ALIGNING CHEMICAL AND PROTEIN LANGUAGE MODELS WITH CONTINUOUS FEEDBACK USING ENERGY RANK ALIGNMENT

Shriram Chennakesavalu, Frank Hu, Sebastian Ibarraran & Grant M. Rotskoff *
Department of Chemistry
Stanford University
Stanford, CA 94305, USA
{shriramc,frankhu,sebastian.ibarraran,rotskoff}@stanford.edu

ABSTRACT

Large, autoregressive models trained on databases of chemical compounds and biomolecules have yielded powerful generators, but we still lack robust strategies for controlled generation. This molecular search problem closely resembles the "alignment" problem for large language models, though for many chemical tasks we have a specific and easily evaluable reward function. Here, we introduce an algorithm called energy rank alignment (ERA) that leverages an explicit reward function to produce a gradient-based objective that we use to optimize autoregressive policies. We deploy this approach to align molecular transformers and protein language models to generate molecules and protein sequences, respectively, with externally specified properties and find that it does so robustly, searching through diverse parts of chemical space. The algorithm is highly scalable, does not require reinforcement learning, and performs well relative to DPO when the number of preference observations per pairing is small.

1 INTRODUCTION

Foundation models strongly reflect the distribution of the data on which they are trained Ouyang et al. (2022), and controlling the outputs to reflect externally imposed preferences is an increasingly important challenge for deployment. The aforementioned task, often called "alignment", requires either careful curation of training data or large sets of human preference data—both options are labor-intensive Casper et al. (2023). Reinforcement learning from human feedback (RLHF), a family of algorithms that employs these human preference datasets, has been widely employed to align instruction and chat models Ouyang et al. (2022); Bai et al. (2022), but it is both expensive to acquire the training data and difficult to carry out in practice Casper et al. (2023). Recent algorithmic developments, such as direct preference optimization (DPO) Rafailov et al. (2023), simplify the alignment framework by making the reward function implicit, but still require human preference data. While these algorithms succeed in constraining outputs, many "alignment"-like tasks require evaluation that would be difficult for human experts, including applications to chemical and biomolecular design.

We formulate a generic alignment algorithm that we call *Energy Rank Alignment* (ERA) that leverages an explicit reward function to guide autoregressive sampling while targeting specific properties or preferences. Unlike reward maximization in RL-based algorithms, the policy that minimizes our objective is designed to sample fluctuations around a maximal reward value to promote sample diversity. Our algorithm enables direct gradient-based optimization of a policy to match the ideal preference distribution and converges asymptotically to an optimal distribution with tuneable entropy and controllable regularization, which we show theoretically. In numerical experiments, we demonstrate that this algorithm successfully aligns a molecule transformer model to identify a highly diverse set of chemicals with properties favored by our choice of reward. Finally, we demon-

*Correpsonding Author

strate that ERA is able to align a protein language model to generate mutated protein sequences with desirable properties according to a computational reward model.

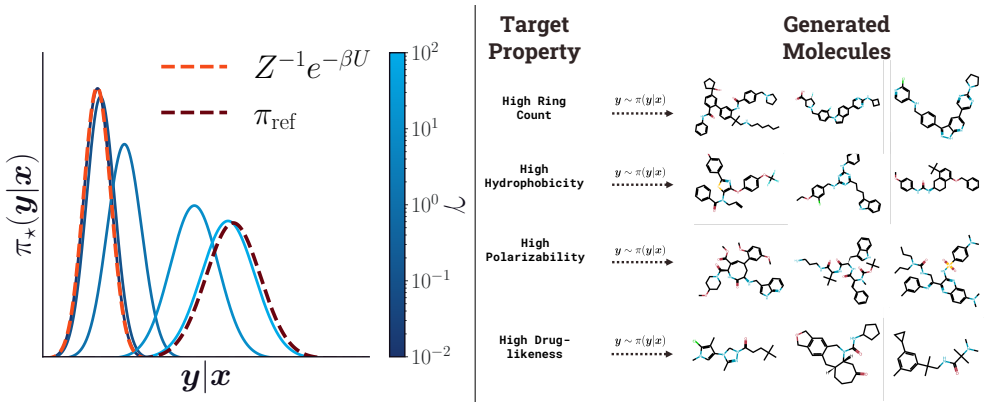

Figure 1: Energy rank alignment (ERA) enables targeting low-energy, high-reward regions with controllable fluctuations. Optimal policy approaches Boltzmann distribution with low regularization ($\gamma \to 0$) and reference policy with high regularization ($\gamma \to \infty$) (left). Aligned models can be used to sample molecules with desired chemical properties (right).

Related Work

- Applications of machine learning to inverse molecular design tasks. Existing approaches use weaker base models or rely on costly and tedious RL workflows Sanchez-Lengeling & Aspuru-Guzik (2018); Gromski et al. (2019); Gómez-Bombarelli et al. (2018); Zhou et al. (2023); Sanchez-Lengeling & Aspuru-Guzik (2018); Chithrananda et al. (2020); Schwaller et al. (2018); Wang et al. (2019); Bagal et al. (2022); Schwaller et al. (2019)
- A number of algorithms for LLM alignment are already in wide use. Our approach uniquely views the alignment procedure as a conditional sampling algorithm and has strong statistical guarantees Ouyang et al. (2022); Schulman et al. (2017); Rafailov et al. (2023).
- Our theoretical findings provide support to observations in the literature regarding existing LLM alignment algorithms, including Azar et al. (2023); Munos et al. (2023); An et al. (2023); Park et al. (2023).

2 Energy Rank Alignment

A policy is a conditional probability distribution $\pi(\cdot|\boldsymbol{x}) : \mathcal{Y} \to \mathbb{R}$; we generate an output $\boldsymbol{y}$ from prompt $\boldsymbol{x}$. The spaces $\mathcal{Y}$ and $\mathcal{X}$ are discrete and finite, corresponding to sequences of tokenized outputs of the model with a maximum length. In alignment tasks, we begin with a pre-trained reference policy π_{ref} and seek to optimize a parametric, trainable policy $\pi_{\boldsymbol{\theta}}$ to adapt the conditional sampling for a particular task or constraint.

Consider a prompt $\boldsymbol{x} \in \mathcal{X}$ and model outputs $\boldsymbol{y}, \boldsymbol{y}' \in \mathcal{Y}$ and a collection of preferences $\mathcal{D} = \{(\boldsymbol{y}_i \succ \boldsymbol{y}_i'; \boldsymbol{x}_i)\}_{i=1}^n$; the notation $\succ$ indicates that $\boldsymbol{y}_i$ is preferred to $\boldsymbol{y}_i'$. The conditional probability that $\boldsymbol{y} \succ \boldsymbol{y}'$ given $\boldsymbol{x}$ can be modeled as a pairwise Boltzmann ranking within the Bradley-Terry model, i.e.,

$$p(\boldsymbol{y} \succ \boldsymbol{y}'|\boldsymbol{x}) = \frac{e^{-\beta U(\boldsymbol{x},\boldsymbol{y})}}{e^{-\beta U(\boldsymbol{x},\boldsymbol{y})} + e^{-\beta U(\boldsymbol{x},\boldsymbol{y}')}} \equiv \sigma\big(\beta U(\boldsymbol{x},\boldsymbol{y}') - \beta U(\boldsymbol{x},\boldsymbol{y})\big). \tag{1}$$

Here $\beta > 0$ is a constant, $\sigma(x) = (1 + e^{-x})^{-1}$ and we refer to $U : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ as an energy function to make clear the connection to statistical physics, but it is the negative reward within the RL framework for alignment.

To impose the preferences we minimize the objective

$$J(\pi) = \mathbb{E}_{\boldsymbol{x}\sim\nu}\left[\int U(\boldsymbol{x},\boldsymbol{y})\mathrm{d}\pi(\boldsymbol{y}|\boldsymbol{x}) + \beta^{-1}\int (1+\gamma)\log\pi(\boldsymbol{y}|\boldsymbol{x}) - \gamma\log(\pi_{\mathrm{ref}}(\boldsymbol{y}|\boldsymbol{x}))\mathrm{d}\pi(\boldsymbol{y}|\boldsymbol{x})\right], \tag{2}$$

where β^{-1} is a parameter controlling the magnitude of the entropic term, γ sets the scale of the Kullback-Leibler regularization compared with the energy term, and ν is a probability distribution over the prompts $\nu \in \mathcal{P}(\mathcal{X})$. A proximal scheme for gradient descent on this objective corresponds to a gradient flow on J Santambrogio (2017); Maas (2011); the functional can be viewed as a free energy, and the corresponding flow is

$$\partial_t \pi_t = \nabla\cdot(\pi_t \nabla\delta_\pi J[\pi_t]), \tag{3}$$

and δ_π denotes the Fréchet derivative with respect to π. Assuming that π_0 has full support on $\mathcal{X}\times\mathcal{Y}$, the optimization converges asymptotically to a stationary policy which satisfies

$$\nabla\delta_\pi J[\pi_\star] = 0 \iff \pi_\star \propto e^{-\frac{\beta}{1+\gamma}U + \frac{\gamma}{\gamma+1}\log\pi_{\mathrm{ref}}}, \tag{4}$$

and this minimizer is globally optimal. In the context of LLM alignment, a representation of the energy function $U : \mathcal{X}\times\mathcal{Y}\to\mathbb{R}$ is learned as a "reward model", though we also consider tasks in which U is an easily evaluated function of the pair $(\boldsymbol{x},\boldsymbol{y})$. The optimal distribution $\pi_\star$ is a Gibbs-Boltzmann measure

$$\pi_\star(\boldsymbol{y}|\boldsymbol{x}) = Z^{-1}(\boldsymbol{x})\exp\left[-\frac{\beta}{1+\gamma}\big(U(\boldsymbol{x},\boldsymbol{y}) - \beta^{-1}\gamma\log\pi_{\mathrm{ref}}(\boldsymbol{y}|\boldsymbol{x})\big)\right] \tag{5}$$

where $Z(\boldsymbol{x})$ is the $\boldsymbol{x}$-dependent normalization constant. This expression makes clear the effect of β: when $\beta\to\infty$ (low temperature), the reward dominates and fluctuations around the maximal reward are small, which could lead to "mode-seeking"; when $\beta\to 0$ (high physical temperature) fluctuations around the maximal reward increase and the regularization term favors proximity to π_{ref}. Similarly, $\gamma\to 0$ recovers a Gibbs-Boltzmann distribution proportional to $e^{-\beta U}$ at inverse temperature β, while $\gamma\to\infty$ is dominated by the reference policy.

3 Experiments

	GSK3β top-100		JNK3 top-100	
	mean score	IntDiv	mean score	IntDiv
ERA	0.996 ± 0.000	**0.219 ± 0.002**	**0.987 ± 0.001**	**0.264 ± 0.005**
MolRL-MGPT	**1.000 ± 0.000**	0.362 ± 0.015	0.961 ± 0.010	0.372 ± 0.025
GFlowNet	0.649 ± 0.072	0.715 ± 0.104	0.437 ± 0.219	0.716 ± 0.145
GraphGA	0.919 ± 0.016	0.365 ± 0.024	0.875 ± 0.025	0.380 ± 0.015
JT-VAE	0.235 ± 0.083	0.770 ± 0.067	0.159 ± 0.040	0.781 ± 0.127
REINVENT	0.965 ± 0.011	0.308 ± 0.035	0.942 ± 0.019	0.368 ± 0.021

Table 1: Mean scores and internal diversities (IntDiv) of experiments on GSK3β and JNK3 tasks averaged across 5 random seeds. For each task, 20K molecules were sampled, and metrics were computed on top-100 scoring *valid, novel and unique* molecules filtered from the initial 20K samples (i.e. molecules not in dataset and molecules not previously sampled). Compared to state-of-the-art methods, ERA samples more diverse molecules with higher predicted docking scores. Results for compared methods are reproduced from Hu et al. (2023).

Unprompted molecular alignment on protein-ligand docking oracles We investigate the performance of ERA in designing compounds that have high predicted docking scores for the kinases JNK3 and GSK3β. For each of these targets, we use an *in silico* oracle that predicts docking scores, ranging from 0 to 1, where a higher value corresponds to stronger predicted score Sun et al. (2017). Using only data from ChemBL, we first carry out a short supervised fine-tuning step on all molecules in ChemBL with an oracle score above 0.5 (7386 molecules for JNK3 and 43381 for GSK3β). Using this fine-tuned model as our reference policy, we then carry out alignnment using ERA (β=100 and γ=0), where we use a comparably high β to target molecules with high activity. We define the energy for this task as the negative logarithm of the oracle score.

From the aligned models, we sample 20000 molecules (see Fig. 5) and tabulate metrics of the top-100 performing molecules (see Table 1). We note that the molecules in the top-100 are filtered to

exclude any molecules that are present in the ChemBL dataset and any repeated molecules. As such, the top-100 selected molecules are both *novel* and *unique*. For GSK3β, our mean score is marginally lower than the best performing method but the diversity in sampled molecules is significantly higher (i.e. lower IntDiv). For JNK3 our mean score is significantly higher than the best performing method *and* the diversity in sampled molecules is higher than any method. The inference costs are notably low for our approach; sampling 20000 molecules and filtering takes only minutes on a single GPU.

We additionally measure sample efficiency using the top-10 AUC metric Gao et al. (2022), which is the area under the curve (AUC) of the mean property value of the top-10 performing molecules versus the number of oracle calls (see Fig. 6 and Table 2). We likewise only include novel, unique, and valid molecules in this analysis; any sampled molecule that is in ChemBL, that has already been sampled, or that is invalid is discarded and additionally does not count towards an oracle call as these are filtered out before oracle evaluation. We observe that we are able to generate novel and unique high-scoring molecules, with high sample efficiency especially in comparison to existing state-of-the-art methods. Ultimately, high sample efficiency is crucial in settings where evaluation is expensive, which will generally be true for most real-world chemical and biological tasks (e.g. wet-lab experiment). Finally, we also perform Glide Standard Precision Friesner et al. (2004) docking on the top-scoring molecules according to the oracles (score of 1.0) against their respective receptors. We observe that the diverse set of sampled molecules exhibit chemically plausable docked poses obtained from a physics-based docking approach (Fig. 2).

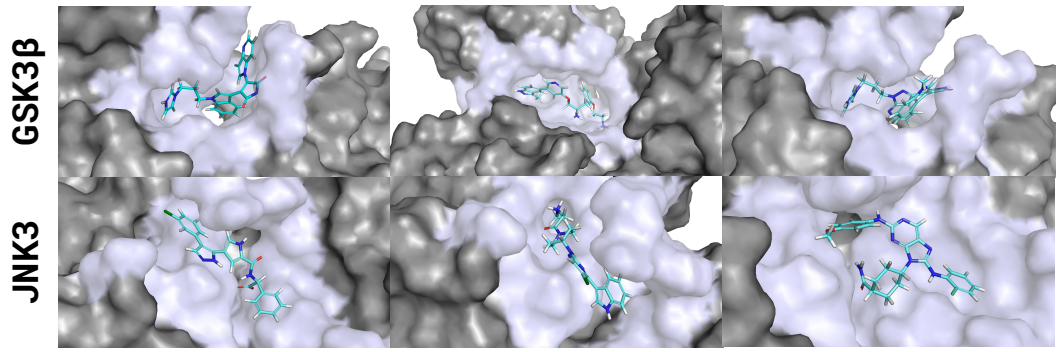

Figure 2: Visualization of three generated ligands docked against the GSK3β kinase target (top) and three generated ligands docked against the JNK3 kinase target (bottom). In each case, these were the three molecules with the best (most negative) Glide Standard Precision docking scores and oracle scores of 1.0.

Prompted multi-property molecular alignment on RDKit oracles Inspired by the task of lead optimization in drug discovery efforts Keserü & Makara (2009), we ask whether we can use ERA to train a molecular generator that can sample a molecule that is both similar to the prompt molecule *and* also exhibits some desired property. First, we fine-tune the pretrained molecular generator to enable prompted molecular generation and use this fine-tuned model as our reference policy for all prompted molecular alignment tasks. This reference policy disproportionately samples molecules that are identical (i.e. a Tanimoto similarity of 1.0) to the prompt molecule (see Fig. 3), so we carry out multi-property alignment on this reference policy to generate molecules that are similar—but not identical—to the prompt molecule and also have a high drug-likeness as measured by the quantitative estimate of drug-likeness (QED). Using ERA, we optimize the reference policy with a generated dataset $\mathcal{D} = \{(\boldsymbol{y}_1^{(i)}, \boldsymbol{x}^{(i)}), (\boldsymbol{y}_2^{(i)}, \boldsymbol{x}^{(i)}), U(\boldsymbol{y}_1^{(i)}, \boldsymbol{x}^{(i)}), U(\boldsymbol{y}_2^{(i)}, \boldsymbol{x}^{(i)})\}_{i=1}^{N}$, where we sample four molecules for each prompt molecule from the reference policy and consider all possible preference pairs for a total of six preference pairs per prompt molecule.

We observe that the per-prompt average QED under the optimized policy for a given prompt is higher than the corresponding average under the reference policy (Fig. 3). Furthermore, we see that we are able to sample a diverse set of molecules that are chemically similar to the prompt molecule, and also chemically valid. We repeat the experiment with a related objective of generating molecules similar to the prompt molecule with a high Wildman-Crippen LogP (hydrophobicity) instead and again observe that we increase the per-prompt average LogP under the optimized policy relative to the

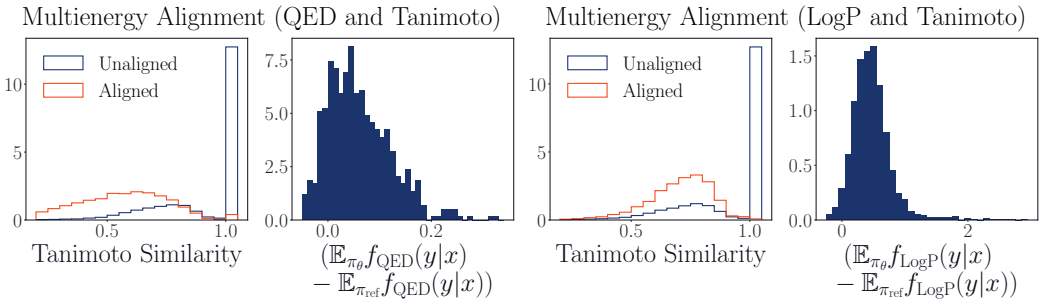

Figure 3: Prompted multi-property molecular generator alignment. From left to right: Tanimoto similarities computed between the prompt and sampled molecules for both aligned and unaligned policies (QED and Tanimoto alignment), per-prompt difference in the average QED under aligned and unaligned policies (QED and Tanimoto alignment), Tanimoto similarities computed between the prompt and sampled molecules for both aligned and unaligned policies (LogP and Tanimoto alignment), and per-prompt difference in the average LogP under aligned and unaligned policies (LogP and Tanimoto alignment). With alignment, we target higher QED and LogP values, while still sampling molecules chemically similar—but not identical to—the prompt molecule.

reference policy without degrading sample diversity and validity. For both of these experiments, we required regularization to the reference policy ($\gamma > 0$). With no regularization, the aligned generator would almost exclusively sample sequences that were chemically invalid ($< 25\%$ chemical validity).

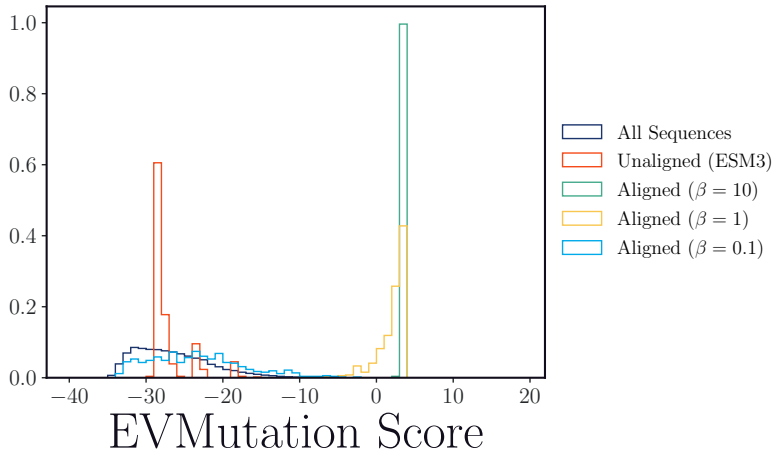

Figure 4: Alignment of ESM3-1.4B with β=0, 0.1, 1.0, 10.0 and γ=0.001 on the task of maximizing EVmutation score. Positions 182, 183, 184, and 186 of the TrpB parent sequence were masked and ESM3-1.4B predicted amino acids at those sites. The distribution of the EVmutation scores for generated sequences shifts significantly as β is increased.

Directed evolution of proteins with ERA We also consider the performance of ERA in a large-molecule setting, namely ML-guided directed evolution of proteins. Directed evolution campaigns aim to optimize a protein sequence toward some desired property of interest via iterative mutagenesis, library screening, and selection of best variants Wang et al. (2021). This has become a widely used methodology in protein engineering but comes with key limitations. The inherently iterative nature of directed evolution campaigns can lead to costly and time-consuming experimental campaigns, and meaningfully understanding the effects of protein mutations on protein activity can often be difficult. These challenges have led to the application of machine learning methods to more efficiently guide directed evolution campaigns Yang et al. (2024) Shanker et al. (2024).

There has been significant recent effort to design and train large protein language models (PLMs) Lin et al. (2023); Hayes et al. (2024). Furthermore, these models have demonstrated remarkable

capabilities across a number of protein tasks Widatalla et al. (2024); Shanker et al. (2024). As such, we decided to use the state-of-the-art ESM3-1.4B Hayes et al. (2024) as our pretrained model, for which we carried out alignment using ERA. Despite the multimodal nature of ESM3, here, we only focus on generating primary-sequence-based representations of proteins.

We consider directed evolution of the β-subunit of tryptophansynthase (TrpB) from *Thermotoga maritima*, an enzyme that catalyzes tryptophan production Buller et al. (2015). Here, we seek to evolve the protein to increase its evolutionary fitness. In this work, we do not have access to experimental validation and so we evaluate the fitness of sequences using the computationally evaluable EVmutation score, an oracle that is predictive of a variant sequence's performance relative to the parent sequence in its native function Hopf et al. (2017).

As in other directed evolution campaigns for the TrpB protein Yang et al. (2023), we consider mutating four different sites to one of the 20 standard amino acids. We randomly sampled 512 mutated sequences, emulating a random mutagenesis experiment. Using ESM3-1.4B as our reference model, we carry out alignment using ERA with various $\beta = (0.1, 1.0, 10.0)$ and $\gamma = 0.001$ and plot the results in Fig. 4. We observe that with higher β, we are able to sample mutants with the highest possible EVmutation score in a single round of alignment. These results are promising for the application of ERA in directed evolution campaigns and future work will focus on the guidance of wet-lab directed evolution campaigns in conjunction with multi-round, on-policy ERA.

REFERENCES

Gaon An, Junhyeok Lee, Xingdong Zuo, Norio Kosaka, Kyung-Min Kim, and Hyun Oh Song. Direct Preference-based Policy Optimization without Reward Modeling. *Advances in Neural Information Processing Systems*, 36:70247–70266, December 2023.

Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. A General Theoretical Paradigm to Understand Learning from Human Preferences, November 2023.

Viraj Bagal, Rishal Aggarwal, P. K. Vinod, and U. Deva Priyakumar. MolGPT: Molecular Generation Using a Transformer-Decoder Model. *Journal of Chemical Information and Modeling*, 62 (9):2064–2076, May 2022. ISSN 1549-9596. doi: 10.1021/acs.jcim.1c00600.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback, April 2022.

Andrew R. Buller, Sabine Brinkmann-Chen, David K. Romney, Michael Herger, Javier Murciano-Calles, and Frances H. Arnold. Directed evolution of the tryptophan synthase β-subunit for stand-alone function recapitulates allosteric activation. *Proceedings of the National Academy of Sciences*, 112(47):14599–14604, 2015. doi: 10.1073/pnas.1516401112. URL https://www.pnas.org/doi/full/10.1073/pnas.1516401112.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Wang, Samuel Marks, Charbel-Raphaël Segerie, Micah Carroll, Andi Peng, Phillip Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J. Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Bıyık, Anca Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. Open problems and fundamental limitations of reinforcement learning from human feedback, September 2023.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction. In *Machine Learning for Molecules Workshop at NeurIPS*, 2020.

Richard A. Friesner, Jay L. Banks, Robert B. Murphy, Thomas A. Halgren, Jasna J. Klicic, Daniel T. Mainz, Matthew P. Repasky, Eric H. Knoll, Mee Shelley, Jason K. Perry, David E. Shaw, Perry Francis, and Peter S. Shenkin. Glide: A new approach for rapid, accurate docking and scoring. 1. method and assessment of docking accuracy. *Journal of Medicinal Chemistry*, 47(7), 2004. doi: https://doi.org/10.1021/jm0306430. URL https://pubs.acs.org/doi/10.1021/jm0306430.

Wenhao Gao, Tianfan Fu, and Jimeng Sun andConnor W. Coley. Sample efficiency matters: A benchmark for practical molecular optimization. In *Advances in Neural Information Processing Systems*, volume 36. Curran Associates, Inc., 2022.

Piotr S. Gromski, Alon B. Henson, Jarosław M. Granda, and Leroy Cronin. How to explore chemical space using algorithms and automation. *Nature Reviews Chemistry*, 3(2):119–128, 2019.

Rafael Gómez-Bombarelli, Jennifer N. Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D. Hirzel, Ryan P. Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Science*, 4(2):268–276, February 2018. ISSN 2374-7943. doi: 10.1021/acscentsci.7b00572.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J. Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q. Tran, Jonathan Deaton, Marius Wiggert, Rohil Badkundri, Irhum Shafkat, Jun Gong, Alexander Derry, Raul S. Molina, Neil Thomas, Yousuf Khan, Chetan Mishra, Carolyn Kim, Liam J. Bartie, Matthew Nemeth, Patrick D. Hsu, Tom Sercu, Salvatore Candido, and Alexander Rives. Simulating 500 million years of evolution with a language model, July 2024.

Thomas A Hopf, John B Ingraham, Frank J Poelwijk, Charlotta P I Schärfe, Michael Springer, Chris Sander, and Debora S Marks. Mutation effects predicted from sequence co-variation. *Nature Biotechnology*, 35(2):128–135, 2017. doi: 10.1038/nbt.3769. URL https://www.nature.com/articles/nbt.3769.

Xiuyuan Hu, Guoqing Liu, Yang Zhao, and Hao Zhang. De novo drug design using reinforcement learning with dynamic vocabulary. In *Advances in Neural Information Processing Systems*, volume 37. Curran Associates, Inc., 2023.

György M. Keserü and Gergely M. Makara. The influence of lead discovery strategies on the properties of drug candidates. *Nature Reviews Drug Discovery*, 8(3):203–212, March 2009. ISSN 1474-1776, 1474-1784. doi: 10.1038/nrd2796.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan Dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, March 2023. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.ade2574. URL https://www.science.org/doi/10.1126/science.ade2574.

Jan Maas. Gradient flows of the entropy for finite Markov chains. *Journal of Functional Analysis*, 261(8):2250–2292, October 2011. ISSN 0022-1236. doi: 10.1016/j.jfa.2011.06.009.

Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Andrea Michi, Marco Selvi, Sertan Girgin, Nikola Momchev, Olivier Bachem, Daniel J. Mankowitz, Doina Precup, and Bilal Piot. Nash Learning from Human Feedback, December 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744. Curran Associates, Inc., 2022.

Ryan Park, Ryan Theisen, Navriti Sahni, Marcel Patek, Anna Cichońska, and Rayees Rahman. Preference Optimization for Molecular Language Models, October 2023.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741. Curran Associates, Inc., 2023.

Benjamin Sanchez-Lengeling and Alán Aspuru-Guzik. Inverse molecular design using machine learning: Generative models for matter engineering. *Science*, 361(6400):360–365, July 2018. doi: 10.1126/science.aat2663.

Filippo Santambrogio. {Euclidean, Metric, and Wasserstein} gradient flows: An overview. *Bulletin of Mathematical Sciences*, 7(1):87–154, April 2017. ISSN 1664-3615. doi: 10.1007/s13373-017-0101-1.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal Policy Optimization Algorithms, August 2017.

Philippe Schwaller, Théophile Gaudin, Dávid Lányi, Costas Bekas, and Teodoro Laino. "Found in Translation": Predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. *Chemical Science*, 9(28):6091–6098, 2018. doi: 10.1039/C8SC02339E.

Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Christopher A. Hunter, Costas Bekas, and Alpha A. Lee. Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction. *ACS Central Science*, 5(9):1572–1583, September 2019. ISSN 2374-7943, 2374-7951. doi: 10.1021/acscentsci.9b00576.

Varun R. Shanker, Theodora U. J. Bruun, Brian L. Hie, and Peter S. Kim. Unsupervised evolution of protein and antibody complexes with a structure-informed language model. *Science*, 385(6704):46–53, 2024. doi: 10.1126/science.adk8946. URL https://www.science.org/doi/full/10.1126/science.adk8946.

Jiangming Sun, Nina Jeliazkov, Vladimir Chupakhin, Jose-Felipe Golib-Dzib, Ola Engkvist, Lars Carlsson, Jörg Wegner, Hugo Ceulemans, Ivan Georgiev, Vedrin Jeliazkov, Nikolay Kochev, Thomas J. Ashby, and Hongming Chen. Excape-db: an integrated large scale dataset facilitating big data analysis in chemogenomics. *Journal of Cheminformatics*, 9(17), 2017. doi: https://doi.org/10.1186/s13321-017-0203-5. URL https://jcheminf.biomedcentral.com/articles/10.1186/s13321-017-0203-5#citeas.

Sheng Wang, Yuzhi Guo, Yuhong Wang, Hongmao Sun, and Junzhou Huang. SMILES-BERT: Large Scale Unsupervised Pre-Training for Molecular Property Prediction. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, BCB '19, pp. 429–436, New York, NY, USA, September 2019. Association for Computing Machinery. ISBN 978-1-4503-6666-3. doi: 10.1145/3307339.3342186.

Yajie Wang, Pu Xue, Mingfeng Cao, Tianhao Yu, Stephan T. Lane, and Huimin Zhao. Directed evolution: Methodologies and applications. *Chemical Reviews*, 121(20):12384–12444, 2021. doi: 10.1021/acs.chemrev.1c00260. URL https://pubs.acs.org/doi/full/10.1021/acs.chemrev.1c00260.

Talal Widatalla, Rafael Rafailov, and Brian Hie. Aligning protein generative models with experimental fitness via direct preference optimization, May 2024.

Jason Yang, Julie Ducharme, Kadina E. Johnston, Francesca-Zhoufan Li, Yisong Yue, and Frances H. Arnold. Decoil: Optimization of degenerate codon libraries for machine learning-assisted protein engineering. *ACS Synthetic Biology*, 12(8):2444–2454, 2023. doi: 10.1021/acssynbio.3c00301. URL https://pubs.acs.org/doi/full/10.1021/acssynbio.3c00301.

Jason Yang, Ravi G. Lal, James C. Bowden, Raul Astudillo, Mikhail A. Hameedi, Sukhvinder Kaur, Matthew Hill, Yisong Yue, and Frances H. Arnold. Active learning-assisted directed evolution, July 2024.

Zhanhui Zhou, Jie Liu, Chao Yang, Jing Shao, Yu Liu, Xiangyu Yue, Wanli Ouyang, and Yu Qiao. Beyond One-Preference-Fits-All Alignment: Multi-Objective Direct Preference Optimization, December 2023.

A APPENDIX

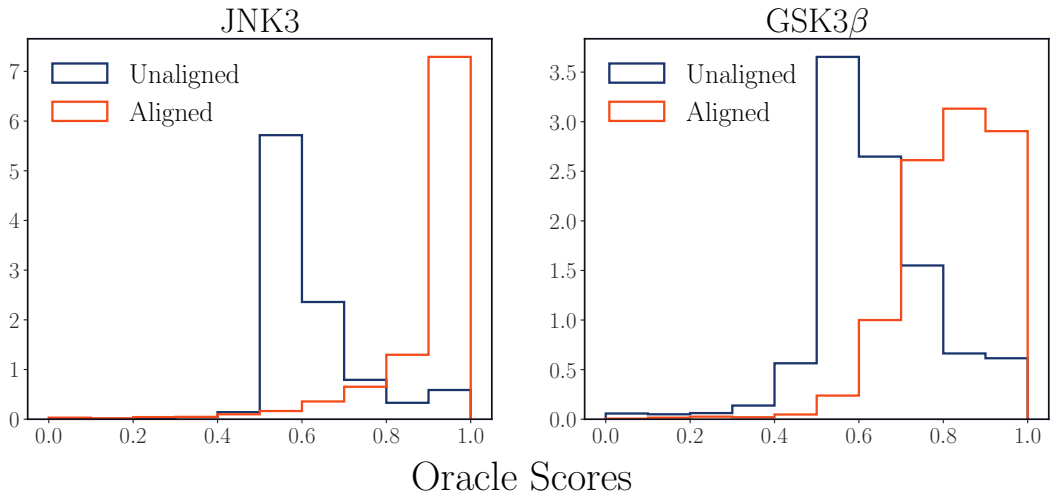

Figure 5: Distribution of GSK3β and JNK3 oracle scores sampled from unaligned reference model and aligned model (β = 100.0, γ = 0.0). 20K molecules were sampled from each model and only oracle scores of valid molecules are plotted.

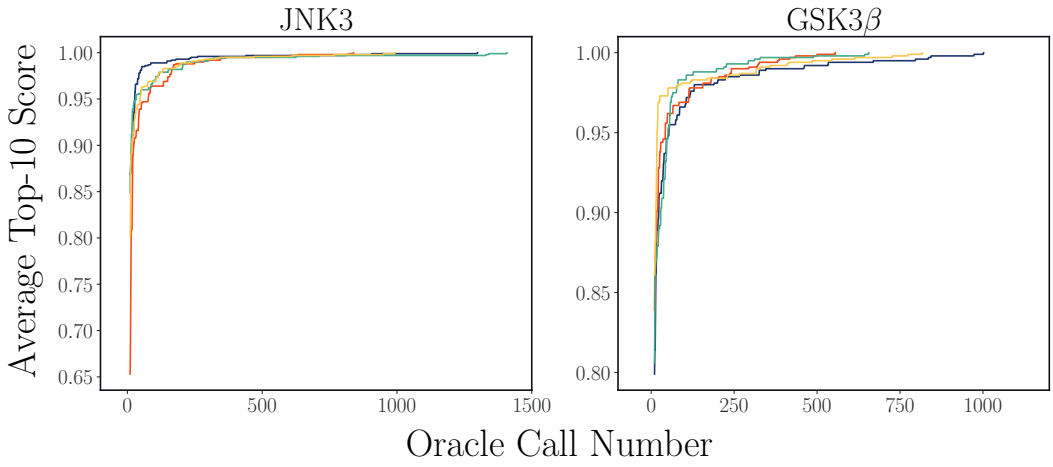

Figure 6: The average score of top-10 performing *valid, novel*, and *unique* molecules as a function of the number of oracle calls made to the aligned models. Scores are computed using the JNK3 and GSK3β oracles, respectively, for five different random seeds. Samples that are invalid, present in the dataset, or already previously sampled are discarded and do not count towards an oracle call.

	GSK3β top-10 AUC	JNK3 top-10 AUC
ERA	**0.985 ± 0.001**	**0.989 ± 0.002**
REINVENT	0.865 ± 0.043	0.783 ± 0.023
GraphGA	0.788 ± 0.070	0.553 ± 0.136

Table 2: Top-10 AUC scores on GSK3β and JNK3 tasks averaged across 5 random seeds. Compared to state-of-the-art methods as reported in Gao et al. (2022), ERA has higher sampling efficiency. Results for compared methods are reproduced from Gao et al. (2022).