Improving Antibody Design with Force-Guided Sampling in Diffusion Models

Paulina Kulytė
University of Cambridge

Francisco Vargas
University of Cambridge

Simon Valentin Mathis
University of Cambridge

Yu Guang Wang
Shanghai Jiao Tong University
University of New South Wales

José Miguel Hernández-Lobato
University of Cambridge

Pietro Liò
University of Cambridge

Abstract

Antibodies, crucial for immune defense, primarily rely on complementarity-determining regions (CDRs) to bind and neutralize antigens, such as viruses. The design of these CDRs determines the antibody's affinity and specificity towards its target. Generative models, particularly denoising diffusion probabilistic models (DDPMs), have shown potential to advance the structure-based design of CDR regions. However, only a limited dataset of bound antibody-antigen structures is available, and generalization to out-of-distribution interfaces remains a challenge. Physics based force-fields, which approximate atomic interactions, offer a coarse but universal source of information to better mold designs to target interfaces. Integrating this foundational information into diffusion models is, therefore, highly desirable. Here, we propose a novel approach to enhance the sampling process of diffusion models by integrating force field energy-based feedback. Our model, DIFFFORCE, employs forces to guide the diffusion sampling process, effectively blending the two distributions. Through extensive experiments, we demonstrate that our method guides the model to sample CDRs with lower energy, enhancing both the structure and sequence of the generated antibodies.

1 Introduction

Antibodies are key therapeutic proteins due to their ability to selectively bind to a variety of disease-causing antigens, including viruses. Antibodies consist of two heavy and two light chains, forming a Y-shaped structure. Critical to their ability to recognize diverse antigens are the six complementarity determining regions (CDRs) located at the tips of this structure. The diversity of antibodies is derived from the extensive combinatorial possibilities of these CDRs. A CDR of length L can theoretically have up to 20^L different amino acid sequences, owing to the 20 types of amino acids that can be placed at each position. Therefore, a key step in developing therapeutic antibodies is designing effective CDRs that specifically bind to target antigens [Kunik et al., 2012, Sela-Culang et al., 2013].

Traditional approaches to antibody design predominantly rely on animal immunization and computational methods. Animal immunization is inherently limited to the production of naturally occurring antibodies and raises ethical concerns [Gray et al., 2020], despite its effectiveness in generating high-affinity antibodies. Traditional *in silico* methods, on the other hand, utilize complex biophysical energy functions [Warszawski et al., 2020, Adolf-Bryfogle et al., 2018] to predict how potential antibodies might interact with their targets. However, they depend on expensive simulations, are prone to convergence to local optima, and possess inherent limitations due to the complex nature of interactions which cannot be efficiently represented by basic statistical functions [Graves et al., 2020]. This situation underscores the need for alternative approaches in antibody design.

AI for New Drug Modalities at NeurIPS 2024.

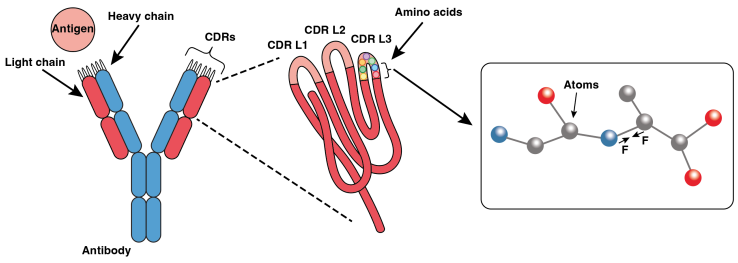

Figure 1: The antigen-binding region comprises six complementarity-determining regions (CDRs). Each CDR is constructed from a variety of amino acids, which are themselves made up of atoms. These atoms are governed by forces, denoted by the symbol F.

Recently, denoising diffusion probabilistic models (DDPMs) have emerged as a powerful technique for learning and sampling from complex, high-dimensional protein distributions [Watson et al., 2023, Yim et al., 2023, Trippe et al., 2023]. In particular, this advancement has shown potential in the structure-based design of CDRs. Recent work [Luo et al., 2022, Martinkus et al., 2023] has demonstrated the capabilities of diffusion models for modeling the CDRs of antibodies at the atomic level, conditioned on the antigen and an antibody framework. However, the available dataset of bound antibody-antigen structures is limited, and generalization to out-of-distribution interfaces remains a challenge. While diffusion models provide accurate approximations within the known distribution, they struggle with out-of-distribution scenarios. This limitation poses a challenge for advancing CDR design as many antibodies generated *in silico* with diffusion models fail to demonstrate functionality *in vitro* [Shanehsazzadeh, 2024, Zeni et al., 2023, Sidhu and Fellouse, 2006].

To address this challenge, we propose DIFFFORCE, a force-guided DDPM sampling method inspired by traditional physics-based simulation techniques such as molecular dynamics (MD). Physics-based force fields, which approximate atomic interactions (as shown in Figure 1), provide a coarse but universal source of information to better align antibody designs with target interfaces. Integrating this foundational data into diffusion models overcomes the limitations of distribution learning, as physics-based models generalize well despite being poor approximators. By combining these approaches, we enhance the ability to model out-of-distribution interfaces as we are guided by force field energy, while the structural *antibody-like* details are left to be determined by the diffusion model. While previous studies have used force field-based functions to refine antibody structures after diffusion generation [Luo et al., 2022], or have trained separate networks to approximate the forces for guiding an unconditional diffusion model [Wang et al., 2024], we are the first to construct a principled method of force-guided DDPM sampling, effectively blending the two distributions. Given a protein complex consisting of an antigen and an antibody framework as input, we first initialize the CDR with arbitrary positions, sequence and orientations. Then, during the sampling stage, we iteratively update the atom positions guided by the gradients of force field energy, which are calculated for the denoised sample approximation. We highlight our main contributions as follows:

- We introduce the first force-guided diffusion model, which utilizes a differentiable force field to guide the sampling process, effectively leveraging the weighted geometric mean of the two distributions. Unlike existing methods, our model does not require to train a separate network for energy approximation or condition the diffusion model on energy.

- We propose a method to approximate the denoised sample of antibody atom coordinates, offering an elegant interpolative interpretation. This enables accurate energy computation, ensuring the precise application of forces during diffusion sampling. We also present an approach for approximating the denoised samples of amino acid types and orientations.

We evaluate our model on the CDR sequence-structure co-design task. We show that our proposed method effectively guides the model to sample CDRs with lower energy, outperforming several state-of-the-art models. We observe that our model generates more favorable structures earlier in the sampling process, leading to an enhanced quality of produced antibody sequences.

2 Related Work

Diffusion Models for Antibody Design Antibody design involves creating the sequence and structure of antibodies that can bind to target antigens. This process differs from general protein design, where sequences are derived from known structures [Dauparas et al., 2022, Ingraham et al., 2019], or structures are predicted based on amino acid sequences [Jumper et al., 2021]. In antibody design, the sequences and structures of the CDRs are usually initially unknown. While various generative models have been proposed to learn such data distribution, diffusion models [Sohl-Dickstein et al., 2015, Dhariwal and Nichol, 2021] have recently gained prominence for their effectiveness in ensuring stable training and achieving good distribution coverage. Diffusion models achieve state-of-the-art performance in antibody design by learning to generate new data through denoising samples from a prior distribution. The DiGress model [Vignac et al., 2023] demonstrated how to utilize a discrete diffusion process for molecules, while the work of DiffAb [Luo et al., 2022] proposed the first diffusion model to perform joint design of sequence and structure of the antibody CDR regions while conditioning on the antigen-antibody complex. AbDiffuser [Martinkus et al., 2023] improved this further by incorporating strong priors and being more memory efficient with side chain generation. However, these models still face challenges in accurately modeling the complex interactions within antigen-antibody interfaces, particularly when dealing with out-of-distribution data.

Guided Generation Guiding generative models to produce specific outcomes is highly desirable for a variety of applications [Ho et al., 2022, Nichol et al., 2023]. To achieve this, two main methods have been proposed, a classifier guidance [Dhariwal and Nichol, 2021, Song et al., 2021b] and a classifier-free guidance [Ho and Salimans, 2021]. Recently, a concurrent work [Wang et al., 2024] introduced a force-guided diffusion model to produce protein conformations aligned with Boltzmann's equilibrium distribution, based on the classifier guidance approach. However, this method requires training an additional network to approximate the intermediate force vector to guide an unconditional model, which can result in inaccurate estimates. In contrast, our method employs a differentiable force field for guided sampling, eliminating the need for a separate energy approximation network and ensuring more accurate energy calculations. Additionally, a loss guidance approach has been proposed [Song et al., 2023], leveraging differentiable loss functions to guide the model without additional training on noisy paired data. Similarly, our approach uses a differentiable force field to guide the sampling.

3 Method

We propose force-guided DIFFFORCE, a diffusion model targeting CDR region generation for antibodies. Building upon the DIFFAB diffusion model introduced in Section 3.1, we present a novel strategy in Section 3.2 that integrates force guidance into the diffusion model's sampling. By employing force to guide the sampling process, DIFFFORCE achieves CDRs with lower energy, leading to an improved structure and ultimately the sequence of the generated antibodies. A visualization of the method is shown in Figure 2.

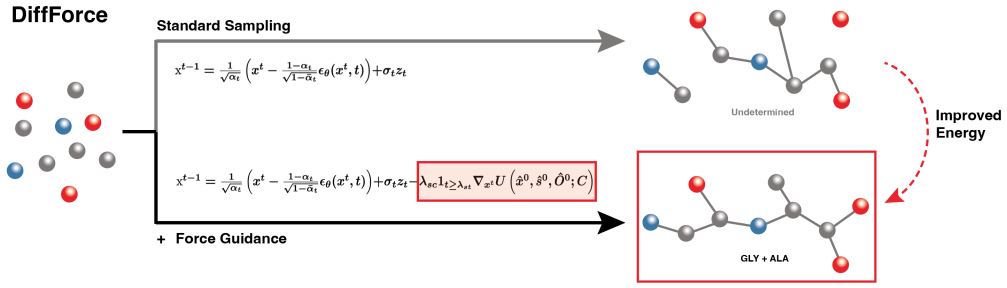

Figure 2: Antibody CDR generation with different sampling strategies. **Upper**: Standard DDPM sampling without force guidance. **Lower**: Incorporating force guidance into sampling, the model generates CDR structures with lower energy. Notation explained in the main text.

3.1 Diffusion Model

Our model builds upon the DIFFAB diffusion model [Luo et al., 2022]. DIFFAB represents each amino acid in an antibody by its type $s_i \in \{A \dots Y\}$, the coordinates of its C_α atom $x_i \in \mathbb{R}^3$, and its orientation $O_i \in SO(3)$. Assuming that the structures of the antigen, the antibody framework, and five other CDRs are known, it designs one CDR loop at a time, denoted as $R = \{(s_j, x_j, O_j) \mid j = l+1, \dots, l+m\}$, given the rest of the antibody-antigen complex $C = \{(s_i, x_i, O_i) \mid i \neq j\}$, which includes a set of five fixed CDRs.

The forward diffusion process from $t = 0$ to T, is Markovian and incrementally adds noise to three different modalities using non-learnable distributions q: The C_α atom positions follow a Gaussian distribution, $q(x_j^t \mid x_j^0)$; amino acid types follow a multinomial distribution, $q(s_j^t \mid s_j^0)$; and the orientations of amino acids follow an isotropic Gaussian distribution, $q(O_j^t \mid O_j^0)$. The backward diffusion process (from $t = T$ to 0), refines each modality back towards the original data distribution. The reverse process is guided by learnable models p_θ, which approximate the posterior distributions at each step using three distinct neural networks (further denoted as F, G, H, respectively) for the three modalities. For more details on the DIFFAB model, see Section 3 of the original paper [Luo et al., 2022], and for additional information on DDPMs, refer to Appendix A.

3.2 Force Guided Antibody Design

3.2.1 Force Field

Molecular dynamics (MD) simulations provide insights into the dynamic behavior of molecular systems by numerically integrating Newton's equations of motion [Chandler et al., 1987] for N particles:

$$m_i \frac{d^2 x_i}{dt^2} = F_i = -\frac{\partial}{\partial x_i} U(x_1, x_2, \dots, x_N), \tag{1}$$

where m_i, x_i, and F_i represent the mass, position, and force on each particle, respectively. The energy $U(x_1, x_2, \dots, x_N)$ is a function of the coordinates of all N particles. By solving Newton's equation, MD simulations approximate the evolution of molecular systems over time.

An MD force field is a parametrised function used to evaluate the energy $U(x_1, x_2, \dots, x_N)$ of a given configuration. For proteins, these forcefields are typically empirical, due to the large system sizes, and their functional forms and parameters are tuned to closely match experimental observations. Common terms include both bonded interactions, such as bond stretching, angle bending, and torsional angles, and non-bonded interactions, like van der Waals forces and electrostatic interactions.

In the context of antibody design, the force field takes a protein P (e.g., set of atom coordinates x) and computes the energy U. By calculating the gradient ∇U, we can determine how U varies with changes in atomic positions. This gradient indicates how to adjust each atom's position to minimize the total energy of the protein structure. Lower energy configurations often correspond to more thermodynamically stable antigen-antibody complexes, which are associated with higher affinity [Ji et al., 2023]. Using the relationship between energy and force, $-\nabla U(x) = F$, we can simulate the equations of motion to evolve this dynamical system according to the energy U.

3.2.2 DiffForce-C_α : Interpolating Between p_{data} and $e^{-\kappa U(x_0;C)}$

For simplicity, we consider the setting where the residues are fixed, and our focus is to guide the C_α atom coordinates with a prescribed force field. Rather than sample unconditionally from the data distribution, we are interested in sampling from the following tilted distribution:

$$\pi_0(x_0) = \frac{p_{\text{data}}(x_0) e^{-\kappa U(x_0;C)}}{\int p_{\text{data}}(x_0) e^{-\kappa U(x_0;C)} \mathrm{d}x_0}, \tag{2}$$

where we use the notation $U(x_0; C)$ to denote that C is fixed throughout simulation. This induces a new distribution we wish to sample from that interpolates between the Boltzmann distribution $e^{-\kappa U(x_0)}$ [1] and $p_{\text{data}}(x_0)$. One way to interpret this is to think of $p_{\text{data}}(x_0)$ as a prior and $e^{-\kappa U(x_0)}$ acting as a likelihood of the form $p(y|x_0)$. Thus $\pi_0(x_0)$ is akin to a posterior of the form $p(x_0|y)$

[1] For brevity we have dropped the conditioning on C.

that is in a way conditioned to make the binding energy small. However, unlike [Song et al., 2023, Komorowska et al., 2024], we do not have an explicit notion of the variable y in this setting. We highlight that [Wang et al., 2024] concurrently explore an akin setting; however, their approach is focused on learning a new modified score while our is focused on approximations during inference.

An alternate and akin approach is to construct π_0 as a log-concave interpolation, as in annealed sampling [Neal, 2001], that is to form the weighted geometric mean $\pi_0 \propto p_{\text{data}}^{1-\beta} \exp(-\kappa U(x_0))^\beta$ for $\beta \in [0, 1]$. This has the interpretation that we are now trying to sample from a distribution that is an interpolation between $p_{\text{data}}(x_0)$ and $e^{-\kappa U(x_0)}$. By leveraging the weighted geometric mean of the distributions, we ensure that if one distribution suggests a particular outcome is extremely unlikely, it influences the other, thus pulling the combined distribution towards more realistic outcomes. This method aligns well with our goal of generating high-quality samples with good binding energies, providing a balanced compromise between the two. In practice, however, we follow Equation 2 as it provides a form that is easier to tune and more in line with prior works on conditioning diffusion models. Due to this connection, we will refer to π_0 as the interpolating distribution. To sample from Equation 2 we estimate the interpolating score $\nabla_{x_t} \ln \pi_t(x_t)$ [Chung et al., 2023]:

$$\nabla_{x_t} \ln \pi_t(x_t) = \nabla \ln \int \pi_0(x_0) p(x_t|x_0) dx_0, \tag{3}$$

$$= \nabla_{x_t} \ln \int e^{-\kappa U(x_0;C)} p_{\text{data}}(x_0) p(x_t|x_0) dx_0, \tag{4}$$

$$= \nabla_{x_t} \ln \int e^{-\kappa U(x_0;C)} p(x_0|x_t) dx_0 + \nabla_{x_t} \ln p(x_t), \tag{5}$$

where $p(x_0|x_t)$ is the transition density of the backwards SDE (the denoising process), which we do not have access to. Following [Komorowska et al., 2024, Chung et al., 2023], we approximate it with a point mass centered at its mean:

$$\int e^{-\kappa U(x_0;C)} p(x_0|x_t) dx_0 \approx \int e^{-\kappa U(x_0;C)} \delta_{\mathbb{E}[x_0|x_t]}(x_0) dx_0 \tag{6}$$

$$= e^{-\kappa U(\mathbb{E}[x_0|x_t];C)}. \tag{7}$$

Then, the approximate interpolating score is given by $\nabla_{x_t} \ln \pi_t(x_t) \approx -\kappa \nabla_{x_t} U(\mathbb{E}[x_0|x_t]) + \nabla_{x_t} \ln p(x_t)$, and we can use Tweedie's formula [Robbins, 1992] to compute $\mathbb{E}[x_0|x_t]$ given we have a good approximation of the score:

$$\mathbb{E}\left[x_0 \mid x_t\right] = \frac{x_t + (1 - \bar{\alpha}_t) \nabla_{x_t} \ln p_t\left(x_t\right)}{\sqrt{\bar{\alpha}_t}} \approx \hat{x}_0(x_t) = \frac{1}{\sqrt{\bar{\alpha}_t}} \left(x_t - \sqrt{1 - \bar{\alpha}_t} \epsilon_\theta(x_t, t)\right). \tag{8}$$

Here, $\bar{\alpha}_t = \prod_{\tau=1}^{t} \alpha_\tau = \prod_{\tau=1}^{t} (1 - \beta_\tau)$, where β_t is the cosine variance schedule for the diffusion model, and ϵ_θ is the standard Gaussian noise added to the x_t predicted by the neural network F. This yields the following sampler, with z_t denoting standard Gaussian:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(x_t, t)\right) + \sigma_t z_t - \kappa \nabla_{x_t} U(\hat{x}_0(x_t)). \tag{9}$$

We now have ingredients to generate an approximate sample from the interpolating distribution π_0.

3.2.3 Force Guidance for Residue Types

We have derived an approach to approximate the C_α atom coordinates at $t = 0$, further denoted as $\hat{x}^0$. However, we also need to devise approximations for the amino acid types and orientations to obtain an estimate for $\mathbb{E}[R^0|R^t]$, which is required to calculate the energy U. Unlike the C_α coordinates, the approximations for amino acid types and orientations do not follow Tweedie's formula. To account for it, we derive an alternative approach to estimate $\hat{s}^0$ and $\hat{O}^0$ using the settings provided.

Amino Acid Types The generative diffusion process for amino acid types, denoted by $p(s_j^{t-1}|R^t, C)$ and defined in [Luo et al., 2022, Equation 3], is designed to approximate the posterior $q(s_j^{t-1}|s_j^t, s_j^0)$. This alignment is quantified using the Kullback–Leibler (KL) divergence, as suggested in [Hoogeboom et al., 2021, Equation 15]:

$$\text{KL}(q(s^{t-1}|s^t, s^0) \| p(s^{t-1}|s^t)) = \text{KL}\left(C(\theta_{\text{post}}(s^t, s^0)) \| C(\theta_{\text{post}}(s^t, \hat{s}^0))\right), \tag{10}$$

where the KL divergence is minimized when the parameterized posterior $\theta_{\text{post}}(s^t, s^0)$ is equivalent to $\theta_{\text{post}}(s^t, \hat{s}^0)$ thus making $\hat{s}^0$ a good predictor for s^0 given we observe s^t. Following this, we can derive the distribution for the posterior sample at timestep $t-1$ as:

$$q(s_j^{t-1}|s_j^t, s_j^0) = \text{Multinomial}\left(\left[\alpha_{\text{type}}^t \cdot \text{onehot}(s_j^t) + (1-\alpha_{\text{type}}^t)\cdot\frac{1}{20}\right] \odot \left[\bar{\alpha}_{\text{type}}^{t-1}\cdot\text{onehot}(s_j^0) + (1-\bar{\alpha}_{\text{type}}^{t-1})\cdot\frac{1}{20}\right]\right). \tag{11}$$

Here $\bar{\alpha}_{\text{type}}^t = \prod_{\tau=1}^t(1-\beta_{\text{type}}^\tau)$ and β_{type}^t is the probability of uniformly resampling another amino acid from among the 20 types. The neural network G is tasked with predicting s_j^0, leveraging the learned distributional characteristics of amino acid types. In order to approximate the denoised sample for amino acid types at $t=0$, namely $\hat{s}^0$, the idea is to utilize only the second term of Equation 11:

$$\hat{s}_j^0 = \bar{\alpha}_{\text{type}}^{t-1}\cdot\text{onehot}(s_j^0) + (1-\bar{\alpha}_{\text{type}}^{t-1})\cdot\frac{1}{20}, \tag{12}$$

where $\hat{s}_j^0$ predicts the amino acid type at $t=0$ for each amino acid j.

Amino Acid Orientations The denoising process for amino acid orientations is captured via SO(3) elements, as described by [Leach et al., 2022] and implemented by [Luo et al., 2022, Equation 11]:

$$p(O_j^{t-1}|R^t, C) = \mathcal{IG}_{\text{SO(3)}}\left(O_j^{t-1}|H(R^t, C)[j], \beta_{\text{ori}}^t\right), \tag{13}$$

where H is a neural network that denoises the orientation matrix for amino acid j, $\mathcal{IG}_{\text{SO(3)}}$ denotes the isotropic Gaussian distribution on SO(3) parameterized by a mean rotation and a scalar variance, β_τ^t is the variance increase with the step t. To obtain the approximation $\hat{O}_j^0$ for amino acid orientation, we propose an approach of iteratively denoising the sample O_j^t for t iterations, where each iteration predicts the sample O_j^{t-1}. Namely, by iteratively applying Equation 13 until timestep t reaches 0 for each amino acid j, we converge to the approximation $\hat{O}_j^0$, effectively reversing the forward diffusion:

$$\hat{O}_j^0(R^t) \approx \tilde{O}_j^0 \sim p(O_j^{t-1}|R^t, C)\prod_{s=1}^{t-1} p(O_j^{s-1}|R^s, C). \tag{14}$$

We have now obtained denoised approximations of atom coordinates, amino acid types, and orientations. This estimate can be utilized in further algorithms to compute the energy U.

3.2.4 Implementation

We derive a novel approach for guiding the sampling of C_α atom coordinates with a prescribed force:

Algorithm 1 DIFFFORCE-C_α Sampling with Force Guidance

$x^T \sim \mathcal{N}(0, I)$
for $t = T, \ldots, 1$ **do**
  $z \sim \mathcal{N}(0, I)$ if $t > 1$, else $z = 0$
  estimate $\hat{x}^0(R^t)$ using Eq 8, $\hat{s}^0(R^t)$ using Eq 12 and $\hat{O}^0(R^t)$ using Eq 14
  $x^{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x^t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x^t, t)\right) + \sigma_t z_t - \lambda_{sc}\mathbb{1}_{t\geq\lambda_{st}}\nabla_{x^t}U\left(\hat{x}^0, \hat{s}^0, \hat{O}^0; C\right)$
  $R^{t-1} = \left(x^{t-1}, s^{t-1}, O^{t-1}\right)$, sample s^{t-1}, O^{t-1} following [Luo et al., 2022]
end for
return x^0, s^0, O^0

We introduce two hyperparameters, force scale (λ_{sc}) and force start (λ_{st}). The λ_{sc} parameter dictates the magnitude of the force, gradually adjusted from 0.0 to λ_{sc} using a linear scheduling strategy. This parameter is applied to normalized forces as detailed in Appendix B. The λ_{st} parameter defines when force application begins, with a value of 0.3 indicating initiation of force at 70% of the sampling.

4 Experiments

We evaluate the effectiveness of the DIFFFORCE model on the CDR sequence-structure co-design task. We demonstrate that 1) force guidance effectively guides the model to sample CDRs with lower energy; 2) using force guidance, DIFFFORCE outperforms current state-of-the-art models by generating high-quality antibody samples, with an emphasis on the CDR H3 region.

4.1 Experimental Setup

Baselines We compare DIFFFORCE against two baseline models, the diffusion model DIFFAB [Luo et al., 2022] and the traditional energy-based method RABD [Adolf-Bryfogle et al., 2018]. Both baseline models are evaluated using default settings. For more details, see Appendix F.

Dataset and Diffusion Model To evaluate our model, we use the SAbDab database [Dunbar et al., 2013] (with Chothia numbering scheme), filtering out complexes with resolutions worse than 4Å and those targeting non-protein antigens. Following [Luo et al., 2022], we cluster antibodies based on 50% H3 sequence identity and select five clusters for the test set, comprising 19 complexes. We use the *codesign_single* pre-trained model from DIFFAB, which generates one CDR region at a time.

Energy Energy of protein structures is evaluated using MadraX [Orlando et al., 2023], which provides the Gibbs free energy (ΔG) of the complex. Unlike other force fields, such as Rosetta [Alford et al., 2017] or FoldX [Schymkowitz et al., 2005], Madrax is fully differentiable. MadraX evaluates several categories of interaction energies, adapting 7 categories from FoldX [Schymkowitz et al., 2005] into a differentiable format. The energy considers the full protein structure, whose reconstruction is described in Appendix D. The energy is reported in *kcal/mol*.

Metrics To validate our model's performance, we use three key metrics; *1) Binding Energy Improvement (IMP)* is calculated as the percentage of designed CDRs that show a reduction (improvement) in free binding energy ($\Delta\Delta$G) compared to the reference CDRs, indicating a stronger interaction with the target antigen. This evaluation uses the InterfaceAnalyzer from Rosetta [Alford et al., 2017]. *2) Root Mean Square Deviation (RMSD)* measures the average spatial discrepancy between the C_α atoms of the generated and reference antibody structures, with a higher RMSD indicating greater structural diversity. *3) Amino Acid Recovery Rate (AAR)* is defined as the overlapping ratio of the generated sequence to the ground truth, evaluating how accurately the generated CDR sequences replicate the reference sequences [Adolf-Bryfogle et al., 2018].

4.2 Results

We evaluate the performance of DIFFFORCE model on the sequence-structure co-design as introduced by [Luo et al., 2022], where the reference CDR is removed from the antibody-antigen complex. The diffusion model is therefore conditioned on antibody framework and antigen. For each antigen-antibody complex, we generate $n = 25$ samples for 3 heavy chain CDRs (HCDRs). We choose to focus on the heavy chain since it typically exhibits greater variability and influences on binding affinity compared to the light chain [López-Requena et al., 2007]. The samples are produced through 100 generative timesteps ($T = 100$), with each sample maintaining the same length as its corresponding reference CDR in the test set. Finally, the generated structures, as well as reference original ones, are relaxed using OpenMM [Eastman et al., 2017] and Rosetta [Alford et al., 2017].

Table 1 shows that DIFFFORCE model recovers all three HCDRs sequences with greater accuracy (higher AAR) than both DIFFAB and RABD. Furthermore, DIFFFORCE achieves improved binding scores (higher IMP) for the H1 and H3 regions. The model exhibits RMSDs comparable to those of DIFFAB. Overall, the most substantial improvement is observed in the H3 region, which can be attributed to its significantly longer sequence and the smaller variability present in the H1 and H2 regions. This length allows for a wider range of adjustments and provides a greater scope for applying force guidance during sampling. This test validates the efficacy of our model in generating high-quality CDRs, with an emphasis on handling the complex CDR H3 region.

Figure 3 presents three generated CDR samples using DIFFFORCE. The binding specificity is determined by the interaction between the antibody's paratope region and the antigen's epitope region [Peng et al., 2014]. The paratope region, comprising the interacting amino acid residues from a

Method	AAR (%) ↑			IMP (%) ↑			RMSD (Å) ↓		
	H1	H2	H3	H1	H2	H3	H1	H2	H3
RABD	22.85	25.50	22.14	43.88	**53.50**	23.25	2.261	1.641	**2.900**
DIFFAB	58.70	49.37	26.08	47.91	30.77	23.59	**1.438**	**1.235**	3.605
DIFFFORCE	**60.78**	**53.51**	**29.52**	**49.45**	36.81	**30.22**	1.561	1.401	3.612

Table 1: Results on CDR sampling. The best result for each metric is highlighted in **bold**.

specific CDR region of an antibody, is highlighted in blue. The epitope, defined as the antigen residues within $<= 5$Å of the CDR, is marked in red. The antibodies target the SARS-CoV-2 RBD antigen, potentially offering a treatment strategy for COVID-19 [Law et al., 2021].

We focus on visualizing the CDR-H3 region, as it is often the most variable part of the antibody, determining its precise binding capability to a wide range of antigens and playing a key role in the immune response to pathogens [Regep et al., 2017]. The antigen-antibody framework is obtained from PDB:7DK2. All three samples show enhanced binding energy ($\Delta\Delta$G) as measured by Rosetta, despite significant structural deviations from the reference. This implies that a larger RMSD in the predicted CDR structure might indicate a viable alternative with enhanced binding capabilities, rather than a flaw in the prediction. Notably, Sample 1, which exhibits the best binding energy, appears to conform the best to the antigen, underscoring the potential advantages of structural deviations.

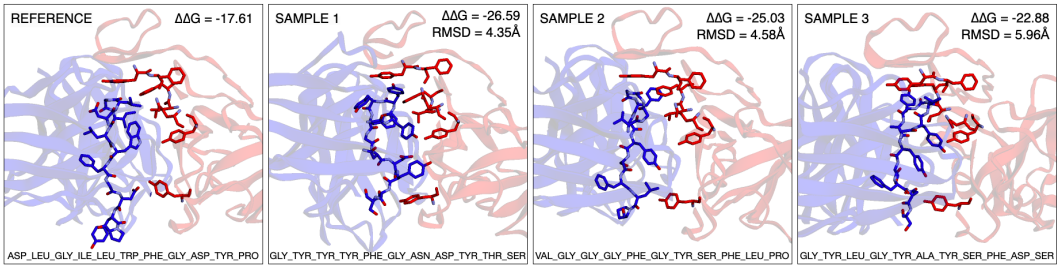

Figure 3: Generated samples for the CDR-H3 region of the PDB:7DK2 antigen-antibody complex. The RMSD, binding energy ($\Delta\Delta G$), and amino acid sequences are reported. The antigen is in red, and the antibody in blue. All samples show improved binding over the reference structure.

4.3 Analysis

We conduct experiments to evaluate DIFFFORCE's performance in generating antibodies, focusing on energy and structure. Our findings highlight two key insights: 1) DIFFFORCE consistently demonstrates improved stability over DIFFAB, indicated by lower energy (Section 4.3.1); 2) DIFFFORCE achieves better structural conformity earlier in the sampling than DIFFAB (Section 4.3.2).

4.3.1 Energy Landscape

In a proof-of-concept study, we demonstrate that the proposed force guidance enhances the efficacy of the DIFFFORCE model. This approach generates CDR conformations with lower energy, indicating increased structural stability compared to DIFFAB. Specifically, we analyze the 7DK2 antigen-antibody complex, focusing on the heavy chain CDR regions H1, H2, and H3, using hyperparameters $\lambda_{sc} = 0.05$ and $\lambda_{st} = 0.3$. We compare the results of $n = 25$ samples, all starting from the same configuration at timestep 70 of the 100-timestep sampling process. The data is smoothed using a 10-period moving average, and energy is measured using MadraX. As shown in Figure 4, the results indicate a decrease in energy for both models, with DIFFFORCE consistently exhibiting lower energy values from $t = 30$ onward. This suggests that the force guidance in DIFFFORCE effectively directs the sampling, leading to more energetically favorable conformations and more stable antigen-antibody interactions. For details on hyperparameter choices and for other complexes, refer to Appendix G.

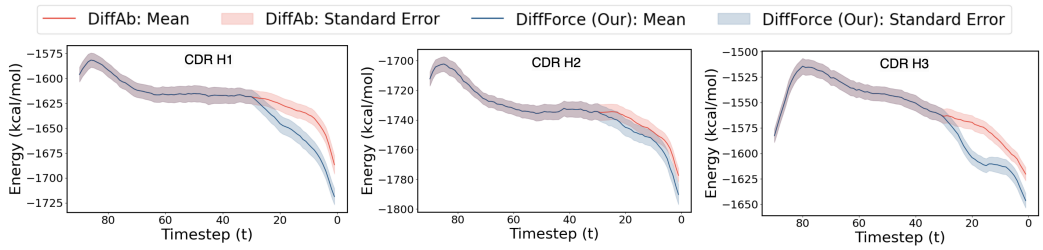

Figure 4: Energy of the PDB:7DK2 antigen-antibody complex's HCDR regions. Mean and standard error are based on $n = 25$ samples. The DIFFFORCE converges to lower energy levels than DIFFAB.

4.3.2 Structural Conformity

To further validate the effectiveness of force guidance, we conduct experiments on the structural conformity of generated antibody samples using the DIFFFORCE and DIFFAB models, focusing on the CDR H3 region of the 7DK2 antigen-antibody complex. We set hyperparameters at $\lambda_{sc} = 0.1$ and $\lambda_{st} = 0.3$, maintaining consistent seed values across both models. Figure 5 compares the models' performance at various sampling stages. Early in the diffusion process, DIFFFORCE consistently produces structures with better atomic coherence, fewer steric clashes, and higher structural connectivity than DIFFAB, particularly noticeable at earlier timesteps (e.g., $t = 15$, $t = 10$), indicating better sample fidelity. Additionally, DIFFFORCE achieves better energy at all sampled timesteps, demonstrating faster convergence to energetically optimal configurations. These empirical results highlight the potential of force guidance in improving the structural outcomes of diffusion-based antibody design, as well as reducing the need for post-generation relaxation.

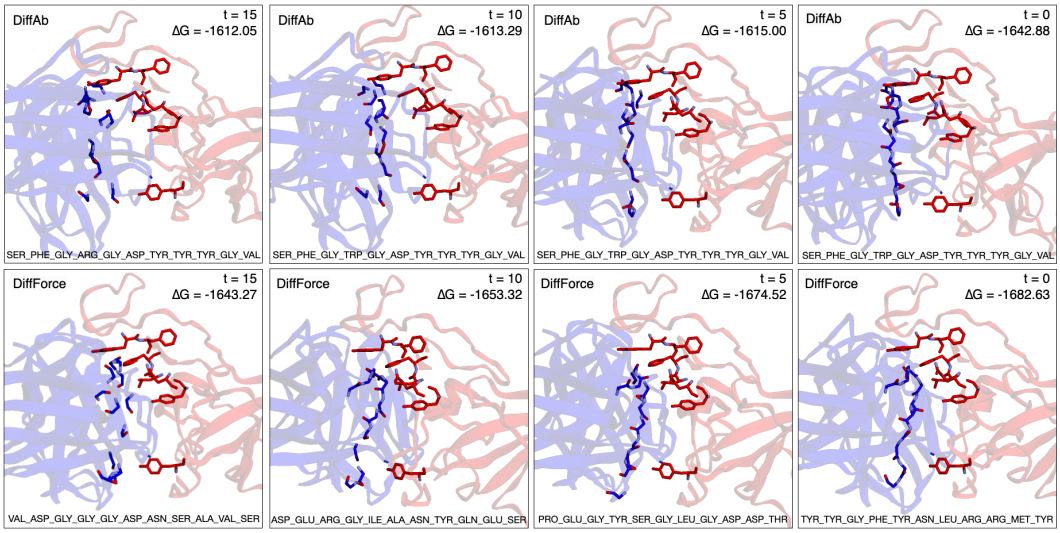

Figure 5: Results for the PDB:7DK2 complex's CDR-H3 region. Samples for DIFFAB (top) and DIFFFORCE (bottom) at timesteps $t = [15, 10, 5, 0]$. The energy and amino acid sequence are reported. DIFFFORCE achieves better structure and lower energy earlier in the sampling.

5 Conclusions and Future Work

Antibodies play a vital role in the immune system by identifying and neutralizing antigens, such as viruses. Inspired by the fact that integrating physics-based force fields with generative models can improve out-of-distribution generalization for antibody design, we introduce DIFFFORCE, a diffusion model that incorporates force guidance into the sampling. Unlike existing methods, our model does not require conditioning a diffusion model on energy or training a separate network to approximate energy. We demonstrate that our model effectively guides the diffusion sampler to generate CDRs

of better energy, outperforming several state-of-the-art models. This results in improved structure earlier in the sampling and enhances the sequences of the generated antibody CDRs.

While DIFFFORCE demonstrates promising results, it focuses on CDR sequence-structure co-design, with future potential in designing antibodies without bound framework structures. Moreover, the generated samples will require wet-lab experiments to confirm efficacy. Despite these and other limitations discussed in Appendix H, our work represents the first attempt to directly integrate a differentiable force field within diffusion sampling, effectively blending two distributions together.

6 Acknowledgments

José Miguel Hernández-Lobato acknowledges support from a Turing AI Fellowship under grant EP/V023756/1. Paulina Kulytė would like to thank Titas Anciukevičius for the constructive discussions and for proofreading the manuscript, and Augusta Fišerytė for the help with illustrations.

References

Jared Adolf-Bryfogle, Oleks Kalyuzhniy, Michael Kubitz, Brian D. Weitzner, Xiaozhen Hu, Yumiko Adachi, William R. Schief, and Roland L. Dunbrack, Jr. Rosettaantibodydesign (rabd): A general framework for computational antibody design. *PLOS Computational Biology*, 14(4):1–38, 04 2018. doi: 10.1371/journal.pcbi.1006112. URL https://doi.org/10.1371/journal.pcbi.1006112.

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):3031–3048, 2017. doi: 10.1021/acs.jctc.7b00125. URL https://doi.org/10.1021/acs.jctc.7b00125. PMID: 28430426.

D. Chandler, D. Wu, and P.C.D. Chandler. *Introduction to Modern Statistical Mechanics*. Oxford University Press, 1987. ISBN 9780195042764. URL https://books.google.co.uk/books?id=ep5yswEACAAJ.

Hyungjin Chung, Jeongsol Kim, Michael Thompson Mccann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=OnD9zGAGT0k.

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, 2022. doi: 10.1126/science.add2187. URL https://www.science.org/doi/abs/10.1126/science.add2187.

Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat GANs on image synthesis. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=AAWuCvzaVt.

James Dunbar, Konrad Krawczyk, Jinwoo Leem, Terry Baker, Angelika Fuchs, Guy Georges, Jiye Shi, and Charlotte M. Deane. SAbDab: the structural antibody database. *Nucleic Acids Research*, 42(D1):D1140–D1146, 11 2013. ISSN 0305-1048. doi: 10.1093/nar/gkt1043. URL https://doi.org/10.1093/nar/gkt1043.

Peter Eastman, Jason Swails, John D. Chodera, Robert T. McGibbon, Yutong Zhao, Kyle A. Beauchamp, Lee-Ping Wang, Andrew C. Simmonett, Matthew P. Harrigan, Chaya D. Stern, Rafal P. Wiewiora, Bernard R. Brooks, and Vijay S. Pande. Openmm 7: Rapid development of high performance algorithms for molecular dynamics. *PLOS Computational Biology*, 13(7):1–17, 07 2017. doi: 10.1371/journal.pcbi.1005659. URL https://doi.org/10.1371/journal.pcbi.1005659.

R. A. Engh and R. Huber. *Structure quality and target parameters*, chapter 18.3, pages 474–484. John Wiley & Sons, Ltd, 2012. ISBN 9780470685754. doi: https://doi.org/10.1107/97809553602060000857. URL https://onlinelibrary.wiley.com/doi/abs/10.1107/97809553602060000857.

Jordan Graves, Jacob Byerly, Eduardo Priego, Naren Makkapati, S. Vince Parish, Brenda Medellin, and Monica Berrondo. A review of deep learning methods for antibodies. *Antibodies*, 9(2), 2020. ISSN 2073-4468. doi: 10.3390/antib9020012. URL https://www.mdpi.com/2073-4468/9/2/12.

Alison C. Gray, Andrew R.M. Bradbury, Achim Knappik, Andreas Plückthun, Carl A.K. Borrebaeck, and Stefan Dübel. Animal-derived-antibody generation faces strict reform in accordance with european union policy on animal use. *Nature Methods*, 17(8):755–756, 2020. ISSN 1548-7091. doi: 10.1038/s41592-020-0906-9.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. URL https://openreview.net/forum?id=qw8AKxfYbI.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

Jonathan Ho, Tim Salimans, Alexey A. Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In *ICLR Workshop on Deep Generative Models for Highly Structured Data*, 2022. URL https://openreview.net/forum?id=BBelR2NdDZ5.

Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. Argmax flows and multinomial diffusion: Learning categorical distributions. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=6nbpPqUCIi7.

John Ingraham, Vikas Garg, Regina Barzilay, and Tommi Jaakkola. Generative models for graph-based protein design. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/f3a4ff4839c56a5f460c88cce3666a2b-Paper.pdf.

Yatong Ji, Jieyi Liu, Chen Wang, Fan Zhang, Xiangyang Xu, and Liang Zhu. Stability improvement of aerobic granular sludge (ags) based on gibbs free energy change (g) of sludge-water interface. *Water Research*, 240:120059, 2023. ISSN 0043-1354. doi: https://doi.org/10.1016/j.watres.2023.120059. URL https://www.sciencedirect.com/science/article/pii/S0043135423004955.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021. URL https://doi.org/10.1038/s41586-021-03819-2.

Urszula Julia Komorowska, Simon V Mathis, Kieran Didi, Francisco Vargas, Pietro Lio, and Mateja Jamnik. Dynamics-informed protein design with structure conditioning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=jZPqf2G9Sw.

Vered Kunik, Bjoern Peters, and Yanay Ofran. Structural consensus among antibodies defines the antigen binding site. *PLOS Computational Biology*, 8(2):1–12, 02 2012. doi: 10.1371/journal.pcbi.1002388. URL https://doi.org/10.1371/journal.pcbi.1002388.

John Lok Man Law, Michael Logan, Michael A. Joyce, Abdolamir Landi, Darren Hockman, Kevin Crawford, Janelle Johnson, Gerald LaChance, Holly A. Saffran, Justin Shields, Eve Hobart, Raelynn Brassard, Elena Arutyunova, Kanti Pabbaraju, Matthew Croxen, Graham Tipples, M. Joanne Lemieux, D. Lorne Tyrrell, and Michael Houghton. Sars-cov-2 recombinant receptor-binding-domain (rbd) induces neutralizing antibodies against variant strains of sars-cov-2 and sars-cov-1. *Vaccine*, 39(40):5769–5779, 2021. ISSN 0264-410X. doi: https://doi.org/10.1016/j.vaccine.2021.08.081. URL https://www.sciencedirect.com/science/article/pii/S0264410X21011294.

Adam Leach, Sebastian M Schmon, Matteo T. Degiacomi, and Chris G. Willcocks. Denoising diffusion probabilistic models on SO(3) for rotational alignment. In *ICLR 2022 Workshop on Geometrical and Topological Representation Learning*, 2022. URL https://openreview.net/forum?id=BY88eBbkpe5.

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 S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 9754–9767. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/3fa7d76a0dc1179f1e98d1bc62403756-Paper-Conference.pdf.

Alejandro López-Requena, Mabel Rodríguez, Cristina Mateo de Acosta, Ernesto Moreno, Yaquelin Puchades, Majela González, Ariel Talavera, Aisel Valle, Tays Hernández, Ana María Vázquez, and Rolando Pérez. Gangliosides, ab1 and ab2 antibodies: Ii. light versus heavy chain: An idiotype-anti-idiotype case study. *Molecular Immunology*, 44(5):1015–1028, 2007. ISSN 0161-5890. doi: https://doi.org/10.1016/j.molimm.2006.03.004. URL https://www.sciencedirect.com/science/article/pii/S016158900600109X.

Karolis Martinkus, Jan Ludwiczak, WEI-CHING LIANG, Julien Lafrance-Vanasse, Isidro Hotzel, Arvind Rajpal, Yan Wu, Kyunghyun Cho, Richard Bonneau, Vladimir Gligorijevic, and Andreas Loukas. Abdiffuser: full-atom generation of in-vitro functioning antibodies. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=7GyYpomkEa.

Radford M Neal. Annealed importance sampling. *Statistics and computing*, 11:125–139, 2001. URL https://doi.org/10.1023/A:1008923215028.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*. PMLR, 2023. URL https://doi.org/10.48550/arXiv.2112.10741.

Gabriele Orlando, Luis Serrano, Joost Schymkowitz, and Frederic Rousseau. Integrating physics in deep learning algorithms: A force field as a pytorch module. *bioRxiv*, 2023. doi: 10.1101/2023.01.12.523724. URL https://www.biorxiv.org/content/early/2023/01/16/2023.01.12.523724.

Hung-Pin Peng, Kuo Hao Lee, Jhih-Wei Jian, and An-Suei Yang. Origins of specificity and affinity in antibody–protein interactions. *Proceedings of the National Academy of Sciences*, 111(26):E2656–E2665, 2014. doi: 10.1073/pnas.1401131111. URL https://www.pnas.org/doi/abs/10.1073/pnas.1401131111.

David Ramírez and Julio Caballero. Is it reliable to use common molecular docking methods for comparing the binding affinities of enantiomer pairs for their protein target? *International Journal of Molecular Sciences*, 17 (4), 2016. ISSN 1422-0067. doi: 10.3390/ijms17040525. URL https://www.mdpi.com/1422-0067/17/4/525.

David Ramírez and Julio Caballero. Is it reliable to take the molecular docking top scoring position as the best solution without considering available structural data? *Molecules*, 23(5), 2018. ISSN 1420-3049. doi: 10.3390/molecules23051038. URL https://www.mdpi.com/1420-3049/23/5/1038.

Cristian Regep, Guy Georges, Jiye Shi, Bojana Popovic, and Charlotte M. Deane. The h3 loop of antibodies shows unique structural characteristics. *Proteins: Structure, Function, and Bioinformatics*, 85(7):1311–1318, 2017. doi: https://doi.org/10.1002/prot.25291. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.25291.

Herbert E Robbins. An empirical bayes approach to statistics. In *Breakthroughs in Statistics: Foundations and basic theory*, pages 388–394. Springer, 1992.

Joost Schymkowitz, Jesper Borg, Francois Stricher, Robby Nys, Frederic Rousseau, and Luis Serrano. The FoldX web server: an online force field. *Nucleic Acids Research*, 33, 07 2005. URL https://doi.org/10.1093/nar/gki387.

Inbal Sela-Culang, Vered Kunik, and Yanay Ofran. The structural basis of antibody-antigen recognition. *Frontiers in Immunology*, 4, 2013. ISSN 1664-3224. doi: 10.3389/fimmu.2013.00302. URL https://www.frontiersin.org/journals/immunology/articles/10.3389/fimmu.2013.00302.

Amir Pouya Shanehsazzadeh. In vitro validated antibody design against multiple therapeutic antigens using generative inverse folding. In *ICLR 2024 Workshop on Generative and Experimental Perspectives for Biomolecular Design*, 2024. URL https://openreview.net/forum?id=awGfgUJdQY.

Sachdev S Sidhu and Frederic A Fellouse. Synthetic therapeutic antibodies. *Nature chemical biology*, 2(12):682–688, 2006. URL https://doi.org/10.1038/nchembio843.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2256–2265, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/sohl-dickstein15.html.

Jiaming Song, Qinsheng Zhang, Hongxu Yin, Morteza Mardani, Ming-Yu Liu, Jan Kautz, Yongxin Chen, and Arash Vahdat. Loss-guided diffusion models for plug-and-play controllable generation. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 32483–32498. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/song23k.html.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=PxTIG12RRHS.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021b. URL https://openreview.net/forum?id=PxTIG12RRHS.

Brian L. Trippe, Jason Yim, Doug Tischer, David Baker, Tamara Broderick, Regina Barzilay, and Tommi S. Jaakkola. Diffusion probabilistic modeling of protein backbones in 3d for the motif-scaffolding problem. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=6TxBxqNME1Y.

Clement Vignac, Igor Krawczuk, Antoine Siraudin, Bohan Wang, Volkan Cevher, and Pascal Frossard. Digress: Discrete denoising diffusion for graph generation. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=UaAD-Nu86WX.

Yan Wang, Lihao Wang, Yuning Shen, Yiqun Wang, Huizhuo Yuan, Yue Wu, and Quanquan Gu. Protein conformation generation via force-guided se (3) diffusion models. *arXiv preprint arXiv:2403.14088*, 2024.

Shira Warszawski, Aliza Borenstein Katz, Rosalie Lipsh, Lev Khmelnitsky, Gili Ben Nissan, Gabriel Javitt, Orly Dym, Tamar Unger, Orli Knop, Shira Albeck, Ron Diskin, Deborah Fass, Michal Sharon, and Sarel J. Fleishman. Correction: Optimizing antibody affinity and stability by the automated design of the variable light-heavy chain interfaces. *PLOS Computational Biology*, 16(10):1–1, 10 2020. doi: 10.1371/journal.pcbi.1008382. URL https://doi.org/10.1371/journal.pcbi.1008382.

Joseph L. Watson, David Juergens, Nathaniel R. Bennett, Brian L. Trippe, Jason Yim, Helen E. Eisenach, Woody Ahern, Andrew J. Borst, Robert J. Ragotte, Lukas F. Milles, Basile I. M. Wicky, Nikita Hanikel, Samuel J. Pellock, Alexis Courbet, William Sheffler, Jue Wang, Preetham Venkatesh, Isaac Sappington, Susana Vázquez Torres, Anna Lauko, Valentin De Bortoli, Emile Mathieu, Sergey Ovchinnikov, Regina Barzilay, Tommi S. Jaakkola, Frank DiMaio, Minkyung Baek, and David Baker. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, aug 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06415-8. URL https://doi.org/10.1038/s41586-023-06415-8.

Jason Yim, Brian L. Trippe, Valentin De Bortoli, Emile Mathieu, Arnaud Doucet, Regina Barzilay, and Tommi Jaakkola. Se(3) diffusion model with application to protein backbone generation. In *International Conference on Machine Learning*. PMLR, 2023. URL https://doi.org/10.48550/arXiv.2302.02277.

Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Sasha Shysheya, Jonathan Crabbé, Lixin Sun, Jake Smith, Ryota Tomioka, and Tian Xie. Mattergen: a generative model for inorganic materials design. December 2023. URL https://www.microsoft.com/en-us/research/publication/mattergen-a-generative-model-for-inorganic-materials-design/.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the contributions made.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper includes the "Limitations" section in the Appendix H.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: All formulas in the paper are be numbered and cross-referenced. In particular the paper cites well established results with already clearly stated assumptions and builds on those.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The code will be released, and detailed instructions for replicating the results are provided in the paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code will be made publicly available.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the experimental setting for each of the experiment done.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The information required to understand statistical significance of experiments is provided in the paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper contains "Compute Details" section in the Appendix J.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The NeurIPS Code of Ethics was reviewed.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Yes, the section on "Broader Impacts" is presented within the Appendix I.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite all the original papers that produced the code package or dataset.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Yes, we provide the details of the code and our model as part of the submission.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

A Denoising Diffusion Probabilistic Models

Denoising diffusion probabilistic models (DDPMs), introduced by Ho et al. [2020], represent a class of generative models that generate data by reversing a diffusion process. This process involves gradually transforming a sample from a simple distribution, like Gaussian noise, into a complex data distribution through learned reverse diffusion steps. The forward process incrementally adds noise to the data over a series of steps, transforming an initial data distribution into a distribution that is approximately Gaussian. This process is designed as a Markov chain, where each state x_t only depends on the immediate previous state x_{t-1}. The transition from x_{t-1} to x_t is defined as:

$$x_t = \sqrt{\alpha_t} x_{t-1} + \sqrt{1-\alpha_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, I). \tag{15}$$

In this equation, α_t (where $0 < \alpha_t \leq 1$) is a predefined variance schedule decreasing over time, which determines the proportion of the original data and noise at each step. The variable ϵ represents isotropic Gaussian noise, introducing randomness into the process. The reverse process aims to reconstruct the original data by sequentially removing the noise added during the forward process. This is achieved by training a neural network to estimate the original data distribution at each previous timestep, effectively learning the reverse of the forward process. The transition from noisy data x_t back to less noisy data x_{t-1} is modeled as:

$$p(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)). \tag{16}$$

Here, $\mu_\theta(x_t, t)$ and $\Sigma_\theta(x_t, t)$ are the mean and covariance of the Gaussian distribution for x_{t-1}, parameterized by a neural network with parameters θ. These parameters are learned during training to minimize the difference between the actual noise and the predicted noise. The training of a DDPM is based on optimizing the variational lower bound, which effectively focuses on predicting the noise ϵ added at each step of the forward process. The loss function is defined as:

$$\mathcal{L}(\theta) = \mathbb{E}_{t,x_0,\epsilon}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|^2\right]. \tag{17}$$

This loss function measures the mean squared error between the actual noise ϵ and the noise estimated by the neural network ϵ_θ. Successful training minimizes this error, enhancing the model's ability to reverse the diffusion process and, thereby, accurately generate samples that resemble the training data.

Song et al. [2021a] state that DDPMs is an example from the larger class of score-based models. They demonstrated that the discrete forward and reverse diffusion processes have their continuous time equivalents, that is, forward Stochastic Differential Equation, namely:

$$dx = -\frac{1}{2}\beta(t)x_t dt + \sqrt{\beta(t)}dw, \tag{18}$$

and it's reverse:

$$dx_t = \left[-\frac{1}{2}\beta(t)x_t - \beta(t)\nabla_x \ln p_t(x_t)\right]dt + \sqrt{\beta(t)}d\bar{w}_t, \tag{19}$$

where the quantity $\nabla_{x_t} \ln p_t(x_t)$ is called the score and is closely related to the noise in DDPM by the equivalence $\nabla_{x_t} \ln p_t(x_t) = -\epsilon_t/\sqrt{1-\bar{\alpha}_t}$. Any model trained to predict the noise can be written in terms of the score, which is an essential property of our work. Whenever we derive some expression with respect to the score, we can use the noise-based formulation for forward and reverse diffusion processes by simply substituting $\epsilon_t = -\sqrt{1-\bar{\alpha}_t}\nabla_{x_t} \ln p_t(x_t)$.

B Normalisation of Forces

Using the relationship between energy and force, we start with the equation:

$$m_i \frac{d^2 x_i}{dt^2} = F_i = -\frac{\partial}{\partial x_i} U(x_1, x_2, \ldots, x_N), \tag{20}$$

where F_i is the force acting on the i-th atom, m_i is the mass of the i-th atom, x_i is the position vector of the i-th atom, and U is the energy as a function of the positions of all N atoms. Let $F = \{f_1, f_2, \ldots, f_N\}$ be a set of 3-dimensional vectors representing the forces acting on N atoms. Each vector $f_i \in \mathbb{R}^3$ consists of the force components along the x, y, and z coordinates for the i-th atom. We rescale each vector f_i such that its magnitude does not exceed a predefined maximum norm while maintaining its direction. The process involves three main steps:

Norm Calculation Compute the Euclidean norm (or L^2 norm) of each vector f_i. The Euclidean norm of f_i is defined as:

$$\|f_i\|_2 = \sqrt{f_{i,x}^2 + f_{i,y}^2 + f_{i,z}^2}, \tag{21}$$

where $f_{i,x}$, $f_{i,y}$, and $f_{i,z}$ represent the components of the i-th vector f_i along the x, y, and z axes, respectively.

Normalization Normalize each vector f_i to obtain a unit vector $\hat{f}_i$ by dividing it by its norm. To avoid division by zero, a small constant $\epsilon = 1e-6$ is added to the norm. The normalization step is described as:

$$\hat{f}_i = \frac{f_i}{\|f_i\|_2 + \epsilon}. \tag{22}$$

Rescaling Multiply each normalized vector $\hat{f}_i$ by a predefined maximum norm (we set the maximum norm to 1):

$$f_{i,\text{rescaled}} = \hat{f}_i \times \text{max\_norm}. \tag{23}$$

The output is a set of rescaled force vectors $F = \{f_{1,\text{rescaled}}, f_{2,\text{rescaled}}, \ldots, f_{N,\text{rescaled}}\}$, where each 3-dimensional vector $f_{i,\text{rescaled}}$ maintains its original direction and has its components within the range of -1 to 1, ensuring stability in the sampling algorithm.

C Physical Interpretation

In molecular dynamics (MD) simulations, the energy of molecular complexes is typically measured in kilocalories per mole (*kcal/mol*). The derivative of energy with respect to spatial position, i.e., the force, is thus expressed in *kcal/mol/Å*, where Å denotes angstroms (10^{-10} meters). This conversion from energy to force is important as it indicates both the magnitude and direction of forces exerted on atoms, facilitating the prediction of atomic movements over time within the simulation environment.

The relationship between the force applied to an atom and the resulting displacement can be understood through the basic kinematic equation:

$$\Delta x = 0.5 \times \left(\frac{F}{m}\right) \times \Delta t^2, \tag{24}$$

where F is the applied force, m is the mass of the atom, and Δt is the duration of the timestep. This equation emphasizes that the displacement (Δx) of an atom is proportional to the applied force and the square of the time interval, and inversely proportional to the atom's mass.

Essentially, this process is similar to a diffusion process on the coordinates, with a mini one-step MD relaxation at every step, where the time-step size is determined by λ_{sc}. The size of the timestep can be inferred from the hyperparameter λ_{sc}. Thus, to simplify simulation calculations, a scaling factor for force, denoted as λ_{sc}, is introduced, representing the term $\frac{0.5 \times \Delta t^2}{m}$ from Equation 24. Assuming the mass of a typical carbon-alpha (C_α) atom remains constant and that normalized forces F range between -1 and 1, the displacement for each simulation timestep can be efficiently computed as:

$$\Delta x = \lambda_{sc} \times F, \tag{25}$$

where λ_{sc} is the hyperparameter that scales the forces. This relation allows re-interpreting our reverse diffusion process as a combination of a reverse DDPM step on the coordinates, coupled with a one-step MD relaxation at every step, where the time-step size is determined by λ_{sc}. The size of the timestep can be inferred from the hyperparameter λ_{sc}.

D Structure Reconstruction

To calculate the energy, it is essential to reconstruct the full antibody-antigen complex C along with the generated CDR region R. This process involves first reconstructing the complete 3D structure of the atoms in the CDR, following the pipeline outlined in [Luo et al., 2022]. The reconstruction begins by determining the coordinates of the N, C, O, and side-chain C_β atoms, which are positioned relative to the C_α location and orientation of each amino acid [Engh and Huber, 2012]. After these core atoms are reconstructed, the remaining side-chain atoms are built using the side-chain packing function in Rosetta [Alford et al., 2017]. Once the CDR region is restored, the full antibody-antigen complex C is reconstructed. With the complete structure (including the antibody with its 6 CDRs and framework, as well as the antigen), the energy of the complex can be calculated. This process is then iteratively performed for $\lambda_{st} \times 100$ timesteps, assuming diffusion occurs over $t = 100$ timesteps. The iteration begins when forces are first applied at λ_{st} and continues through the sampling process until the final timestep $t = 0$.

E Algorithms

The following subsections describe two additional algorithms that were initially considered alongside our primary method. However, due to the more promising initial results of the main method, we discontinued further experimentation with these alternatives.

E.1 Algorithm 2: Sampling with Force Gradients of x^t

The initial sampling procedure is detailed in Algorithm 2 below.

Algorithm 2 DIFFFORCE-C_α Sampling with Force Guidance

$x^T \sim \mathcal{N}(0, I)$
for $t = T, \ldots, 1$ **do**
  $z \sim \mathcal{N}(0, I)$ if $t > 1$, else $z = 0$
  $x^{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(x^t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(x^t, t) \right) + \sigma_t z - \lambda_{sc} \mathbb{1}_{t \geq \lambda_{st}} \nabla_{x^t} U\left(x^t, s^t, O^t; C\right)$
  $R^{t-1} = \left(x^{t-1}, s^{t-1}, O^{t-1}\right)$, sample s^{t-1}, O^{t-1} following [Luo et al., 2022]
end for
return x^0, s^0, O^0

E.2 Algorithm 3: Sampling with Force Gradients of x^0 via Approximation

Another sampling procedure that was initially considered is detailed in Algorithm 3.

Algorithm 3 DIFFFORCE-C_α Sampling with Force Guidance

$x^T \sim \mathcal{N}(0, I)$
for $t = T, \ldots, 1$ **do**
  $z \sim \mathcal{N}(0, I)$ if $t > 1$, else $z = 0$
  estimate $\hat{x}^0(R^t)$ using Eq 8, $\hat{s}^0(R^t)$ using Eq 12 and $\hat{O}^0(R^t)$ using Eq 14
  $x^{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(x^t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(x^t, t) \right) + \sigma_t z - \lambda_{sc} \mathbb{1}_{t \geq \lambda_{st}} \nabla_{x^0} U\left(\hat{x}^0, \hat{s}^0, \hat{O}^0; C\right)$
  $R^{t-1} = \left(x^{t-1}, s^{t-1}, O^{t-1}\right)$, sample s^{t-1}, O^{t-1} following [Luo et al., 2022]
end for
return x^0, s^0, O^0

F Details of Baselines

DiffAb DIFFAB [Luo et al., 2022] models CDR sequences and structures using a diffusion model. This approach represents the first use of deep learning to integrate antigen 3D structures into antibody

sequence-structure design, thereby enhancing specificity and efficacy. Results for DIFFAB are obtained by our own experiments.

RAbD The RosettaAntibodyDesign (RABD) [Adolf-Bryfogle et al., 2018] is a computational tool for antibody design that utilizes Rosetta energy functions. It employs a Monte Carlo plus minimization (MCM) approach, wherein changes in antibody sequence and structure are randomly sampled and optimized through energy minimization to enhance target specificity [Adolf-Bryfogle et al., 2018]. Results for RABD are taken from a recent study Luo et al. [2022].

G Energy Landscape: Additional Details and Plots

The selection of hyperparameters $\lambda_{sc} = 0.05$ and $\lambda_{st} = 0.3$ for our study was guided by ablation studies examining the influence of the force start and force scale parameters in the DIFFFORCE model. The results are detailed in Figure 6, where the displayed values represent the mean. For example, to calculate the IMP metric for $\lambda_{st} = 0.1$, we averaged the samples of three HCDR regions: H1, H2, and H3. For each region, we computed the mean derived from $n = 25$ samples for the following combinations: $\lambda_{sc} = 0.01, \lambda_{st} = 0.1$; $\lambda_{sc} = 0.05, \lambda_{st} = 0.1$; and $\lambda_{sc} = 0.1, \lambda_{st} = 0.1$, across 19 test complexes.

The choice of $\lambda_{st} = 0.3$ was determined by averaging the optimal performance metrics for IMP and AAR, which peaked at $\lambda_{st} = 0.5$, and for RMSD, which was lowest at $\lambda_{st} = 0.1$. Similarly, $\lambda_{sc} = 0.05$ was selected because it provided the best outcomes for IMP and AAR at $\lambda_{sc} = 0.1$, while maintaining a lower RMSD value at $\lambda_{sc} = 0.01$. This ensured that multiple key metrics were optimized simultaneously. We argue that activating the forces earlier during sampling would enhance AAR and IMP metrics but result in longer sample generation times.

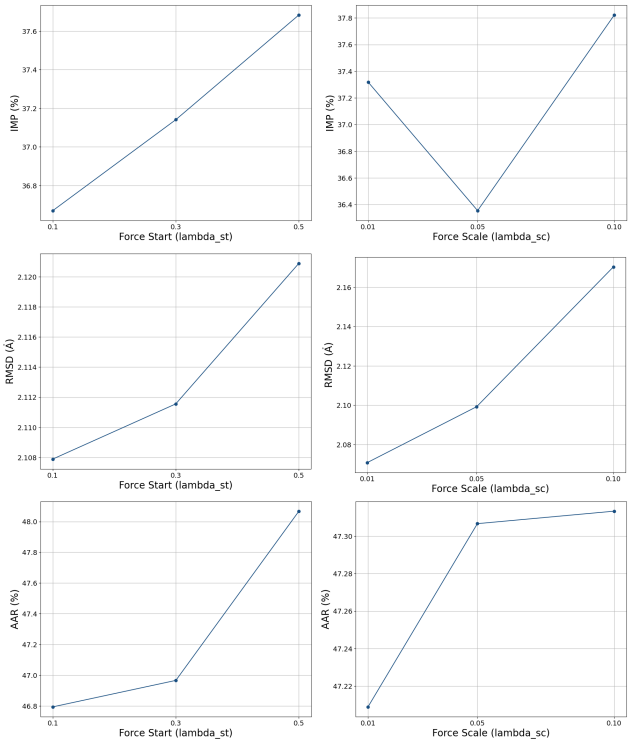

Figure 6: Ablation study showing the impact of the force start (λ_{st}) and force scale (λ_{sc}) hyperparameters on DIFFFORCE performance. The displayed values represent the mean. For IMP and AAR metrics, optimal results are obtained by activating forces early in the sampling process ($\lambda_{st} = 0.5$, 50% into sampling) with a higher force scale ($\lambda_{sc} = 0.1$). Conversely, for RMSD, better performance is achieved by activating forces later ($\lambda_{st} = 0.1$, 90% into sampling) with a lower force scale ($\lambda_{sc} = 0.01$).

Figure 7 provides an additional example from the experiment, analyzing three antigen-antibody complexes—PDB:7CHF, PDB:7CHE, and PDB:5TLK. The focus is on the heavy chain CDR regions, namely CDR-H1, CDR-H2, and CDR-H3. The figure demonstrates that the DIFFFORCE model, guided with force, generates antibody conformations with lower energy, indicating increased structural stability compared to DIFFAB.

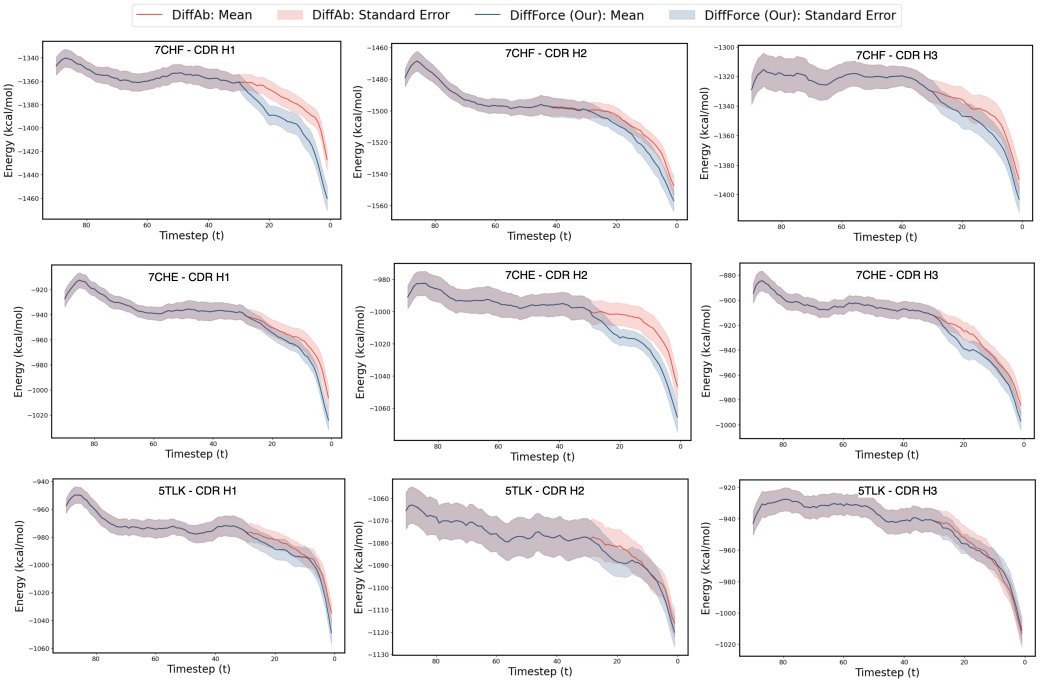

Figure 7: Energy landscape analysis of three antigen-antibody complexes—PDB:7CHF, PDB:7CHE, and PDB:5TLK—focused on the heavy chain CDR regions (CDR-H1, CDR-H2, and CDR-H3). The mean and standard error were derived from $n = 25$ samples across 25 seeds. The DIFFFORCE model, guided with force, converges to lower energy levels compared to DIFFAB.

H Limitations

Computational Cost The iterative use of the MadraX library [Orlando et al., 2023] for force guidance during sampling is time-consuming due to the calculations involved. This process mimics molecular dynamics (MD) simulations to continuously update atomic positions based on various forces, including bond forces, electrostatic interactions, van der Waals forces, and solvent interactions. Each iteration estimates atomic movements, similar to gradient descent optimization steps. Thus, the computational demands should be considered when employing this approach.

Reliability of Energy Function In our study, we utilized the Rosetta energy function [Alford et al., 2017] to evaluate the binding effectiveness of designed antibodies to their target antigens, a common metric in antibody design. Despite the integral role of Rosetta, along with tools such as FoldX [Schymkowitz et al., 2005], in simulating protein interactions, their reliability remains a subject of concern. These computational tools have been documented to exhibit inaccuracies when replicating experimental results, often due to the oversimplified models of complex molecular interactions they utilize [Ramírez and Caballero, 2016, 2018]. This underscores the necessity for ongoing refinement of these computational methods.

Evaluation Metrics In the field of antibody design, Amino Acid Recovery (AAR) and Root Mean Square Deviation (RMSD) are commonly used as evaluation metrics. However, these metrics have inherent limitations that may compromise the accurate assessment of an antibody's functional efficacy. AAR may not always reliably reflect the functional performance of the generated antibody sequences,

while RMSD primarily assesses the alignment of backbone atoms and overlooks the side chains, which are crucial for the specificity and strength of antigen-antibody interactions. These limitations underscore the need for the development of more comprehensive evaluation metrics.

I Broader Impacts

Integrating force guidance within the diffusion sampling process for antibody design can significantly accelerate therapeutic antibody discovery, with broad implications in fields like protein engineering. This advancement enables more precise modeling of atomic systems, enhancing predictions of protein stability, function, and interactions, crucial for designing enzymes and other biologically relevant molecules. However, potential societal drawbacks exist, particularly in dual-use applications. While aimed at therapeutic advancements, this approach could be misused to design harmful biological agents, raising ethical concerns and underscoring the need for regulations to ensure responsible use for societal benefit.

J Compute Details

The sampling phase was performed using four NVIDIA A100-SXM-80GB GPUs. The relaxation stage was executed on an Intel(R) Xeon(R) Gold 6142 CPU @ 2.60GHz, equipped with 48 virtual cores and 256GB of RAM.

K Source Code

The code will be made publicly available.