Improving the Efficacy of Test-Time Steering in Masked Diffusion Models with Parallel Tempering

Po-Yi Lu [1 2] **Hsuan-Tien Lin** [3 4] **Shih-Hsin Wang** [5]

Abstract

Masked Diffusion Models (MDMs) provide expressive generative priors for discrete biological sequences such as proteins and DNA. However, many downstream tasks require steering these models at inference time to optimize arbitrary, external reward functions. Existing test-time steering methods face a fundamental exploration-exploitation trade-off in complicated reward landscapes: they either collapse into suboptimal or unrealistic modes or require massive sampling budgets to find rare, high-reward states. We address this tension with Parallel Tempering for MDMs (PT-MDM). To adapt parallel tempering to MDM's generation, we tightly couple the reward temperature to the model's sequence remasking fraction, thereby enabling both global exploration and local exploitation without compromising sequence plausibility. Experiments on inverse protein folding show that PT-MDM improves Pred-ddG and success rate over the evaluated test-time baselines while retaining low scRMSD, and approaches fine-tuned DRAKES on these metrics without training. In regulatory DNA design, PT-MDM achieves the highest Pred-Activity among the evaluated training-free methods.

1. Introduction

Masked Diffusion Models (MDMs) provide expressive generative priors for discrete objects such as text, proteins, DNA, and other structured sequences (Nie et al., 2026; de Groot et al., 2025; Shi et al., 2024). Nevertheless, in many downstream problems, matching the data distribution is only a starting point to ensure sample fidelity, and generated candidates must also optimize external objectives such as human preferences, biological activity, structural constraints, or task-specific reward models.

The goal can be naturally formalized as a regularized reward optimization problem: seeking a distribution q that maximizes the expected reward while remaining close to the pretrained MDM prior p to retain sample fidelity:

$$\max_q \mathbb{E}_{\boldsymbol{x}\sim q}[R(\boldsymbol{x})] - \frac{1}{\beta} D_{\mathrm{KL}}(q \,\|\, p), \tag{1}$$

where $R(\boldsymbol{x})$ denotes the reward function, and the temperature $\frac{1}{\beta}$ controls how strongly q is constrained to remain close to the prior p, as quantified by the KL divergence D_{KL}.

The solution to Eq. 1 is known as the reward-tilted distribution (Korbak et al., 2022; Gheshlaghi Azar et al., 2024; Go et al., 2023; Rafailov et al., 2023):

$$\tilde{p}(\boldsymbol{x}) = \frac{1}{Z} p(\boldsymbol{x}) \exp(\beta R(\boldsymbol{x})), \tag{2}$$

where $Z := \sum p(\boldsymbol{x}) \exp(\beta R(\boldsymbol{x}))$ is the normalizing constant.

A natural approach to sample from the reward-tilted distribution $\tilde{p}(\boldsymbol{x})$ is to fine-tune the model so that its induced distribution approximates the reward-tilted target $\tilde{p}$, e.g., DRAKES (Wang et al., 2025) or DDPP (Rector-Brooks et al., 2025a). While effective for a fixed objective, fine-tuning couples model parameters tightly to a particular reward and typically requires retraining whenever the reward function changes. This limits their applicability in settings where objectives are frequently updated or changed (Yang et al., 2026). These settings, therefore, demand methods that can adapt immediately to changing rewards without repeated optimization of the base model.

Test-time steering addresses this need by keeping the pretrained MDM fixed and instead modifying the sampling dynamics at inference time to approximately sample from $\tilde{p}$ (Schiff et al., 2025). Despite this flexibility, effective test-time steering remains challenging because the reward-

[1]Graduate Institute of Networking and Multimedia, National Taiwan University, Taipei, Taiwan [2]Department of Computer Science and Engineering, Texas A&M University, College Station, Texas, USA [3]Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan [4]Artificial Intelligence Center of Research Excellence, National Taiwan University [5]Department of Mathematics, University of Utah, Salt Lake City, Utah, USA. Correspondence to: Shih-Hsin Wang <shwang@math.utah.edu>.

Accepted at the 2026 Workshop on Generative and Agentic AI for Biology (ICML 2026)

tilted distribution is often sharp, rugged, and highly multimodal (Oh et al., 2026). High-reward solutions may lie in narrow regions assigned low probability under the pretrained prior, while overly aggressive reward optimization can drive samples away from realistic sequences. Thus, successful steering must balance two competing goals: broad exploration to discover diverse reward modes, and local refinement to concentrate samples within those modes. In discrete sequence spaces, this difficulty is also a transition-design problem: a sampler must decide how many tokens to revise, how to revise them while staying close to the pretrained model manifold, and how to move between global exploration and local improvement without collapsing into a local reward mode.

Existing particle-based methods expose this tension. Best-of-K explores broadly by drawing independent samples and selecting the highest-reward candidate, but uses reward information only at the final selection stage (Mudgal et al., 2024). As a result, it often requires a large sampling budget to find high-reward regions when they are rare under the prior, and even when such regions are occasionally reached, Best-of-K lacks any mechanism to refine samples further once they are identified (Wu et al., 2023). Sequential Monte Carlo (SMC) methods incorporate reward information during generation by filtering and resampling particles. However, because these methods proceed through a forward particle-filtering process, resampling can reduce diversity and lock the particle population onto early-discovered local modes. Consequently, current methods often improve either exploration or refinement, but struggle to maintain both persistent search and post-discovery improvement in complex discrete landscapes (He et al., 2026).

We take a different view: the remask-and-denoise mechanism of MDMs is not merely a way to generate samples, but a structured proposal mechanism for reward-guided sampling in discrete space. Each transition can be decomposed into two components. First, a masking ratio determines the scale of the proposed edit by choosing how much of the current sequence to reopen. Second, a denoising model reconstructs the masked positions conditioned on the remaining context, producing a prior-aware candidate without changing the pretrained model. This decomposition is *time-decoupled*: the edit scale can vary across replicas or sampling stages, while the same reward-independent denoising model supplies the reconstruction kernel throughout the procedure. Figure 1 illustrates the resulting reward-structure trade-off on two representative inverse-folding targets.

1.1. Our contributions

In this work, we develop a Parallel Tempering (PT)-based framework for test-time steering of MDMs (Earl & Deem, 2005). Rather than relying on a single inverse temperature β of reward guidance, PT-MDM maintains a collection of replicas evolving at different reward temperatures $\{\beta_k\}_{k=1}^{K}$ with $\beta_K = \beta$, as illustrated in Figure 2. High-temperature (small β_k) replicas stay close to the pretrained model and promote broad exploration over sequence space, while low-temperature (large β_k) replicas place stronger emphasis on the reward to refine high-quality candidates. Periodic replica exchange enables information sharing across temperatures, allowing discoveries made in exploratory regimes to propagate to more exploitative ones, thereby mitigating the collapse often observed in single-chain reward-guided sampling.

To support both broad exploration and fine-grained refinement in PT, we couple the PT temperature ladder to the MDM remasking ladder. At each candidate update within replica k, a fraction m_k of tokens is remasked and resampled by the same pretrained denoiser to form a model-aligned transition $\boldsymbol{x} \to \boldsymbol{x}_{\text{masked}} \to \boldsymbol{x}'$. Larger masking fractions induce broader moves that explore distant regions of sequence space, whereas smaller fractions produce targeted local edits for refinement. Thus, the reward temperature β_k controls the energy pressure of a replica, while the masking ratio m_k controls its discrete step size. This time-decoupled proposal family is the nontrivial interface between PT and MDMs: PT supplies energy-based exploration and exchange, while masked diffusion supplies a native prior-aware transition kernel in discrete space.

As illustrated in Figure 1, PT-MDM achieves a more favorable reward-structure trade-off in inverse protein folding: it attains substantially higher Pred-ddG rewards while keeping reasonable scRMSD, indicating that the reward improvement does not come from severe structural drift. In contrast, Best-of-K and SMC can preserve plausible structures, but their rewards remain limited and do not reach the same high-reward regime as PT-MDM. This supports our central motivation: effective test-time steering must provide a mechanism for discovering and refining rare, high-reward candidates.

In summary, our contributions are threefold. First, we formulate test-time steering for MDMs as a discrete transition-design problem, highlighting the local-trap limitation of forward particle-filtering methods such as SMC. Second, we introduce a time-decoupled PT-MDM transition that factors each update into a temperature-dependent masking ratio and a shared reward-independent denoising model. Third, we show empirically that this mechanism improves reward steering on protein and DNA generation while preserving prior-aware sample quality, approaching fine-tuned reward performance on key metrics without any parameter updates.

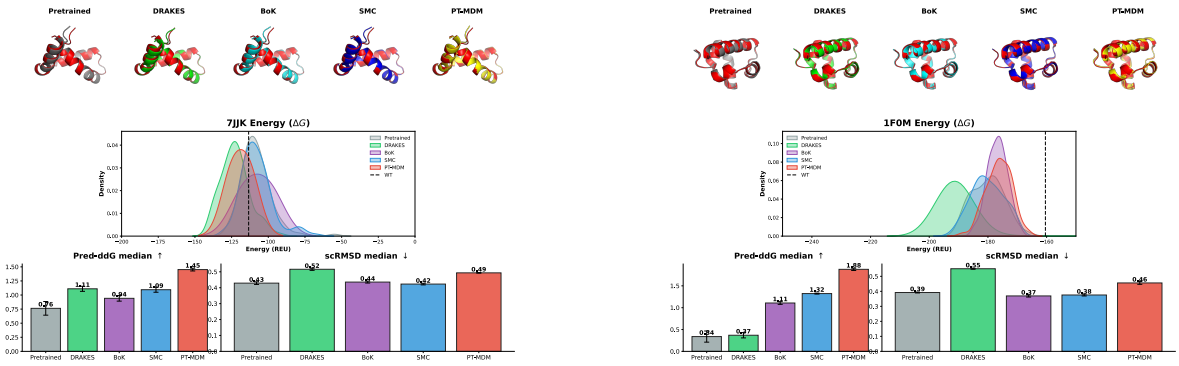

Figure 1. **Protein inverse-folding contrast on representative targets.** The panels compare generated protein designs for 7JJK (left) and 1F0M (right), illustrating how test-time steering changes the reward-structure trade-off relative to the target backbone. PT steers samples toward higher Pred-ddG while retaining structural fidelity through the pretrained denoising model.

1.2. Related work

Remasking-based discrete diffusion models. Masked diffusion models introduce a remasking-and-denoising paradigm for discrete generation, enabling iterative refinement through partial corruption and reconstruction. Methods such as LLaDA (Nie et al., 2026) and recent remasking-based scaling approaches (Wang et al., 2026; Schiff et al., 2026; Rector-Brooks et al., 2025b) demonstrate that selective remasking improves controllability and sample quality at inference time. While these methods provide a natural mechanism for local refinement, they do not explicitly address global exploration in multimodal reward landscapes.

Test-time steering for diffusion models. Continuous diffusion steering methods commonly inject reward, classifier, or guidance information into reverse-time transitions, often exploiting gradients or intermediate noisy states to bias the trajectory toward desired samples (Guo et al., 2026; Jain et al., 2026). In discrete diffusion, however, there is no natural infinitesimal move in token space, so the central challenge is to define a reward-aware state transition that can revise a discrete sample while preserving prior plausibility (Schiff et al., 2025; Avrahami & Nachmani, 2026). Recent PT-based methods such as CREPE (He et al., 2026) run parallel diffusion trajectories with replica exchange across timesteps, improving diversity along the sampling path but without explicitly separating exploration from reward-driven exploitation. In contrast, PT-MDM uses the remask-and-denoise interface itself as the discrete proposal mechanism: reward temperatures control energy pressure, remasking fractions control proposal scale, and replica exchange transfers states across these regimes.

2. Background

Masked Diffusion Models. Masked Diffusion Models (MDMs) (Ou et al., 2025) define a generative model over discrete sequences $\boldsymbol{x} = (x_1, \ldots, x_L) \in \mathcal{V}^L$, where $\mathcal{V}$ is a finite vocabulary, i.e., the collection of all possible discrete tokens, and L is the sequence length. Unlike continuous diffusion models that inject Gaussian noise, MDMs operate through *masking* (Sahoo et al., 2024). A special token [MASK] is introduced, and generation is learned by recovering clean sequences from partially masked inputs (Ou et al., 2025; Zheng et al., 2025).

Formally, let $\boldsymbol{x}^{(0)} \sim p_{\text{data}}$ denote a clean sequence. A forward corruption process constructs progressively noisier states $\{\boldsymbol{x}^{(t)}\}_{t=1}^{T}$ via stochastic masking:

$$\boldsymbol{x}^{(t)} = \mathcal{M}_t(\boldsymbol{x}^{(0)}), \qquad t \in \{1, \ldots, T\}, \tag{3}$$

where $\mathcal{M}_t$ increases the masking level with t. A standard formulation uses an increasing masking schedule $\sigma_t \in [0, 1]$ with $\sigma_1 = 1$, where σ_t denotes the probability that a token is masked at step t (Sahoo et al., 2024; Nie et al., 2026). For each position i, we define

$$x_i^{(t)} = \begin{cases} [\text{MASK}] & \text{if } m_{t,i} = 1, \\ x_i^{(0)} & \text{otherwise,} \end{cases} \tag{4}$$

where

$$m_{t,i} \sim \text{Bernoulli}(\sigma_t). \tag{5}$$

This construction ensures that $\boldsymbol{x}^{(t)}$ preserves a subset of the original context while progressively increasing the proportion of masked tokens as t grows.

The reverse process is parameterized by a denoising model p_θ that predicts clean tokens conditioned on the masked

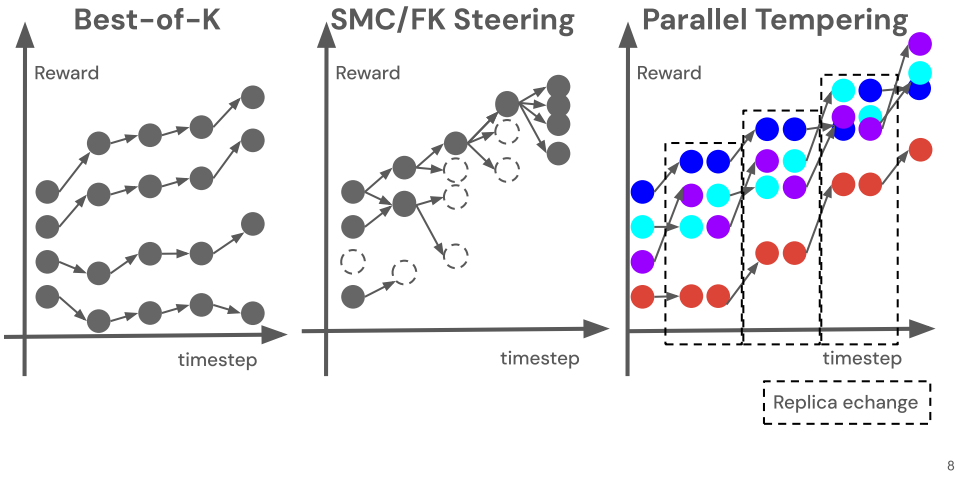

Figure 2. **Comparison of Best-of-K, SMC, and PT-MDM for $K = 4$.** Best-of-K explores with independent samples but uses reward information only for final selection, while SMC can exploit intermediate rewards but may concentrate particles around early high-reward modes. PT-MDM uses high-temperature replicas for broad exploration, low-temperature replicas for reward-driven exploitation, and replica exchange to transfer discoveries across temperatures, increasing the chance of finding and refining higher-reward candidates.

sequence:

$$p_\theta(\boldsymbol{x}^{(0)} \mid \boldsymbol{x}^{(t)}) = \prod_{i \in \mathcal{M}(\boldsymbol{x}^{(t)})} p_\theta(x_i^{(0)} \mid \boldsymbol{x}^{(t)}), \tag{6}$$

where $\mathcal{M}(\boldsymbol{x}^{(t)})$ denotes the masked positions. Training minimizes the expected cross-entropy over masked tokens (Ou et al., 2025):

$$\mathcal{L}_{\text{MDM}}(\theta) = \mathbb{E}_{t, \boldsymbol{x}^{(0)}, \boldsymbol{x}^{(t)}} \left[- \sum_{i \in \mathcal{M}(\boldsymbol{x}^{(t)})} \log p_\theta(x_i^{(0)} \mid \boldsymbol{x}^{(t)}) \right]. \tag{7}$$

Remask-and-Denoise Generation. Sampling from an MDM p_θ is started from a fully masked sequence $\boldsymbol{x}_{\text{full}} = [\text{MASK}, \ldots, \text{MASK}]$, with the initial state sampled as $\hat{\boldsymbol{x}}^{(T)} \sim p_\theta(\cdot \mid \boldsymbol{x}_{\text{full}})$.

Generation then proceeds backward from $t = T$ to 1 through an iterative remask-and-denoise process (Nie et al., 2026; Wang et al., 2026). Given the current state $\hat{\boldsymbol{x}}^{(t)}$, a remasking operator first produces a partially masked intermediate state:

$$\hat{\boldsymbol{x}}_{\text{mask}}^{(t)} \sim \rho(\cdot \mid \hat{\boldsymbol{x}}^{(t)}, m), \tag{8}$$

where ρ selects $\lfloor mL \rfloor$ tokens to reset to [MASK] and $m \in [0, 1]$ controls the expected proportion of tokens to be remasked. Conditioned on this input, the denoiser samples the next state:

$$\hat{\boldsymbol{x}}^{(t-1)} \sim p_\theta(\cdot \mid \hat{\boldsymbol{x}}_{\text{mask}}^{(t)}). \tag{9}$$

Together, these define the one-step transition kernel

$$g_\theta(\hat{\boldsymbol{x}}^{(t-1)} \mid \hat{\boldsymbol{x}}^{(t)}) = \sum_{\boldsymbol{x}_{\text{mask}}^{(t)}} p_\theta(\hat{\boldsymbol{x}}^{(t-1)} \mid \hat{\boldsymbol{x}}_{\text{mask}}^{(t)}) \rho(\hat{\boldsymbol{x}}_{\text{mask}}^{(t)} \mid \hat{\boldsymbol{x}}^{(t)}, m). \tag{10}$$

Notice that the generated states $\{\hat{\boldsymbol{x}}^{(t)}\}_{t=0}^{T}$ are clean sequences except in implementations that allow residual mask tokens at intermediate steps. In typical implementations, the final output $\hat{\boldsymbol{x}}^{(0)}$ is required to be a clean sequence with no remaining mask tokens.

For the remasking distribution $\rho(\hat{\boldsymbol{x}}_{\text{mask}}^{(t)} \mid \hat{\boldsymbol{x}}^{(t)}, m)$, we consider two representative choices:

1. **Uniform:** A subset of tokens is selected uniformly at random by sampling an index set

$$\mathcal{S} \sim \text{Unif}\big(\{S \subseteq \mathcal{I} : |\mathcal{S}| = \lfloor mL \rfloor\}\big), \tag{11}$$

and defining the masked sequence by

$$\hat{x}_{\text{mask},i} = \begin{cases} [\text{MASK}] & \text{if } i \in \mathcal{S}, \\ \hat{x}_i^{(t)} & \text{otherwise.} \end{cases} \tag{12}$$

2. **Low-Confidence (Nie et al., 2026):** Tokens are selected based on the predictive density of p_θ. Let

$$c_i = p_\theta(x_i^{(t)} = \hat{x}_i^{(t)} \mid \hat{\boldsymbol{x}}_{\text{mask}}^{(t-1)}), \tag{13}$$

denote the confidence of token i given $\hat{\boldsymbol{x}}_{\text{mask}}^{(t-1)}$. We define the index set $\mathcal{S}$ as the indices of the $\lfloor mL \rfloor$ smallest confidence values c_i and mask all positions in $\mathcal{S}$.

3. Method

We introduce Parallel-Tempering for MDMs (PT-MDM), a test-time steering method inspired by the PT algorithm (Earl & Deem, 2005) and built around the native transition structure of masked diffusion. The method jointly enables *global exploration* and *fine-grained refinement* within a unified sampling framework. The goal is not merely to run multiple MDM samplers in parallel, but to construct a reward-aware transition in discrete space that can both escape local traps and refine high-reward candidates. PT-MDM maintains a set of interacting replicas operating at different reward temperatures, allowing simultaneous exploration of diverse high-reward regions and refinement under stronger reward pressure. PT's exchange mechanism mitigates the exploration-exploitation tension by propagating high-reward discoveries from exploratory replicas to exploitative ones, while injecting diverse states in the opposite direction to prevent mode collapse.

To operationalize this idea in masked diffusion models, we introduce a time-decoupled remask-and-denoise proposal. Each replica is assigned a remasking fraction that controls the expected proportion of tokens reset at each update, while all replicas share the same reward-independent denoising model p_θ. The remasking fraction acts as a *discrete step size* in sequence space: higher values induce larger structural edits for broad exploration, while lower values restrict updates to localized refinements around high-reward candidates. By coupling this step size to the PT temperature ladder, we obtain a single mechanism that simultaneously governs the scale of local proposals and the strength of reward pressure across replicas. The full procedure is summarized in Algorithm 1.

Temperature ladder and target distributions. We maintain K replicas indexed by increasing inverse temperatures

$$0 \le \beta_1 < \beta_2 < \cdots < \beta_K = \beta, \tag{14}$$

where $\beta_K = \beta$ is the user-specified target reward temperature of interest. Each replica targets the tempered distribution

$$p_{\beta_k}(\boldsymbol{x}) = \frac{1}{Z_{\beta_k}} p_\theta(\boldsymbol{x}) \exp(\beta_k R(\boldsymbol{x})). \tag{15}$$

In particular, small β_k favors exploration under the prior, while large β_k increasingly emphasizes reward preference.

Local updates via temperature-scaled remasking. To construct local proposals within each replica, we adapt the standard remask-and-denoise procedure (Section 2) by introducing temperature-scaled masking fractions $\{m_k\}_{k=1}^K$ such that $1 \ge m_1 > m_2 > \cdots > m_K > 0$.

Given the current sequence $\boldsymbol{x}^{[k]}$ in replica k, a candidate state is generated via:

$$\boldsymbol{x}^{[k]}_{\text{mask}} \sim \rho(\cdot \mid \boldsymbol{x}^{[k]}, m_k), \qquad \tilde{\boldsymbol{x}}^{[k]} \sim p_\theta(\cdot \mid \boldsymbol{x}^{[k]}_{\text{mask}}), \tag{16}$$

where ρ remasks exactly $\lfloor m_k L \rfloor$ tokens of the length-L sequence. This local transition is time-decoupled in the sense that the proposal scale is controlled by the replica-specific masking ratio m_k, while the unmasking model p_θ remains shared and reward-independent across the entire ladder. The masking fraction m_k is tightly coupled to the replica's inverse temperature β_k: high-temperature replicas (small β_k) use large m_k to encourage aggressive, global exploration of the sequence space, while low-temperature replicas (large β_k) use small m_k for conservative, fine-grained refinement. This unified mechanism directly aligns the physical size of each proposal edit with the replica's intended role in the exploration-exploitation trade-off.

Within-replica acceptance. We decide whether to accept the proposal $\tilde{\boldsymbol{x}}^{[k]}$ as the next state of replica k using a Metropolis-Hastings update. Write $\Delta R_k := R(\tilde{\boldsymbol{x}}^{[k]}) - R(\boldsymbol{x}^{[k]})$. The exact acceptance probability is

$$\begin{aligned} &A_k^{\text{MH}}(\boldsymbol{x}^{[k]}, \tilde{\boldsymbol{x}}^{[k]}) \\ &= \min\left\{1, \frac{p_\theta(\tilde{\boldsymbol{x}}^{[k]})}{p_\theta(\boldsymbol{x}^{[k]})} \frac{g_k(\boldsymbol{x}^{[k]} \mid \tilde{\boldsymbol{x}}^{[k]})}{g_k(\tilde{\boldsymbol{x}}^{[k]} \mid \boldsymbol{x}^{[k]})} \exp(\beta_k \Delta R_k)\right\}. \end{aligned} \tag{17}$$

However, for masked diffusion proposals, the model ratio p_θ is intractable and the proposal density $g_k(\boldsymbol{x}^{[k]} \mid \tilde{\boldsymbol{x}}^{[k]})$ is computationally expensive, as it requires marginalizing over all possible masking configurations. We therefore adopt a simplified proxy:

$$A_k(\boldsymbol{x}^{[k]}, \tilde{\boldsymbol{x}}^{[k]}) = \min\{1, \exp(\beta_k \Delta R_k)\}. \tag{18}$$

In particular, we can show that when remasking follows the uniform masking scheme discussed in Section 2, the exact acceptance probability reduces exactly to the simplified proxy:

Proposition 3.1. *Assume the remasking distribution ρ selects a mask of fixed size $\lfloor m_k L \rfloor$ uniformly at random. Then the exact Metropolis-Hastings acceptance probability $A_k^{\text{MH}}(\boldsymbol{x}^{[k]}, \tilde{\boldsymbol{x}}^{[k]})$ for the remask-and-denoise proposal kernel g_k simplifies exactly to:*

$$\begin{aligned} A_k^{\text{MH}}(\boldsymbol{x}^{[k]}, \tilde{\boldsymbol{x}}^{[k]}) &= A_k(\boldsymbol{x}^{[k]}, \tilde{\boldsymbol{x}}^{[k]}) \\ &= \min\{1, \exp(\beta_k \Delta R_k)\}. \end{aligned} \tag{19}$$

The proof is provided in Appendix A. This result guarantees that under uniform remasking, our simplified acceptance rule exactly preserves detailed balance with respect to the tempered target distribution. While low-confidence remasking introduces asymmetries that, in practice, break this guarantee, our experiments show that the simplified acceptance

rule remains an effective proxy, producing stable sampling and strong empirical performance.

Replica exchange for global exploration. To enable communication across replicas, we periodically attempt swaps between adjacent replicas. This mechanism allows high-reward candidates discovered by exploratory (low-β) replicas to propagate to exploitative (high-β) replicas, while those same exploitative replicas receive diverse states that help them escape local modes. For replicas k and $k+1$, we propose the exchange

$$(\boldsymbol{x}^{[k]}, \boldsymbol{x}^{[k+1]}) \to (\boldsymbol{x}^{[k+1]}, \boldsymbol{x}^{[k]}), \tag{20}$$

accepted with probability. Let $\Delta R_{k,k+1} := R(\boldsymbol{x}^{[k]}) - R(\boldsymbol{x}^{[k+1]})$.

$$A_{\text{swap}} = \min\left\{1, \frac{p_{\beta_k}(\boldsymbol{x}^{[k+1]}) p_{\beta_{k+1}}(\boldsymbol{x}^{[k]})}{p_{\beta_k}(\boldsymbol{x}^{[k]}) p_{\beta_{k+1}}(\boldsymbol{x}^{[k+1]})}\right\}. \tag{21}$$

Because the shared prior cancels, this simplifies to

$$A_{\text{swap}} = \min\left\{1, \exp\left((\beta_{k+1} - \beta_k)\Delta R_{k,k+1}\right)\right\}. \tag{22}$$

Especially, the prior p_θ cancels exactly in the ratio, so the swap criterion depends only on reward differences, making it both theoretically exact and cheap to evaluate in practice.

4. Experiments

We evaluate PT-MDM across inverse protein folding and regulatory DNA sequence design.

4.1. Experimental setup

We compare PT-MDM against test-time steering baselines and, for biological sequence design, reward-fine-tuned DRAKES variants as references for performance and over-optimization. PT-MDM reports both uniform remasking, which matches Proposition 3.1, and low-confidence remasking, which is a practical refinement heuristic. We use K as a method-specific search-width parameter. The sample-efficiency comparison matches denoising-model function evaluations (NFE) and the number of final evaluated sequences, but does not normalize reward-oracle calls; Appendix B gives the baseline, metric, runtime, and compute-accounting details.

4.2. Protein sequence design

Dataset, settings, and metrics. We follow the DRAKES inverse-folding protocol (Wang et al., 2025), using the pre-trained protein MDM as a fixed generative prior and evaluating test-time methods without generator updates. We report Pred-ddG, structural fidelity (scRMSD), validity thresholds, and the combined success rate.

Algorithm 1 Discrete Parallel Tempering for Masked Diffusion Models

Input: Pretrained MDM p_θ, reward function $R(\boldsymbol{x})$, number of replicas K, number of iterations N, swap interval τ_{swap}.
Input: Temperature ladder $0 \le \beta_1 < \beta_2 < \cdots < \beta_K = \beta$.
Input: Masking-aggressiveness ladder $m_1 > m_2 > \cdots > m_K$.
Initialize: Generate initial clean sequences $\boldsymbol{x}_0^{[k]} \sim p_\theta$ from fully masked sequences for all $k = 1, \ldots, K$.
for $n = 1$ to N **do**
  # 1. Within-replica remask-and-denoise updates
  for $k = 1$ to K **do**
    Remask: Construct $\boldsymbol{x}_{\text{mask}}^{[k]}$ by remasking a fraction m_k of positions in $\boldsymbol{x}_{n-1}^{[k]}$.
    Denoise: Propose $\tilde{\boldsymbol{x}}^{[k]} \sim p_\theta(\cdot \mid \boldsymbol{x}_{\text{mask}}^{[k]})$.
    $\Delta R_k \leftarrow R(\tilde{\boldsymbol{x}}^{[k]}) - R(\boldsymbol{x}_{n-1}^{[k]})$.
    $A_k \leftarrow \min\{1, \exp(\beta_k \Delta R_k)\}$.
    Sample $u \sim \text{Uniform}(0, 1)$.
    if $u < A_k$ **then**
      $\boldsymbol{x}_n^{[k]} \leftarrow \tilde{\boldsymbol{x}}^{[k]}$ {Accept}
    else
      $\boldsymbol{x}_n^{[k]} \leftarrow \boldsymbol{x}_{n-1}^{[k]}$ {Reject}
    end if
  end for
  # 2. Replica exchange
  if $n \bmod \tau_{\text{swap}} = 0$ **then**
    for $k = 1$ to $K - 1$ **do**
      $\Delta R_{k,k+1} \leftarrow R(\boldsymbol{x}_n^{[k]}) - R(\boldsymbol{x}_n^{[k+1]})$.
      $A_{\text{swap}} \leftarrow \min\{1, \exp((\beta_{k+1} - \beta_k)\Delta R_{k,k+1})\}$.
      Sample $u \sim \text{Uniform}(0, 1)$.
      if $u < A_{\text{swap}}$ **then**
        Swap states: $\boldsymbol{x}_n^{[k]} \leftrightarrow \boldsymbol{x}_n^{[k+1]}$.
      end if
    end for
  end if
end for
Output: Final steered sample $\boldsymbol{x}_{\text{PT}} = \boldsymbol{x}_N^{[k]}$ from the coldest replica.

Results. Table 1 summarizes results on inverse protein folding. PT-MDM substantially improves Pred-ddG over existing test-time steering methods while maintaining high structural validity. Although DRAKES w/o KL obtains high Pred-ddG, it suffers from severe structural degradation, as reflected by its much larger scRMSD and low fraction of structurally valid sequences. This confirms that aggressive reward optimization without sufficient distributional constraints can lead to over-optimization. Among training-free test-time methods, PT-MDM is the strongest reward optimizer in this benchmark, while the fine-tuned DRAKES

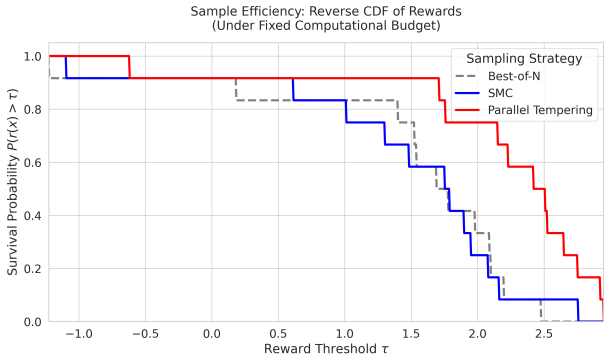

Figure 3. Discovery efficiency on inverse protein folding. The reverse-CDF curves compare the survival probability of generated sequences exceeding each Pred-ddG reward threshold under matched denoising-model NFE.

variants serve as reference points for the trade-off between reward optimization and structural fidelity. Low-confidence remasking technically breaks detailed balance, but on protein design it produces nearly the same Pred-ddG, scRMSD, and success rate as uniform remasking. We therefore treat it as an empirical heuristic rather than a uniformly better proposal. Appendix C visualizes the reward-structure trade-off directly.

Investigations. Figure 3 compares the high-reward tail under matched denoising-model NFE and the same number of final evaluated sequences. PT-MDM maintains a substantially heavier reverse-CDF tail than both Best-of-K and SMC, indicating that it converts the same denoising budget into candidates with high Pred-ddG more efficiently. Figure 1 and Appendix C provide complementary representative and aggregate evidence for the joint Pred-ddG/scRMSD trade-off, but they do not condition scRMSD on the reverse-CDF tail thresholds.

4.3. DNA sequence design

Dataset, settings, and metrics. We follow the DRAKES regulatory enhancer protocol (Wang et al., 2025), using the pretrained DNA MDM as a fixed prior over fixed-length sequences. We report Pred-Activity, ATAC-Acc, 3-mer and JASPAR correlations, approximate log-likelihood, and a diversity diagnostic.

Results. Table 2 summarizes the regulatory DNA design results. DRAKES w/o KL achieves high Pred-Activity but shows weaker agreement with natural sequence statistics than KL-regularized DRAKES, illustrating that reward improvement alone does not establish biological plausibility. PT-MDM achieves the strongest Pred-Activity among training-free methods and competitive 3-mer and JASPAR correlations, together with higher approximate log-likelihood than the fine-tuned baselines. However, its ATAC-Acc is 52.6%, compared with 92.5% for DRAKES. Because ATAC-Acc is computed by an independent accessibility classifier, this gap is consistent with reward over-optimization of Pred-Activity. Metric mismatch remains a possible explanation, but the current evaluation cannot distinguish it from reward hacking without additional biological validation. Overall, PT-MDM improves the optimized Pred-Activity reward and approximate log-likelihood without additional training, but its lower ATAC-Acc than DRAKES prevents a task-wide claim of better biological fidelity.

4.4. Ablation study for PT-MDM on protein

Component ablations. Table 3 compares three configuration axes on inverse protein folding. Relative to full PT, weak initialization reduces mean Pred-ddG from 0.918 to 0.886 and success rate from 71.19% to 70.29%, whereas the $K{=}1$ configuration reduces them to 0.702 and 63.76%, respectively. Because $K{=}1$ simultaneously removes the multi-temperature ladder, its temperature-dependent remasking ladder, and replica exchange, this result supports the combined multi-replica configuration but does not identify which individual mechanism causes the improvement.

Theory versus practice in remasking. The remasking ablation separates the sampler's theoretical validity from the proposal's practical efficiency. Uniform remasking is the theoretical reference point: under Proposition 3.1, it makes the simplified acceptance rule exact for the fixed-size remask-and-denoise proposal. Low-confidence remasking is instead a heuristic that biases edits toward uncertain positions. On protein design, the two variants are nearly identical in Pred-ddG (0.918 versus 0.919) and success rate (71.19% versus 71.16%), and the uniform variant still outperforms SMC and Best-of-K on both metrics. On DNA design, low-confidence remasking improves Pred-Activity (6.854 versus 5.876) and ATAC-Acc (52.6% versus 20.4%), but has lower approximate log-likelihood (-204.541 versus -189.199). This contrast suggests a task-dependent trade-off: confidence-guided edits may identify more reward-relevant positions in DNA, whereas the protein denoiser may already provide similarly effective proposals under either strategy. The current aggregate metrics support this hypothesis but do not establish its mechanism.

5. Conclusion

In this work, we present Parallel Tempering for Masked Diffusion Models (PT-MDM), a test-time framework designed to steer MDMs toward high-reward regions without reward-specific fine-tuning. By coupling the reward temperatures of parallel replicas to their sequence remasking fractions, PT-MDM addresses the exploration-exploitation

Table 1. (Protein sequence design) Model performance on inverse protein folding. Method* indicates our implementation method; other methods are reported as reference results from DRAKES (Wang et al., 2025).

Group	Method	Pred-ddG ↑	%(Pred-ddG>0) ↑	scRMSD ↓	%(scRMSD<2) ↑	Success Rate ↑
Reference	Pretrained	-0.544 ± 0.037	36.6 ± 1.0	0.849 ± 0.013	90.9 ± 0.6	34.4 ± 0.5
Fine-tuned	DRAKES w/o KL	**1.108 ± 0.004**	**100.0 ± 0.0**	7.307 ± 0.054	34.1 ± 0.2	34.1 ± 0.2
	DRAKES	1.095 ± 0.026	86.4 ± 0.2	0.918 ± 0.006	91.8 ± 0.5	**78.6 ± 0.7**
Test-time guidance	CG	-0.561 ± 0.045	36.9 ± 1.1	0.839 ± 0.012	90.9 ± 0.6	34.7 ± 0.9
	SMC	0.659 ± 0.044	68.5 ± 3.1	0.841 ± 0.006	93.8 ± 0.4	63.6 ± 4.0
	TDS	0.674 ± 0.086	68.2 ± 2.4	0.834 ± 0.001	94.4 ± 1.2	62.9 ± 2.8
	Best-of-K (K=10)*	0.361 ± 0.031	59.29 ± 0.25	0.528 ± 0.004	93.86 ± 0.19	56.23 ± 0.43
Ours	PT-MDM (K=10, low-confidence)*	0.918 ± 0.003	74.38 ± 0.49	0.532 ± 0.002	**94.76 ± 0.03**	71.19 ± 0.42
	PT-MDM (K=10, uniform)*	0.919 ± 0.015	74.80 ± 0.55	**0.523 ± 0.003**	94.60 ± 0.72	71.16 ± 0.61

Table 2. (DNA sequence design) Model performance on regulatory DNA sequence design. Fine-tuned methods are DRAKES reference results, while Method* denotes our implementation or reproduction.

Group	Method	Pred-Activity ↑	ATAC-Acc (%) ↑	3-mer Corr ↑	JASPAR Corr ↑	App-Log-Lik ↑
Reference	Pretrained	0.17 ± 0.04	1.5 ± 0.2	-0.061 ± 0.034	0.249 ± 0.015	-261 ± 0.6
Fine-tuned	DRAKES w/o KL	6.44 ± 0.04	82.5 ± 2.8	0.307 ± 0.001	0.557 ± 0.015	-281 ± 0.6
	DRAKES	5.61 ± 0.07	**92.5 ± 0.6**	**0.887 ± 0.002**	**0.911 ± 0.002**	-264 ± 0.6
Test-time guidance	CG	3.30 ± 0.00	0.0 ± 0.0	-0.065 ± 0.001	0.212 ± 0.035	-266 ± 0.6
	SMC	4.15 ± 0.33	39.9 ± 8.7	0.840 ± 0.045	0.756 ± 0.068	-259 ± 2.5
	TDS	4.64 ± 0.21	45.3 ± 16.4	0.848 ± 0.008	0.846 ± 0.044	-257 ± 1.5
	Best-of-K (K=10)*	1.787 ± 0.077	7.6 ± 1.0	0.530 ± 0.033	0.603 ± 0.022	-254.399 ± 0.701
Ours	PT-MDM (K=10, low-confidence)*	**6.854 ± 0.023**	52.6 ± 0.5	0.870 ± 0.003	0.818 ± 0.005	-204.541 ± 0.278
	PT-MDM (K=10, uniform)*	5.876 ± 0.017	20.4 ± 3.2	0.848 ± 0.005	0.787 ± 0.009	**-189.199 ± 0.978**

trade-off in multimodal reward landscapes. The MDM remask-and-denoise process provides a proposal family for discrete reward-guided sampling: a temperature-dependent masking ratio controls proposal scale, while a shared reward-independent denoising model supplies prior-aware reconstructions. On inverse protein folding, PT-MDM improves Pred-ddG over the evaluated test-time baselines while maintaining structural validity. On regulatory DNA design, it achieves the highest Pred-Activity among the evaluated training-free methods, although the ATAC-Acc gap indicates that reward improvement does not by itself establish chromatin accessibility.

Despite its strong empirical performance, we identify a few limitations that motivate future work. For example, the current temperature-scaled proposal mechanism selects tokens for remasking largely independently. This risks disrupting higher-order interactions or coordinated motifs, such as long-range amino acid contacts in folded proteins. We find that the pretrained MDM's dense conditioning context naturally helps preserve these global structures during the denoising step. Consequently, sequence validity remains high. Nevertheless, future work exploring token-level, motif-aware, or structure-informed remasking strategies could enable correlated proposals that accelerate reward improvement. In addition, maintaining K parallel replicas introduces an approximately K-fold increase in memory usage relative to a single chain, which may become a practical bottleneck for very large models or long sequences. Adaptive temperature ladders and dynamic remasking schedules may help mitigate this cost by allocating computation more efficiently while improving exploration. Finally, extending PT-MDM to hybrid continuous-discrete diffusion models could broaden its applicability across a wider range of scientific design and discovery problems.

Acknowledgments

We thank the anonymous reviewers and members of CLLab for their constructive feedback. This work is partially supported by the National Science and Technology Council in Taiwan via NSTC 113-2634-F-002-008 and 114-2221-E-002-102-MY3, the NTU AI Center of Research Excellence within the Taiwan Centers of Excellence, and the NTU Center for Data Intelligence via NTU-114L900901. We thank the National Center for High-performance Computing (NCHC) of the National Institutes of Applied Research (NIAR) in Taiwan for providing computational and storage resources. H.-T. Lin is honored to be supported by the Leap Fellowship of the Foundation for the Advancement of Outstanding Scholarship in Taiwan since 2025.

Table 3. Protein inverse-folding ablation of PT-MDM components at $K=10$. The full PT baseline uses $T=50$ low-confidence remasking; each ablation changes one axis.

Axis	Configuration	Pred-ddG ↑	%(Pred-ddG>0) ↑	scRMSD ↓	%(scRMSD<2) ↑	Success Rate ↑
Full method	Full PT ($T=50$, low-conf, $K=10$)	0.918 ± 0.003	74.38 ± 0.49	0.532 ± 0.002	$\mathbf{94.76 \pm 0.03}$	$\mathbf{71.19 \pm 0.42}$
Initialization quality	Weak init ($T=5$)	0.886 ± 0.008	74.50 ± 0.11	$\mathbf{0.518 \pm 0.002}$	94.10 ± 0.66	70.29 ± 0.61
Remasking strategy	Uniform remask	$\mathbf{0.919 \pm 0.015}$	$\mathbf{74.80 \pm 0.55}$	0.523 ± 0.003	94.60 ± 0.72	71.16 ± 0.61
Temperature ladder	Single-chain ($K=1$)	0.702 ± 0.024	67.40 ± 0.67	0.521 ± 0.001	94.44 ± 0.21	63.76 ± 0.76

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning, with applications to generative modeling for biological sequence design. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here. As with any reward-driven generative method applied to biology, practitioners should track biological plausibility and reward-model over-optimization when using these methods to inform downstream experimental decisions.

References

Avrahami, E. and Nachmani, E. Ilrr: Inference-time steering method for masked diffusion language models. *arXiv preprint arXiv:2601.21647*, 2026.

de Groot, L., Kuiper, R. J. A., and Bagheri, A. A survey of discrete diffusion for text and genomic sequence generation. In *The 37th Benelux Conference on Artificial Intelligence and the 34th Belgian Dutch Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=ubOzVt7oMw.

Earl, D. J. and Deem, M. W. Parallel tempering: Theory, applications, and new perspectives. *Physical Chemistry Chemical Physics*, 7(23):3910–3916, 2005.

Gheshlaghi Azar, M., Daniel Guo, Z., Piot, B., Munos, R., Rowland, M., Valko, M., and Calandriello, D. A general theoretical paradigm to understand learning from human preferences. In Dasgupta, S., Mandt, S., and Li, Y. (eds.), *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 4447–4455. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/gheshlaghi-azar24a.html.

Go, D., Korbak, T., Kruszewski, G., Rozen, J., Ryu, N., and Dymetman, M. Aligning language models with preferences through f-divergence minimization. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Guo, Y., Yang, Y., Yuan, H., and Wang, M. Training-free guidance beyond differentiability: Scalable path steering with tree search in diffusion and flow models. *Advances in Neural Information Processing Systems*, 38:73343–73384, 2026.

He, J., Jeha, P., Potaptchik, P., Zhang, L., Hernández-Lobato, J. M., Du, Y., Syed, S., and Vargas, F. CREPE: Controlling diffusion with REPlica exchange. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=uOZRWcbiZl.

Jain, V., Sareen, K., Pedramfar, M., and Ravanbakhsh, S. Diffusion tree sampling: Scalable inference-time alignment of diffusion models. *Advances in Neural Information Processing Systems*, 38:145576–145615, 2026.

Korbak, T., Elsahar, H., Kruszewski, G., and Dymetman, M. On reinforcement learning and distribution matching for fine-tuning language models with no catastrophic forgetting. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=XvI6h-s4un.

Mudgal, S., Lee, J., Ganapathy, H., Li, Y., Wang, T., Huang, Y., Chen, Z., Cheng, H.-T., Collins, M., Strohman, T., Chen, J., Beutel, A., and Beirami, A. Controlled decoding from language models. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=bVIcZb7Qa0.

Nie, S., Zhu, F., You, Z., Zhang, X., Ou, J., Hu, J., ZHOU, J., Lin, Y., Wen, J.-R., and Li, C. Large language diffusion models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=KnqiC0znVF.

Oh, M., Kim, G., and Lee, S. Parallel tempering initial sampling in inference-time reward alignment. *arXiv preprint arXiv:2605.30991*, 2026.

Ou, J., Nie, S., Xue, K., Zhu, F., Sun, J., Li, Z., and Li, C. Your absorbing discrete diffusion secretly models the

conditional distributions of clean data. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=sMyXP8Tanm.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=HPuSIXJaa9.

Rector-Brooks, J., Hasan, M., Peng, Z., Liu, C., Mittal, S., Dziri, N., Bronstein, M., Chatterjee, P., Tong, A., and Bose, J. Steering masked discrete diffusion models via discrete denoising posterior prediction. In *International Conference on Learning Representations*, volume 2025, pp. 25383–25414, 2025a.

Rector-Brooks, J., Hasan, M., Peng, Z., Liu, C.-H., Mittal, S., Dziri, N., Bronstein, M. M., Chatterjee, P., Tong, A., and Bose, J. Steering masked discrete diffusion models via discrete denoising posterior prediction. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=Ombm8S40zN.

Sahoo, S. S., Arriola, M., Gokaslan, A., Marroquin, E. M., Rush, A. M., Schiff, Y., Chiu, J. T., and Kuleshov, V. Simple and effective masked diffusion language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=L4uaAR4ArM.

Schiff, Y., Sahoo, S. S., Phung, H., Wang, G., Boshar, S., Dalla-torre, H., de Almeida, B. P., Rush, A. M., PIERROT, T., and Kuleshov, V. Simple guidance mechanisms for discrete diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=i5MrJ6g5G1.

Schiff, Y., Belhasin, O., Uziel, R., Wang, G., Arriola, M., Turok, G., Elad, M., and Kuleshov, V. Learn from your mistakes: Self-correcting masked diffusion models, 2026. URL https://arxiv.org/abs/2602.11590.

Shi, J., Han, K., Wang, Z., Doucet, A., and Titsias, M. Simplified and generalized masked diffusion for discrete data. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=xcqSOfHt4g.

Wang, C., Uehara, M., He, Y., Wang, A., Lal, A., Jaakkola, T., Levine, S., Regev, A., Hanchen, and Biancalani, T. Fine-tuning discrete diffusion models via reward optimization with applications to DNA and protein design. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=G328D1xt4W.

Wang, G., Schiff, Y., Sahoo, S. S., and Kuleshov, V. Remasking discrete diffusion models with inference-time scaling. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=IJryQAOy0p.

Wu, L., Trippe, B. L., Naesseth, C. A., Cunningham, J. P., and Blei, D. Practical and asymptotically exact conditional sampling in diffusion models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=eWKqr1zcRv.

Yang, J., Chu, W., Khalil, D., Astudillo, R., Wittmann, B. J., Arnold, F. H., and Yue, Y. Steering generative models with experimental data for protein fitness optimization. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=Ice2BHIumz.

Zheng, K., Chen, Y., Mao, H., Liu, M.-Y., Zhu, J., and Zhang, Q. Masked diffusion models are secretly time-agnostic masked models and exploit inaccurate categorical sampling. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=CTC7CmirNr.

A. Proof of Proposition 3.1

Proof. Let $\boldsymbol{x}^{[k]} = (x_1^{[k]}, \dots, x_L^{[k]})$ and $\tilde{\boldsymbol{x}}^{[k]} = (\tilde{x}_1^{[k]}, \dots, \tilde{x}_L^{[k]})$ denote the current and proposed discrete sequences of length L, respectively. For any subset of indices $A \subseteq \{1, \dots, L\}$, we let $\boldsymbol{x}_A^{[k]}$ denote the sub-sequence of tokens at the positions specified by A, and let $\boldsymbol{x}_{\setminus A}^{[k]}$ denote the unmasked context, namely the sub-sequence of tokens at the complementary positions $\{1, \dots, L\} \setminus A$.

Let $M \subset \{1, \dots, L\}$ denote the set of indices chosen to be masked. Under the assumption of uniform remasking with a fixed masking fraction m_k, the size of the mask is strictly defined as $|M| = \lfloor m_k L \rfloor$. The probability of selecting any specific mask M of this size is a combinatorial constant independent of the actual token values in the sequence:

$$\rho(\boldsymbol{x}_{\setminus M}^{[k]} \mid \boldsymbol{x}^{[k]}, m_k) = \rho(\boldsymbol{x}_{\setminus M}^{[k]} \mid \tilde{\boldsymbol{x}}^{[k]}, m_k) = \frac{1}{\binom{L}{\lfloor m_k L \rfloor}} \equiv C, \tag{23}$$

where we slightly abuse notation and use $\boldsymbol{x}_{\setminus M}^{[k]}$ to denote the remasked version of $\boldsymbol{x}^{[k]}$.

Next, let $D = \{i \in \{1, \dots, L\} : x_i^{[k]} \neq \tilde{x}_i^{[k]}\}$ denote the set of indices where the current and proposed sequences differ. For the forward proposal kernel $g_k(\tilde{\boldsymbol{x}}^{[k]} \mid \boldsymbol{x}^{[k]})$ to have a non-zero probability of generating $\tilde{\boldsymbol{x}}^{[k]}$, the sampled mask M must cover all differing tokens; otherwise, the unmasked tokens would prevent the transition. Therefore, we require $D \subseteq M$. Consequently, for any valid mask M, the unmasked context remains identical between the two sequences: $\boldsymbol{x}_{\setminus M}^{[k]} = \tilde{\boldsymbol{x}}_{\setminus M}^{[k]}$.

Using this property, we can write the forward proposal kernel as a sum over all valid masks that contain the differing tokens:

$$\begin{aligned} g_k(\tilde{\boldsymbol{x}}^{[k]} \mid \boldsymbol{x}^{[k]}) &= \sum_{M \supseteq D} p_\theta(\tilde{\boldsymbol{x}}_M^{[k]} \mid \boldsymbol{x}_{\setminus M}^{[k]}) \rho(\boldsymbol{x}_{\setminus M}^{[k]} \mid \boldsymbol{x}^{[k]}, m_k) \\ &= C \sum_{M \supseteq D} p_\theta(\tilde{\boldsymbol{x}}_M^{[k]} \mid \boldsymbol{x}_{\setminus M}^{[k]}). \end{aligned} \tag{24}$$

By the standard definition of conditional probability, the joint probability of the proposed sequence under the pre-trained model factorizes into the masked and unmasked components: $p_\theta(\tilde{\boldsymbol{x}}^{[k]}) = p_\theta(\tilde{\boldsymbol{x}}_M^{[k]} \mid \tilde{\boldsymbol{x}}_{\setminus M}^{[k]}) p_\theta(\tilde{\boldsymbol{x}}_{\setminus M}^{[k]})$. Rearranging this expression yields $p_\theta(\tilde{\boldsymbol{x}}_M^{[k]} \mid \tilde{\boldsymbol{x}}_{\setminus M}^{[k]}) = \frac{p_\theta(\tilde{\boldsymbol{x}}^{[k]})}{p_\theta(\tilde{\boldsymbol{x}}_{\setminus M}^{[k]})}$. Substituting this ratio into the forward proposal, and applying the established equality $\boldsymbol{x}_{\setminus M}^{[k]} = \tilde{\boldsymbol{x}}_{\setminus M}^{[k]}$, we obtain:

$$\begin{aligned} g_k(\tilde{\boldsymbol{x}}^{[k]} \mid \boldsymbol{x}^{[k]}) &= C \sum_{M \supseteq D} \frac{p_\theta(\tilde{\boldsymbol{x}}^{[k]})}{p_\theta(\boldsymbol{x}_{\setminus M}^{[k]})} \\ &= C\, p_\theta(\tilde{\boldsymbol{x}}^{[k]}) \sum_{M \supseteq D} \frac{1}{p_\theta(\boldsymbol{x}_{\setminus M}^{[k]})} \equiv C\, p_\theta(\tilde{\boldsymbol{x}}^{[k]})\, C', \end{aligned} \tag{25}$$

where $C' = \sum_{M \supseteq D} [p_\theta(\boldsymbol{x}_{\setminus M}^{[k]})]^{-1}$ is constant across all valid masks.

Applying the exact same logic and derivation to the reverse proposal kernel yields a symmetric result:

$$\begin{aligned} g_k(\boldsymbol{x}^{[k]} \mid \tilde{\boldsymbol{x}}^{[k]}) &= C \sum_{M \supseteq D} \frac{p_\theta(\boldsymbol{x}^{[k]})}{p_\theta(\tilde{\boldsymbol{x}}_{\setminus M}^{[k]})} \\ &= C\, p_\theta(\boldsymbol{x}^{[k]}) \sum_{M \supseteq D} \frac{1}{p_\theta(\boldsymbol{x}_{\setminus M}^{[k]})} = C\, p_\theta(\boldsymbol{x}^{[k]})\, C'. \end{aligned} \tag{26}$$

We now substitute these symmetric proposal functions into the model-ratio and proposal-ratio term of the exact Metropolis-Hastings acceptance probability:

$$\frac{p_\theta(\tilde{\boldsymbol{x}}^{[k]})\, g_k(\boldsymbol{x}^{[k]} \mid \tilde{\boldsymbol{x}}^{[k]})}{p_\theta(\boldsymbol{x}^{[k]})\, g_k(\tilde{\boldsymbol{x}}^{[k]} \mid \boldsymbol{x}^{[k]})} = \frac{p_\theta(\tilde{\boldsymbol{x}}^{[k]}) \big[C\, p_\theta(\boldsymbol{x}^{[k]})\, C'\big]}{p_\theta(\boldsymbol{x}^{[k]}) \big[C\, p_\theta(\tilde{\boldsymbol{x}}^{[k]})\, C'\big]} = 1. \tag{27}$$

Because the model priors and the forward/reverse proposal probabilities perfectly cancel one another, the exact acceptance probability simplifies purely to the energy difference defined by the reward-tilted target distribution:

$$A_k^{\mathrm{MH}}(\boldsymbol{x}^{[k]}, \tilde{\boldsymbol{x}}^{[k]}) = \min\left\{1, \exp\left(\beta_k \big[R(\tilde{\boldsymbol{x}}^{[k]}) - R(\boldsymbol{x}^{[k]})\big]\right)\right\}. \tag{28}$$

This concludes the proof. □

B. Experimental protocol details

Baselines and remasking variants. For biological sequence design, we report reward-fine-tuned DRAKES and DRAKES without KL regularization as references for assessing performance and over-optimization, rather than as direct test-time competitors. We compare with classifier guidance (CG), sequential Monte Carlo (SMC), and twisted diffusion sampling (TDS) as reported in DRAKES, and we additionally implement Best-of-K sampling. For PT-MDM, uniform remasking matches the detailed-balance setting of Proposition 3.1, while low-confidence remasking targets uncertain tokens for stronger local refinement.

Terminology. We use *candidate sequence* for any completed biological sequence that is ultimately evaluated by

the reward and validation metrics. For Best-of-K, the K objects are independently generated candidate sequences. For SMC, the K objects are particles that are reweighted and resampled during generation. For PT, the K objects are persistent replicas, each associated with a single inverse temperature on the temperature ladder. Equivalently, a PT replica is the MCMC trajectory at a fixed inverse temperature; we reserve the term particle for SMC.

Search width and compute budget. The symbol K denotes the internal search width across Best-of-K, SMC, and PT, but its algorithmic role is method-specific. The same value of K therefore does not necessarily imply the same computational cost. For sample-efficiency comparisons, we normalize by the total number of denoising model forward evaluations, which we report as the number of function evaluations (NFE). For PT, every denoising forward pass performed by any replica at any MCMC proposal step is included in the reported NFE count. Reward-oracle evaluations can differ across methods and are not matched in our NFE-normalized analysis. The comparison therefore isolates denoising-model compute rather than total evaluation cost.

Protein metrics. Following DRAKES, we report *Pred-ddG* (predicted $\Delta\Delta G$); *scRMSD*, the side-chain RMSD measuring structural fidelity; the fraction of sequences with positive Pred-ddG, denoted %(Pred-ddG> 0); and the fraction of structurally valid sequences, denoted $\%(\mathrm{scRMSD} < 2)$. The success rate is the fraction of generated sequences satisfying both Pred-ddG> 0 and $\mathrm{scRMSD} < 2$.

DNA metrics. *Pred-Activity* measures predicted enhancer activity under a held-out reward oracle. *ATAC-Acc* evaluates whether generated sequences are classified as chromatin-accessible enhancers by an independent accessibility classifier. *3-mer Corr* is the Pearson correlation between 3-mer frequencies of generated sequences and high-activity natural sequences, and *JASPAR Corr* is the correlation between transcription-factor motif occurrence profiles and those of high-activity natural sequences. *App-Log-Lik* is the approximate sequence log-likelihood under the pretrained MDM, computed as the sequence-level ELBO and averaged over generated sequences. All generated DNA sequences have the same fixed length, making approximate log-likelihood comparisons directly comparable across methods.

Reverse-CDF analysis. For sample-efficiency analysis, we compare the empirical reverse CDF of protein Pred-ddG rewards, $\widehat{P}(r(x) > \tau)$, under matched denoising-model NFE and the same number of final evaluated sequences. We estimate the survival probability at threshold τ as

$$\widehat{P}(r(x) > \tau) = \frac{1}{N_{\mathrm{eval}}} \sum_{i=1}^{N_{\mathrm{eval}}} \mathbf{1}\{r(x_i) > \tau\},$$

where N_{eval} is identical across Best-of-K, SMC, and PT-MDM. This prevents high-reward-tail comparisons from being confounded by different numbers of evaluated samples.

NFE accounting. Let T_{gen} be the number of denoising steps used to generate a completed sequence, T_{init} the number of denoising steps used to initialize PT, and T_{prop} the number of denoising-model forward evaluations used by one remask-and-denoise proposal. The following table summarizes how the same search-width symbol K maps to denoising-model NFE for each method. Reward-oracle calls are not included, so the NFE comparison measures denoising-model compute rather than total evaluation cost.

Experimental settings Protein and DNA aggregate results are reported as means and standard deviations over three independent seeds. Tables 1 and 2 distinguish our implementations from reference values reported by DRAKES; we do not interpret cross-method seed variation as a formal significance test.

Computational costs All experiments were run on NVIDIA RTX 6000 Ada GPUs. Per-seed runtime and GPU memory are approximately: Full PT baseline ($K = 10$), 35 minutes and 20 GB; Best-of-K, 30 minutes and 20 GB; uniform remasking ($K = 10$), 35 minutes and 20 GB; baseline without guidance, 5 minutes and 5 GB; and SMC, 35 minutes and 20 GB.

C. Protein reward-structure trade-off

Figure 4 visualizes the inverse-folding trade-off by plotting Pred-ddG against scRMSD. The desired region contains sequences with positive Pred-ddG and low structural deviation. PT-MDM shifts generated samples toward this high-reward, structurally plausible region, without the severe increase in scRMSD observed under reward-only optimization.

Table 4. Denoising-model NFE accounting for Best-of-K, SMC, and PT. The formulas are written per final reported candidate; multiplying by N_{eval} gives the total budget for evaluating N_{eval} final candidates.

Method	**Role of K**	**Denoising-model NFE per final candidate**	**Reward-oracle evaluations**
Best-of-K	K independent completed candidates are sampled, then the best candidate is selected by terminal reward.	$K\,T_{\text{gen}}$	Typically K terminal reward evaluations per selected candidate.
SMC	K particles are propagated through the denoising process with intermediate reweighting and resampling.	$K\,T_{\text{gen}}$ for a standard one-forward-pass-per-particle-per-step implementation. Extra corrector or proposal calls should be added explicitly.	Depends on the number of reward-weighting stages; report separately from NFE.
PT	K persistent replicas evolve on an inverse-temperature ladder, with each replica making MCMC proposals and optional replica exchanges.	$T_{\text{init}} + K\,T_{\text{prop}}$ when one initialized sequence is shared across replicas. Replica swaps add no denoising-model NFE.	Depends on how often rewards are computed for MH acceptance and swap decisions; report separately from NFE.

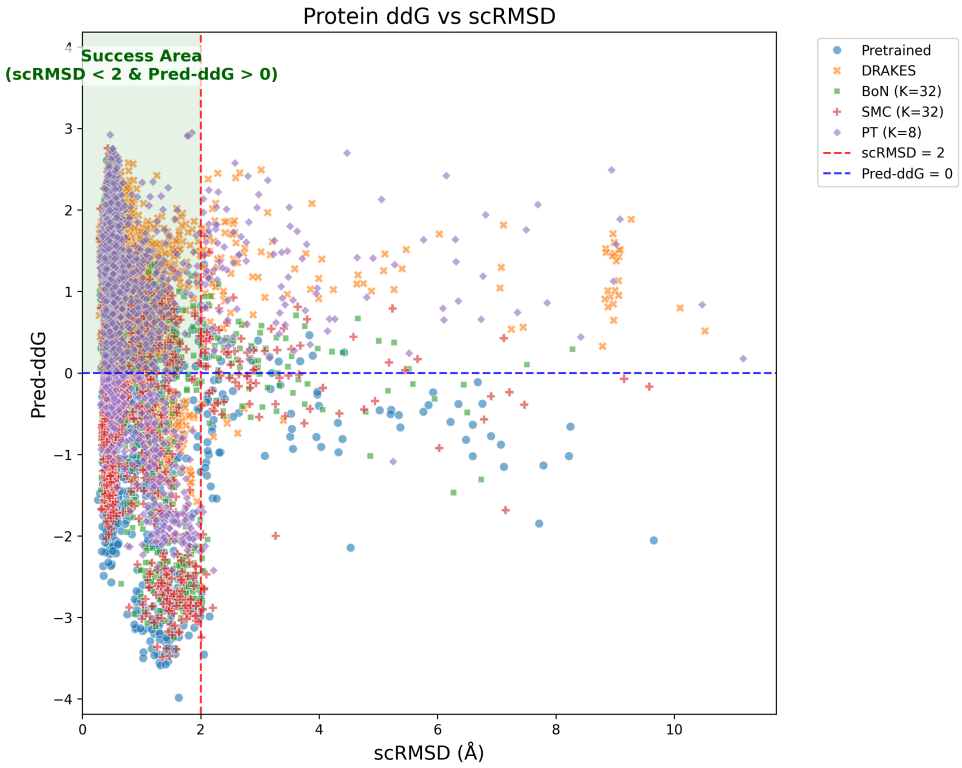

Figure 4. Protein reward-structure trade-off. Pred-ddG is plotted against scRMSD to show whether methods improve the optimized reward while preserving the target fold.