioning + replacement			0.778	0.803	0.594	0.407
Ablate replacement	✓		0.787	0.810	0.613	0.387
MCFOLD	✓	✓	0.789	0.813	0.609	0.377
n			462	462	181	66

Table 4: Interface quality measured by DockQ.

Method	Cond.	Replace	Prot.-prot.	> 0.23	> 0.49	Nuc.-prot.
Ablate conditioning + replacement			0.559	75.4%	62.9%	0.377
Ablate replacement	✓		0.574	79.1%	64.7%	0.366
MCFOLD	✓	✓	0.567	78.0%	65.3%	0.359
n			177			

B ADDITIONAL RESULTS

B.1 MULTISTATE ATOMIC MOTIF ENZYME CURATION (MAME V1.0)

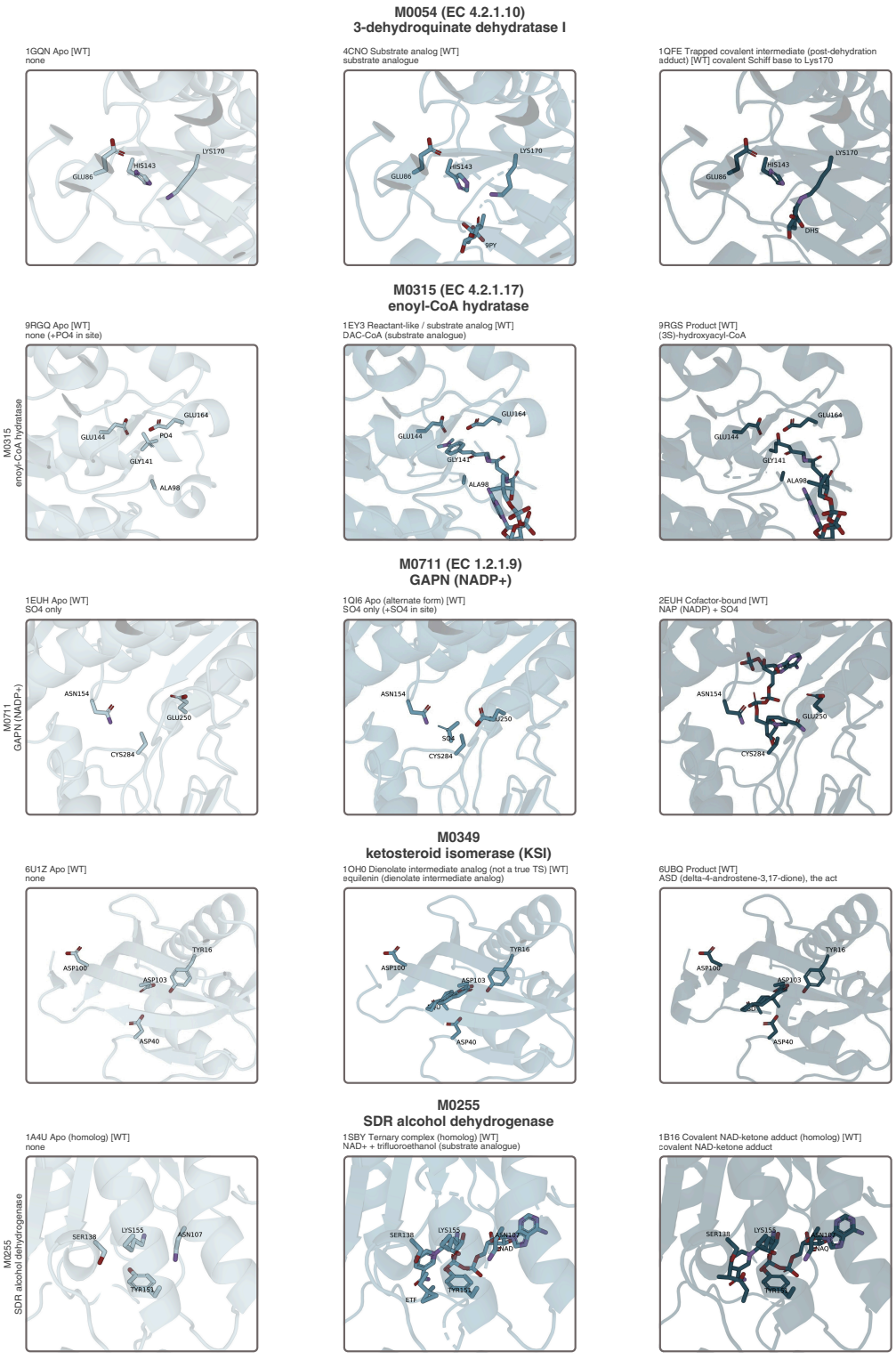


Figure 8: Initial curated set of multistate motifs for MAME.

We constructed an initial multistate extension of the Atomic Motif Enzyme (AME) benchmark (Ahern et al., 2025)) by identifying experimentally determined structures that capture distinct functional states of the same enzyme or a closely related homolog. Starting from M-CSA catalytic annotations, we searched PDB and associated structural literature for alternate apo, ligand-bound, product, and trapped intermediate states. Catalytic residues were mapped between structures using sequence and structural correspondence. For each state, the protein motif was defined as all heavy side chain atoms of the M-CSA catalytic residues, rather than the random connected subsets of atoms used in the original AME. Structures were superimposed using backbone atoms around the active site, allowing changes in catalytic side chain geometry to be measured independently of global protein motion.

This procedure yielded five initial multistate motifs. Figure 8 and Table 5 summarize the curated structures and their sources. M0054 spans apo, substrate-analog, and trapped covalent-intermediate states; M0315 spans apo, substrate-analog, and product states; and M0711 contains two experimentally observed apo conformations together with an NADP-bound holo state. M0349 captures apo KSI, a dienolate intermediate analog, and the bound product, while M0255 includes apo, reactant-like, and covalent-intermediate states from a closely related homolog. Together, these examples provide experimentally grounded multistate motifs spanning distinct points along catalytic processes, with the homolog-derived M0255 states interpreted more cautiously.

Table 5: Experimental structures used to construct the initial MAME multistate motifs.

M-CSA	Enzyme	EC	PDB	State	Reference
M0054	Type-I DHQase	4.2.1.10	1GQN	Apo	Lee et al. (2002)
			4CNO	Substrate analog	Maneiro et al. (2014)
			1QFE	Trapped covalent int.	Gourley et al. (1999)
M0315	Enoyl-CoA hydratase	4.2.1.17	9RGQ	Apo	Dalwani et al. (2026)
			1EY3	Substrate analog	Bahnson et al. (2002)
			9RGS	Product	Dalwani et al. (2026)
M0711	NADP-dependent GAPN	1.2.1.9	1EUH	Apo	Cobessi et al. (1999)
			1QI6	Alternative apo	Cobessi et al. (2000)
			2EUH	NADP-bound holo	Cobessi et al. (1999)
M0349	Δ^5-3-ketosteroid isomerase	5.3.3.1	6U1Z	Apo	Yabukarski et al. (2020)
			1OH0	Intermediate analog	Kim et al. (1997)
			6UBQ	Product	Yabukarski et al. (2020)
M0255	SDR alcohol dehydrogenase	1.1.1.1	1A4U	Apo	Benach et al. (1998)
			1SBY	Reactant analog	Benach et al. (2005)
			1B16	Covalent intermediate	Benach et al. (1999)

The current set emphasizes experimentally supported states for which active-site geometry can be defined directly from structure, primarily spanning apo, ligand-bound, and product states. M0054 additionally provides a trapped covalent intermediate, offering a structural snapshot of chemistry between reactant and product. True transition-state geometries are generally not directly accessible from ordinary structural experiments, so they are not included in the current benchmark. More generally, however, MUGS can condition on computationally derived transition-state or intermediate geometries, for example from QM reaction-path calculations.

This curation provides a starting point that can be expanded in several directions. Additional experimental states can be incorporated by retrieving homologous structures that capture distinct catalytic or functional states using the same mapping and alignment procedure. For reactions without suitable experimental structures, mechanism intermediate geometries could instead be constructed from M-CSA chemistry and refined with molecular potentials and semiempirical methods such as UMA or xTB (Wood et al., 2026; Bannwarth et al., 2019).

B.2 SERINE HYDROLASE MOTIFS

The serine hydrolase design pipeline of Lauko et al. (2025) provides a useful setting for testing multistate design directly. Their pipeline constructs and scaffolds a single state catalytic geometry with RFdiffusion, then evaluates whether the resulting design remains preorganized in other states along the reaction coordinate using PLACER. Thus, multistate compatibility enters primarily as a downstream filtering criterion rather than as a constraint during backbone and sequence generation.

To test whether MUGS can instead model multiple catalytic states directly during design, we construct multistate motifs from *super*, an experimentally validated serine hydrolase from Lauko et al. (2025). *super* catalyzes multiple turnover hydrolysis of 4MU-butyrate and was subsequently validated by X-ray crystallography with close agreement between the designed and experimental geometries. Following their reaction coordinate validation, we model five states of the hydrolysis cycle:

- **Apo**: unbound Ser128–OH
- **Substrate**: noncovalently bound 4MU-butyrate
- **TI1**: first tetrahedral Ser128 adduct
- **AEI**: acyl–Ser128 intermediate
- **TI2**: second tetrahedral Ser128 adduct

From the generated PLACER ensembles, we select representative TI1, AEI, and TI2 structures that best satisfy the catalytic geometry criteria and use them as proxy motifs for multistate conditioning. Apo and substrate-bound states are held out for evaluation, since each additional conditioning state requires another fold per Gibbs iteration and may provide diminishing information for sequence design. The experimentally validated *super* reference achieves a worst-state functional (catalytic triad) motif RMSD of 1.59 Å across all five states, which we use as the success threshold for this benchmark.

B.3 MuGS MULTISTATE RECOVERY

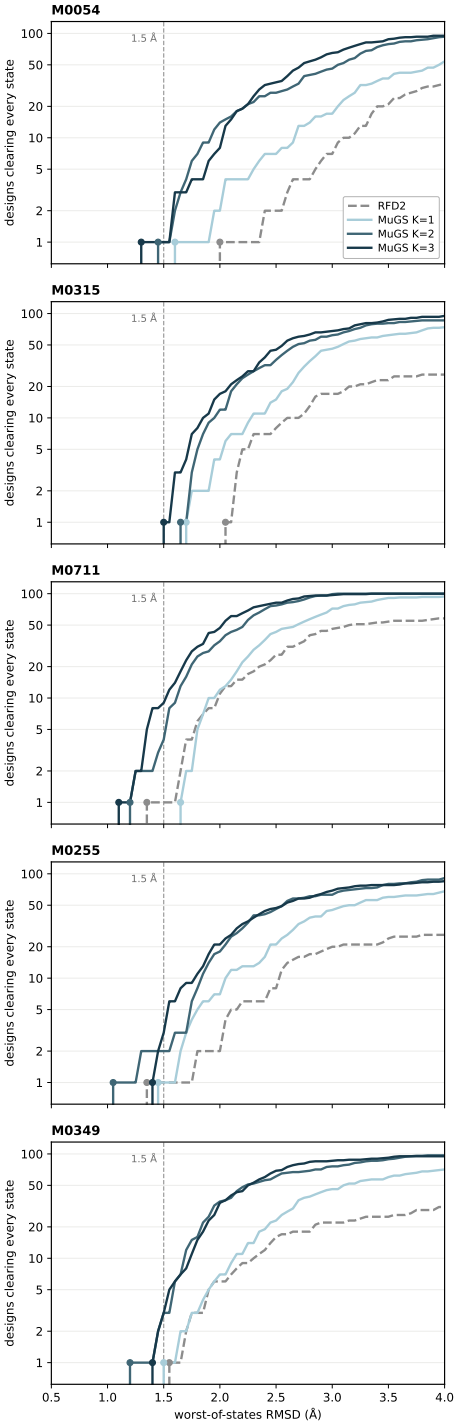


Figure 9: Multistate scaffolding performance on curated set of motifs (MAME). Designs (of 100) passing each RMSD threshold in ALL three experimental states; dots mark each run's first success.