Improving few-shot learning-based protein engineering with evolutionary sampling

M. Zaki Jawaid
EpiCRISPR Biotechnologies
zaki.jawaid@epic-bio.com

Aayushma Gautam
EpiCRISPR Biotechnologies
aayushma.gautam@epic-bio.com

T. Blair Gainous
EpiCRISPR Biotechnologies
blair.gainous@epic-bio.com

Daniel O. Hart
EpiCRISPR Biotechnologies
dan.hart@epic-bio.com

Robin W. Yeo[†]
EpiCRISPR Biotechnologies
robin.yeo@epic-bio.com

Timothy P. Daley [*] [†]
timy.pdaley@gmail.com

Abstract

Designing novel functional proteins remains a slow and expensive process due to a variety of protein engineering challenges; in particular, the number of protein variants that can be experimentally tested in a given assay pales in comparison to the vastness of the overall sequence space, resulting in low hit rates and expensive wet lab testing cycles. ML-guided protein engineering promises to accelerate this process through computational screening of proposed variants *in silico*. However, exploring the prohibitively large protein sequence space presents a significant challenge for the design of novel functional proteins using ML-guided protein engineering. Here, we propose using evolutionary Monte Carlo search (EMCS) to efficiently explore the fitness landscape and accelerate novel protein design. As a proof-of-concept, we use our approach to design a library of peptides predicted to be functionally capable of transcriptional activation and then experimentally screen them, resulting in a dramatically improved hit rate compared to existing methods. Our method can be easily adapted to other protein engineering and design problems, particularly where the cost associated with obtaining labeled data is significantly high. We have provided open source code for our method at https://github.com/SuperSecretBioTech/evolutionary_monte_carlo_search.

1 Introduction

The design and optimization of proteins with specific functionality is a long-sought pursuit in protein engineering. Since proteins are composed of sequences of amino acids which ultimately dictate their structure and function, the protein engineering problem can be reformulated as finding the optimal mapping from amino acid sequence s of length L to biological function $f : s \to f(s)$, where we call f the fitness function. Finding the optimum of f can be seen as a high-dimensional discrete combinatorial optimization problem [1]. The enormous size of the protein sequence space, together with the presence of sensitive and sporadic high fitness regions in the fitness landscape [2], makes novel protein design extremely challenging.

[*] Formerly affiliated with EpiCrispr Biotechnologies
[†] Co-corresponding Author

NeurIPS 2023 Generative AI and Biology Workshop.

The traditional experimental approach involves high-throughput, iterative laboratory methods such as directed evolution [3, 4], deep mutational scans [5], and semi-rational design [6]. However, these methods typically require multiple rounds of engineering and analysis, making them tedious, expensive, and time-consuming [7]. Furthermore, the number of variants capable of being tested in even the most advanced laboratories ($\approx 10^5$ to 10^6) is miniscule in comparison to the size of the total sequence space; additionally, high-throughput screening can be challenging to implement for some classes of proteins [8].

In the past decade, the application of machine learning methods to protein engineering problems has been massively successful [9]. In this context, machine learning models are trained to learn the sequence-to-function map and then used to propose new sequences that maximize the predicted fitness.

In recent years, methods such as generative models have been proposed to tackle this problem, including deep generative networks [10–15], generative adversarial networks [16, 17] and diffusion models [18–20]. In these cases the exploration problem is trivial, as the model produces an embedding in a low-dimensional space where sampling is computationally inexpensive. However, generative approaches typically require huge amounts of training data and a large number of positive examples to ensure that the model embeddings are meaningful and so that they do not simply memorize positive examples, an issue that has been widely observed to happen in image GANs [21, 22]. Given the relatively small number of sequences in our training data (≈ 3.4 x 10^4) and the extreme paucity of positive examples (173), we anticipated our small and skewed training data would would prove insufficient for a generative modeling approach. On the other hand, transfer learning of large protein language models (LPLMs) has shown success in modeling and designing novel proteins with fitness functions trained on small numbers of positive hits [8, 23–25].

Model-guided fitness landscape exploration remains an understudied problem in the context of protein engineering [26, 27]. Novel methods such as importance-weighted expectation maximization [15] and GFlowNets [27] have been proposed to tackle this problem, however, the algorithm of choice for exploring a machine learning based fitness landscape is Markov Chain Monte Carlo (MCMC) sampling [8]. In this work, we focus on improving MCMC for fitness landscape exploration.

Evolutionary Monte Carlo (EMC) [28, 29] is an advanced sampling method that draws inspiration from genetic recombination as well as physics-based MCMC techniques. While EMC has previously been used for a variety of sampling tasks [28, 30–33], its potential as an exploratory algorithm for protein design remains unexplored. In this paper, we modify EMC as a search tool for exploring the complex fitness landscape of protein sequences capable of gene regulation, which we call EMC Search (EMCS). EMCS is much less computationally intensive than gradient-based approaches and Gibbs sampling. We further expect it to benefit from faster convergence and to provide a more comprehensive and efficient exploration of the fitness landscape by allowing for interpolation at the molecular level between chains. We think one of the primary strengths of EMCS combined with a LPLM-based fitness function is the ability of the LPLM to implicitly identify biological domains critical to protein function and for EMCS to interpolate between molecules and combine domains from distinct chains to form higher fitness proteins (Fig 1).

In this study we propose a design strategy for generating novel protein sequences using a few-shot transfer learning-based approach. We then experimentally validate our proposed method in the lab by screening a library of proposed sequences for their transcriptional activation ability, demonstrating that our approach is capable of improving discovery rates. Though here we have applied our method to the design of small gene activator proteins, we anticipate that our method will be generally applicable to a diverse range of problems in the field of protein engineering.

2 Model and Search

Training Transfer Learning-Based Fitness Models We performed and independently validated a high-throughput screen in which 85 amino acid (85aa) peptides were experimentally screened for their ability to activate a synthetic genetic locus using a nuclease-inactivated Cas platform for transcriptional activation [34].

Using this screening platform, we identified 173 gene activators ("positive hits") from a training set of 34217 protein sequences (0.51% hit rate). Using these data, we sought to train a machine learning

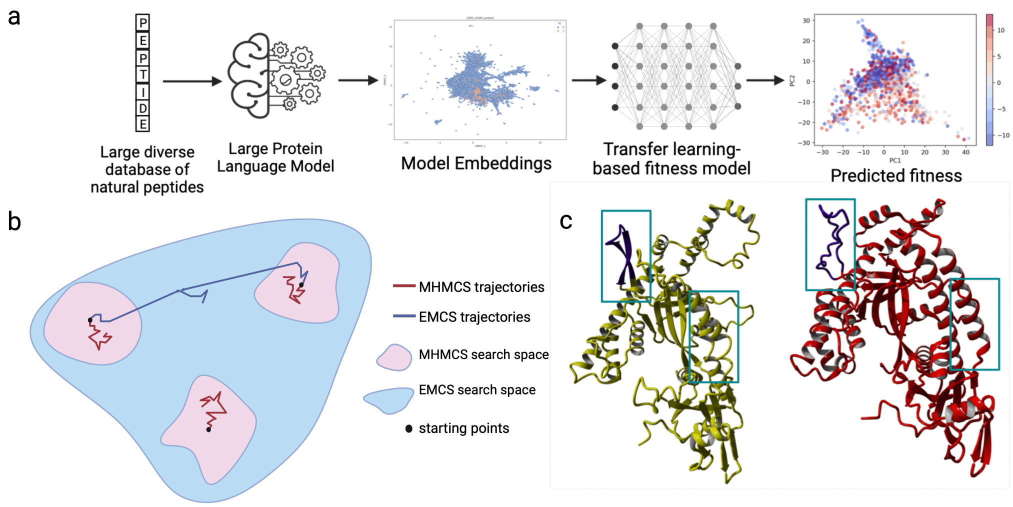

Figure 1: **a.** Transfer learning approach for predicting gene activators from sequence using LPLM embeddings. **b.** Fitness landscape diagram representing the effective search spaces of MHMCS (pink) and EMCS (blue). When initialized at positive hits, MHMCS is constrained to locally search near the starting molecule's other high fitness sequences, while EMCS interpolates between multiple starting molecules with varying resolution to optimize the search and escape deep local optima. **c.** EMCS can evaluate multiple protein sequences simultaneously while allowing for favorable protein domains to aggregate. In this example, an exchange among the boxed regions between the proteins can potentially be evaluated in one iteration, while the same qualitative evaluation in MHMCS would take multiple iterations.

model capable of predicting proteins capable of gene activation from sequence alone. An in-depth discussion of the screening platform together with additional information about the final training dataset is provided in our prior work [34].

We compared OneHot encoding with transfer learning using a 650 million parameter LPLM (ESM-2 model) [35] as input features for two models: an XGBoost model, where we flatten the features by taking the mean, and a CNN model. As a consequence of mean-featurization, the XGBoost model learned protein sequences' global features whereas we anticipated that the CNN model would be capable of identifying local features. For both XGBoost and CNN models, we found that transfer learning significantly improved prediction when compared to OneHot encoding (Table S1,S2).

Metropolis-Hasting Monte Carlo Search (MHMCS) The MHMCS algorithm operates by proposing a low number of mutations to modify the current molecule and then evaluating the new molecule's fitness; if fitness improves, the proposal is accepted, while, if fitness decreases, the proposal is accepted with probability weighted by the ratio of the proposed fitness to the current fitness. Algorithm S1 details our implementation of MHMCS with our choice of default parameters in Table S3.

Evolutionary Monte Carlo Search (EMCS) Evolutionary Monte Carlo Search (EMCS) extends traditional Metropolis-Hastings Monte Carlo Search (MHMCS) by introducing genetic crossover events in a parallel tempering setup [29, 36]. In parallel tempering, multiple MHMCS chains are run simultaneously at different temperatures and are swapped at two randomly chosen temperatures after a predetermined number of iterations. The primary advantage of parallel tempering is that it allows MHMCS to occur over a larger search radius without sacrificing resolution. EMCS builds upon parallel tempering by adding genetic crossover events (domain swapping through chain interpolation). This allows for an even larger search radius (Fig. 1), while also adding the possibility of aggregation of favorable protein domains, which we hypothesize is critical to exploit the small number of positive hits in our training data. Algorithm S2 details our implementation of EMCS with our choice of default parameters in Table S3. We provide a more in-depth discussion of the EMCS algorithm in the supplementary material.

Search Method	Initialization	# Sequences	# Positive Hits	Hit Percentage
EMCS	known	410	94	22.9%
EMCS	random	390	39	10%
MHMCS	random	200	2	1 %
HTS	Biological Origins	34217	173	0.51%

Table 1: Positive hit results for the ensemble model. Initialization column denotes the sampling algorithm's starting sequence as either randomly initialized ("Random"), or known positive hit ("Known"). HTS: High throughput screening.

3 Results

The protein fitness landscape is known to be highly sensitive, multi-peaked, and rugged [1, 2], reflecting the possibility that a complete loss of function can arise due to a relatively small number of point mutations (e.g. mutations in catalytic domains, mutations that cause misfolding, ...). The complexity of this space presents obvious challenges for efficient exploration. Here, as a proof-of-concept, we compare how EMCS and MHMCS respectively explore the discrete fitness landscape of 85aa proteins capable of gene activation, and evaluate prediction success rates, sequence diversity, and convergence speeds.

Experimental Screening For experimental validation, we used EMCS and MHMCS to design novel proteins using all three of our models (XGBoost, CNN, ensemble) using parameters defined in Table S3. Together, we used EMCS and MHMCS to design 4600 novel sequences (Table S4) that are largely distinct from the sequence space occupied by the original training data, confirming that both model-guided sampling techniques are capable of proposing diverse novel proteins (Fig. S3). We then experimentally assayed the peptides for their ability to activate a genetic locus (full details of experimental design can be found in the supplementary material, Fig. S1-S2). In total, we identified 357 positive hits (7.59% hit rate), peptide sequences capable of activating a synthetic gene reporter significantly over background fluorescence. In contrast, the initial screen had a hit rate of only 0.51%. If we use the latter number as a proxy for the fraction of naturally occurring 85aa peptide sequences that are capable of gene activation, then our approach increased the baseline hit rate by ≈ 15-fold. In fact, the best model-guided sampling technique (ensemble model + EMCS from known hits), increased the hit rate ≈ 45-fold (Table 1, Table S5, S6) by this metric.

Sequence Diversity To compare sequence proposals between EMCS and MHMCS, we performed an *in silico* sampling experiment where we explore the fitness landscape a minimum of 1000 times with each algorithm using identical and controlled initial conditions, including ablation studies. A unique advantage of EMCS is its ability to identify novel high fitness sequences even when initialized from sequences that were known positive hits (and thus already in a high fitness neighborhood). When initialized from known positive hits, the final edit distances of sequences discovered by EMCS are 1.5 - 3x higher compared to those discovered by MHMCS using a similar temperature regime (Table S8-S9, Fig. S5). Consistently, using entropy as a measure of information change, we in-silico experiments show that the average entropy change per iteration in EMCS is $\approx$ 3-fold higher than that of MHMCS (using default parameters - Fig. S4, Table S7).

Convergence When initialized at random sequences, EMCS converges 1.25 - 5x faster than MHMCS (depending on choice of temperature and crossover rates, as shown in Fig. S4, Table. S10) likely due to the algorithm's increased versatility over MHMCS.

4 Discussion

In this work, we propose a two-step machine learning and sampling approach for protein engineering problems where training data is limited and positive hits are rare. Our method involves leveraging Large Protein Language Models (LPLMs) with transfer learning to estimate a fitness landscape, and then efficiently sampling the fitness landscape with Evolutionary Monte Carlo Search (EMCS) to propose novel high fitness protein sequences. As a proof-of-concept, we apply this approach to design small gene activators and demonstrate that it is capable of successfully generating novel and diverse protein sequences with dramatically higher experimental validation rates when compared to a more traditional sampling method (MHMCS) or baseline discovery from high-throughput screening.

Finally, though our proof-of-concept involved the design of relatively small proteins, we anticipate that our approach will generalize to arbitrarily lengthy peptides and generalize especially well to protein engineering problems involving larger proteins with multiple well characterized domains. While we aim to extend our approach to the application of larger proteins, our sampling algorithm will first need to be modified and optimized as random swaps within larger proteins are increasingly likely to result in low fitness predictions due to the presence of longer conserved domains.

5 Competing Interests Statement

The authors are affiliated with EpiCRISPR Biotechnologies as employees and equity holders. Several authors are inventors on provisional patent applications for the small gene activators described in this work.

6 Acknowledgements

The authors would like to acknowledge Professor Wing Hung Wong for his helpful advice and comments, Chris Still and Spencer Lopp (Biological Data Science, EpiCRISPR Biotechnologies) for assistance with code review and data analysis pipelines, as well as Xiao Yang and Giovanni Carosso (Technology Development, EpiCRISPR Biotechnologies) for useful conversations. The authors would also like to acknowledge Gladstone Genomics Core for next generation library sequencing.

References

[1] Daniel M Weinreich, Richard A Watson, and Lin Chao. Perspective: sign epistasis and genetic costraint on evolutionary trajectories. *Evolution*, 59(6):1165–1174, 2005.

[2] Zhizhou Ren, Jiahan Li, Fan Ding, Yuan Zhou, Jianzhu Ma, and Jian Peng. Proximal exploration for model-guided protein sequence design. In *International Conference on Machine Learning*, pages 18520–18536. PMLR, 2022.

[3] Frances H. Arnold. Design by directed evolution. *Accounts of Chemical Research*, 31(3):125–131, February 1998. doi: 10.1021/ar960017f. URL `https://doi.org/10.1021/ar960017f`.

[4] Michael S. Packer and David R. Liu. Methods for the directed evolution of proteins. *Nature Reviews Genetics*, 16(7):379–394, June 2015. doi: 10.1038/nrg3927. URL `https://doi.org/10.1038/nrg3927`.

[5] Douglas M Fowler and Stanley Fields. Deep mutational scanning: a new style of protein science. *Nature Methods*, 11(8):801–807, July 2014. doi: 10.1038/nmeth.3027. URL `https://doi.org/10.1038/nmeth.3027`.

[6] Roberto A Chica, Nicolas Doucet, and Joelle N Pelletier. Semi-rational approaches to engineering enzyme activity: combining the benefits of directed evolution and rational design. *Current Opinion in Biotechnology*, 16(4):378–384, August 2005. doi: 10.1016/j.copbio.2005.06.004. URL `https://doi.org/10.1016/j.copbio.2005.06.004`.

[7] Tuck Wong, Daria Zhurina, and Ulrich Schwaneberg. The diversity challenge in directed protein evolution. *Combinatorial Chemistry &amp High Throughput Screening*, 9(4):271–288, May 2006. doi: 10.2174/138620706776843192. URL `https://doi.org/10.2174/138620706776843192`.

[8] Surojit Biswas, Grigory Khimulya, Ethan C. Alley, Kevin M. Esvelt, and George M. Church. Low-n protein engineering with data-efficient deep learning. *Nature Methods*, 18(4):389–396, April 2021. doi: 10.1038/s41592-021-01100-y. URL `https://doi.org/10.1038/s41592-021-01100-y`.

[9] Yuting Xu, Deeptak Verma, Robert P. Sheridan, Andy Liaw, Junshui Ma, Nicholas M. Marshall, John McIntosh, Edward C. Sherer, Vladimir Svetnik, and Jennifer M. Johnston. Deep dive into machine learning models for protein engineering. *Journal of Chemical Information and Modeling*, 60(6):2773–2790, April 2020. doi: 10.1021/acs.jcim.0c00073. URL `https://doi.org/10.1021/acs.jcim.0c00073`.

[10] Robert Verkuil, Ori Kabeli, Yilun Du, Basile I. M. Wicky, Lukas F. Milles, Justas Dauparas, David Baker, Sergey Ovchinnikov, Tom Sercu, and Alexander Rives. Language models generalize beyond natural proteins. December 2022. doi: 10.1101/2022.12.21.521521. URL `https://doi.org/10.1101/2022.12.21.521521`.

[11] Andrew Giessel, Athanasios Dousis, Kanchana Ravichandran, Kevin Smith, Sreyoshi Sur, Iain McFadyen, Wei Zheng, and Stuart Licht. Therapeutic enzyme engineering using a generative neural network. *Scientific Reports*, 12(1), January 2022. doi: 10.1038/s41598-022-05195-x. URL `https://doi.org/10.1038/s41598-022-05195-x`.

[12] Namrata Anand and Possu Huang. Generative modeling for protein structures. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL `https://proceedings.neurips.cc/paper_files/paper/2018/file/afa299a4d1d8c52e75dd8a24c3ce534f-Paper.pdf`.

[13] 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`.

[14] Alex Hawkins-Hooker, Florence Depardieu, Sebastien Baur, Guillaume Couairon, Arthur Chen, and David Bikard. Generating functional protein variants with variational autoencoders. *PLOS Computational Biology*, 17(2):e1008736, February 2021. doi: 10.1371/journal.pcbi.1008736. URL `https://doi.org/10.1371/journal.pcbi.1008736`.

[15] Zhenqiao Song and Lei Li. Importance weighted expectation-maximization for protein sequence design. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 32349–32364. PMLR, 2023. URL `https://proceedings.mlr.press/v202/song23g.html`.

[16] Donatas Repecka, Vykintas Jauniskis, Laurynas Karpus, Elzbieta Rembeza, Irmantas Rokaitis, Jan Zrimec, Simona Poviloniene, Audrius Laurynenas, Sandra Viknander, Wissam Abuajwa, Otto Savolainen, Rolandas Meskys, Martin K. M. Engqvist, and Aleksej Zelezniak. Expanding functional protein sequence spaces using generative adversarial networks. *Nature Machine Intelligence*, 3(4):324–333, March 2021. doi: 10.1038/s42256-021-00310-5. URL https://doi.org/10.1038/s42256-021-00310-5.

[17] Allison Rossetto and Wenjin Zhou. GANDALF. In *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*. ACM, September 2020. doi: 10.1145/3388440.3412487. URL https://doi.org/10.1145/3388440.3412487.

[18] 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. Broadly applicable and accurate protein design by integrating structure prediction networks and diffusion generative models. *bioRxiv*, pages 2022–12, 2022.

[19] 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, July 2023. doi: 10.1038/s41586-023-06415-8. URL https://doi.org/10.1038/s41586-023-06415-8.

[20] Minkai Xu, Alexander S Powers, Ron O. Dror, Stefano Ermon, and Jure Leskovec. Geometric latent diffusion models for 3D molecule 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 38592–38610. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/xu23n.html.

[21] Casey Meehan, Kamalika Chaudhuri, and Sanjoy Dasgupta. A non-parametric test to detect data-copying in generative models. In *International Conference on Artificial Intelligence and Statistics*, 2020.

[22] Ishaan Gulrajani, Colin Raffel, and Luke Metz. Towards gan benchmarks which require generalization. *arXiv preprint arXiv:2001.03653*, 2020.

[23] Ali Madani, Ben Krause, Eric R Greene, Subu Subramanian, Benjamin P Mohr, James M Holton, Jose Luis Olmos Jr, Caiming Xiong, Zachary Z Sun, Richard Socher, et al. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, pages 1–8, 2023.

[24] Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Xi Chen, John Canny, Pieter Abbeel, and Yun S Song. Evaluating protein transfer learning with tape. In *Advances in Neural Information Processing Systems*, 2019.

[25] Christian Dallago, Jody Mou, Kadina E. Johnston, Bruce J. Wittmann, Nicholas Bhattacharya, Samuel Goldman, Ali Madani, and Kevin K. Yang. FLIP: Benchmark tasks in fitness landscape inference for proteins. November 2021. doi: 10.1101/2021.11.09.467890. URL https://doi.org/10.1101/2021.11.09.467890.

[26] Sam Sinai and Eric D Kelsic. A primer on model-guided exploration of fitness landscapes for biological sequence design, 2020. URL https://arxiv.org/abs/2010.10614.

[27] Moksh Jain, Emmanuel Bengio, Alex-Hernandez Garcia, Jarrid Rector-Brooks, Bonaventure F. P. Dossou, Chanakya Ekbote, Jie Fu, Tianyu Zhang, Micheal Kilgour, Dinghuai Zhang, Lena Simine, Payel Das, and Yoshua Bengio. Biological sequence design with gflownets, 2022. URL https://arxiv.org/abs/2203.04115.

[28] Faming Liang and Wing Hung Wong. Real-parameter evolutionary monte carlo with applications to bayesian mixture models. *Journal of the American Statistical Association*, 96(454):653–666, June 2001. doi: 10.1198/016214501753168325. URL https://doi.org/10.1198/016214501753168325.

[29] Faming Liang and Wing Hung Wong. Evolutionary monte carlo for protein folding simulations. *The Journal of Chemical Physics*, 115(7):3374–3380, August 2001. doi: 10.1063/1.1387478. URL https://doi.org/10.1063/1.1387478.

[30] Luigi Spezia, Andy Vinten, Roberta Paroli, and Marc Stutter. An evolutionary monte carlo method for the analysis of turbidity high-frequency time series through markov switching autoregressive models. *Environmetrics*, 32(8), July 2021. doi: 10.1002/env.2695. URL https://doi.org/10.1002/env.2695.

[31] Arnaud Dufays. Evolutionary sequential monte carlo samplers for change-point models. *Econometrics*, 4(4):12, March 2016. doi: 10.3390/econometrics4010012. URL https://doi.org/10.3390/econometrics4010012.

[32] Gopi Goswami, Jun S Liu, and Wing H Wong. Evolutionary monte carlo methods for clustering. *Journal of Computational and Graphical Statistics*, 16(4):855–876, December 2007. doi: 10.1198/106186007x255072. URL https://doi.org/10.1198/106186007x255072.

[33] Byoung-Tak Zhang and Dong-Yeon Cho. System identification using evolutionary markov chain monte carlo. *Journal of Systems Architecture*, 47(7):587–599, July 2001. doi: 10.1016/s1383-7621(01)00017-0. URL https://doi.org/10.1016/s1383-7621(01)00017-0.

[34] Giovanni A. Carosso, Robin W. Yeo, T. Blair Gainous, M. Zaki Jawaid, Xiao Yang, Vincent Cutillas, Lei Stanley Qi, Timothy P. Daley, and Daniel O. Hart. Discovery and engineering of hypercompact epigenetic modulators for durable gene activation. *bioRxiv*, 2023. doi: 10.1101/2023.06.02.543492. URL https://www.biorxiv.org/content/early/2023/06/15/2023.06.02.543492.

[35] 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. doi: 10.1126/science.ade2574. URL https://doi.org/10.1126/science.ade2574.

[36] Faming Liang and Wing Hung Wong. Evolutionary monte carlo: applications to $\mathcal{C}_p$ model sampling and change point problem. *Statistica sinica*, pages 317–342, 2000.