TLDR: RL-guided latent diffusion for *de novo* TCR design

Soo-Jeong Kim
EMBL-EBI, University of Cambridge
Cambridge, UK
soojeong@ebi.ac.uk

Isidro Cortes-Ciriano
EMBL-EBI
Cambridge, UK
icortes@ebi.ac.uk

Abstract

T cell receptor (TCR) design has been hindered by biased datasets that limit generalisation to novel epitopes. Consequently, existing models are restricted to only designing TCR's short binding region, CDR3β, for a small set of well-studied epitopes. We introduce TLDR, a framework that mitigates this challenge by fine-tuning an epitope-conditioned diffusion model with reinforcement learning guided by structural and biophysical constraints. By incentivising models with rewards that reflect more complex properties, we (i) generate plausible TCR candidates for under-represented and unseen epitopes, (ii) guide generation into novel regions of the design space and (iii) reduce dependence on the training data for supervision. To the best of our knowledge, TLDR is the first model capable of designing full-length, paired TCRs. Crucially, TLDR matches the CDR3β-only baselines on binding specificity tasks on seen epitopes and outperforms them on unseen tasks.

1 Introduction

T cell receptors (TCRs) are central to the adaptive immunity, enabling the recognition and elimination of infected and malignant cells (Mason, 1998). The specificity with which TCRs bind presented epitopes underlies their effectiveness, thus the design of *de novo* TCRs promises a new generation of personalised immunotherapy (Arellano et al., 2016; Chung et al., 2024; Zhao et al., 2024).

Recent breakthroughs in generative protein design have largely been confined to protein families supported by deep homology and strong structural priors (Ferruz et al., 2022; Jumper et al., 2021; Watson et al., 2023). However, TCRs, like many other therapeutically relevant proteins, fall outside of this regime: they have high combinatorial diversity, weak evolutionary constraints and complex many-to-many binding interactions. The existing dataset remains small and heavily imbalanced in comparison (Bagaev et al., 2020; Vita et al., 2019; Wei et al., 2025).

Nonetheless, related works on epitope-specific TCR design have largely been framed as a seq2seq task, where performance scales with paired data availability. Consequently, previous methods have poor generalisation and have confined their sequence design to the CDR3β hypervariable loop, the main binding region, for a few well-characterised epitopes (Li et al., 2023; Seo and Rhee, 2025).

Generating proteins as diverse as TCRs requires an approach that can effectively search this complex design space. Here, we explore the question: *could reinforcement learning (RL) guided by biological constraints mitigate generative models from the limitations of small and biased datasets?* A successful model would explore a vastly larger sequence space while ensuring that the candidates remain plausible, thereby offering a generalisable route to novel and functional TCRs.

TLDR (TCR design via Latent Diffusion fine-tuned with Reinforcement Learning) addresses the key bottleneck of data sparsity in epitope-specific $\alpha\beta$TCR design, and further extends sequence generation to full-length, paired-chain TCRs. TLDR integrates conditional latent diffusion with

Machine Learning for Structural Biology Workshop

policy-based optimization, fine-tuning generation with structural, sequence, and biophysical-based rewards. On binding specificity benchmarks, TLDR achieves competitive performance on dominant epitopes and outperforms existing approaches on underrepresented and unseen epitopes.

Our primary contributions are as follows:

1. We introduce, to the best of our knowledge, the first model for *de novo* epitope-specific design of full-length, paired-chain TCRs supporting dominant, underrepresented, and unseen epitopes.
2. We propose a framework for fine-tuning conditional latent diffusion models with reinforcement learning for both single- and multi-objective optimization
3. We present TLDR as a proof-of-concept for protein design in data sparse regimes.

2 Related Work

Reinforcement learning for protein design RL has gained traction for protein design across a range of tasks, including optimizing binding affinity, biophysical properties, antimicrobial activity, and inverse folding (Angermueller et al., 2019; Sternke and Karpiak, 2023; Chen et al., 2023; Xu et al., 2023). While prior work has shown that RL can steer models toward task-specific goals, its potential to improve generalization under data sparsity remains underexplored. Recent antibody design models (Mille-Fragoso et al., 2025; Swanson et al., 2025) backpropagated related biological objectives, but RL-based latent steering may provide a more generalizable framework for data-limited protein classes.

Diffusion modelling for protein design Diffusion models (Sohl-Dickstein et al., 2015; Song et al., 2020; Song and Ermon, 2019; Ho et al., 2020) for protein sequence design have predominantly been explored on discrete space (Alamdari et al., 2023; Su et al., 2025). More recent work by Wang et al. (2024) has enabled RL-guided discrete diffusion. In parallel, continuous latent diffusion models based on protein language model embeddings have been shown to support flexible sequence generation (Zhang et al., 2025; Lu et al., 2025; Meshchaninov et al., 2024), yet remain underexplored in the context of RL.

Reinforcement learning in diffusion models In text and vision, RL has been widely adopted to fine-tune continuous diffusion models, aligning them with human feedback and preferences (Fan et al., 2023; Black et al., 2023; Zekri and Boullé, 2025; Prabhudesai et al., 2023). We build on this line of work by extending RL-based fine-tuning to conditional latent diffusion for biological sequence design on both single- and multi-objective optimization.

3 Method

3.1 Stage 1: TCR reconstruction training

The goal of the TCR autoencoder is to train a decoder that can map a learned TCR embedding back to the interpretable sequence space. Here, we use ESM models as encoders (Lin et al., 2023; Hayes et al., 2025). The decoder reconstructs full TCR sequences using a conditional autoregressive distribution:

$$p_\beta(x \mid z) = \prod_{i=1}^{L} p_\beta(x_i \mid x_{<i}, z),$$

where $x = (x_1, \ldots, x_L)$ denotes the TCR amino acid sequence, z the ESM latent TCR embedding, and β the parameters of the autoencoder. Note that the latent embedding conditions the sequence generation.

3.2 Stage 2: Conditional latent diffusion modelling

We trained a conditional Gaussian latent diffusion model to generate TCR latent representations conditioned on an epitope embedding c , which was obtained by encoding the epitope peptide sequence with an encoder initialised from the ESM-C model. Starting from Gaussian noise $z^{(0)} \sim \mathcal{N}(0, I_d)$, the

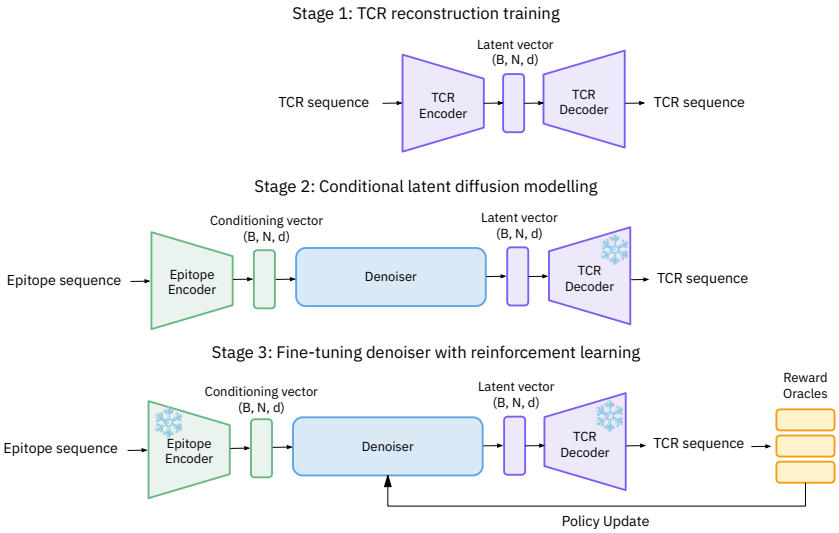

Figure 1: TLDR framework is trained in three stages, where batch size is B, sequence length is N, and model dimension is d. See Appendix A.1 for training and implementation details.

model learns to denoise toward TCR latents $z^{(1)}$ learned by the autoencoder, yielding the conditional distribution $p_\theta(z^{(1)} \mid c)$. At an interpolation step $t \sim \mathcal{U}(0,1)$, a noisy sample $z^{(t)}$ is constructed, and the denoiser $\mathcal{F}$ is tasked with predicting the denoised TCR latent from $(z^{(t)}, t, c)$. For a training batch of size N, the objective is a mean squared error between predicted $z_i'^{(1)}$ and true latents $z_i^{(1)}$:

$$\mathcal{L} = \frac{1}{(1-t)^2} \cdot \frac{1}{N} \sum_{i=1}^{N} \left\| z_i^{(1)} - z_i'^{(1)} \right\|^2 .$$

After training, epitope embeddings can be transformed via diffusion into TCR latents, which the decoder D of the autoencoder reconstructs into TCR sequences.

3.3 Stage 3: Finetuning denoiser with reinforcement learning

The goal here is to fine-tune the denoiser to generate TCR sequences with higher biological plausibility and functional relevance. We formulated the reverse denoising process as a Markov decision process (MDP) with T steps. At each step $t \in \{T, T-1, \ldots, 1\}$, the state is $s_t = (c, t, z_t)$, the action is the next denoised sample $a_t = z_{t-1}$, and the policy is the Gaussian transition distribution: $\pi_\theta(a_t \mid s_t) = p_\theta(z_{t-1} \mid z_t, c)$.

A trajectory corresponds to the reverse diffusion chain in the TCR latent space, starting from pure noise $z_T \sim \mathcal{N}(0, I_d)$, yielding the sequence $\{z_T, z_{T-1}, \ldots, z_0\}$. We define the reward only at the final step $(t = 0)$, where z_0 is decoded into a full-length TCR by decoder D. Thus the reward is $R(z_0) = r(D(z_0), c)$. Here, maximising the expected return is equivalent to optimising the denoising block to generate higher-reward TCR sequences. We evaluate rewards along three axes: (i) sequence (perplexity), (ii) structure (pLDDT, pTM from ESMFold (Lin et al., 2023)), and (iii) biophysical plausibility (amino acid property annotations (UniProt Consortium, 2018)). Full definitions are provided in Appendix A.4.

We adopted the Denoising Diffusion Reinforcement Learning (DDRL) objective from Black et al. (2023):

$$\nabla_\theta J_{\text{DDRL}} = \mathbb{E}\left[\sum_{t=1}^{T} \frac{p_\theta(z_{t-1} \mid z_t, c)}{p_{\theta_{\text{old}}}(z_{t-1} \mid z_t, c)} \nabla_\theta \log p_\theta(z_{t-1} \mid z_t, c)\, r(D(z_0), c) \right],$$

where trajectories are sampled under parameters of previous policy θ_{old} to update the current parameters θ. For model architecture, training and implementation details, see Appendix A.1.

4 Results

4.1 TLDR generates highly plausible and novel full-length epitope-specific TCRs

Table 1: Comparison of predicted structural quality and sequence-derived properties for natural versus generated TCR sequences. Structural model confidence is evaluated using pLDDT and pTM scores from AlphaFold2 (Jumper et al., 2021), and sequence-level properties include hydrophobicity, net charge at pH 7.0, and amino acid composition (Appendix A.2). All values report the mean $\pm$ standard deviation. TLDR generates proteins with properties within the natural held-out TCR range.

Source	pLDDT	pTM	Hydrophobicity	Sequence composition	Repertoire diversity
Natural	0.56 ± 0.29	0.49 ± 0.30	-0.45 ± 0.10	0.94 ± 0.01	0.78
TLDR	0.54 ± 0.24	0.41 ± 0.24	-0.46 ± 0.07	0.94 ± 0.02	0.83

We investigated the ability of TLDR to generate feasible paired $\alpha\beta$TCR chains. The goal was to learn the distribution of the natural TCR space in order to generate realistic yet novel and diverse protein sequences. Accordingly, we evaluated structural and biophysical properties alongside sequence diversity and novelty (see Appendix A.2 for detailed metrics). Note that because TLDR was the first model capable of generating beyond the CDR3β region, we could not directly benchmark against existing models; instead, we evaluated against natural TCRs.

TLDR generated samples have lower pLDDT and pTM mean values to natural TCRs, but remain within the natural distribution (Table 1). The biophysical scores confirmed that the generated sequences occupy the same range as natural TCRs. We also found that the generated sequence repertoire exhibited similar diversity to the natural TCR.

4.2 RL fine-tuning enables high-fidelity TCR generation for unseen epitopes

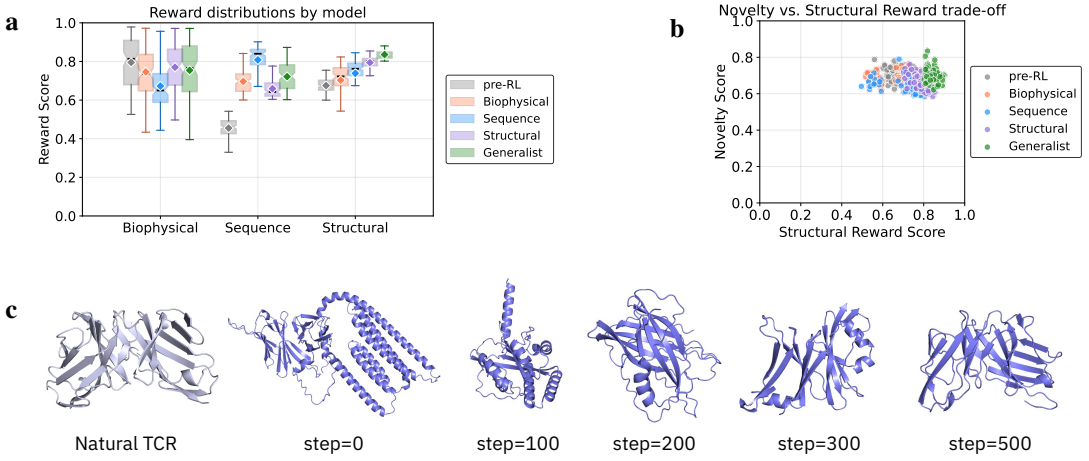

Figure 2: RL fine-tuned models for unseen epitopes. **Top**: a) Reward transferability between specialist (sequence, structural, biophysical) and generalist (structural and biophysical) policies, showing b) trade-off between reward and novelty. **Bottom**: c) Reward progression over training for the zero-shot epitope EEAAGIGIL, with the corresponding full-length TCR 3D structure visualised with AlphaFold3 (Abramson et al., 2024). Overall, the structural policy alone achieves favorable performance and novelty.

To investigate the capacity of RL to mitigate epitope imbalance and sparsity, we trained four distinct optimisation policies for unseen epitope-conditioned TCR design. A single-objective, sequence-based policy optimized directly on perplexity; and three multi-objective policies guided by: (i) structural metrics, (ii) biophysical properties and (iii) a generalist combining the structural and biophysical objectives.

First, we evaluated whether specialist and generalist policies exhibited transfer across rewards beyond those used for training. Specifically, we computed rewards across all types for sequences generated by each policy and compared them against an untrained baseline. All policies outperformed the untrained baseline on both sequence and structural rewards, indicating gains on metrics outside of their direct optimisation scope. Poor performance exhibited by the biophysical specialist across reward types indicated that biophysical properties alone are insufficient as a reward, likely due to the reward primarily reflecting amino acid frequency biases. Interestingly, the generalist, which was trained on biophysical alongside structural rewards, achieved higher structural scores than the structural specialist. Overall, the structural specialist achieved the strongest transfer, and achieved a favourable trade-off between structural plausibility and novelty. 3D visualisation across training confirmed that an increase of reward resulted in genuine improvements. Thus we adopted the structural specialist for following evaluation and benchmarking.

4.3 TLDR generalises epitope-specific design to unseen epitopes

Table 2: Benchmark of TLDR against CDRβ-only baseline models, GRATCR (Zhou et al., 2025) and TCR-T5 (Karthikeyan et al., 2025), on novelty, diversity and binding specificity. The latter is reported as the Wasserstein distance ratio between NetTCR2.0 (Montemurro et al., 2021) score distributions to binding and non-binding TCRs. TLDR generates sequences with greater novelty and diversity while producing binding candidates.

Model	CDR3β Novelty ↑	CDR3β Diversity ↑	W-Dist. Ratio (Seen) Binders/Non-binders ↓	W-Dist. Ratio (Unseen) Binders/Non-binders ↓
GRATCR	0.215	0.12	1.90	2.32
TCR-T5	0.261	0.57	**0.36**	4.52
TLDR	**0.431**	0.60	0.59	**0.61**

As no generative models for full-length TCRs exist, we benchmarked our model against the epitope-specific CDR3β generation models GRATCR (Zhou et al., 2025) and TCR-T5 (Karthikeyan et al., 2025). For direct comparison, CDR3β sequences were extracted from TLDR outputs using conserved anchor positions. All models were evaluated on binding specificity, novelty, and diversity metrics to assess performance while diagnosing potential mode collapse and overfitting.

To select an oracle for binding specificity, we tested three state-of-the-art external classifiers (Appendix A.3, Extended Figure 3), and selected NetTCR2.0 (Montemurro et al., 2021), which exhibited the highest recall in our setup (Appendix A.3). Specificity was quantified by measuring the distributional similarity between the predicted scores of generated sequences and those of known binding and non-binding TCRs, using the Wasserstein distance ratio (Appendix A.3).

Our results showed that generated sequences from TLDR exhibited score distributions closer to the true binding CDR3β sequences than to non-binding, indicating that our approach captures binding-relevant features of CDR3β despite operating at the full-sequence level (Table 2). We report poor discriminative abilities along with low novelty and diversity for GRATCR, indicating potential mode collapse. TCR-T5 was able to explore larger sequence space and excelled at generating binding CDR3β sequences for seen epitopes. However, it failed to generalise to unseen epitopes. Our model, TLDR, was able to have the highest performance to unseen epitopes, and also achieved the most stable performance between seen and unseen epitopes.

5 Conclusion

We introduced TLDR, a framework for *de novo* generation of full-length, paired TCRs - including for unseen epitopes - using reward-based fine-tuning of latent diffusion models to guide sequences toward biologically plausible and natural properties. TLDR demonstrates strong generalisation but remains limited by reliance on external oracles, though ESMFold's structural priors were sufficient to recover plausible candidates. We also propose an evaluation framework that accounts for sequence novelty and the known limitations of current evaluation models. More broadly, TDLR serves as a proof-of-concept offering a way to explore data sparse yet functionally desirable regions of the sequence space. Future work should integrate MHC for a more complete biological model.

References

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, 2024.

Sarah Alamdari, Nitya Thakkar, Rianne Van Den Berg, Neil Tenenholtz, Bob Strome, Alan Moses, Alex Xijie Lu, Nicolo Fusi, Ava Pardis Amini, and Kevin K Yang. Protein generation with evolutionary diffusion: sequence is all you need. *BioRxiv*, pages 2023–09, 2023.

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.

Benjamine Arellano, David J Graber, and Charles L Sentman. Regulatory t cell-based therapies for autoimmunity. *Discovery medicine*, 22(119):73, 2016.

Dmitry V Bagaev, Renske MA Vroomans, Jerome Samir, Ulrik Stervbo, Cristina Rius, Garry Dolton, Alexander Greenshields-Watson, Meriem Attaf, Evgeny S Egorov, Ivan V Zvyagin, et al. Vdjdb in 2019: database extension, new analysis infrastructure and a t-cell receptor motif compendium. *Nucleic Acids Research*, 48(D1):D1057–D1062, 2020.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023.

Ziqi Chen, Baoyi Zhang, Hongyu Guo, Prashant Emani, Trevor Clancy, Chongming Jiang, Mark Gerstein, Xia Ning, Chao Cheng, and Martin Renqiang Min. Binding peptide generation for mhc class i proteins with deep reinforcement learning. *Bioinformatics*, 39(2):btad055, 2023.

James B Chung, Jennifer N Brudno, Dominic Borie, and James N Kochenderfer. Chimeric antigen receptor t cell therapy for autoimmune disease. *Nature Reviews Immunology*, 24(11):830–845, 2024.

Yasha Ektefaie, Olivia Viessmann, Siddharth Narayanan, Drew Dresser, J Mark Kim, and Armen Mkrtchyan. Reinforcement learning on structure-conditioned categorical diffusion for protein inverse folding. *arXiv preprint arXiv:2410.17173*, 2024.

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36:79858–79885, 2023.

Noelia Ferruz, Steffen Schmidt, and Birte Höcker. Protgpt2 is a deep unsupervised language model for protein design. *Nature communications*, 13(1):4348, 2022.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, et al. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, 2025.

James M Heather, Matthew J Spindler, Marta Herrero Alonso, Yifang Ivana Shui, David G Millar, David S Johnson, Mark Cobbold, and Aaron N Hata. Stitchr: stitching coding tcr nucleotide sequences from v/j/cdr3 information. *Nucleic acids research*, 50(12):e68–e68, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

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*, pages 9786–9801. PMLR, 2022.

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.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the nineteenth international conference on machine learning*, pages 267–274, 2002.

Dhuvarakesh Karthikeyan, Sarah N Bennett, Amy G Reynolds, Benjamin G Vincent, and Alex Rubinsteyn. Conditional generation of real antigen-specific t cell receptor sequences. *Nature Machine Intelligence*, pages 1–16, 2025.

Jack Kyte and Russell F Doolittle. A simple method for displaying the hydropathic character of a protein. *Journal of molecular biology*, 157(1):105–132, 1982.

Tianxiao Li, Hongyu Guo, Filippo Grazioli, Mark Gerstein, and Martin Renqiang Min. Disentangled wasserstein autoencoder for t-cell receptor engineering. *Advances in Neural Information Processing Systems*, 36:73604–73632, 2023.

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):1123–1130, 2023.

Amy X Lu, Wilson Yan, Sarah A Robinson, Simon Kelow, Kevin K Yang, Vladimir Gligorijevic, Kyunghyun Cho, Richard Bonneau, Pieter Abbeel, and Nathan C Frey. All-atom protein generation with latent diffusion. In *ICLR 2025 Workshop on Generative and Experimental Perspectives for Biomolecular Design*, 2025.

Don Mason. A very high level of crossreactivity is an essential feature of the t-cell receptor. *Immunology Today*, 19(9):395–404, 1998. ISSN 0167-5699. doi: https://doi.org/10.1016/S0167-5699(98)01299-7. URL `https://www.sciencedirect.com/science/article/pii/S0167569998012997`.

Viacheslav Meshchaninov, Pavel Strashnov, Andrey Shevtsov, Fedor Nikolaev, Nikita Ivanisenko, Olga Kardymon, and Dmitry Vetrov. Diffusion on language model encodings for protein sequence generation. *arXiv preprint arXiv:2403.03726*, 2024.

Luis S Mille-Fragoso, John N Wang, Claudia L Driscoll, Haoyu Dai, Talal Widatalla, Xiaowei Zhang, Brian L Hie, and Xiaojing J Gao. Efficient generation of epitope-targeted de novo antibodies with germinal. *bioRxiv*, pages 2025–09, 2025.

Alessandro Montemurro, Viktoria Schuster, Helle Rus Povlsen, Amalie Kai Bentzen, Vanessa Jurtz, William D Chronister, Austin Crinklaw, Sine R Hadrup, Ole Winther, Bjoern Peters, et al. Nettcr-2.0 enables accurate prediction of tcr-peptide binding by using paired tcrα and β sequence data. *Communications biology*, 4(1):1060, 2021.

Sean Nolan, Marissa Vignali, Mark Klinger, Jennifer N Dines, Ian M Kaplan, Emily Svejnoha, Tracy Craft, Katie Boland, Mitchell W Pesesky, Rachel M Gittelman, et al. A large-scale database of t-cell receptor beta sequences and binding associations from natural and synthetic exposure to sars-cov-2. *Frontiers in Immunology*, 16:1488851, 2025.

Mihir Prabhudesai, Anirudh Goyal, Deepak Pathak, and Katerina Fragkiadaki. Aligning text-to-image diffusion models with reward backpropagation. 2023.

Matthew IJ Raybould, Alexander Greenshields-Watson, Parth Agarwal, Broncio Aguilar-Sanjuan, Tobias H Olsen, Oliver M Turnbull, Nele P Quast, and Charlotte M Deane. The observed t cell receptor space database enables paired-chain repertoire mining, coherence analysis and language modelling. *bioRxiv*, pages 2024–05, 2024.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Se Yeon Seo and Je-Keun Rhee. Tcr-epidiff: solving dual challenges of tcr generation and binding prediction. *Bioinformatics*, 41(Supplement_1):i125–i132, 2025.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. pmlr, 2015.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Martin Steinegger and Johannes Söding. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11):1026–1028, 2017.

Matt Sternke and Joel Karpiak. Proteinrl: Reinforcement learning with generative protein language models for property-directed sequence design. In *NeurIPS 2023 Generative AI and Biology (GenBio) Workshop*, 2023.

Xingyu Su, Xiner Li, Masatoshi Uehara, Sunwoo Kim, Yulai Zhao, Gabriele Scalia, Ehsan Hajiramezanali, Tommaso Biancalani, Degui Zhi, and Shuiwang Ji. Iterative distillation for reward-guided fine-tuning of diffusion models in biomolecular design. *arXiv preprint arXiv:2507.00445*, 2025.

Erik Swanson, Michael Nichols, Supriya Ravichandran, and Pierce Ogden. mber: Controllable de novo antibody design with million-scale experimental screening. *bioRxiv*, pages 2025–09, 2025.

The UniProt Consortium. Uniprot: the universal protein knowledgebase. *Nucleic acids research*, 46 (5):2699–2699, 2018.

Randi Vita, Swapnil Mahajan, James A Overton, Sandeep Kumar Dhanda, Sheridan Martini, Jason R Cantrell, Daniel K Wheeler, Alessandro Sette, and Bjoern Peters. The immune epitope database (iedb): 2018 update. *Nucleic acids research*, 47(D1):D339–D343, 2019.

Chenyu Wang, Masatoshi Uehara, Yichun He, Amy Wang, Tommaso Biancalani, Avantika Lal, Tommi Jaakkola, Sergey Levine, Hanchen Wang, and Aviv Regev. Fine-tuning discrete diffusion models via reward optimization with applications to dna and protein design. *arXiv preprint arXiv:2410.13643*, 2024.

Andrew Waterhouse, Martino Bertoni, Stefan Bienert, Gabriel Studer, Gerardo Tauriello, Rafal Gumienny, Florian T Heer, Tjaart A P de Beer, Christine Rempfer, Lorenza Bordoli, et al. Swiss-model: homology modelling of protein structures and complexes. *Nucleic acids research*, 46(W1): W296–W303, 2018.

Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.

Mengmeng Wei, Jingcheng Wu, Shengzuo Bai, Yuxuan Zhou, Yichang Chen, Xue Zhang, Wenyi Zhao, Ying Chi, Gang Pan, Feng Zhu, et al. Trait: A comprehensive database for t-cell receptor–antigen interactions. *Genomics, Proteomics & Bioinformatics*, page qzaf033, 2025.

Kevin E Wu, Kathryn Yost, Bence Daniel, Julia Belk, Yu Xia, Takeshi Egawa, Ansuman Satpathy, Howard Chang, and James Zou. Tcr-bert: learning the grammar of t-cell receptors for flexible antigen-binding analyses. In *Machine Learning in Computational Biology*, pages 194–229. PMLR, 2024.

Xiaopeng Xu, Tiantian Xu, Juexiao Zhou, Xingyu Liao, Ruochi Zhang, Yu Wang, Lu Zhang, and Xin Gao. Ab-gen: antibody library design with generative pre-trained transformer and deep reinforcement learning. *Genomics, Proteomics & Bioinformatics*, 21(5):1043–1053, 2023.

Jason Yim, Andrew Campbell, Andrew YK Foong, Michael Gastegger, José Jiménez-Luna, Sarah Lewis, Victor Garcia Satorras, Bastiaan S Veeling, Regina Barzilay, Tommi Jaakkola, et al. Fast protein backbone generation with se (3) flow matching. *arXiv preprint arXiv:2310.05297*, 2023.

Oussama Zekri and Nicolas Boullé. Fine-tuning discrete diffusion models with policy gradient methods. *arXiv preprint arXiv:2502.01384*, 2025.

Pengfei Zhang, Seojin Bang, Michael Cai, and Heewook Lee. Context-aware amino acid embedding advances analysis of tcr-epitope interactions. *bioRxiv*, pages 2023–04, 2023.

Sitao Zhang, Zixuan Jiang, Rundong Huang, Wenting Huang, Siyuan Peng, Shaoxun Mo, Letao Zhu, Peiheng Li, Ziyi Zhang, Emily Pan, et al. Pro-ldm: A conditional latent diffusion model for protein sequence design and functional optimization. *Advanced Science*, page e02723, 2025.

Xiang Zhao, Shuai Shao, and Lanxin Hu. The recent advancement of tcr-t cell therapies for cancer treatment: Tcr-t cell therapies for cancer treatment. *Acta biochimica et biophysica Sinica*, 56(5): 663, 2024.

Zhenghong Zhou, Junwei Chen, Shenggeng Lin, Liang Hong, Dong-Qing Wei, and Yi Xiong. Gratcr: epitope-specific t cell receptor sequence generation with data-efficient pre-trained models. *IEEE Journal of Biomedical and Health Informatics*, 2025.

A Appendix

A.1 Model architecture, training and datasets

Model architecture For the first stage TCR autoencoder, we employed ESM-C (300m) (Lin et al., 2023) model as the encoder, augmented with a single trainable Transformer layer to project into the latent space. The decoder is a 6-layer Transformer (hidden dimension 1280, 8 attention heads). The second stage diffusion model also uses the ESM-C (300m) (Lin et al., 2023) as the encoder for the epitope peptide sequence. The denoiser has a hidden dimension 1280, 8 attention heads, 6 layers, thus the diffusion block with 184M parameters in total.

Training and hyperparamters We trained the first-stage autoencoder, then the second stage denoiser with a constant learning rate of $1\mathrm{e}{-4}$ and a batch size of 256. Both models are trained for at least 150 epochs on 1 A100 GPU. After pre-training, we froze the denoiser and decoder weights, and then fine-tuned the model with RL by attaching LoRA adaptors (rank$=8$, $\alpha = 16$, dropout$=0.1$) (Hu et al., 2022) to the DiT blocks of the denoiser. Training was done using the AdamW optimizer.

For diffusion, 0.1 KL divergence penalty and a 15% drop out rate were used during training. We also incorporated self-conditioning (Yim et al., 2023). During training, the noisy latent $z^{(t)}$ is obtained by perturbing the clean TCR latent $z^{(0)}$ from the autoencoder's encoder using a cosine noise schedule. For inference, we used 150 integration steps and classifier free guidance scale of 0.15.

In the third stage, fine-tuning was performed for 30 epochs, using 64 batches per epoch with 10 generated sequences per epitope sample. The PPO ratio clipping was set to 0.3, following the trust-region principle of proximal policy optimization. For stability, we applied gradient and advantage clipping at 0.5 and 0.7 respectively. Following the method of (Black et al., 2023), a policy gradient estimator (Schulman et al., 2017) and importance sampling ratio (Kakade and Langford, 2002) were implemented.

Datasets We trained the autoencoder on the Observed T cell receptor Space database (Raybould et al., 2024) which contains 5.35 million full-length, human paired $\alpha\beta$ TCR sequences. For the denoiser, 23,544 TCR-epitope interaction data was taken from TRAIT database (Wei et al., 2025), from which full-length TCRs were reconstructed using stitchr (Heather et al., 2022). Binary classifiers were trained on the MIRA dataset (Nolan et al., 2025).

A.2 Evaluation metrics - Sequence level

Natural TCR sequences shown here were not used during training; they originate from the held-out test set, which was constructed to ensure a maximum of 60% sequence identity to the training set, as computed using MMseqs2 (Steinegger and Söding, 2017).

Perplexity A standard sequence modeling metric, defined as the exponential of the negative average log-likelihood. Lower perplexity indicates that the model assigns higher probability to the sequences.

Structural validity We assess structural validity using AlphaFold2 outputs (Jumper et al., 2021). In particular, we use pLDDT score as a residue-level confidence measure and the pTM score as a global confidence metric. High pLDDT values indicate that the local structure is predicted with high confidence, while high pTM values suggest that the overall fold is reliable. Together, these metrics serve as proxies for whether a generated sequence is likely to adopt a stable and functional protein structure.

Novelty To ensure that the generated sequences are not overfitting to the training set, we follow the novelty formulation from Jain et al. (2022) for each sequence x to a reference template set:

$$\mathrm{Nov}(x) = \min_{s_i \in D_0} d(x, s_i),$$

where D_0 is the set of template sequences and $d(\cdot, \cdot)$ is the Levenshtein distance normalized by the length of the longer sequence. Intuitively, this score captures how different a proposed sequence is from those already known.

For a dataset D, we summarize novelty as:

$$\text{Novelty}(D) = \frac{1}{|D|} \sum_{(x_i,y_i)\in D} \min_{s_j \in D_0} d(x_i, s_j).$$

Sequence diversity We quantify sequence diversity as given by Ektefaie et al. (2024) - the average pairwise Hamming distance, d_H, in amino acids across a set of generated sequences. Given a collection of M sequences $\{\hat{S}^1, \ldots, \hat{S}^M\}$, each of length N, diversity is defined as

$$\text{DIVERSITY}(\{\hat{S}^1, \ldots, \hat{S}^M\}) \;=\; \frac{2}{NM(M-1)} \sum_{j=1}^{M} \sum_{k=1}^{j-1} \sum_{i=1}^{N} \mathbf{1}\big[\hat{S}^j[i] \neq \hat{S}^k[i]\big],$$

where $\mathbf{1}[\cdot]$ is the indicator function that equals 1 if the amino acids differ at position i and 0 otherwise.

Hydrophobicity The grand average of hydropathy (GRAVY) score was used to quantify the overall hydrophobicity of designed sequences. For a sequence of length N, the GRAVY score is defined as

$$\text{GRAVY}(\hat{S}) \;=\; \frac{1}{N} \sum_{i=1}^{N} H(a_i),$$

where a_i denotes the amino acid at position i and $H(\cdot)$ is the hydropathy index according to the Kyte-Doolittle scale (Kyte and Doolittle, 1982).

Net charge The net charge of sequences at pH 7.0 was calculated using the SwissProt/ExPASy algorithm (Waterhouse et al., 2018), which assigns residue-specific pK_a values to ionizable groups. The total net charge is obtained as the sum over all groups:

$$Q(\text{pH}) \;=\; \sum_{x\in\text{basic}} \frac{+1}{1+10^{(\text{pH}-pK_a(x))}} \;-\; \sum_{y\in\text{acidic}} \frac{1}{1+10^{(pK_a(y)-\text{pH})}}.$$

Amino acid composition For each sequence, we computed the empirical amino-acid composition $p \in \Delta^{19}$ and compared it to a Swiss-Prot background composition $q \in \Delta^{19}$ using the Jensen–Shannon divergence (JSD). Let $m = \frac{1}{2}(p+q)$. The JSD is

$$\text{JSD}(p\|q) = \tfrac{1}{2}\,\text{KL}(p\|m) + \tfrac{1}{2}\,\text{KL}(q\|m),$$

with KL the Kullback–Leibler divergence. To ensure numerical stability, we applied a small additive smoothing ε (followed by renormalization) to both p and q.

A.3 Evaluation metrics - Binding specificity

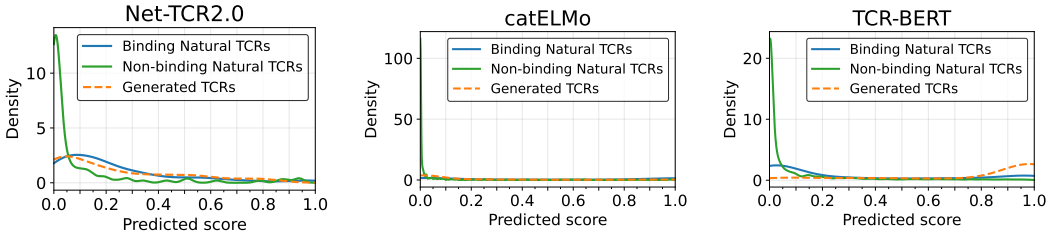

Figure 3: Predicted binding scores of generated TCRs compared to ground truth binders and non-binders, using from left to right: NetTCR 2.0 (Montemurro et al., 2021), catELMo (Zhang et al., 2023), and TCRBERT (Wu et al., 2024). The score distributions of generated TCRs closely align with those of true binders across the classifiers.

Binding specificity To evaluate the binding specificity of our generated sequences *in silico*, we employed three state-of-the-art predictors: NetTCR 2.0 (Montemurro et al., 2021), catELMo (Zhang

et al., 2023), and TCR-BERT (Wu et al., 2024). This evaluation task presents a methodological paradox: the state-of-the-art predictors required to validate our generated sequences themselves exhibit poor generalization to the novel binders we aim to create. This limitation stems from the same fundamental data sparsity that our generative model is designed to overcome.

Therefore, we first set out to establish a performance baseline. Since our generated sequences share approximately 60% of sequence identity with the training templates, we retrained the classifiers on a dataset split by the same 60% sequence identity. This ensures the classifier's test set mirrors the novelty of our generated sequences relative to its training data. We found that NetTCR2.0 (Montemurro et al., 2021) offered the highest recall, making it the most suitable for our benchmark. To mitigate over-reliance on the low recall rates of these classifiers on unseen sequences, we avoid treating their outputs as definitive class labels.

Instead, our primary metric is the Wasserstein distance. By comparing the distribution of predicted scores from our generated repertoire against the distributions of known binding and non-binding TCR sequences, we can quantify its resemblance to true binders without over-relying on individual predictions.

A.4 Rewards

We trained two classes of models: three specialised models and one generalist. The specialist policies employed a linear scalarisation approach. Here, objectives within a given modality were combined into a single, scalar reward using a fixed, weighted sum. The generalist was trained with preference conditioning: raw scores for all five objectives — pTM, pLDDT, hydrophobicity, charge, and composition — were first normalized to a consistent [0,1] scale relative to the property distributions observed in natural TCRs. The normalized scores then served as inputs for preference-conditioned policy training, allowing the model to approximate the Pareto front in settings where the properties are potentially in conflict.

Sequence To penalize high sequence quality values beyond a specified threshold, we defined a sharp, decreasing sigmoid function:

$$R(q) = \frac{1}{1 + \exp\big(\beta\,(q - c)\big)},$$

where q is the sequence quality score, c is the sigmoid center (here $c = 1.03$), and β controls the steepness (here $\beta = 40$). This formulation produces values close to 1.0 when $q < c$, but rapidly decays toward 0 as q exceeds the center.

Structure To combine the structural confidence metrics into a single scalar reward, we defined a piecewise function $R(p_{\text{TM}}, p_{\text{LDDT}})$ based on the pTM (p_{TM}) and pLDDT (p_{LDDT}) values from ESMFold (Lin et al., 2023). A weighted baseline score was computed as

$$b = \alpha\, p_{\text{TM}} + (1 - \alpha)\, p_{\text{LDDT}}, \qquad b_{\text{thr}} = \alpha\, p_{\text{TM}}^{\text{thr}} + (1 - \alpha)\, p_{\text{LDDT}}^{\text{thr}},$$

where $\alpha \in [0, 1]$ controls the relative contribution of the two metrics, and $p_{\text{TM}}^{\text{thr}}, p_{\text{LDDT}}^{\text{thr}}$ denote threshold values. These were defined as the lower standard deviation bounds from ESMFold predictions on 4,000 natural TCR sequences. To penalize low-confidence regions, we introduced a gating term

$$g = \min\left(\min\left(1, \frac{p_{\text{TM}}}{p_{\text{TM}}^{\text{thr}}}\right),\ \min\left(1, \frac{p_{\text{LDDT}}}{p_{\text{LDDT}}^{\text{thr}}}\right)\right).$$

The final reward was then defined piecewise as

$$R(p_{\text{TM}}, p_{\text{LDDT}}) = \begin{cases} 0.6\, g, & \text{if } p_{\text{TM}} < p_{\text{TM}}^{\text{thr}} \ \text{ or } \ p_{\text{LDDT}} < p_{\text{LDDT}}^{\text{thr}}, \\ 0.6 + 0.3\sqrt{\dfrac{b - b_{\text{thr}}}{\max(\varepsilon,\ \text{anchor} - b_{\text{thr}})}}, & \text{if } b \le \text{anchor}, \\ 0.9 + 0.1\sqrt{\dfrac{b - \text{anchor}}{\max(\varepsilon,\ 1 - \text{anchor})}}, & \text{if } b > \text{anchor}, \end{cases}$$

where anchor $\in (0, 1)$ is a reference point and the perturbation $\varepsilon \ll 1$ prevents division by zero. This formulation ensured smooth reward scaling: low-confidence predictions were down-weighted, intermediate-confidence predictions were rewarded proportionally to their proximity to the anchor, and high-confidence predictions asymptotically saturated toward 1.0.

Biophysical properties The biophysical reward was computed as a normalized weighted average of the individual component rewards (hydrophobicity, net charge, and amino-acid composition). Specifically, for a sequence s we defined

$$R_{\mathrm{bio}}(s) \;=\; \frac{w_{\mathrm{hyd}}\, R_{\mathrm{hyd}}(s) \;+\; w_{\mathrm{chg}}\, R_{\mathrm{chg}}(s) \;+\; w_{\mathrm{cmp}}\, R_{\mathrm{cmp}}(s)}{w_{\mathrm{hyd}} + w_{\mathrm{chg}} + w_{\mathrm{cmp}}},$$

where $w_{\mathrm{hyd}}, w_{\mathrm{chg}}, w_{\mathrm{cmp}}$ are nonnegative weights (set to 1.0 by default). Each component reward R_{hyd}, R_{chg}, and R_{cmp} is described in Appendix A.2.

To obtain the biophysical component rewards R_{hyd} and R_{chg}, we mapped the sequence hydrophobicity and net charge through Gaussian kernels. These kernels were centered at -0.5 and -1.5, respectively; these target centers were selected to align with the empirical biophysical distributions observed in natural TCR sequences. For R_{cmp}, we quantified the deviation of the generated sequence's distribution, p, from the natural background distribution, q. We converted the Jensen-Shannon Divergence (JSD) between these distributions into a similarity score via an exponential decay function.