Entropy-Regularized Diffusion-Policies in Offline Reinforcement Learning for Antibody Sequence Design

Anonymous authors
Paper under double-blind review

Abstract

The discovery of therapeutic antibodies is traditionally performed through wet lab screening, which is costly and time-consuming. Generative models offer a data-driven alternative, however such methods become unreliable outside the training distribution. We present Sequential Diffusion + Q-Learning (SeqDiff+QL), which formulates antibody sequence design as a constrained offline Reinforcement Learning (RL) problem, enforcing proximity to the training distribution. SeqDiff+QL employs an entropy-regularized diffusion policy that, through policy improvement, is trained sequentially generate Complementarity Determining Region (CDR) sequences with higher predicted binding affinity based on a variety of training distributions. Our novel entropy regularization thereby promotes diverse candidate generation, while the integration of biophysical priors through contrastive Variational Autoencoder (VAE) latent representations improves the stability of the generative process. The framework can learn from heterogeneous sequence sources across different training distributions. Using the Absolut! simulator and Rosetta energy function as affinity evaluation oracles, we show that SeqDiff+QL produces candidate sequences with improved predicted affinity across multiple target antigens while maintaining diversity.

1 Introduction

Antibodies are a class of proteins with great potential for treating diseases such as cancer (Kaplon et al., 2023; Norman et al., 2020; Robert et al., 2022). However, the discovery of therapeutic antibodies in classical wet lab experiments is constrained by high costs and low throughput (Angermueller et al., 2019; Shanehsazzadeh et al., 2024; Vamathevan et al., 2019). Computational approaches to antibody design, such as generative models, have emerged as a powerful tool for addressing these challenges, offering a data-driven approach, significantly reducing the time and resources required (Shanehsazzadeh et al., 2024). In prior works, specific requirements in the field of protein design have been identified, translating to *algorithmic requirements* for computational approaches to protein design.

Objective-driven design: The ability to optimize desired properties such as target specificity and binding affinity instead of generating outputs similar to training data (Gruver et al., 2024; Shanehsazzadeh et al., 2024; Vamathevan et al., 2019). **Diversity**: Approaches should employ diverse, non-deterministic design schemes to account for the diverse nature of real-world antibody-antigen interaction (Jain et al., 2022), non-deterministic behavior of proteins (Wang et al., 2025), and the limited ability of in silico simulations to mimic real-world interaction, and overall low success rates in real-world evaluation (Vamathevan et al., 2019). **Offline Learning:** Applicability in the offline setting, which enables training on and improving upon pre-collected offline data without repeated online access to evaluation methods, as such interactions are infeasible due to the respective time and cost involved (Shanehsazzadeh et al., 2024).

A promising class of algorithms is diffusion or flow-matching models, which recently have received considerable attention due to their ability to model multimodal data distributions and generate diverse and high-quality data (Murphy, 2023). Their versatility makes them applicable to numerous tasks in the realm of protein design, including protein structure prediction (Anand & Achim, 2022; Wang et al., 2025), protein-protein docking (Ketata et al., 2023), as well as protein sequence and structure (co-)design (Alamdari et al.,

2023; Campbell et al., 2024; Chen et al., 2024b; Jin et al., 2022; Kim et al., 2024; Luo et al., 2022; Martinkus et al., 2023; Verma et al., 2023; Watson et al., 2023; Zhou et al., 2024).

However, basic diffusion only models a given data distribution and does not optimize for a desired objective, such as binding affinity to an antigen, leading to the development of many guiding methods (Dhariwal & Nichol, 2021; Ho & Salimans, 2022; Park et al., 2025; Wang et al., 2023; Zheng et al., 2023). While some methods, such as Direct Preference Optimization (DPO) (Wallace et al., 2023) and Noise Contrastive Alignment (NCA) (Chen et al., 2024a), reweight samples within the data distribution according to a preference or reward, RL based methods, due to the maximization of a learned Q-function, can explicitly guide the model towards higher-reward regions of sequence space. However, learned affinity estimates are unreliable far from the training distribution, so unconstrained optimization may produce sequences with high predicted scores but low real-world evaluation performance. Therefore, optimization must balance improvement with proximity to the empirical data distribution (Fujimoto & Gu, 2021; Levine et al., 2020).

In this work, we approach the task of designing antigen-specific antibodies with maximized binding affinity. Specifically, we focus on antibody sequence design, in contrast to antibody structure design or sequence structure co-design. Similar to Angermüller et al. (2020) and Jain et al. (2022), we choose a stepwise approach to antibody sequence, one amino acid (AA) at a time, and formulate the task as an Markov Decision Process facilitating the use of RL methods to maximize affinity. We propose a new offline RL algorithm, which improves predicted affinity through policy improvement while constraining the policy to the proximity of the data distribution and maintaining diversity through entropy regularization, thereby satisfying the algorithmic requirements stated above. We evaluate our method using the Absolut! simulator (Robert et al., 2022) and the Rosetta energy function (Alford et al., 2017; Simons et al., 1997) and show:

- Continuous diffusion policies employing RL enable constrained policy improvement for antibody sequence design under a data-distribution constraint
- A entropy regularization allows simultaneous affinity improvement and generation of diverse candidate sequences, often improving over non-regularized agents
- Incorporating biophysical priors into the generative process using contrastive latent representations can be beneficial if priors align with evaluation properties
- That learned Q-functions can be used to identify high-affinity sequences after the generative process
- The method supports offline learning from heterogeneous sequence distributions, including random sequences, sequences generated by methods found in literature (Dauparas et al., 2022; Vogt et al., 2023; Watson et al., 2023), and murine antibody sequences

2 Background

In this section, we provide the background on antibody sequence design, RL, and latent diffusion models.

2.1 Antibody Sequence Design as an RL task

Antibodies are a class of proteins, consisting of a sequence of AAs, utilized by the immune system to recognize and bind foreign molecules (antigens) with high specificity (Norman et al., 2020; Robert et al., 2022). Due to their favorable binding properties, they have become the leading class of new drugs developed (Lu et al., 2020; Norman et al., 2020). When designing an antibody sequence, each of the L designed positions of an antibody sequence can be assigned one of the 20 natural AAs, resulting in a search space of 20^L candidates. While antibodies consist of hundreds of AAs (Chiu et al., 2019), the variable CDRs contain the majority of antigen-binding AAs (Norman et al., 2020). Furthermore, the third CDR of the heavy chain (CDRH3) has the largest influence on the antibodies' specificity (Xu & Davis, 2000). Thus, the design of CDRs and especially the CDRH3 is a frequently chosen sequence design task (Cowen-Rivers et al., 2022; Khan et al., 2022; Liu et al., 2020; Luo et al., 2022; Zhou et al., 2024). We choose to maximize the binding affinity of the antibody to the antigen as our primary objective, and measures such as novelty and diversity of generated

candidates as auxiliary objectives. In this work, we evaluate two design tasks. On the one hand, we design an antibody CDRH3 region of length $L = 11$, evaluated for its binding affinity towards an antigen using the Absolut! simulator. On the other hand, we design the CDR1, CDR2, and CDR3 sequences of a single-domain antibody, with a total length of $L = 28$ AAs evaluated using the Rosetta energy function (Alford et al., 2017; Simons et al., 1997). See Figure 1(a) for a visualization of such CDR regions and the design task. The vast resulting search spaces of up to $20^{28} \approx 2.68 \cdot 10^{36}$ candidates preclude an exhaustive search, thereby underscoring the potential impact of computational antibody design in therapeutics development.

In RL, tasks are typically formulated as MDPs. We define a deterministic MDP as a tuple $\langle S, S_0, A, P, R \rangle$, where S is the set of possible states, $S_0 \subseteq S$ is the set of initial states, A is the set of actions, $P(s, a) : S \times A \mapsto S$ is a deterministic transition function, and $R(s, a) : S \times A \mapsto \mathbb{R}$ is a deterministic reward function.

Like Jain et al. (2022) and Angermueller et al. (2019), we choose to frame the task of designing discrete AA sequences as a stepwise generation process where the AAs are placed in the sequence one after the other. Thus, we define the set of states S as the set of all possible AA sequences up to length L, including the empty sequence. We define the set S_0 as an empty AA sequence. The set A is then defined as the set of 20 natural AAs. Note that we use the symbol a throughout this work to refer to the RL action and the AA it represents. Consequently, we define $P(s, a) = s \parallel a$ as the concatenation of the sequence generated thus far with the next AA, extending the sequence by one more AA. To prevent variable-size representations for s, we use padding tokens and a fixed sequence length L. The reward function $R(s, a)$ is defined corresponding to the predicted free energy using the Absolut! simulator or Rosetta energy function. As sequences of length shorter than L AAs can not be evaluated in our tasks, the reward function is sparse, returning the predicted free energy for sequences of length L and a reward of 0 for all shorter sequences.

The objective in RL is to learn a policy π that maximizes the expected sum of rewards. The action-value function Q represents this expected sum starting from a given state s_t. As the search space of CDR sequences is huge, we estimate π and Q with function approximations π_θ and $Q_\phi(s, a)$, parameterized by θ and ϕ respectively. We define $Q_\phi(s_t, a_t) := \mathbb{E}_{\pi_\theta}[R(s_t, a_t) + \sum_{i=1}^{L-1-t} R(s_{t+i}, a_{t+i}) | a_{t+i} \sim \pi_\theta(a_{t+i}|s_{t+i})]$. An optimal policy π thus selects the action a that maximizes Q for each state s. In our method, the policy $\pi_\theta(a_t|s_t)$ will be implemented as a continuous diffusion policy. Thus, the action $a_t \sim \pi_\theta(a_t|s_t)$ will not be a categorical AA but instead a two-dimensional continuous representation $\boldsymbol{a}_t$ encoding the respective AA. To create such a two-dimensional representation, we train a VAE to encode the 20 AAs into a two-dimensional latent space. We freeze the VAE during the training of the policy. Keep in mind that we only use two-dimensional VAE latent representations for actions, not for states s.

In line with the algorithmic requirements identified above, we focus on the offline RL setting, where the agent is trained using only a pre-collected dataset, which we consider well-suited for the antibody design task, as continuous interactive access to a wet lab is not always feasible (Jain et al., 2022).

2.2 Diffusion Models

Diffusion Models are a class of generative models that learn to generate data by iteratively denoising samples from a Gaussian noise distribution. They employ a *forward process*, or *diffusion process*, to gradually corrupt observed data into noisy data and learn a *reverse process*, or *denoising process*, to undo the corruption. A trained model can then be used to generate high-quality data from noise (Murphy, 2023).

In this work, we are dealing with both diffusion steps $n \in \{0, ..., N\}$ and RL time steps $t \in \{0, ..., T\}$. To facilitate clarity, we will use superscripts for diffusion steps and subscripts for time steps. Diffusion probabilistic models (Ho et al., 2020; Sohl-Dickstein et al., 2015) are a class of latent variable models defined as $p_\theta(x^0) := \int p_\theta(x^{0:N}) dx^{1:N}$. Here, $x^1, ..., x^N$ are latent variables of the same dimensionality as the data sample x^0 drawn from the observed data distribution $q(x^0)$. In our setting, these data samples are embeddings $\boldsymbol{a}$ of AAs drawn from a VAE latent space. The forward process gradually adds Gaussian noise to x^0 according to a noise schedule $\beta^1, ..., \beta^N$, over N steps (Ho et al., 2020). In particular, the forward process is defined as $q(x^{1:N}|x^0) := \prod_{n=1}^{N} q(x^n|x^{n-1})$, with a single step transition $q(x^n|x^{n-1}) := \mathcal{N}(x^n; \sqrt{1-\beta^n} x^{n-1}, \beta^n \mathbf{I})$.

The reverse process is the joint distribution $p_\theta(x^{0:N})$ defined as a Markov chain starting at $p(x^N) = \mathcal{N}(x^N; 0, \mathbf{I})$ given as $p_\theta(x^{0:N}) := p(x^N) \prod_{n=1}^{N} p_\theta(x^{n-1}|x^n)$, with a learned Gaussian transition $p_\theta(x^{n-1}|x^n)$.

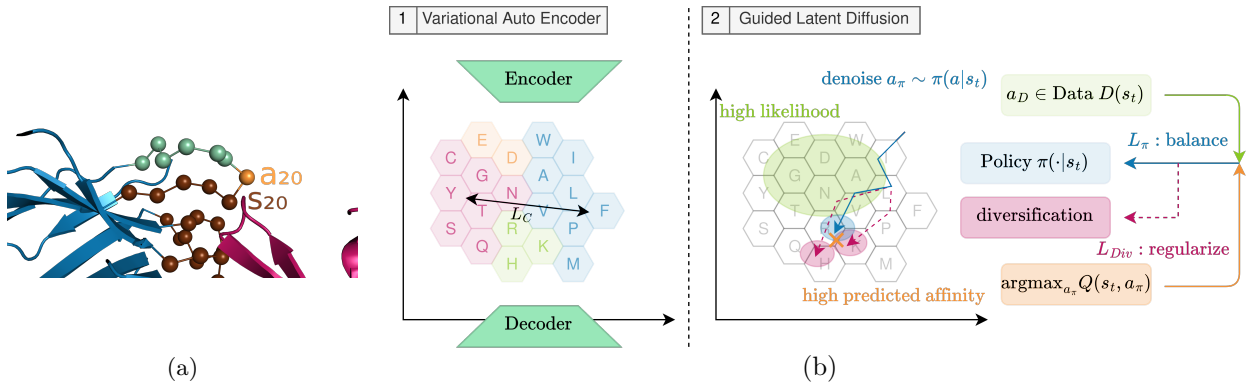
The objective of training p_θ is to maximize the expected log-likelihood of the data, given by the evidence lower bound (ELBO) $\mathbb{E}_q[\log \frac{p_\theta(x^{0:N})}{q(x^{1:N}|x^0)}]$. In essence, the objective is the reconstruction of a sample x^0 from a corresponding noisy sample x^N. This is achieved by training a noise model ϵ_θ to predict the noise introduced at each diffusion step (Ho et al., 2020; Murphy, 2023). Consequently, the loss for the diffusion model given a dataset D can be simplified to $L(\theta) = \mathbb{E}_{n\sim\text{Unif}(1,N),\epsilon\sim\mathcal{N}(0,\mathbf{I}),x^0\sim D}[||\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}^n}x^0 + \sqrt{1-\bar{\alpha}^n}\epsilon, n)||^2]$ where $\alpha^n := 1 - \beta^n$ and $\bar{\alpha}^n := \prod_{i=1}^n \alpha^i$.

3 Related Work

The field of antibody design has been studied at the level of amino acid sequences, 3D structures, and joint sequence–structure co-design (Alamdari et al., 2023; Chen et al., 2024b; Gruver et al., 2024; Li et al., 2023; Watson et al., 2023; Campbell et al., 2024; Jin et al., 2022; Kim et al., 2024; Luo et al., 2022; Martinkus et al., 2023; Verma et al., 2023; Zhou et al., 2024) and let to the development of antibodies with confirmed real-world applicability (Dauparas et al., 2022; Gruver et al., 2024; Li et al., 2023; Vázquez Torres et al., 2025). An overview of this broad field can be found in recent work by Tang et al. (2024). In our work, we will focus on the task of antibody sequence design. Sequence design approaches differ by their interaction regime. The online regime allows repeated interaction with evaluation feedback and is well-suited for methods such as Bayesian optimization and online RL (Cowen-Rivers et al., 2022; Khan et al., 2022; Vogt et al., 2023). Active learning approaches iteratively retrain models with newly evaluated sequences. In this setting, ensembles of evolutionary algorithms (Angermüller et al., 2020), RL algorithms (Angermueller et al., 2019), and Generative Flow Networks (GFlowNets) (Jain et al., 2022) have been employed as generative models. In contrast, offline methods train once on a fixed dataset and optimize without further interaction (Chen et al., 2024b; Gruver et al., 2024; Jain et al., 2022; Li et al., 2023). We adopt this last setting, which we estimate to be most suited given the high cost and time involved when evaluating designed antibody sequences (Shanehsazzadeh et al., 2024), and present contributions in this regime here.

In their approach, Chen et al. (2024b) utilize a continuous diffusion model to generate entire antimicrobial peptide (AMP) sequences in an ESM-2 (Lin et al., 2023) latent space. They demonstrated that generated peptides exhibited similar physicochemical properties to natural peptides and aligned closely with respect to AA diversity, which highlights the expressive power of their method. However, they do not employ any technique to facilitate objective-driven design. In contrast, Gruver et al. (2024) employ discrete diffusion, whereby sequences are directly sampled in the discrete sequence space. Such discrete diffusion models are not suited for naive gradient-based guidance, as the categorical sampling at each intermediate step prevents gradient propagation. To facilitate guidance, Gruver et al. (2024) propose sharing some of the diffusion models' hidden layers with a learned value function. Guidance is then applied only to the shared continuous latent space of the diffusion model and value model, by utilizing the gradient of the value model for optimization (Gruver et al., 2024). Li et al. (2023) fine-tune pre-trained language models on experimental data to predict binding affinity and uncertainty. Subsequently, Bayesian optimization is used on the learned model to design large and diverse libraries of high-affinity single-chain variable fragments. In our work, we combine and apply recent advances in diffusion models and offline RL to the protein sequence design task. Diffusion models, which are capable of modeling complex multi-modal distributions (Celik et al., 2025; Park et al., 2025; Ho et al., 2020; Wang et al., 2023), appear well suited for the complex circumstances underlying AA sequence design and diverse design schemes required for real-world applicability (Jain et al., 2022; Vamathevan et al., 2019; Wang et al., 2025). In the offline RL setting, it is often necessary to constrain the learned policy to the proximity of the data distribution to prevent exploiting erroneously high estimated actions. Fujimoto & Gu (2021) showed that combining a behavior cloning (BC) loss term and policy improvement is an effective approach in many offline RL domains. In their work, Wang et al. (2023) extend this core idea to diffusion policies, alleviating the limitation of a deterministic policy class. Besides diffusion policies, training policies to achieve diverse outputs has long been of interest, especially in the maximum entropy framework (Celik et al., 2025; Haarnoja et al., 2018). Here, we introduce entropy regularization for diffusion policies to increase the diversity of sampled sequences. Note that, while building upon ideas from maximum entropy RL research, our method is not a maximum entropy RL method, as we do not combine the entropy term with the Q-value to be maximized. Instead, we regularize the entropy of the diffusion policy for a single action selection.

4 Sequential Diffusion-Policies for Antibody Sequence Design



(a) (b)

Figure 1: (a) Antibody Sequence Design Task: For each of the L residues in the antibody CDRs, shown as beads, an AA has to be assigned to minimize the free energy between the antibody (blue and beads) and the antigen (red). We formulate the task as an iterative MDP where the AA for the current residue (shown in yellow) has to be assigned conditioned on all residues designed so far (shown in brown). Starting from an empty sequence s_0, an RL policy π_θ iteratively designs each AA a_t given the previous AAs s_t until the entire sequence s_{L-1} is designed. (b) Antibody Sequence Design Agent: (b.1) We use a VAE to encode AAs into a two-dimensional latent space. Optionally, a contrastive loss L_C (Section 4.1) can be used to group AAs based on biophysical priors, such as their side chain properties. (b.2) During inference, we sample a two-dimensional latent AA representation $\boldsymbol{a}_t \sim \pi_\theta(\boldsymbol{a}_t|s_t)$ from the diffusion policy given an incomplete AA sequence and decode it using the frozen VAE. During training, L_π (Section 4.2) balances the policy π_θ between generating AAs with high likelihood given the training dataset D and AAs that maximize a learned Q-function Q^π, which predicts sequence affinity to a given antigen. An optional entropy maximization L_{Div} (Section 4.3) can be used to increase the diversity of generated AAs.

The objective of our method, SeqDiff+QL, summarized in Figure 1, is to generate high-affinity CDR sequences while remaining in the proximity of the empirical antibody sequence distribution defined by a dataset D of sequence-affinity pairs. We approach the design of a sequence s in a stepwise manner. Conditioned on the incomplete sequence s_t, the policy $\pi_\phi(\boldsymbol{a}_t|s_t)$ is used to iteratively generate AAs which are concatenated with s_t to extend the sequence. As the generated sequences should be novel and diverse, we represent the policy π_ϕ using a continuous latent diffusion model, which can model complex distributions and generate diverse, high-quality samples (Ho et al., 2020; Murphy, 2023; Wang et al., 2023).

4.1 Encoding Biophysical Properties through Continuous Amino Acid Representations

To apply continuous diffusion policies to the task of discrete AA generation, we encode each AA as a two-dimensional latent embedding $\boldsymbol{a}$ using a VAE (Kingma & Welling, 2014). To train the VAE, each AA a is represented as a one-hot vector and mapped to a two-dimensional latent $\boldsymbol{z} = e_\omega(a) \sim \mathcal{N}(\mu^a_\omega, \sigma^a_\omega)$ using the encoder network e_ω. The decoder network $d_\psi(\boldsymbol{z})$ then maps the latent vector $\boldsymbol{z}$ back to a probability distribution over discrete AAs. The VAE is trained end-to-end by minimizing the Binary Cross Entropy (BCE) loss between the input a and the decoder's output. Additionally, the distribution of latent variables $\boldsymbol{z}$ is regularized to minimize the Kullback–Leibler (KL) divergence D_{KL} to the Gaussian distribution $D_{KL}(\mathcal{N}(\mu^a_\omega, \sigma^a_\omega)||\mathcal{N}(0, \mathbf{I}))$. This promotes a dense and continuous latent space while preventing discrepancies between the VAE latent space and the Gaussian noise $x^N \sim \mathcal{N}(x^N; 0, \mathbf{I})$ used in the diffusion process.

In an arbitrarily organized latent space, small deviations in the non-deterministic diffusion process can cause large functional changes in the generated AA. To alleviate this, we induce biophysical priors in the generative process such that AAs are grouped by functional similarities induced by their side-chain polarity, as visualized in Figure 1(b). See Section A.6 in the appendix for details on our chosen grouping. To incorporate the grouping of AAs by biophysical properties, we added a supervised contrastive loss to the VAE training objective (Khosla et al., 2020). Specifically, the contrastive loss is given by

$L_C(a) = -log[\frac{1}{|G(a)|}\sum_{p\in G(a)}\frac{\exp(\boldsymbol{z}_a\cdot\boldsymbol{z}_p)/\tau}{\sum_{a'\in A\setminus a}\exp(\boldsymbol{z}_a\cdot\boldsymbol{z}_{a'})/\tau}]$, where A is the set of all AAs, $G(a)$ represents the subset of AAs belonging to the same functional group as a, cosine similarity over latent representations $\boldsymbol{z}$ is represented by $\cdot$, and τ is a temperature hyperparameter. This loss maximizes the similarity between AAs in the same group and minimizes it between groups. The loss function of the VAE is then given as $L(a) = \text{BCE}(a, d_\psi(e_\omega(a))) + \text{KL}(\mathcal{N}(\mu^a_\omega, \sigma^a_\omega)||\mathcal{N}(0, \mathbf{I})) + L_C(a)$. Note that the VAE is pre-trained and kept frozen during the training of the diffusion process.

4.2 Guiding Diffusion Policies using Reinforcement Learning

Our training dataset includes up to 20 possible AAs for extending an incomplete sequence and is highly multimodal. This further motivates the use of continuous diffusion models as generative RL policies, which proved more effective than other training paradigms when dealing with multimodal data (Park et al., 2025; Wang et al., 2023). Thereby, the diffusion policy π_θ is trained to achieve a balance between two objectives: generating latent vectors representing AAs with high likelihood given a dataset D and generating AAs maximizing a learned Q-function.

The loss function corresponding to the first objective, referred to as the BC loss, is a slight adaptation of the standard loss function for continuous diffusion models given in Section 2.2. In particular, as we generate sequences stepwise, one AA a after the other, we condition the diffusion model on the sequence s of AAs generated so far. The resulting BC loss is given by $L_{BC}(\theta) = \mathbb{E}_{n\sim\text{Unif}(1,N),\epsilon\sim\mathcal{N}(0,\mathbf{I}),(s,a)\sim D}[||\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}^n}\boldsymbol{a} + \sqrt{1-\bar{\alpha}^n}\epsilon, s, n)||^2]$. Simply put, this loss function trains the model to reconstruct the latent representation $\boldsymbol{a}$ for the next a in a sequence from a noisy sample conditioned on an incomplete sequence s from the dataset D.

Using policy improvement, the diffusion policy can be optimized toward high-affinity sequences while remaining close to the data distribution. We follow prior work (Park et al., 2025; Wang et al., 2023) and maximize the Q-function $Q(s, \boldsymbol{a}^0)$ given an incomplete sequence s and a latent action $\boldsymbol{a}^0$ generated by the policy π_θ. As $\boldsymbol{a}^0$ is generated using the reverse process of the diffusion model π_θ, the gradient of $Q_\phi(s, \boldsymbol{a}^0)$ with respect to $\boldsymbol{a}^0$ is propagated through the diffusion model's reverse process, thereby guiding the selection of actions with a high Q-value given the current state s. In contrast to guided discrete diffusion (Gruver et al., 2024), continuous diffusion policies directly support this gradient propagation without any architectural modifications. By combining the L_{BC} with Q-maximization, we define the policy loss L_π as $L_\pi(\theta) = (1-\eta)\cdot L_{BC}(\theta) - \eta\cdot\mathbb{E}_{s\sim D,\boldsymbol{a}^0\sim\pi_\theta(\boldsymbol{a}^0|s)}[Q_\phi(s,\boldsymbol{a}^0)]$. The combination of the likelihood term L_{BC} and Q-maximization can be interpreted as regularized policy improvement in an offline setting: the diffusion likelihood term keeps the policy close to the empirical sequence distribution, effectively acting as a soft trust region around the behavior data, while the value term drives affinity improvement. A similar combination has proven effective in many offline RL domains (Fujimoto & Gu, 2021). Furthermore, using a Q-function for guidance in sequence design is particularly promising, given its ability to stitch together improved sequences from suboptimal ones and its excellent performance in states requiring specific actions (Kumar et al., 2022).

The Q-function Q_ϕ, implemented as clipped double Q-learning (Fujimoto et al., 2018), is trained to minimize the so-called TD-error: $L_{Q_i} = \mathbb{E}_{(s_t,a_t,s_{t+1})\sim D,\boldsymbol{a}^0_{t+1}\sim\pi'_\theta(\boldsymbol{a}^0_{t+1}|s_{t+1})}[||R(s_t,a_t) + \min_{j=1,2} Q_{\phi'_j}(s_{t+1},\boldsymbol{a}^0_{t+1}) - Q_{\phi_i}(s_t,\boldsymbol{a}_t)||^2]$, where subscripts t indicate the trajectory index (AA position). To prevent overestimation, we implement the Q-function as a categorical distribution as proposed by Farebrother et al. (2024). During training, the diffusion policy π_θ and the Q-function Q_ϕ are updated alternately. We refer to our algorithm, without any improvements presented in the following sections, as SeqDiff+QL. However, this algorithm is lacking a mechanism to control the diversity of generated sequence distributions, which we describe in the following section.

4.3 Increasing Diversity of generated Sequences

The ability to generate a diverse set of candidate sequences is of high importance for biological screening, due to the diverse nature of antibody-antigen interaction as well as the limited ability of in silico simulations to mimic real-world interaction (Jain et al., 2022). Finetuning Diffusion Models with RL can, however, lead to reduced diversity and mode collapse (Barceló et al., 2024).

To counteract this, we add a entropy regularization to increase the diversity of samples generated. Specifically, we introduce the auxiliary loss $L_{Div}(\theta) = \mathbb{E}_{p_\theta}[\log \frac{p_\theta(\boldsymbol{a}^{0:N}|s)}{q(\boldsymbol{a}^{1:N}|\boldsymbol{a}^0)}]$, where $\boldsymbol{a}^0$ is sampled from $p_\theta(\boldsymbol{a}^0|s)$. This auxiliary loss maximizes a lower bound on the marginal entropy of the reverse process p_θ Celik et al. (2025), by minimizing the KL divergence $D_{KL}(p_\theta(\boldsymbol{a}^{0:N}|s)||q(\boldsymbol{a}^{1:N}))$ between forward and learned backward diffusion process representing the policy $\pi_\theta(\boldsymbol{a}|s)$. Integrating this into the policy loss as $L_{\pi+Div}(\theta) = L_\pi(\theta)+\rho\cdot L_{Div}(\theta)$ allows increasing diversity in addition to preserving the trust-region constrained induced by the dataset likelihood.

4.4 Identifying promising generated Sequences

To reduce cost, it is advantageous to select the most promising generated candidates before real-world evaluation. Vázquez Torres et al. (2025) use Alphafold2 and Rosetta metrics for this task. Recall from Section 2.1 that for a sequence of length L, the Q-value $Q_\phi(s_{L-1}, \boldsymbol{a}_{L-1})$ of s_{L-1} and the last amino acid $\boldsymbol{a}_{L-1}$ is trained to predict the sequence's affinity. Thus, we use the learned Q-values as a principled scoring function, sorting generated sequences by their predicted free energy and identifying high-affinity sequences without additional training or auxiliary methods. This can be used as a post-generation selection scheme.

5 Experiments

In this section, we present experimental results across two distinct benchmark environments and multiple relevant data distributions. We design our experiments to reflect realistic variations in desired sequence length, variations in training data distribution, and variations in evaluation procedures. All experiments are carried out over five seeds. Note that the datasets and source code will be made publicly available.

5.1 Antibody Sequence Design Tasks

There are multiple tools available to estimate binding affinity, or the inversely proportional free energy, which could be used to create antibody sequence design tasks. However, many of them are available only as web servers, come with high resource demands, and provide no, outdated, or only partial code release (Abbasi et al., 2020; Jain et al., 2022; Li et al., 2024; Myung et al., 2021; Romero-Molina et al., 2022; Xue et al., 2016; Yang et al., 2023). For our analysis, we chose two tools, Absolut! (Robert et al., 2022) and Rosetta (Simons et al., 1997), which can be installed locally and are relatively lightweight, taking 6.2 and 10.2 seconds per sample evaluation on consumer hardware, respectively. We create two antibody sequence design tasks, where the goal is to maximize the affinity / minimize free energy estimated using Absolut! or Rosetta, respectively. Here, we give a short description of the task details. See Section A.1 in the appendix for an in-depth description of both tasks.

Absolut! The goal is to generate sequences of length $N = 11$ AAs, representing the CDRH3 region of an antibody, to maximize binding affinity based on a 3D lattice-discretized representation of the CDRH3 and a chosen antigen (SARS-CoV Spike Receptor-Binding Domain) and the Miyazawa-Jernigan energy potential (Miyazawa & Jernigan, 1999). We curated three diverse sequence distribution datasets, referred to as *random*, *natural*, and *expert*, reflecting distributions as they could occur in real-world applications. Dataset sizes range from 2167 to 2753 unique sequences.

Rosetta We use the Rosetta Energy Function *REF15* (Alford et al., 2017) to estimate the free energy between a nanobody (single-domain antibody) and the B-cell maturation antigen based on their 3D structure. The goal is to redesign all CDRs of the nanobody (CDR1, CDR2, and CDR3), to minimize free energy. This leads to a total of $N = 28$ AAs to be designed. We curated a random dataset, comprising 2448 sequences, and an expert dataset. To create the expert sequences, we recreated the computational sequence design pipeline as utilized by Dauparas et al. (2022) and Vázquez Torres et al. (2025), which was used to generate sequences successfully verified in multiple real-world experiments (Dauparas et al., 2022; Vázquez Torres et al., 2025). In this approach, we first sampled 1000 CDR backbone structures using RFdiffusion (Watson et al., 2023) conditioned on the nanobody and antigen structure. We then applied Protein-MPNN (Dauparas et al., 2022) to sample a total of 2483 unique sequences likely to fold into the respective structures.

5.2 Evaluation Process

The main goal in the antibody design task, and thus also the reward agents are trained to maximize, is the binding affinity of an antibody towards an antigen as described above. Similar to Jain et al. (2022), we evaluate this via the mean free energy of the top 100 candidates generated by each method. We choose to evaluate the top 100 instead of the entire set of generated sequences, as finding a few or even a single good binder often suffices, and it is not of primary interest to optimize the mean affinity of generated sequences.

Besides the primary task of maximizing the binding affinity towards a target, it is of large interest to generate novel and diverse candidate sequences. To evaluate these properties, we utilize the definition of diversity and novelty proposed by Jain et al. (2022): $Diversity(D_{gen}) := \frac{\sum_{x_i \in D_{gen}} \sum_{x_j \in D_{gen} \setminus \{x_i\}} d(x_i, x_j)}{|D_{gen}|(|D_{gen}|-1)}$ and $Novelty(D_{gen}) := \frac{\sum_{x_i \in D_{gen}} \min_{x_j \in D} d(x_i, x_j)}{|D_{gen}|}$, where D represents the training dataset and D_{gen} the dataset of generated sequences, while $d(\cdot, \cdot)$ is the Levenshtein distance quantifying the difference between two sequences (Miller et al., 2013). These auxiliary measures provide insight into the average number of pointwise mutations in the sequence relative to other sequences in the generated dataset D_{gen} (diversity) and their closest relative in the original dataset D (novelty).

In our experiments, we run the generative process of all evaluated methods until we receive 500 unique novel sequences or reach a maximal budget of 6144 generated sequences. We then compute novelty and diversity over the set of unique and novel sequences generated, and additionally count the number of duplicates in the generative process (either with sequences in the dataset or with previously generated sequences by the same method). This way, we can observe when an algorithm can be used to generate a diverse but very limited number of unique sequences.

5.3 Baselines and Implementation

We introduce a set of baselines containing classical RL algorithms, constructing sequences step by step, and Diffusion Models, generating sequences simultaneously. To facilitate a fair comparison, we use the same network architecture for shared components in all algorithms. Here, we present only a brief overview. For more details, see Section A.5 in the appendix. The baselines consist of:

Behavior Cloning (BC): Sequential BC policy, trained using cross-entropy loss to estimate the next action (AA distribution) conditioned on the current state. Suited for diverse generation and offline learning, but not for objective-driven design.

Behavior cloning + Q-learning (BC+QL): Sequential policy, combining BC with Q-learning (QL) to balance staying close to the training dataset and maximizing the affinity. This combination was previously successfully applied in continuous action settings and showed remarkable performance in the offline RL setting (Goecks et al., 2019; Fujimoto & Gu, 2021; Nair et al., 2021). We extend it to the discrete action setting. Suited for diverse, objective-driven design and offline learning.

Simultaneous Diffusion (SimDiff): Simultaneous BC diffusion policy, implemented using Denoising Diffusion Probabilistic Model (DDPM) (Ho et al., 2020), generating all AAs in the sequence at once. Suited for diverse generation and offline learning, but not for objective-driven design.

Sequential Diffusion (SeqDiff): Sequential BC diffusion policy, which models the training data distribution by sequentially sampling AAs. Equivalent to only using the L_{BC} component of our method SeqDiff+QL. Suited for diverse generation and offline learning, but not for objective-driven design.

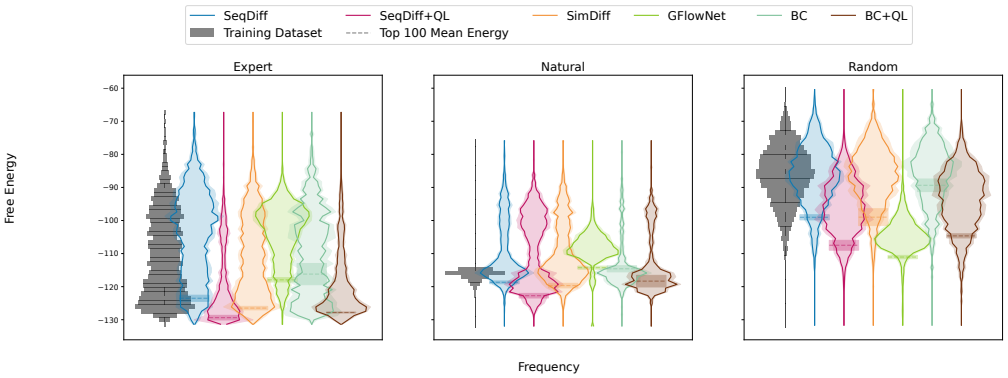
GFlowNet: Sequential policy, combining generative flows with model-based data generation. Proposed by Jain et al. (2022) to design biological sequences.

We carried out a hyperparameter search to identify suitable parameters for our method and all baselines. For details, see Section A.4 in the appendix.

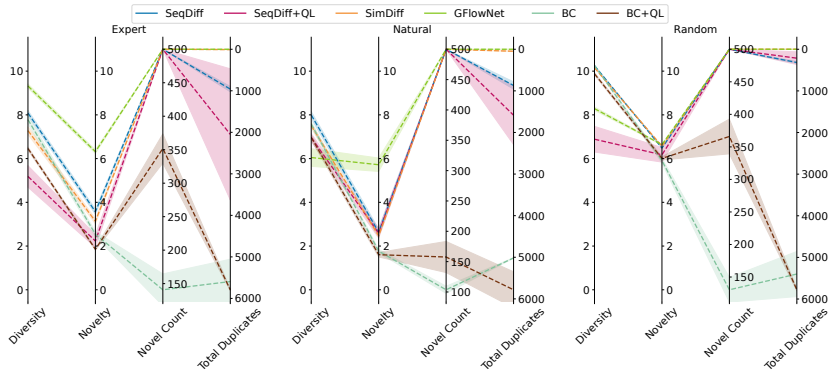
5.4 Results

We first present an analysis of our method without modifications, such as entropy regularization and biophysical priors, and baselines in both the Absolut! and Rosetta environments, followed by an ablation study for all modifications. We choose a visual presentation in this section, for tabular results and significance tests see Section A.2 in the appendix.

5.4.1 Absolut!



(a)



(b)

Figure 2: (a) Free energy distributions of the top 100 unique, novel generated sequences in the Absolut! task. Dataset distributions are visualized as histograms. Lower free energy is preferred. (b) Auxiliary Metrics: Diversity, Novelty, Novel Count and Total Duplicates, defined in Section 5.2, in the Absolut! task.

In Figure 2, we visualize the free energy distributions and auxiliary metrics on the Absolut! task.

For the methods without objective-driven design, we observe that BC and SimDiff do model the energy distribution of the expert and random dataset well, while SimDiff exhibits reduced performance on the natural dataset. Further, SeqDiff exhibits reduced modeling capability compared to these methods on the expert and natural data distribution, generating an increased amount of low-affinity sequences. This can be attributed to a higher randomness in the generative process, which is also supported by the increased diversity and novelty. This shows that while modeling of the distribution via sequential design, as indicated by BC, and via diffusion models, as indicated by SimDiff, the combination of both leads to decreased modeling capability.

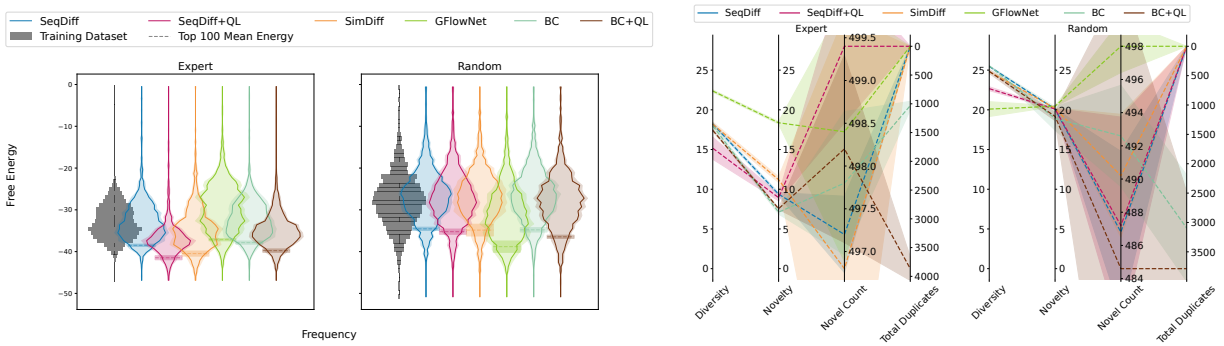
For methods employing an objective-driven design, we observe strong improvements in mean affinity of the top 100 sequences in BC+QL and SeqDiff+QL trained on the expert and random datasets, compared to their non-QL counterparts. Thereby, the mean affinity in the top 100 sequences is significantly higher for SeqDiff+QL than for all other methods, except for the random dataset. In contrast, we observe that

GFlowNet does not generate many high-affinity sequences on the expert and natural dataset but excels on the random dataset, where it achieves significantly higher affinity scores than all other methods. Note that to achieve good performance on the natural dataset, the η hyperparameter of our method SeqDiff+QL had to be decreased, highlighting the importance of dataset constraints in this setting. When trying to increase the focus on high-affinity sequences in GFlowNet via the corresponding hyperparameter, we observed an overfitting to model-bias leading to decreased diversity but not improved energy.

When analyzing the auxiliary metrics diversity, novelty, and the number of novel generated sequences and duplicated sequences, we observe that BC and BC+QL while generating a set of novel, unique sequences with comparable novelty and diversity generate significantly more duplicated sequences (either with the training dataset or previously generated sequences) and exceed the generative budget of 6144 sequences before reaching the desired amount of 500 novel, unique sequences. This indicates that while these non-diffusion methods can be adapted to perform objective-driven design, their usability to generate extended sets of novel and diverse sequences is limited. GFlowNet, on the other hand, did not create any duplicated sequences and exhibited high diversity and novelty. We further observe in BC+QL and SeqDiff+QL that the use of QL leads to a decrease in diversity and novelty. We attribute this to a decline in multi-modality, which has been previously observed for RL-based finetuning of diffusion models (Barceló et al., 2024).

Note that while we follow Jain et al. (2022) in the analysis of novelty and diversity, we observe some limitations. We observed that high-affinity sequences (as a subsample of random sequences) in the Absolut! environment generally share a similar pattern/mode and are thus less diverse per se. In other words, it is easy to be diverse (e.g., by randomly sampling sequences), but it gets harder when optimizing the affinity. A reduced diversity for a high-performing agent can thus partially be attributed to this effect, and diversity scores should always be analyzed with the corresponding energy scores in mind. See Section A.7 in the appendix for more details.

5.4.2 Rosetta



(a) (b)

Figure 3: (a) Free energy distributions of the top 100 unique, novel generated sequences in the Rosetta task. Dataset distributions are visualized as histograms. Lower free energy is preferred. (b) Auxiliary Metrics: Diversity, Novelty, Novel Count and Total Duplicates, defined in Section 5.2, in the Rosetta task.

In Figure 3, we visualize the free energy distributions and auxiliary metrics on the Rosetta task. Similar to results in Section 5.4.1, we observe an improved affinity through QL in comparison to non-QL counterparts and a significantly higher performance of SeqDiff+QL than all other methods on the expert dataset. Further, we again observe a low performance of GFlowNet on the expert dataset but high performance on the random dataset, where it achieves significantly better results. In contrast to the Absolut! task, we also observe a comparably small improvement through QL for SeqDiff+QL on the random dataset.

With respect to auxiliary metrics, we again generally observe a decrease in diversity and novelty for QL-based agents. Also GFlowNet exhibited a reduced novelty on the random dataset. On the Absolut! dataset, we did not observe any issues with duplicates being generated, which can be attributed to a higher length

of generated sequences ($L = 28$ vs $L = 11$ in Absolut!). Note that due to outlier removal from Rosetta evaluation, see Section A.1 for details, the number of granted sequences can be slightly below 500.

Lastly, we observe that most methods top 100 as well as the overall mean affinity improve upon the expert dataset. This shows potential to use such objective-driven methods to improve upon candidate distributions generated using RFdiffusion and Protein-MPNN, which currently represent a gold standard for protein engineering (Dauparas et al., 2022; Vázquez Torres et al., 2025).

5.4.3 Entropy Regularization, Biophysical Priors and Q-value based Filtering

In the following, we present ablation studies to show the effect of the introduced entropy regularization, the addition of biophysical priors, and identifying high-affinity sequences through learned Q-functions.

Entropy Regularization In our previous experiments on Absolut! and Rosetta tasks, we observed that QL led to increased affinity but reduced diversity. A possible cause for this effect is a reduced diversity in the generative process. In this section, we will analyze the effect of entropy regularization as a way to counteract this phenomenon. In Figure 4 we present the Pareto front with respect to affinity and diversity of

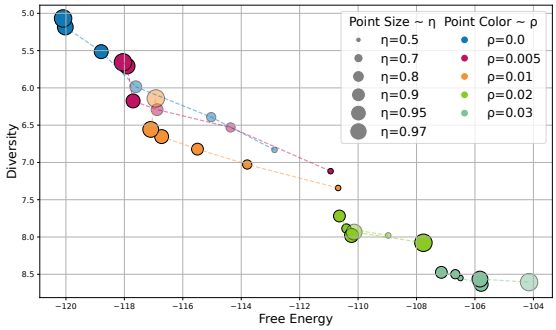
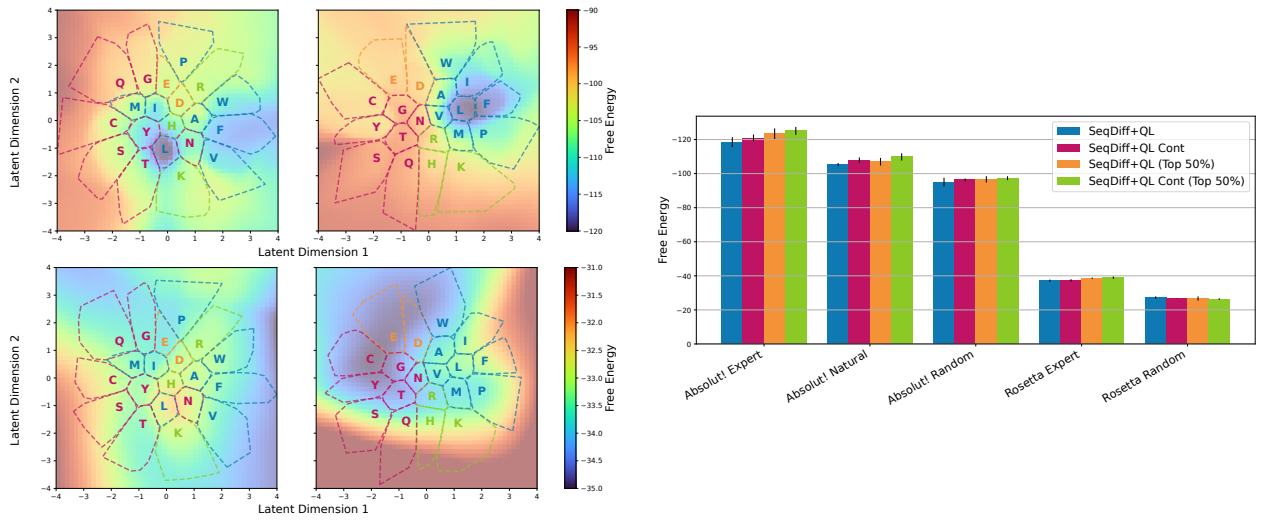


Figure 4: Pareto front over η and ρ on the Absolut! expert dataset. Agents using entropy regularization are Pareto-dominant over non-regularized configurations on large areas of the energy-spectrum, giving both higher affinity and diversity. Very high affinity scores can only be reached with non-regularized settings with high weight on Q-guidance.

novel generated sequences on the Absolut! expert dataset. We observe that there exists a range of very high affinity, which is only reachable with high η and no entropy regularization ($\rho = 0.0$, shown in blue). However, as in the previous experiments, these come with the downside of having reduced diversity. By utilizing our introduced entropy regularization, we can reach a set of configurations (shown in red and orange) which are Pareto-dominant over configurations without regularization in the high-affinity range, giving both a higher affinity and diversity compared to non-regularized agents. As a point of reference, a free energy below -114.83 is better than 99.9% of the 6.9 million murine CDRH3 sequences tested in the Absolut! publication (Robert et al., 2022). This shows that entropy regularization effectively allows for counteracting the decrease in diversity through guidance, even for high-affinity ranges.

Biophysical Priors and Q-value based filtering As the integration of biophysical priors is independent of the Q-learning mechanism, and Q-value-based filtering affects the entire sequence distribution, we carry out this analysis on a fixed $\eta = 0.9$ setting and analyze the mean energy of all generated sequences instead of the top 100 sequences. Note that the same effect is visible using other configurations and the top 100 sequences. In Figure 5(a), we visualize the effect of integrating biophysical priors into the latent space via a contrastive loss. We observe that AAs become clustered according to their group, representing their side-chain properties. Especially in the Absolut! environment, this helps to group high-valued hydrophobic AAs.

As a result, visualized in Figure 5(b), the mean affinity of generated novel sequences increases on all Absolut! environments. In the Rosetta environment, the positive effect is not observable. This can be attributed to



(a) (b)

Figure 5: (a) Learned latent spaces in the Absolut! expert (top) and Rosetta expert (bottom) task. Convex hulls indicate areas encoding for a given amino acid. The left plot shows a random encoding, while the right plot highlights the effect of contrastive losses conditioned on amino acid properties. Heatmaps indicate learned Q-values. (b) The effect of integrating biophysical priors into the VAE latent space (Cont) and filtering the top 50% of sequences according to learned Q-values (Top 50%) in our SeqDiff+QL algorithm.

our chosen prior coinciding with the Miyazawa-Jernigan energy potential (Miyazawa & Jernigan, 1999) used in the Absolut! reward function, but less so with the Rosetta energy function. See Section A.7 for a detailed analysis.

Filtering generated sequences according to their estimated Q-values mostly allows identifying good sequences, as indicated by the higher affinity of the top 50% sequences sorted by estimated Q-values, shown in Figure 5(b). These observations, like in the previous experiments, do not hold for the random Rosetta dataset. For this particular dataset, we observe a very low correlation between learned Q-values and real binding values (p=0.015). We hypothesize that this could be due to low signal in the random data and low coverage of only 2448 sequences, while designing 28 residues and thus covering a design space of 20^{28} sequences. Lastly, we observe that for all evaluations except for the one on random Rosetta data, the incorporation of priors and the filtering are compatible, leading to even higher affinity scores if used in combination.

6 Conclusion

We presented SeqDiff+QL a novel diffusion-based RL method for antibody sequence design, and evaluated it on multiple antibody sequence design tasks with varying training data distributions, sequence lengths, and evaluation functions using the Absolut! and Rosetta software. We showed that our method is applicable for diverse, objective-driven design of novel high-affinity sequences, and significantly outperforms a diverse set of baseline methods, comprising classical RL methods, diffusion methods, and GFlowNet on a majority of tasks. The comparison to GFlowNet on random data distribution highlights room for improvement of our method, which potentially could be alleviated using a model-based approach similar to GFlowNet. For all objective-driven methods, we observed an increase in affinity of generated sequences, but also a decrease in sequence diversity. We showed that our proposed entropy regularization for diffusion agents can help alleviate this problem and improve upon non-regularized agents. We further demonstrated how learned Q-values can be used to identify promising candidates in the set of generated sequences and that biophysical priors in the diffusion process can improve the affinity of generated sequences if the priors coincide with those present in the evaluation method. In conclusion, methods such as ours have the potential to have great implications for real-world biological sequence design, where the objective-driven design of diverse binders from pre-collected data could reduce cost in a field constrained by cost-intensive processes.

References

Wajid Arshad Abbasi, Adiba Yaseen, Fahad Ul Hassan, Saiqa Andleeb, and Fayyaz Ul Amir Afsar Minhas. ISLAND: in-silico proteins binding affinity prediction using sequence information. *BioData Mining*, 13(1), 2020.

Sarah Alamdari, Nitya Thakkar, Rianne van den Berg, Alex X. Lu, Nicolo Fusi, Ava P. Amini, and Kevin K. Yang. Protein generation with evolutionary diffusion: sequence is all you need, 2023.

Rebecca F. Alford, Andrew Leaver-Fay, Jeliazko R. Jeliazkov, Matthew J. O'Meara, Frank P. DiMaio, Hahnbeom Park, Maxim V. Shapovalov, P. Douglas Renfrew, Vikram K. Mulligan, Kalli Kappel, Jason W. Labonte, Michael S. Pacella, Richard Bonneau, Philip Bradley, Roland L. Jr. Dunbrack, Rhiju Das, David Baker, Brian Kuhlman, Tanja Kortemme, and Jeffrey J. Gray. The rosetta all-atom energy function for macromolecular modeling and design. *Journal of Chemical Theory and Computation*, 13(6), 2017.

Namrata Anand and Tudor Achim. Protein structure and sequence generation with equivariant denoising diffusion probabilistic models. *arXiv preprint arXiv:2205.15019*, 2022.

Christof Angermueller, David Dohan, David Belanger, Ramya Deshpande, Kevin Murphy, and Lucy Colwell. Model-based reinforcement learning for biological sequence design. In *International conference on learning representations*, 2019.

Christof Angermüller, David Belanger, Andreea Gane, Zelda Mariet, David Dohan, Kevin Murphy, Lucy J. Colwell, and D. Sculley. Population-based black-box optimization for biological sequence design. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*. PMLR, 2020.

Roberto Barceló, Cristóbal Alcázar, and Felipe Tobar. Avoiding mode collapse in diffusion models fine-tuned with reinforcement learning, October 2024.

Emmanuel Bengio, Moksh Jain, Maksym Korablyov, Doina Precup, and Yoshua Bengio. Flow Network based Generative Models for Non-Iterative Diverse Candidate Generation. November 2021.

Andrew Campbell, Jason Yim, Regina Barzilay, Tom Rainforth, and Tommi S. Jaakkola. Generative flows on discrete state-spaces: Enabling multimodal flows with applications to protein co-design. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024.

Onur Celik, Zechu Li, Denis Blessing, Ge Li, Daniel Palenicek, Jan Peters, Georgia Chalvatzaki, and Gerhard Neumann. DIME:Diffusion-Based Maximum Entropy Reinforcement Learning, June 2025.

Huayu Chen, Guande He, Lifan Yuan, Ganqu Cui, Hang Su, and Jun Zhu. Noise Contrastive Alignment of Language Models with Explicit Rewards, October 2024a.

Tianlai Chen, Pranay Vure, Rishab Pulugurta, and Pranam Chatterjee. Amp-diffusion: Integrating latent diffusion with protein language models for antimicrobial peptide generation. *bioRxiv*, 2024b.

Mark L. Chiu, Dennis R. Goulet, Alexey Teplyakov, and Gary L. Gilliland. Antibody structure and function: The basis for engineering therapeutics. *Antibodies*, 8(4), 2019.

Alexander I Cowen-Rivers, Philip John Gorinski, Aivar Sootla, Asif Khan, Liu Furui, Jun Wang, Jan Peters, and Haitham Bou Ammar. Structured q-learning for antibody design. *arXiv preprint arXiv:2209.04698*, 2022.

J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, A. Courbet, R. J. de Haas, N. Bethel, P. J. Y. Leung, T. F. Huddy, S. Pellock, D. Tischer, F. Chan, B. Koepnick, H. Nguyen, A. Kang, B. Sankaran, A. K. Bera, N. P. King, and D. Baker. Robust deep learning–based protein sequence design using ProteinMPNN. *Science*, 378(6615):49–56, October 2022.

Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat gans on image synthesis. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, 2021.

Jesse Farebrother, Jordi Orbay, Quan Vuong, Adrien Ali Taiga, Yevgen Chebotar, Ted Xiao, Alex Irpan, Sergey Levine, Pablo Samuel Castro, Aleksandra Faust, Aviral Kumar, and Rishabh Agarwal. Stop Regressing: Training Value Functions via Classification for Scalable Deep RL. June 2024.

Scott Fujimoto and Shixiang Shane Gu. A minimalist approach to offline reinforcement learning. *Advances in neural information processing systems*, 34, 2021.

Scott Fujimoto, Herke Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *International conference on machine learning*. PMLR, 2018.

R. Garrett and Charles M. Grisham. *Biochemistry*. Brooks/Cole, Cengage Learning, 4th ed edition, 2010. ISBN 978-0-495-10935-8.

Vinicius G. Goecks, Gregory M. Gremillion, Vernon J. Lawhern, J. Valasek, and Nicholas R. Waytowich. Integrating Behavior Cloning and Reinforcement Learning for Improved Performance in Sparse Reward Environments. October 2019.

Nate Gruver, Samuel Stanton, Nathan Frey, Tim GJ Rudner, Isidro Hotzel, Julien Lafrance-Vanasse, Arvind Rajpal, Kyunghyun Cho, and Andrew G Wilson. Protein design with guided discrete diffusion. *Advances in Neural Information Processing Systems*, 36, 2024.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pp. 1861–1870. Pmlr, 2018.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *CoRR*, abs/2207.12598, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33, 2020.

Moksh Jain, Emmanuel Bengio, Alex Hernandez-Garcia, Jarrid Rector-Brooks, Bonaventure FP Dossou, Chanakya Ajit Ekbote, Jie Fu, Tianyu Zhang, Michael Kilgour, Dinghuai Zhang, et al. Biological sequence design with gflownets. In *International Conference on Machine Learning*. PMLR, 2022.

Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Antibody-antigen docking and design via hierarchical equivariant refinement, 2022.

Hélène Kaplon, Silvia Crescioli, Alicia Chenoweth, Jyothsna Visweswaraiah, and Janice M Reichert. Antibodies to watch in 2023. In *MAbs*, volume 15. Taylor & Francis, 2023.

Mohamed Amine Ketata, Cedrik Laue, Ruslan Mammadov, Hannes Stärk, Menghua Wu, Gabriele Corso, Céline Marquet, Regina Barzilay, and Tommi S Jaakkola. Diffdock-pp: Rigid protein-protein docking with diffusion models. *arXiv preprint arXiv:2304.03889*, 2023.

Mohammad Asif Khan, Alexander I. Cowen-Rivers, Derrick-Goh-Xin Deik, Antoine Grosnit, Kamil Dreczkowski, Philippe A. Robert, Victor Greiff, Rasul Tutunov, Dany Bou-Ammar, Jun Wang, and Haitham Bou-Ammar. Antbo: Towards real-world automated antibody design with combinatorial bayesian optimisation. *CoRR*, abs/2201.12570, 2022.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Nayoung Kim, Minsu Kim, Sungsoo Ahn, and Jinkyoo Park. Decoupled sequence and structure generation for realistic antibody design, 2024.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

Aviral Kumar, Joey Hong, Anikait Singh, and Sergey Levine. Should i run offline reinforcement learning or behavioral cloning? In *International Conference on Learning Representations*, 2022.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Lin Li, Esther Gupta, John Spaeth, Leslie Shing, Rafael Jaimes, Emily Engelhart, Randolph Lopez, Rajmonda S. Caceres, Tristan Bepler, and Matthew E. Walsh. Machine learning optimization of candidate antibody yields highly diverse sub-nanomolar affinity antibody libraries. *Nature Communications*, 14(1), 2023.

Minghui Li, Yao Shi, Shengqing Hu, Shengshan Hu, Peijin Guo, Wei Wan, Leo Yu Zhang, Shirui Pan, Jizhou Li, Lichao Sun, and Xiaoli Lan. Mvsf-ab: Accurate antibody-antigen binding affinity prediction via multi-view sequence feature learning. *Bioinformatics*, 10 2024.

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

Ge Liu, Haoyang Zeng, Jonas Mueller, Brandon Carter, Ziheng Wang, Jonas Schilz, Geraldine Horny, Michael E Birnbaum, Stefan Ewert, and David K Gifford. Antibody complementarity determining region design using high-capacity machine learning. *Bioinformatics*, 36(7), 2020.

Ruei-Min Lu, Yu-Chyi Hwang, I-Ju Liu, Chi-Chiu Lee, Han-Zen Tsai, Hsin-Jung Li, and Han-Chung Wu. Development of therapeutic antibodies for the treatment of diseases. *Journal of biomedical science*, 27, 2020.

Shitong Luo, Yufeng Su, Xingang Peng, Sheng Wang, Jian Peng, and Jianzhu Ma. Antigen-specific antibody design and optimization with diffusion-based generative models for protein structures. In *Advances in Neural Information Processing Systems*, 2022.

Karolis Martinkus, Jan Ludwiczak, Wei-Ching Liang, Julien Lafrance-Vanasse, Isidro Hötzel, Arvind Rajpal, Yan Wu, Kyunghyun Cho, Richard Bonneau, Vladimir Gligorijevic, and Andreas Loukas. Abdiffuser: full-atom generation of in-vitro functioning antibodies. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

Frederic P Miller, Agnes F Vandome, and John McBrewster. *Levenshtein Distance*. Alphascript Publishing, January 2013.

S. Miyazawa and R. L. Jernigan. An empirical energy potential with a reference state for protein fold and sequence recognition. *Proteins*, 36(3):357–369, August 1999.

Kevin P Murphy. *Probabilistic machine learning: Advanced topics*. MIT press, 2023.

Yoochan Myung, Douglas E V Pires, and David B Ascher. Csm-ab: graph-based antibody–antigen binding affinity prediction and docking scoring function. *Bioinformatics*, 38(4), 11 2021.

Ashvin Nair, Abhishek Gupta, Murtaza Dalal, and Sergey Levine. AWAC: Accelerating Online Reinforcement Learning with Offline Datasets, April 2021.

Richard A Norman, Francesco Ambrosetti, Alexandre MJJ Bonvin, Lucy J Colwell, Sebastian Kelm, Sandeep Kumar, and Konrad Krawczyk. Computational approaches to therapeutic antibody design: established methods and emerging trends. *Briefings in bioinformatics*, 21(5), 2020.

Seohong Park, Qiyang Li, and Sergey Levine. Flow Q-Learning, February 2025.

Philippe A Robert, Rahmad Akbar, Robert Frank, Milena Pavlović, Michael Widrich, Igor Snapkov, Andrei Slabodkin, Maria Chernigovskaya, Lonneke Scheffer, Eva Smorodina, et al. Unconstrained generation of synthetic antibody–antigen structures to guide machine learning methodology for antibody specificity prediction. *Nature Computational Science*, 2(12), 2022.

Sandra Romero-Molina, Yasser B. Ruiz-Blanco, Joel Mieres-Perez, Mirja Harms, Jan Münch, Michael Ehrmann, and Elsa Sanchez-Garcia. Ppi-affinity: A web tool for the prediction and optimization of protein–peptide and protein–protein binding affinity. *Journal of Proteome Research*, 21(8), 2022.

Amir Shanehsazzadeh, Matt McPartlon, George Kasun, Andrea K. Steiger, John M. Sutton, Edriss Yassine, Cailen McCloskey, Robel Haile, Richard Shuai, Julian Alverio, Goran Rakocevic, Simon Levine, Jovan Cejovic, Jahir M. Gutierrez, Alex Morehead, Oleksii Dubrovskyi, Chelsea Chung, Breanna K. Luton, Nicolas Diaz, Christa Kohnert, Rebecca Consbruck, Hayley Carter, Chase LaCombe, Itti Bist, Phetsamay Vilaychack, Zahra Anderson, Lichen Xiu, Paul Bringas, Kimberly Alarcon, Bailey Knight, Macey Radach, Katherine Bateman, Gaelin Kopec-Belliveau, Dalton Chapman, Joshua Bennett, Abigail B. Ventura, Gustavo M. Canales, Muttappa Gowda, Kerianne A. Jackson, Rodante Caguiat, Amber Brown, Douglas Ganini da Silva, Zheyuan Guo, Shaheed Abdulhaqq, Lillian R. Klug, Miles Gander, Engin Yapici, Joshua Meier, and Sharrol Bachas. Unlocking de novo antibody design with generative artificial intelligence. *bioRxiv*, 2024.

K. T. Simons, C. Kooperberg, E. Huang, and D. Baker. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *Journal of Molecular Biology*, 268(1):209–225, April 1997.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*. PMLR, 2015.

Xiangru Tang, Howard Dai, Elizabeth Knight, Fang Wu, Yunyang Li, Tianxiao Li, and Mark Gerstein. A survey of generative AI for *de novo* drug design: new frontiers in molecule and protein generation. *Briefings Bioinform.*, 25(4), 2024.

Jessica Vamathevan, Dominic Clark, Paul Czodrowski, Ian Dunham, Edgardo Ferran, George Lee, Bin Li, Anant Madabhushi, Parantu Shah, Michaela Spitzer, and Shanrong Zhao. Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery*, 18(6):463–477, June 2019.

Yogesh Verma, Markus Heinonen, and Vikas Garg. Abode: Ab initio antibody design using conjoined odes. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*. PMLR, 2023.

Yannick Vogt, Mehdi Naouar, Maria Kalweit, Christoph Cornelius Miething, Justus Duyster, Roland Mertelsmann, Gabriel Kalweit, and Joschka Boedecker. Stable online and offline reinforcement learning for antibody cdrh3 design, 2023.

Susana Vázquez Torres, Melisa Benard Valle, Stephen P. Mackessy, Stefanie K. Menzies, Nicholas R. Casewell, Shirin Ahmadi, Nick J. Burlet, Edin Muratspahić, Isaac Sappington, Max D. Overath, Esperanza Rivera-de Torre, Jann Ledergerber, Andreas H. Laustsen, Kim Boddum, Asim K. Bera, Alex Kang, Evans Brackenbrough, Iara A. Cardoso, Edouard P. Crittenden, Rebecca J. Edge, Justin Decarreau, Robert J. Ragotte, Arvind S. Pillai, Mohamad Abedi, Hannah L. Han, Stacey R. Gerben, Analisa Murray, Rebecca Skotheim, Lynda Stuart, Lance Stewart, Thomas J. A. Fryer, Timothy P. Jenkins, and David Baker. De novo designed proteins neutralize lethal snake venom toxins. *Nature*, pp. 1–7, January 2025.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion Model Alignment Using Direct Preference Optimization, November 2023.

Yuyang Wang, Jiarui Lu, Navdeep Jaitly, Josh Susskind, and Miguel Angel Bautista. SimpleFold: Folding Proteins is Simpler than You Think, September 2025.

Zhendong Wang, Jonathan J. Hunt, and Mingyuan Zhou. Diffusion policies as an expressive policy class for offline reinforcement learning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

Joseph L. Watson, David Juergens, Nathaniel R. Bennett, Brian L. Trippe, Jason Yim, Helen E. Eisenach, Woody Ahern, Andrew J. Borst, Robert J. Ragotte, Lukas F. Milles, Basile I. M. Wicky, Nikita Hanikel, Samuel J. Pellock, Alexis Courbet, William Sheffler, Jue Wang, Preetham Venkatesh, Isaac Sappington, Susana Vázquez Torres, Anna Lauko, Valentin De Bortoli, Emile Mathieu, Sergey Ovchinnikov, Regina Barzilay, Tommi S. Jaakkola, Frank DiMaio, Minkyung Baek, and David Baker. De novo design of protein structure and function with RFdiffusion. *Nature*, 620(7976), 2023.

John L Xu and Mark M Davis. Diversity in the cdr3 region of vh is sufficient for most antibody specificities. *Immunity*, 13(1), 2000.

Li C. Xue, João Pglm Rodrigues, Panagiotis L. Kastritis, Alexandre Mjj Bonvin, and Anna Vangone. Prodigy: a web server for predicting the binding affinity of protein–protein complexes. *Bioinformatics*, 32(23), 08 2016.

Yong Xiao Yang, Jin Yan Huang, Pan Wang, and Bao Ting Zhu. Area-affinity: A web server for machine learning-based prediction of protein–protein and antibody–protein antigen binding affinities. *Journal of Chemical Information and Modeling*, 63(11), 2023.

Qinqing Zheng, Matt Le, Neta Shaul, Yaron Lipman, Aditya Grover, and Ricky T. Q. Chen. Guided flows for generative modeling and decision making. *CoRR*, abs/2311.13443, 2023.

Xiangxin Zhou, Dongyu Xue, Ruizhe Chen, Zaixiang Zheng, Liang Wang, and Quanquan Gu. Antigen-specific antibody design via direct energy-based preference optimization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

A Appendix

A.1 Evaluation Tasks

A.1.1 Absolut!

The Absolut! software can be used to estimate the free energy of a sequence of length $L = 11$ AAs, representing the CDRH3 region of an antibody, based on a 3D lattice-discretized representation of the CDRH3 and the antigen and the Miyazawa-Jernigan energy potential (Miyazawa & Jernigan, 1999). The lattice-based representation allows for faster approximation, leading to a computation time of 6.2 seconds for a single sample. For the Absolut! design task, we choose the SARS-CoV Spike Receptor-Binding Domain (PDB 2DD8_S) as the antigen, which in previous publications was one of the most challenging targets evaluated (Cowen-Rivers et al., 2022; Vogt et al., 2023).

We curated three training sequence distributions with corresponding energy values reflecting distributions as they could occur in real-world applications. The first distribution comprises a set of 2500 randomly generated sequences. The second set contains 2753 murine CDRH3 sequences, which were categorized as good but not exceptional binders in the Absolut! publication Robert et al. (2022) (specifically the top 0.01% to 0.1% of 6.9 million tested murine CDRH3 sequences. The final distribution, comprising 2167 sequences, was gathered during the exploration phase of an online Q-learning agent, similar to those described by Vogt et al. (2023). We refer to the three datasets as *random*, *natural*, and *expert*. These datasets reflect scenarios

that could occur in an application of our method on distributions from random essays, enrichment data, and active learning AI pipelines.

A.1.2 Rosetta

The Rosetta software (Simons et al., 1997) is an established software for many biomolecular modeling tasks. To create the Rosetta design task, we utilize the Rosetta Energy Function *REF15* (Alford et al., 2017) to estimate the free energy in a nanobody-BCMA complex structure (PDB: 8HXQ). The comparably small antibody and antigen in this complex allow a relatively fast evaluation (10.2 seconds per sample) despite the overall high cost of the Rosetta software. The goal of this task is to design all three CDRs sequences in the nanobody such that the free energy is minimized. To evaluate the energy of a designed sequence, we execute the following steps. First, we relax the original complex using Rosetta's FastRelax protocol to reach a low-energy state. Using Rosetta's MutateResidue protocol, we then mutate the AAs in the complex to the newly designed sequences. To remove any clashes introduced throughout the mutation, we apply FastRelax again to any mutated residue and any residues within a 10 Å neighborhood. Lastly, we utilize Rosetta's InterfaceAnalyzerMover to compute the binding affinity (measured as dG_separated). Typically, this roughly leads to estimated free energy values of -1 to -50 Rosetta Energy Unitss (REUs) for 2500 random sequences. However, in $\sim$2% of evaluated sequences, the estimated binding energy is far outside those we typically observe for this complex (see Figure 3 for distributions) with a sharp drop at the 98th percentile. These values then can reach up to hundreds or thousands of REUs. As these values are not reproducible across multiple runs of the FastRelax protocol across different seeds, we assume that they occur due to FastRelax failing to find a local minimum. Thus, we discard sequences with values below 0 in the training dataset and during evaluation to reduce noise.

In the Rosetta task, we curated two sequence distributions with respective energy values. In the random dataset, we used 2448 random unique CDR sequences and respective affinity measures. For the expert dataset, we recreated the computational sequence design pipeline as utilized by Dauparas et al. (2022) and Vázquez Torres et al. (2025), which was used to generate sequences successfully verified in multiple real-world experiments. In this approach, we first sampled 1000 CDR backbone structures using RFdiffusion (Watson et al., 2023) conditioned on the nanobody and antigen structure. We then applied Protein-MPNN (Dauparas et al., 2022) to sample a total of 2483 unique sequences likely to fold into the respective structures. We evaluate the generated sequences using the same steps outlined above to prevent mismatches between the training and evaluation analyses.

A.2 Tabular Results

A.3 Absolut!

In Table 1 we show evaluation metrics of our method and all baselines on three data distributions in the Absolut! environment.

A.3.1 Rosetta

In Table 2 we show evaluation metrics of our method and all baselines on two data distributions in the Rosetta environment.

A.4 Implementation Details

A.4.1 Hyperparameter search

For all methods we implemented and task configurations we tested the following hyperparameter settings, given that the hyperparameter was applicable:

- $\eta \in [0.1, 0.3, 0.5, 0.7, 0.9, 0.95]$
- training duration $\in [50, 100, 150, 200]$

	Method	Absolut! Expert	Absolut! Natural	Absolut! Random
Free Energy & Top 100	Dataset	-110.53 ± 0.00	-116.46 ± 0.00	-86.21 ± 0.00
	SeqDiff	-104.95 ± 1.62 / -123.51 ± 1.03	-109.48 ± 1.44 / -118.69 ± 0.47	-86.49 ± 0.72 / -99.03 ± 0.87
	SeqDiff+QL	**-120.02 ± 3.16** / **-129.37 ± 0.78**	-110.00 ± 1.13 / **-122.73 ± 0.77**	-95.80 ± 2.13 / -107.48 ± 1.67
	SimDiff	-111.30 ± 1.18 / -126.48 ± 0.58	-110.00 ± 1.71 / -119.68 ± 0.77	-86.76 ± 2.42 / -98.99 ± 2.85
	GFlowNet	-103.87 ± 0.41 / -117.94 ± 0.73	-108.25 ± 0.42 / -114.25 ± 0.51	**-104.62 ± 0.28** / **-111.05 ± 0.56**
	BC	-110.02 ± 0.99 / -116.20 ± 3.59	**-113.73 ± 0.49** / -114.59 ± 1.04	-86.23 ± 0.71 / -89.36 ± 2.13
	BC+QL	**-119.58 ± 0.45** / -127.82 ± 0.15	**-113.80 ± 0.96** / -118.33 ± 1.96	-93.46 ± 0.25 / -104.64 ± 0.70
Diversity & Novelty	Dataset	7.72 ± 0.00	7.38 ± 0.00	10.27 ± 0.00
	SeqDiff	8.08 / 3.57	8.00 / 2.63	10.25 / 6.48
	SeqDiff+QL	5.18 / 2.23	6.97 / 2.58	6.89 / 6.17
	SimDiff	7.28 / 3.13	7.51 / 2.47	10.15 / 6.57
	GFlowNet	9.32 / 6.32	6.05 / 5.72	8.28 / 6.64
	BC	7.78 / 2.54	7.59 / 1.77	10.26 / 5.91
	BC+QL	6.46 / 1.85	6.97 / 1.60	9.88 / 6.01
Novel & Duplicated	SeqDiff	500 ± 0 / 959 ± 65	500 ± 0 / 870 ± 87	500 ± 0 / 318 ± 25
	SeqDiff+QL	500 ± 0 / 2049 ± 1579	500 ± 0 / 1585 ± 706	500 ± 0 / 214 ± 153
	SimDiff	500 ± 0 / 9 ± 4	500 ± 0 / 48 ± 8	500 ± 0 / 0 ± 0
	GFlowNet	500 ± 0 / 0 ± 0	500 ± 0 / 0 ± 0	500 ± 0 / 0 ± 0
	BC	141 ± 24 / 5594 ± 538	104 ± 6 / 5016 ± 6	130 ± 19 / 5399 ± 542
	BC+QL	352 ± 22 / 5792 ± 22	158 ± 26 / 5782 ± 433	366 ± 27 / 5778 ± 27

Table 1: Summary of results across different methods and metrics on the Absolut! task. Best results are underlined, results that are not significantly worse than the best are bold.

	Method	Rosetta Expert	Rosetta Random
Free Energy & Top 100	Dataset	-32.75 ± 0.00	-27.67 ± 0.00
	SeqDiff	-33.12 ± 0.55 / -38.49 ± 0.30	-26.64 ± 0.39 / -34.54 ± 0.43
	SeqDiff+QL	**-37.10 ± 0.79** / **-41.47 ± 0.53**	-27.28 ± 0.76 / -35.23 ± 0.70
	SimDiff	-34.75 ± 0.69 / -40.47 ± 0.68	-26.98 ± 1.31 / -34.85 ± 1.57
	GFlowNet	-29.64 ± 0.31 / -37.14 ± 0.32	**-30.88 ± 1.36** / **-38.81 ± 1.51**
	BC	-32.46 ± 0.46 / -37.84 ± 0.42	-26.72 ± 0.44 / -34.79 ± 0.58
	BC+QL	-35.03 ± 0.16 / -39.76 ± 0.45	-28.06 ± 0.48 / -36.46 ± 0.43
Diversity & Novelty	Dataset	18.05 ± 0.00	25.49 ± 0.00
	SeqDiff	18.22 / 9.38	25.46 / 20.10
	SeqDiff+QL	15.16 / 8.94	22.66 / 20.10
	SimDiff	18.19 / 11.21	24.81 / 20.20
	GFlowNet	22.37 / 18.37	20.09 / 20.46
	BC	17.98 / 7.11	25.54 / 18.91
	BC+QL	17.42 / 7.55	24.84 / 19.17
Novel & Duplicated	SeqDiff	497 ± 0 / 0 ± 0	487 ± 5 / 0 ± 0
	SeqDiff+QL	499 ± 1 / 0 ± 0	487 ± 7 / 0 ± 0
	SimDiff	497 ± 3 / 0 ± 0	490 ± 4 / 0 ± 0
	GFlowNet	498 ± 2 / 0 ± 0	498 ± 2 / 0 ± 0
	BC	498 ± 1 / 1032 ± 75	493 ± 3 / 3098 ± 905
	BC+QL	498 ± 1 / 3866 ± 205	485 ± 17 / 3787 ± 1303

Table 2: Summary of results across different methods and metrics on the Rosetta task. Best results are underlined, results that are not significantly worse than the best are bold.

We found that for the narrow natural dataset on Absolut! a higher focus on BC was necessary to prevent distribution shift.

We then selected the following hyperparameters: Training episodes (with 1000 gradient updates each):

- 200 for all Absolut! tasks
- 50 for Rosetta expert task, except for BC+QL, where 200 training epochs let to better results
- 200 for Rosetta random task

η was used for SeqDiff+QL and BC+QL:

- Absolut! expert 0.95 for both

- Absolut! natural 0.5 for both
- Absolut! random 0.95 for SeqDiff+QL and 0.9 BC+QL
- Rosetta expert 0.9 for SeqDiff+QL and 0.95 BC+QL
- Rosetta random 0.9 for SeqDiff+QL and 0.95 BC+QL

For the GFlowNet implementation, we utilized the original code released by Jain et al. (2022) and did a hyperparameter search over training duration and reward exponent (β in the respective publication). For the reward exponent, we tested $\beta \in [3, 5, 10, 25, 50]$ and found:

- $\beta = 3$ to work best on Absolut! expert
- $\beta = 3$ to work best on Absolut! natural
- $\beta = 5$ to work best on Absolut! random
- $\beta = 10$ to work best on both Rosetta tasks

and a training duration of 50 epochs to be best for all tasks except Absolut! random, where 150 epochs led to slightly better results. Note that a higher β for GFlowNet can slightly reduce the mean free energy, but leads to significantly increased duplicates and diversity loss, as the policies start overfitting to model-bias (e.g., only generating sequences with high counts of AA F.

A.4.2 Additional Details

To reduce the effect of the latent space's structure on the reported results, we share the pre-trained VAE between all datasets for a given seed. Due to the large computational burden, we chose $N = 10$ diffusion steps for our experiments. We follow (Wang et al., 2023) for the choice of β noise schedule to train our diffusion model.

In contrast to the implementation by Wang et al. (2023), we do not generate 50 actions using the Diffusion Model per step and sample the final action via a softmax distribution over the respective Q-weights. Instead, we directly take the action sampled from the diffusion model.

When conditioning a policy π on a state s_t, we transform the state into a token-representation using an embedding layer instead of the learned VAE representation we use for AAs.

For all tested methods, normalize all rewards in the training distribution between 0 and 1 to prevent reward shifts between datasets and evaluation schemes.

A.5 Baseline Algorithms

A.5.1 BC and BC+QL

Inspired by prior work (Fujimoto & Gu, 2021; Nair et al., 2021; Goecks et al., 2019), we create a stochastic actor-critic agent balancing behavior cloning and QL in a categorical action setting. This combination, which we refer to as BC+QL, was previously successfully applied in continuous action settings and showed remarkable performance in the offline RL setting. The actor is implemented using a simple feed-forward network π_θ predicting a probability vector $\boldsymbol{p}$ containing all 20 possible next AAs given the current sequence as $\boldsymbol{p}_t = \pi_\theta(s_t)$. Thus, we iteratively sample from the stochastic policy to generate entire sequences. The behavior cloning part is trained using cross-entropy loss to model the dataset distribution $L_{BC} = \mathbb{E}_{(s_t,a_t)\sim D, \boldsymbol{p}=\pi_\theta(s_t)}[-\sum_{k=1}^{20} y_k \log \boldsymbol{p}_k, y_k = 1[k = a_t]]$. This loss alone is used to create the BC agent.

For policy optimization, we add a Q-function $Q_\phi(s_t)$ estimation the action values for all 20 AAs given the incomplete sequence s_t and optimize Q_{ϕ_i} to minimize $L_{Q_i} = \mathbb{E}_{(s_t,a_t,s_{t+1})\sim D}[||R(s_t, a_t) + \min_{j=1,2}(Q_{\phi'_j}(s_{t+1})\pi'_\theta(s_{t+1})) - Q_{\phi_i}(s_t, a_t)||^2]$, and policy π_θ to minimize $(1-\eta)L_{BC}(\theta) - \eta\mathbb{E}_{s\sim D}[Q_\phi(s)\pi_\theta(s_t)]$,

where η can be tuned to balance the loss terms. This algorithm, referred to as BC+QL, despite its simplicity, fulfills the requirements of being non-deterministic, suited for offline learning, and capable of objective-driven design, thus serving as a minimalist and lightweight baseline.

A.5.2 SimDiff

Alternatively to our chosen incremental approach to sequence design, the entire sequence could be designed at once using a diffusion model. We utilize a DDPM (Ho et al., 2020) for this purpose and refer to the agent as SimDiff. Specifically, we use latent diffusion by representing all sequences of length L in the training dataset in a concatenation of their amino acid's 2D latent representations using the pretrained VAE, $\boldsymbol{z}_{cat}$. $\boldsymbol{z}_{cat} = \text{concat}(\boldsymbol{z}_0, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_{L-1})\,, \qquad \boldsymbol{z}_l = e_\omega(a)$, where e_ω is trained as described in Section 4.1. After sampling, we decode the discrete AAs using the VAE. While this baseline is not capable of objective-driven design, it can be used to estimate a potential performance gap between unguided incremental and unguided simultaneous diffusion.

A.5.3 SeqDiff

In SeqDiff+QL we employ a BC loss to constrain the agent to stay close to the training data distribution. If we only keep this loss, by setting $\eta = 0$, we create an algorithm, SeqDiff, which models the training data distribution by sequentially sampling AAs. This algorithm is suited for diverse generation and offline training, but not for objective-driven design. Using this baseline, we can analyze a potential gap between sequential and simultaneous generation using diffusion models.

A.5.4 GFlowNet

Introduced by Jain et al. (2022) for biological sequence design, this algorithm uses Generative Flow Networks (Bengio et al., 2021) in combination with a learned proxy reward model. Specifically, the proxy reward model is implemented as an ensemble with mean μ and an uncertainty as the standard deviation σ of the ensemble. The model is trained to maximize the upper confidence bound as $\mu + 0.1\sigma$.

A.6 Amino Acid Groups

Our grouping of AAs is taken from the classification by Garrett & Grisham (2010) and given as follows:

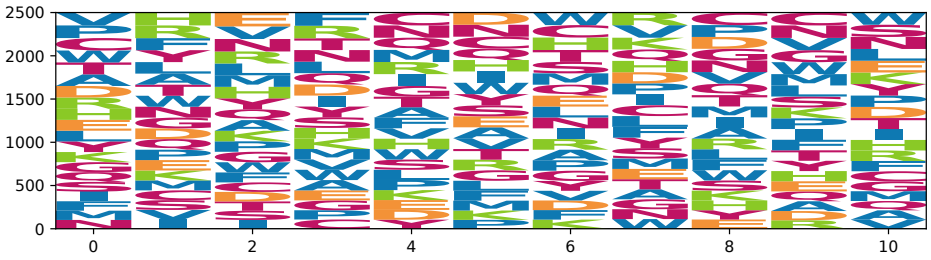
- Nonpolar (hydrophobic) AAs: L, P, A, V, M, W, F, I
- Polar, uncharged AAs: G, S, N, Q, T, C, Y
- Acidic AAs: D, E
- Basic AAs: K, R, H

Note that we chose this specific grouping not because we are convinced it bears an advantage, but rather because it was the most prominent grouping we found in the literature.

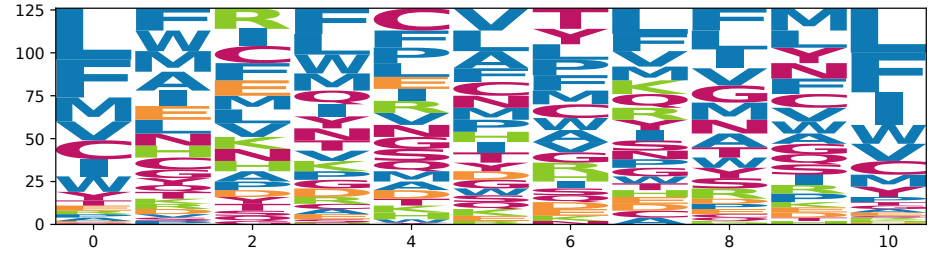
A.7 Bias in evaluation Environments

To examine why our induced priors have a positive effect on all Absolut! environments, but not the Rosetta environments, and why diversity of sequences in the Absolut! environment decreases with increased affinity, we analyzed the shift in AA distribution between all sequences in the random datasets and the top 5% sequences in the random dataset. In the Absolut! dataset, the fraction of hydrophobic AAs (colored blue in Figure 6) increased from 39.77% to 55.63%, indicating a bias towards such AAs. This is further emphasized by the natural dataset, which represents the top 0.01% to 0.1% of 6.9 million tested murine CDRH3 sequences in the Absolut! publication Robert et al. (2022). Here, hydrophobic AAs make up 60.31% of the dataset. In the Rosetta dataset, while we observe that certain positions favor specific AA groups, we observe no general bias towards a single AA group. However, we can observe that in the Rosetta task, position 9 seems to favor

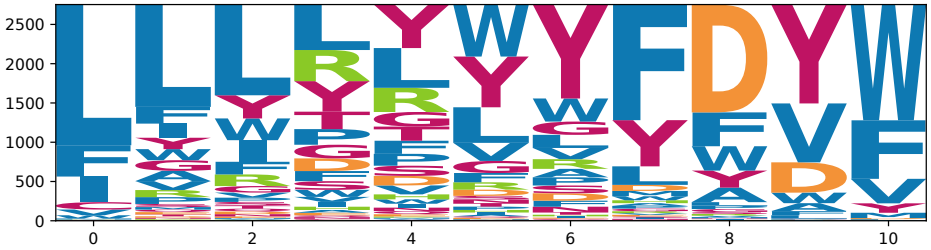
polar, uncharged AAs and position 14 seems to favor hydrophobic AAs, indicating a relevance for binding the antigen. This shows that the assumed prior was thus only present in one of our environments.



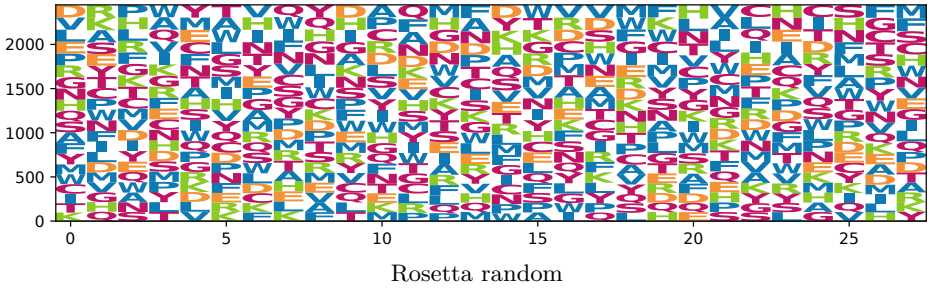
Absolut! random



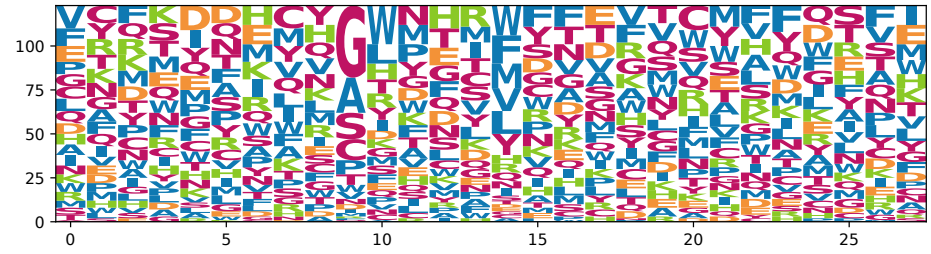
Absolut! random top 5%.



Absolut! natural training dataset.



Rosetta random



Rosetta random top 5%

Figure 6: Logo plots for different datasets. Higher affinity datasets in the Absolut! task exhibit a significantly higher fraction of hydrophobic AAs. In the Rosetta task, such tendencies are less observable, and focus on a few positions (e.g., index 9 and 14).