SYNEVO: TOWARDS SYNTHETIC EVOLUTION OF BIOMOLECULES VIA ALIGNING PROTEIN LANGUAGE MODELS TO BIOLOGICAL HARDWARE

Maria Artigues-Lleixà[1,2], Eduard Suñé[1**], Filippo Stocco[1,2], Noelia Ferruz[2,1] & Marc Güell[1,3] ***
[**]These authors contributed equally to this work
[1]Department of Medicine and Life Sciences, Universitat Pompeu Fabra, Barcelona, Spain
[2]Centre for Genomic Regulation, the Barcelona Institute of Science and Technology, Barcelona, Spain
[3]ICREA, Institució Catalana de Recerca I Estudis Avançats, Barcelona, Spain
{maria.artigues, eduard.sune, marc.guell}@upf.edu

ABSTRACT

Applications of biomolecular systems span gene writing, drug discovery, and environmental remediation. Despite their potential, biodesign remains slow and labor-intensive, often relying on trial and error. Recent advances in high-throughput sequencing, automated synthesis, and generative AI offer new opportunities but remain fragmented. We propose SYNEVO: an AI-driven, closed-loop system integrating automated protein design, and real-time experimental feedback to iteratively optimize biomolecular function. SYNEVO does not use template-based DNA replication, enabling a constraint-free generation of new genotypes, which departs from conventional evolution. We aim to validate our platform studying Zinc Finger proteins, a versatile class of DNA-binding proteins with significant therapeutic potential. Preliminary results showed that, by iteratively generating large libraries with autoregressive protein language models and experimentally testing their phenotypes, we optimized sequence selection. The measured features were fed back into the model via Reinforcement Learning to maximize protein enrichment scores, achieving a progressive improvement of generated phenotypes. This method, compared with Directed Evolution, shows higher efficiency in sampling high fitness protein candidates, and broader exploration of the sequence space. Potentially, by continuously refining its designs with minimal human intervention, this approach will accelerate protein engineering and provide a scalable solution for engineering new biomolecules with broad use across biotechnology and synthetic biology.

1 INTRODUCTION

Current biodesign processes are often laborious, time-consuming, and heavily rely on continuous experimental testing. To increase biomolecular fitness toward a desired goal, multiple characteristics need to be optimized and balanced, such as expressibility, activity and solubility. Historically, two main approaches have driven protein engineering: rational design and directed evolution. The first approach is strictly dependent on detailed structural knowledge of a particular family of proteins (Sievers et al. (2011)), given by the expertise of the designer or preexisting literature. The latter, although has greatly accelerated protein evolution campaigns, is still dependent on methods to perform an optimal sampling of the vast multidimensional sequence space, often counting on the available biological diversity (Packer & Liu (2015)), or bootstrapping from a single starting sequence. Neither method allows for an exhaustive exploration of the protein sequence space and struggles to meet the growing demand for faster, more efficient, and controllable protein engineering pipelines.

Recent advances in machine learning (ML) applied to biology have fundamentally reshaped the field, unlocking unprecedented possibilities. Specifically, protein language models (pLMs) have demonstrated exceptional capabilities across a range of applications, including protein structure

*Corresponding author

prediction, design, or prediction of mutational effects Weissenow & Rost (2025). Among them, autoregressive pLMs, trained to predict the next amino acid, can serve as a powerful tool to generate protein sequences, sampling from the learned distribution of immense training datasets (Ferruz & Höcker (2022)). Even though pLMs allow the exploration of a high-quality subspace of the possible genotypes, optimizing or conditionally generating for specific properties remains a challenge (Ertelt et al. (2025)). Effectively controlling the sampling of the model from a specific part of the sequence space, would enable the pLM to preferentially generate sequences from those rare regions that are often enriched in desired phenotypes. Noticeably, the frequent production of such rare events would result in a reduction of the exploration-exploitation dilemma.

However, many current ML tools for protein engineering operate as standalone platforms, lacking seamless integration into experimental workflows. This disconnection misses an extraordinary opportunity to exploit the synergic potential of artificial intelligence (AI) and experimental methodologies. In nature, goal-directed behaviors are shaped by rewards and punishments. Just as toddlers learn physics by tossing objects and observing the effects, Reinforcement Learning (RL) in ML operates similarly: an agent (model) interacts with its environment, refining decisions based on feedback. RL has been successfully applied to a wide range of challenges, from mastering 40 classic Atari games and Go (Schrittwieser et al. (2020)) to fine-tuning Large Language Models (LLMs) (Gemini et al. (2024)). These successes have established RL as the standard approach when training AI models (DeepSeek-AI et al. (2025); Ouyang et al. (2022)). However, its application to protein design remains significantly more complex. Unlike games or language models, where feedback mechanisms are explicit, proteins follow an intricate evolutionary grammar that remains poorly understood. As a result, assessing sequence quality directly is highly challenging as current methods, whether machine learning-based or not, often fail to reliably predict protein fitness. Thus, experimental validation becomes the necessary and primary bottleneck of any current protein engineering campaign, but provides an excellent direct reward for the generative models to improve upon.

Some automatic systems have been proposed that integrate sequence exploration with experimental testing to generate a fitness landscape of a desired protein (Rapp et al. (2024)). These self-driving laboratories, computational agents that interact with experimental setup, are automated in a closed loop so the system can keep iterating over the sequence space. These systems often are limited on the sequences they can experimentally explore each new cycle. This may be due to its variability generation, being combinatorial assembly or mutation, both somewhat depending on the previous generation. The number of variants tested each round depends on the processability of each system, making plate-based ones rely on the number of compartmentalized variants that can be tested.

Here, we introduce SYNEVO, a synthetic evolution machine designed to bridge the critical gap between experimental and ML worlds. SYNEVO is designed to seamlessly integrate generative models with biological automated hardware to accelerate protein engineering. The system consists of two key components: (i) an input/output (I/O) module for DNA synthesis, phenotypic selection and variant reading linked with (ii) an intelligent agent for sequence design and iterative learning through RL from continuous experimentally aware data provided, which is based on its proposed designs **(fig. 1a)**. Having established the system, we ought to establish our proof of concept by designing proteins that bind to target DNA sequences.

In particular, we synthesise an oligo pool library of designed protein variants that are translated and barcoded into a cell free environment where they undergo several rounds of target binding selection, whose enrichment can be obtained directly via Next Generation Sequencing (NGS). The resulting genotype-phenotype pairs are passed as feedback to the generative models using RL, which then will produce a second, possibly better-performing round of designs for testing. In this framework, we test several strategies to design the sequences, including deep mutational scanning, homologous search and generative models, for a total of over 6,000 sequences synthesized and tested in two rounds.

We further benchmarked the exploratory capabilities of SYNEVO versus DE by simulating ten rounds for each method and systematically comparing their outputs and performance under fair conditions. SYNEVO overcomes the limitations of rational design and directed evolution through the integration of machine learning. By leveraging cutting-edge methodologies from both computational and experimental biology, SYNEVO enables a "super-Darwinian" evolutionary framework, accelerating protein engineering and design efforts.

2 RESULTS

2.1 PROOF OF CONCEPT: *in vitro* SYNTHETIC EVOLUTION OF ZINC FINGER SCAFFOLDS

To test our workflow we started on-DNA binding proteins, using Zinc Fingers (ZnF) as proof of concept. ZnF are small protein structural motifs (80-100 aa) stabilized by zinc ions that facilitate DNA, RNA, or protein binding (Pavletich & Pabo (1991)). They play key roles in gene regulation, genome editing, and protein-protein interactions, with applications in fields such as synthetic biology and therapeutic gene editing (Urnov et al. (2010)). Specifically, these proteins are able to establish electrostatic interactions with the big grooves of the DNA with higher affinity and precision (Elrod-Erickson & Pabo (1999)). The objective of this first campaign is to diversify the scaffold of one classic ZnF, Zif268 (WT), while maintaining its specific binding to the WT target DNA.

First we explored the potential of SYNEVO using *in vitro* results to reinforce the model. We designed an oligo pool library of protein variants that are translated and barcoded into a cell free environment where they undergo several rounds of selection for binding affinity. Later, we use Next Generation Sequencing (NGS) to characterize the variants' abundance in each round and calculate their enrichment score. This score is later leveraged to refine the generation pLMs with ProtRL, a framework to implement Direct Preference Optimization (DPO), a RL method, on pLMs, and obtain a second library hypothesized to have better experimentally characteristics **(fig. 1a)**.

For the protein barcoding technology, the variants are inserted into a CIS display DNA construct, consisting of library variants fused with DNA replication initiator protein A (RepA) that binds exclusively to its template cDNA **(fig. S1)** (Odegrip et al. (2004)). This method proved to be efficient in the selection and enrichment of transcription factors in other works (Qi et al. (2024)). An *E. coli* lysate is used for the translation and selection in a pool of the assembled protein library. The selection of binding protein variants consists in a pulldown assay with the Zif268 DNA target bait. This CIS display setup allows for the recovery of the cDNA of protein variants directly from its barcode after selection, making possible several continuous rounds of enrichment and selection from the initial library **(fig. S2)**.

For testing the *in vitro* potential of this approach, we performed one whole cycle of SYNEVO consisting of two libraries: the first contained AI-generated and natural proteins and the second contained exclusively proteins generated with ZymCTRL refined with ProtRL and the experimental outcomes of the first library.

For the 1st library we used several protein language models (pLMs), including zero-shot (ESM-IF, ESM3 (Hsu et al. (2022); Hayes et al. (2025)) and fine-tuned (Progen2, ZymCTRL (Nijkamp et al. (2023); Munsamy et al.) models, to generate 2965 sequences. In contrast, non-AI methods consisted of a Deep Mutational Scan (DMS) of Zif268 (WT) and a structural homology search using Foldseek, which yielded 1740 and 477 sequences respectively. For the second library, we generated sequences at different epochs of the ProtRL alignment of ZymCTRL with the experimental results of the 1st library **(table S1)** (**Methods**). All the sequences present in the library passed a very stringent filtering process that took in consideration: (1) containing three C2H2 ZnF domains, (2) conservation of the 12 specificity residues of WT to bind its DNA target **(fig. 2b)**, (3) preservation of the WT fold (TM-score > 0.5) and (4) positive charged fraction of at least 0.3 (WT) **(fig. 1b)**.

Both libraries cover a broad range of the sequence space, reaching 40% sequence identity, even after the strict filtering process **(fig. 1c)**. Taking a look at the tendencies throughout the experimental selection, a tiny fraction of the first library shows a positive enrichment tendency, all of them below the WT protein Zif268 (Figure 1d). Regarding the second library, even though 4th rounds of selection were performed, due to technical difficulties, we only could get data from the first two rounds of selection. Still, we see a higher number of proteins with positive enrichment in respect to the first library, this time even surpassing WT values **(fig. 1d)**. Nevertheless, considering the ratios of positively enriched proteins in the first two rounds, we see a 15-fold increase in performance of the second library compared to the first **(fig. 1e)**. We are currently replicating the experiments to obtain stronger validation of the observed trend. Overall, these preliminary results point to ProtRL being able to align pLM with protein engineering campaign goals, also with experimental data.

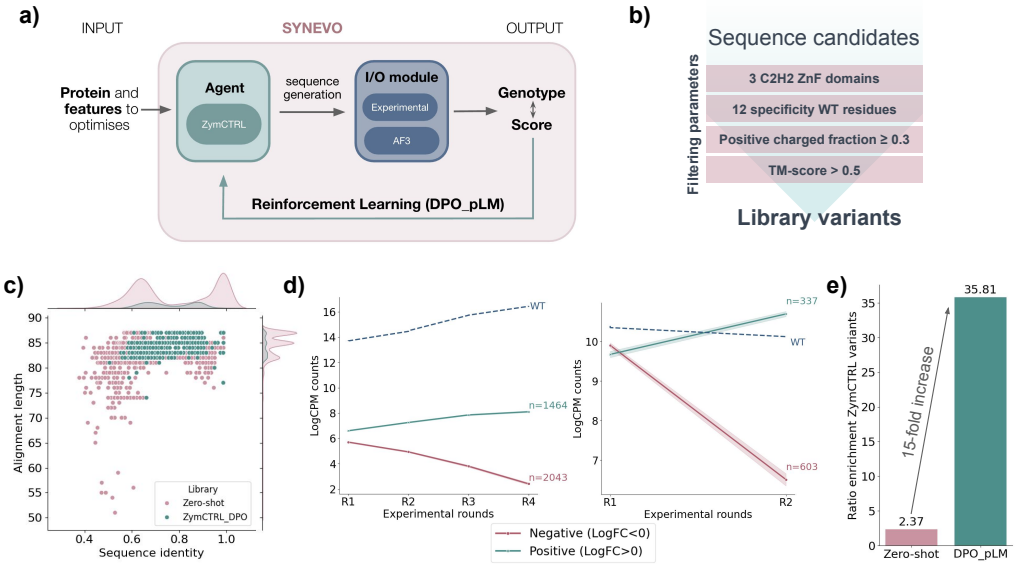

Figure 1: **(a)** General diagram of SYNEVO approach. **(b)** Parameters used in filtering sequences to build the two libraries. **(c)** Distribution of proteins in the sequence space, represented as the sequence identity of each variant against the WT and the length of aligned residues. **(d)** Linear regression of LogCPM counts of variants in each round of selection divided for positive (Log Fold Change > 0) and negative enrichment (LogFC < 0). Numbers plotted represent the amount of variants that show the behaviour across the whole library. **(e)** Ratio of ZymCTRL positive enriched proteins over the total generated from the model in both libraries.

2.2 *In silico* SYNTHETIC EVOLUTION

We explored the potential of AlphaFold3 (AF3) (Abramson et al. (2024)) to co-fold macromolecules, assess interactions, and evolve designs *in silico*. This approach emerged when we observed a correlation between experimental data of ZnF affinity (Kd value) and the results of AF3 contact probabilities of co-folding of ZnF and target DNA **(fig. S3)**. AF3 predictions generate a contact probability matrix of dimensions num_tokens × num_tokens, where each value represents the probability that two tokens are at least 8 Å from each other. We thus have applied a RL campaign with ProtRL starting from ZymCTRL, to align the model to the features that are also considered in the filtering. Again, the ZnF finetuned ZymCTRL (**Methods**) was used as a starting point for the reinforcement learning campaign.

Over iterative cycles, we generated 200 sequences, scored them using a complex multi-objective function, and refined the model using ProtRL. Specifically, the scoring function was taking in consideration the 1) target TM-score when superimposing with WT, 2) WT-like electrostatic charge, 3) preservation of 12 specificity residues, and 4) maximized AF3 contact probabilities between the ZnF domain and target DNA (**Methods**). The latter objective was set to maximise the specificity of the protein to recognize a specific DNA pattern, and was computed subtracting the probabilities of contact at the expected nucleotides **(fig. 2d)** with the ones that ZnF should not interact with. Remarkably, ProtRL is able to optimize for such complex reward functions, while keeping a low sequence identity with Zif268, but increasing the values of the other objectives over time. We are planning to experimentally test these *in silico*-evolved sequences, marking the next phase of validation and refinement.

2.3 COMPARISON OF SYNEVO AND SIMULATED DIRECTED EVOLUTION

In an effort to directly compare the presented SYNEVO approach against Directed Evolution (DE), a simulation of the latter was conducted. In this simulation, we performed 10 iterations of SYNEVO

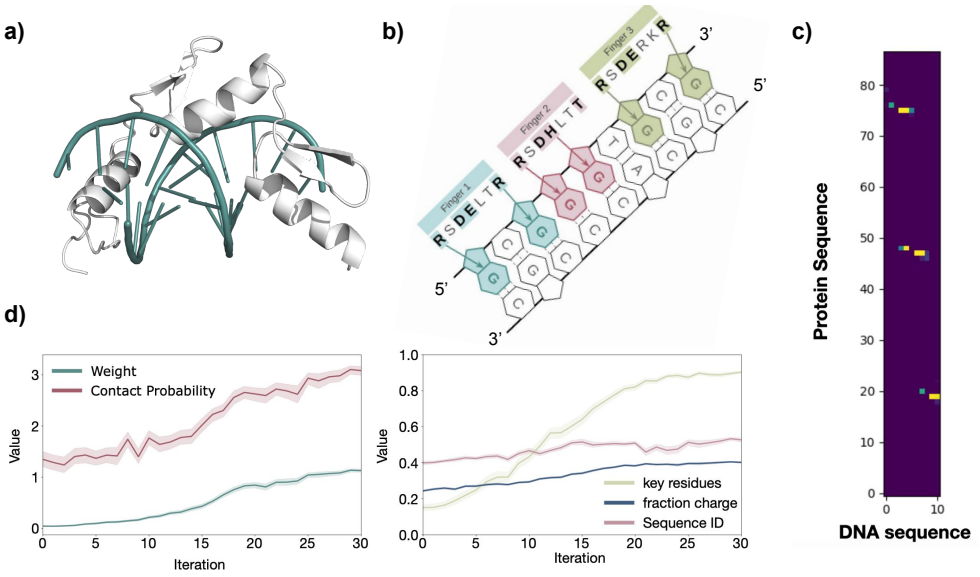

Figure 2: **(a)** Zinc Finger binding to the DNA target. **(b)** Zinc Finger (ZnF) interaction with target DNA. **(c)** Contact probability matrix, where higher probabilities are represented by brighter colors. **(d)** Optimization of the multi-objective function over time.

with the same parameters and reward function as in our in silico SYNEVO experiments (***In silico* synthetic evolution**). While we acknowledge the limitations of the AF3-based scoring function, its consistent application across both methods allows for a fair comparison of their ability to generate diverse and high-performing variants. For the simulation of DE, we generated 2000 initial cDNA copies of the reference protein, Zif268 and applied one mutation at each cDNA, following the mutation rate of one commercial kit of error-prone PCR and also considering its biases of type of mutation (**Methods**). We then translated and evaluated the variants with the same phenotype score used for SYNEVO and chose the 200 top performers. We then generated 10 copies of each to have a population of 2000 cDNAs and iterated this process 10 times.

As per the results, it can be appreciated that both methods are able to generate diversity from the WT protein **(fig. 3a)**, although SYNEVO, whose genotype network is unconstrained, is able to diverge most from WT values, showing in sequence identity and phenotype **(fig. 3a,c)**. To appreciate the amount of diversity generated across methods and rounds, we calculated the distance between proteins in each round to the WT after performing a Multiple Sequence Alignment (**Methods**). This calculation allows us to see the trend of both methods, in which SYNEVO is able to accumulate more diversity in fewer rounds when compared to DE **(fig. 3b)**. To evaluate sequence space exploration, we visualized the pairwise distances among all generated proteins using Multidimensional Scaling (MDS). This representation highlights the broader and more uniform coverage of sequence space achieved by SYNEVO compared to DE, while still yielding the best high-fitness variants **(fig. 3d,e)**.

3 Discussion

Novelty in the biosphere has been driven by imperfect template-based nucleic acid replication, generating a network of genotypes where each new variant arises as a minor modification of its predecessor. While directed evolution has accelerated this process, existing techniques remain constrained by this paradigm. SYNEVO conceptually departs from template-based DNA replication, producing genotypic networks without parent-child relationships and unconstrained genotype distances, enabling a broader and more exploratory search space for molecular design. This also is combined

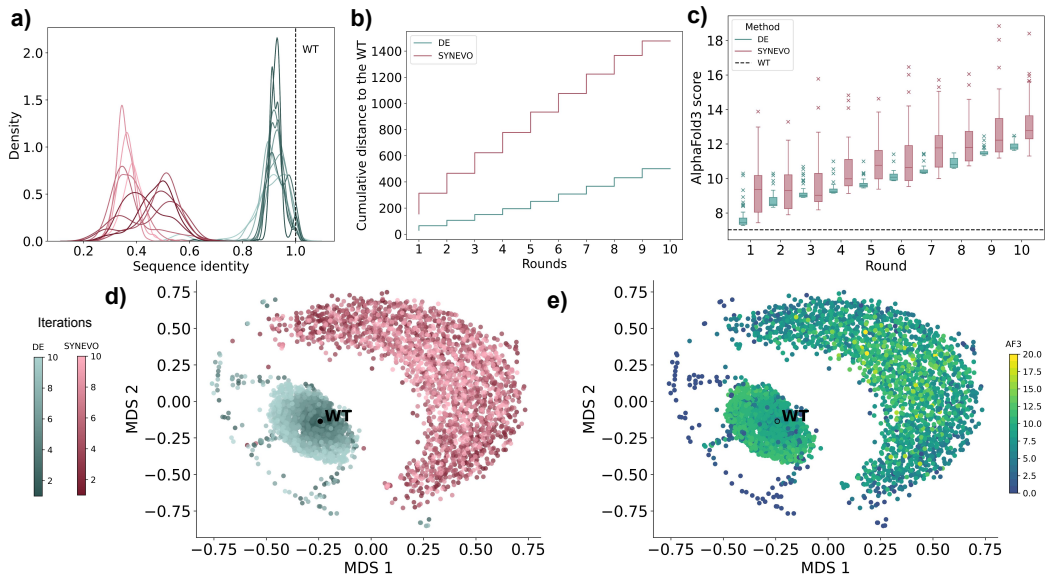

Figure 3: **(a)** Sequence identity distribution across rounds of SYNEVO and DE. **(b)** Accumulated diversity across rounds of SYNEVO and DE, calculated as the distance between all proteins in a round against the WT Zif268. **(c)** AlphaFold3 score of the top 50 best performing variants in all rounds of SYNEVO and DE. **(d)** Multidimensional Scaling (MDS) of the matrix of distances when performing a MSA of all proteins of SYNEVO and DE, colored by method and round of generation. **(e)** MDS of the matrix of distances of all proteins of SYNEVO and DE, colored by AlphaFold3 score.

with the potential of RL to learn from negative results, allowing SYNEVO to couple genotype emergence with evolutionary pressure, adjusting the designs at each step.

We envision SYNEVO as an autonomous system that integrates sequence design, testing, learning, and iterative optimization to align generative pLMs with desired functional features. This paradigm is built upon two key pillars: (i) a high-throughput experimental platform capable of generating high-quality genotype-phenotype correlations that can be automated end to end, and (ii) a reinforcement learning framework that continuously refines model generations based on experimental feedback. By seamlessly bridging computational design and laboratory validation, SYNEVO unifies dry and wet lab processes, addressing critical gaps in protein engineering workflows and accelerating the overall process.

We applied SYNEVO to ZnF proteins, aiming to evolve variants with enhanced enrichment scores throughout multiple rounds of selection and across sequence space. In both *in silico* and *in vitro* experiments, SYNEVO proved to be efficient in the alignment of protein engineering expectatives with pLMs' generated sequences. In the case of the *in silico* experiment, the optimized model improved the ratio of generation of sequences with desired characteristics over the iterations. As per the *in vitro* experiments, we are able to see a clear difference in the enrichment tendencies of both tested libraries. Most of the AI-generated proteins present in the first library were lost throughout rounds of selection **(table S1)** and underperformed in comparison to the WT, while after just one optimization round with ProtRL, preliminary data suggests that we are able to recover a higher percentage of proteins with positive enrichment that surpass WT values. Furthermore, SYNEVO is capable of yielding more genotypes with high-value phenotypes than simulated DE, while exploring more sequence space. Overall, SYNEVO has demonstrated its high-throughput capabilities, as it can be highly multiplexed and automatized. In the specific case of ZnF, we figure we can reach a turnover of 5 phenotypes per second (**Methods**). Still, we acknowledge that this data is preliminary and more experimentation is needed to validate this experimental methodology. Nevertheless, these preliminary results pinpoint SYNEVO as a faster and easier way to travel the path of protein optimization.

ACKNOWLEDGMENTS

This project has received funding from the European Research Council (ERC) under the European Union's Horizon Europe research and innovation programme (grant agreement no. 101123888) and from Agencia Estatal de Investigación y el Mecanismo de Recuperación y Resiliencia de la Unión Europea PRTR-NextGenerationUE (EUR2023-143459 by MCIN/AEI/10.13039/501100011033). M.A.L acknowledges support from an FI Fellowship (AGAUR-Catalan Government) co-funded by the European Social Fund (Award 2024 FI-3 00065).

REFERENCES

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, Sebastian W Bodenstein, David A Evans, Chia-Chun Hung, Michael O'Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvilė Žemgulytė, Eirini Arvaniti, Charles Beattie, Ottavia Bertolli, Alex Bridgland, Alexey Cherepanov, Miles Congreve, Alexander I Cowen-Rivers, Andrew Cowie, Michael Figurnov, Fabian B Fuchs, Hannah Gladman, Rishub Jain, Yousuf A Khan, Caroline M R Low, Kuba Perlin, Anna Potapenko, Pascal Savy, Sukhdeep Singh, Adrian Stecula, Ashok Thillaisundaram, Catherine Tong, Sergei Yakneen, Ellen D Zhong, Michal Zielinski, Augustin Žídek, Victor Bapst, Pushmeet Kohli, Max Jaderberg, Demis Hassabis, and John M Jumper. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630(8016):493–500, June 2024.

Simon Andrews, Felix Krueger, Anne Segonds-Pichon, Laura Biggins, Christel Krueger, and Steven Wingett. FastQC. Babraham Institute, January 2012.

Alex N Nguyen Ba, Katherine R Lawrence, Artur Rego-Costa, Shreyas Gopalakrishnan, Daniel Temko, Franziska Michor, and Michael M Desai. Barcoded bulk qtl mapping reveals highly polygenic and epistatic architecture of complex traits in yeast. *Elife*, 11:e73983, 2022.

D. Charif and J.R. Lobry. SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In U. Bastolla, M. Porto, H.E. Roman, and M. Vendruscolo (eds.), *Structural approaches to sequence evolution: Molecules, networks, populations*, Biological and Medical Physics, Biomedical Engineering, pp. 207–232. Springer Verlag, New York, 2007. ISBN : 978-3-540-35305-8.

Yunshun Chen, Lizhong Chen, Aaron T L Lun, Pedro L Baldoni, and Gordon K Smyth. edger v4: powerful differential analysis of sequencing data with expanded functionality and improved support for small counts and larger datasets. *Nucleic Acids Research*, 53(2):gkaf018, 01 2025. ISSN 1362-4962. doi: 10.1093/nar/gkaf018. URL https://doi.org/10.1093/nar/gkaf018.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z F Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J L Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R J Chen, R L Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S S Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W L Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X Q Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y K Li, Y Q Wang, Y X Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y X Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z Z Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen

Zhang. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. 2025.

M Elrod-Erickson and C O Pabo. Binding studies with mutants of zif268. contribution of individual side chains to binding affinity and specificity in the zif268 zinc finger-DNA complex. *J Biol Chem*, 274(27):19281–19285, July 1999.

Moritz Ertelt, Rocco Moretti, Jens Meiler, and Clara T. Schoeder. Self-supervised machine learning methods for protein design improve sampling but not the identification of high-fitness variants. *Science Advances*, 11(7):eadr7338, 2025. doi: 10.1126/sciadv.adr7338. URL `https://www.science.org/doi/abs/10.1126/sciadv.adr7338`.

Noelia Ferruz and Birte Höcker. Controllable protein design with language models. *Nature Machine Intelligence*, 4(6):521–532, June 2022.

Team Gemini, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah Liu, Andras Orban, Fabian Güra, Hao Zhou, Xinying Song, Aurelien Boffy, Harish Ganapathy, Steven Zheng, HyunJeong Choe, Ágoston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Merey, Martin Baeuml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Gaurav Singh Tomar, Evan Senter, Martin Chadwick, Ilya Kornakov, Nithya Attaluri, Iñaki Iturrate, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Xavier Garcia, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Ravi Addanki, Antoine Miech, Annie Louis, Denis Teplyashin, Geoff Brown, Elliot Catt, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaliy Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal,

Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Sidharth Mudgal, Romina Stella, Kevin Brooks, Gautam Vasudevan, Chenxi Liu, Mainak Chain, Nivedita Melinkeri, Aaron Cohen, Venus Wang, Kristie Seymore, Sergey Zubkov, Rahul Goel, Summer Yue, Sai Krishnakumaran, Brian Albert, Nate Hurley, Motoki Sano, Anhad Mohananey, Jonah Joughin, Egor Filonov, Tomasz Kepa, Yomna Eldawy, Jiawern Lim, Rahul Rishi, Shirin Badiezadegan, Taylor Bos, Jerry Chang, Sanil Jain, Sri Gayatri Sundara Padmanabhan, Subha Puttagunta, Kalpesh Krishna, Leslie Baker, Norbert Kalb, Vamsi Bedapudi, Adam Kurzrok, Shuntong Lei, Anthony Yu, Oren Litvin, Xiang Zhou, Zhichun Wu, Sam Sobell, Andrea Siciliano, Alan Papir, Robby Neale, Jonas Bragagnolo, Tej Toor, Tina Chen, Valentin Anklin, Feiran Wang, Richie Feng, Milad Gholami, Kevin Ling, Lijuan Liu, Jules Walter, Hamid Moghaddam, Arun Kishore, Jakub Adamek, Tyler Mercado, Jonathan Mallinson, Siddhinita Wandekar, Stephen Cagle, Eran Ofek, Guillermo Garrido, Clemens Lombriser, Maksim Mukha, Botu Sun, Hafeezul Rahman Mohammad, Josip Matak, Yadi Qian, Vikas Peswani, Pawel Janus, Quan Yuan, Leif Schelin, Oana David, Ankur Garg, Yifan He, Oleksii Duzhyi, Anton Älgmyr, Timothée Lottaz, Qi Li, Vikas Yadav, Luyao Xu, Alex Chinien, Rakesh Shivanna, Aleksandr Chuklin, Josie Li, Carrie Spadine, Travis Wolfe, Kareem Mohamed, Subhabrata Das, Zihang Dai, Kyle He, Daniel von Dincklage, Shyam Upadhyay, Akanksha Maurya, Luyan Chi, Sebastian Krause, Khalid Salama, Pam G Rabinovitch, Pavan Kumar Reddy M, Aarush Selvan, Mikhail Dektiarev, Golnaz Ghiasi, Erdem Guven, Himanshu Gupta, Boyi Liu, Deepak Sharma, Idan Heimlich Shtacher, Shachi Paul, Oscar Akerlund, François-Xavier Aubet, Terry Huang, Chen Zhu, Eric Zhu, Elico Teixeira, Matthew Fritze, Francesco Bertolini, Liana-Eleonora Marinescu, Martin Bölle, Dominik Paulus, Khyatti Gupta, Tejasi Latkar, Max Chang, Jason Sanders, Roopa Wilson, Xuewei Wu, Yi-Xuan Tan, Lam Nguyen Thiet, Tulsee Doshi,

Sid Lall, Swaroop Mishra, Wanming Chen, Thang Luong, Seth Benjamin, Jasmine Lee, Ewa Andrejczuk, Dominik Rabiej, Vipul Ranjan, Krzysztof Styrc, Pengcheng Yin, Jon Simon, Malcolm Rose Harriott, Mudit Bansal, Alexei Robsky, Geoff Bacon, David Greene, Daniil Mirylenka, Chen Zhou, Obaid Sarvana, Abhimanyu Goyal, Samuel Andermatt, Patrick Siegler, Ben Horn, Assaf Israel, Francesco Pongetti, Chih-Wei "Louis" Chen, Marco Selvatici, Pedro Silva, Kathie Wang, Jackson Tolins, Kelvin Guu, Roey Yogev, Xiaochen Cai, Alessandro Agostini, Maulik Shah, Hung Nguyen, Noah Ó Donnaile, Sébastien Pereira, Linda Friso, Adam Stambler, Adam Kurzrok, Chenkai Kuang, Yan Romanikhin, Mark Geller, ZJ Yan, Kane Jang, Cheng-Chun Lee, Wojciech Fica, Eric Malmi, Qijun Tan, Dan Banica, Daniel Balle, Ryan Pham, Yanping Huang, Diana Avram, Hongzhi Shi, Jasjot Singh, Chris Hidey, Niharika Ahuja, Pranab Saxena, Dan Dooley, Srividya Pranavi Potharaju, Eileen O'Neill, Anand Gokulchandran, Ryan Foley, Kai Zhao, Mike Dusenberry, Yuan Liu, Pulkit Mehta, Ragha Kotikalapudi, Chalence Safranek-Shrader, Andrew Goodman, Joshua Kessinger, Eran Globen, Prateek Kolhar, Chris Gorgolewski, Ali Ibrahim, Yang Song, Ali Eichenbaum, Thomas Brovelli, Sahitya Potluri, Preethi Lahoti, Cip Baetu, Ali Ghorbani, Charles Chen, Andy Crawford, Shalini Pal, Mukund Sridhar, Petru Gurita, Asier Mujika, Igor Petrovski, Pierre-Louis Cedoz, Chenmei Li, Shiyuan Chen, Niccolò Dal Santo, Siddharth Goyal, Jitesh Punjabi, Karthik Kappaganthu, Chester Kwak, Pallavi LV, Sarmishta Velury, Himadri Choudhury, Jamie Hall, Premal Shah, Ricardo Figueira, Matt Thomas, Minjie Lu, Ting Zhou, Chintu Kumar, Thomas Jurdi, Sharat Chikkerur, Yenai Ma, Adams Yu, Soo Kwak, Victor Ähdel, Sujeevan Rajayogam, Travis Choma, Fei Liu, Aditya Barua, Colin Ji, Ji Ho Park, Vincent Hellendoorn, Alex Bailey, Taylan Bilal, Huanjie Zhou, Mehrdad Khatir, Charles Sutton, Wojciech Rzadkowski, Fiona Macintosh, Konstantin Shagin, Paul Medina, Chen Liang, Jinjing Zhou, Pararth Shah, Yingying Bi, Attila Dankovics, Shipra Banga, Sabine Lehmann, Marissa Bredesen, Zifan Lin, John Eric Hoffmann, Jonathan Lai, Raynald Chung, Kai Yang, Nihal Balani, Arthur Bražinskas, Andrei Sozanschi, Matthew Hayes, Héctor Fernández Alcalde, Peter Makarov, Will Chen, Antonio Stella, Liselotte Snijders, Michael Mandl, Ante Kärrman, Paweł Nowak, Xinyi Wu, Alex Dyck, Krishnan Vaidyanathan, Raghavender R, Jessica Mallet, Mitch Rudominer, Eric Johnston, Sushil Mittal, Akhil Udathu, Janara Christensen, Vishal Verma, Zach Irving, Andreas Santucci, Gamaleldin Elsayed, Elnaz Davoodi, Marin Georgiev, Ian Tenney, Nan Hua, Geoffrey Cideron, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Dylan Scandinaro, Heinrich Jiang, Jasper Snoek, Mukund Sundararajan, Xuezhi Wang, Zack Ontiveros, Itay Karo, Jeremy Cole, Vinu Rajashekhar, Lara Tumeh, Eyal Ben-David, Rishub Jain, Jonathan Uesato, Romina Datta, Oskar Bunyan, Shimu Wu, John Zhang, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Jane Park, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Geoffrey Irving, Edward Loper, Michael Fink, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Ivan Petrychenko, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Evan Palmer, Paul Suganthan, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Ginger Perng, Elena Allica Abellan, Mingyang Zhang, Ishita Dasgupta, Nate Kushman, Ivo Penchev, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Daniel Andor, Pedro Valenzuela, Minnie Lui, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Ken Franko, Anna Bulanova, Rémi Leblond, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Mark Omernick, Colton Bishop, Rachel Sterneck, Rohan Jain, Jiawei Xia, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Daniel J. Mankowitz, Alex Polozov, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Matthieu Geist, Ser tan Girgin, Hui Li,

Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Kathy Wu, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Saaber Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Yeqing Li, Nir Levine, Ariel Stolovich, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Charlie Deck, Hyo Lee, Zonglin Li, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Sho Arora, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Lynette Webb, Sahil Dua, Dong Li, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Evgenii Eltyshev, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Christof Angermueller, Xiaowei Li, Anoop Sinha, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Denny Zhou, Komal Jalan, Dinghua Li, Blake Hechtman, Parker Schuh, Milad Nasr, Kieran Milan, Vladimir Mikulik, Juliana Franco, Tim Green, Nam Nguyen, Joe Kelley, Aroma Mahendru, Andrea Hu, Joshua Howland, Ben Vargas, Jeffrey Hui, Kshitij Bansal, Vikram Rao, Rakesh Ghiya, Emma Wang, Ke Ye, Jean Michel Sarr, Melanie Moranski Preston, Madeleine Elish, Steve Li, Aakash Kaku, Jigar Gupta, Ice Pasupat, Da-Cheng Juan, Milan Someswar, Tejvi M., Xinyun Chen, Aida Amini, Alex Fabrikant, Eric Chu, Xuanyi Dong, Amruta Muthal, Senaka Buthpitiya, Sarthak Jauhari, Nan Hua, Urvashi Khandelwal, Ayal Hitron, Jie Ren, Larissa Rinaldi, Shahar Drath, Avigail Dabush, Nan-Jiang Jiang, Harshal Godhia, Uli Sachs, Anthony Chen, Yicheng Fan, Hagai Taitelbaum, Hila Noga, Zhuyun Dai, James Wang, Chen Liang, Jenny Hamer, Chun-Sung Ferng, Chenel Elkind, Aviel Atias, Paulina Lee, Vít Listík, Mathias Carlen, Jan van de Kerkhof, Marcin Pikus, Krunoslav Zaher, Paul Müller, Sasha Zykova, Richard Stefanec, Vitaly Gatsko, Christoph Hirnschall, Ashwin Sethi, Xingyu Federico Xu, Chetan Ahuja, Beth Tsai, Anca Stefanoiu, Bo Feng, Keshav Dhandhania, Manish Katyal, Akshay Gupta, Atharva Parulekar, Divya Pitta, Jing Zhao, Vivaan Bhatia, Yashodha Bhavnani, Omar Alhadlaq, Xiaolin Li, Peter Danenberg, Dennis Tu, Alex Pine, Vera Filippova, Abhipso Ghosh, Ben Limonchik, Bhargava Urala, Chaitanya Krishna Lanka, Derik Clive, Yi Sun, Edward Li, Hao Wu, Kevin Hongtongsak, Ianna Li, Kalind Thakkar, Kuanysh Omarov, Kushal Majmundar, Michael Alverson, Michael Kucharski, Mohak Patel, Mudit Jain, Maksim Zabelin, Paolo Pelagatti, Rohan Kohli, Saurabh Kumar, Joseph Kim, Swetha Sankar, Vineet Shah, Lakshmi Ramachandruni, Xiangkai Zeng, Ben Bariach, Laura Weidinger, Tu Vu, Alek Andreev, Antoine He, Kevin Hui, Sheleem Kashem, Amar Subramanya, Sissie Hsiao,

Demis Hassabis, Koray Kavukcuoglu, Adam Sadovsky, Quoc Le, Trevor Strohman, Yonghui Wu, Slav Petrov, Jeffrey Dean, and Oriol Vinyals. Gemini: A family of highly capable multimodal models, 2024. URL https://arxiv.org/abs/2312.11805.

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 A 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. *Science*, pp. eads0018, January 2025.

Robert Gentleman Saikat DebRoy Hervé Pagès, Patrick Aboyoun. Biostrings: Efficient manipulation of biological strings. Bioconductor, 2024. URL https://doi.org/10.18129/B9.bioc.Biostrings.

Chloe Hsu, Robert Verkuil, Jason Liu, Zeming Lin, B Hie, Tom Sercu, Adam Lerer, and Alexander Rives. Learning inverse folding from millions of predicted structures. *bioRxiv*, 162:8946–8970, September 2022.

Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1):10–12, 2011. ISSN 2226-6089. doi: 10.14806/ej.17.1.200. URL https://journal.embnet.org/index.php/embnetjournal/article/view/200.

Geraldene Munsamy, Sebastian Lindner, Philipp Lorenz, and Noelia Ferruz. ZymCTRL: a conditional language model for the controllable generation of artificial enzymes.

Erik Nijkamp, Jeffrey A Ruffolo, Eli N Weinstein, Nikhil Naik, and Ali Madani. ProGen2: Exploring the boundaries of protein language models. *Cell Syst.*, 14(11):968–978.e3, November 2023.

Richard Odegrip, David Coomber, Bill Eldridge, Rosemarie Hederer, Philip A Kuhlman, Christopher Ullman, Kevin FitzGerald, and Duncan McGregor. CIS display: In vitro selection of peptides from libraries of protein-DNA complexes. *Proc Natl Acad Sci U S A*, 101(9):2806–2810, February 2004.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L 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 Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. 2022.

Michael S Packer and David R Liu. Methods for the directed evolution of proteins. *Nat. Rev. Genet.*, 16(7):379–394, July 2015.

Nikola P Pavletich and Carl O Pabo. Zinc finger-dna recognition: crystal structure of a zif268-dna complex at 2.1 å. *Science*, 252(5007):809–817, 1991.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Lin Qi, Emily Bennett, and Mark Isalan. A directed evolution protocol for engineering minimal transcription factors, based on CIS display. *Methods Mol. Biol.*, 2774:1–13, 2024.

Jacob T. Rapp, Bennett J. Bremer, and Philip A. Romero. Self-driving laboratories to autonomously navigate the protein fitness landscape. *Nature Chemical Engineering*, 1(1):97–107, Jan 2024. ISSN 2948-1198. doi: 10.1038/s44286-023-00002-4. URL https://doi.org/10.1038/s44286-023-00002-4.

Matthew E. Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47–e47, 01 2015. ISSN 0305-1048. doi: 10.1093/nar/gkv007. URL https://doi.org/10.1093/nar/gkv007.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy Lillicrap, and David Silver. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609, Dec 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-03051-4. URL https://doi.org/10.1038/s41586-020-03051-4.

Fabian Sievers and Desmond G. Higgins. Clustal omega. *Current Protocols in Bioinformatics*, 48(1):3.13.1–3.13.16, 2014. doi: https://doi.org/10.1002/0471250953.bi0313s48. URL https://currentprotocols.onlinelibrary.wiley.com/doi/abs/10.1002/0471250953.bi0313s48.

Stuart A Sievers, John Karanicolas, Howard W Chang, Anni Zhao, Lin Jiang, Onofrio Zirafi, Jason T Stevens, Jan Münch, David Baker, and David Eisenberg. Structure-based design of non-natural amino-acid inhibitors of amyloid fibril formation. *Nature*, 475(7354):96–100, June 2011.

Martin Steinegger and Johannes Söding. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11):1026–1028, Nov 2017. ISSN 1546-1696. doi: 10.1038/nbt.3988. URL https://doi.org/10.1038/nbt.3988.

Filippo Stocco, Maria Artigues-Lleixa, Andrea Hunklinger, Talal Widatalla, Marc Guell, and Noelia Ferruz. Guiding generative protein language models with reinforcement learning. 2024.

Fyodor D. Urnov, Edward J. Rebar, Michael C. Holmes, H. Steve Zhang, and Philip D. Gregory. Genome editing with engineered zinc finger nucleases. *Nature Reviews Genetics*, 11(9):636–646, Sep 2010. ISSN 1471-0064. doi: 10.1038/nrg2842. URL https://doi.org/10.1038/nrg2842.

Michel van Kempen, Stephanie S. Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron L. M. Gilchrist, Johannes Söding, and Martin Steinegger. Fast and accurate protein structure search with foldseek. *Nature Biotechnology*, 42(2):243–246, Feb 2024. ISSN 1546-1696. doi: 10.1038/s41587-023-01773-0. URL https://doi.org/10.1038/s41587-023-01773-0.

Konstantin Weissenow and Burkhard Rost. Are protein language models the new universal key? *Current Opinion in Structural Biology*, 91:102997, 2025. ISSN 0959-440X. doi: https://doi.org/10.1016/j.sbi.2025.102997. URL https://www.sciencedirect.com/science/article/pii/S0959440X25000156.

Jiajie Zhang, Kassian Kobert, Tomáš Flouri, and Alexandros Stamatakis. Pear: a fast and accurate illumina paired-end read merger. *Bioinformatics*, 30(5):614–620, 10 2013. ISSN 1367-4803. doi: 10.1093/bioinformatics/btt593. URL https://doi.org/10.1093/bioinformatics/btt593.

A MATERIALS AND METHODS

A.1 PLMS FINETUNING AND LIBRARY DESIGN

Foldseek web server (van Kempen et al. (2024)) was used to search for Zif268 structural homologs, a dataset of around 4000 proteins that were leveraged to fine tune both ProGen2 (Nijkamp et al. (2023)) and ZymCTRL (Munsamy et al.). For ProGen2, sequences were generated using the following parameters: temperature=0.5, max_length=90, top_p=0.95. ZymCTRL is a conditional autoregressive protein language model (pLM) trained on the known enzyme space, where each sequence is paired with its corresponding Enzyme Commission (EC) number as a functional label. Due to ZymCTRL's architecture, we appended an E.C. number (1.3.3.18, not present in ZymCTRL's training dataset) to the fine tuning sequences that later allowed us to prompt and generate sequences with the following parameters: top_k=9, repetition_penalty=1.2. Sequences from ESM3 (Hayes et al. (2025)) were generated using as prompt the first 8 amino acids of Zif268 and the following parameters: num_steps=8, temperature=0.5. For ESM-IF (Hsu et al. (2022)), we used the Zif268's experimental structure (PDB:1ZAA) as backbone to generate new sequences using temperatures ranging from 10^{-6} to 2.

After all sequences were generated, a filtering process took part, except for the DMS variants. Only sequences that contained three C2H2 ZnF domains were considered. All the sequences were pairwise aligned to Zif268 and those without intact conservation of the 12 residues key for its target DNA specificity (4 positions in each alpha helix) –as described in Elrod-Erickson & Pabo (1999)– were discarded. Furthermore, to maintain DNA binding, proteins without positive charged fraction equal or higher than Zif268 were filtered out. The protein libraries were codon optimized and flanking sequences for assembly were added. Then synthesized as 300bp DNA oligo pools from Integrated DNA Technologies and Twist Biosciences, for the 1st and 2nd library, respectively.

A.2 EXPERIMENTAL SETUP

A.2.1 DNA PREPARATION

Protocols were carried out independently in a Opentrons Flex (Opentrons, USA) pipetting workstation with the modules necessary for each step. Oligo pools synthesized containing the library of variants were PCR amplified following manufacturers instructions and assembled into the expression vector containing the CIS-Display elements (Qi et al. (2024)) via Golden Gate assembly. From the assembly product, expression vectors are linearized and amplified using KAPA HiFi HotStart ReadyMix (Roche, Cat. No. KK2602), resulting in a cloning-free library assembly. Template library is purified using Ampure magnetic beads.

Proves used as bait for the pulldown assay of proteins are prepared using isothermal amplification (Ba et al. (2022)) of a binding site containing oligo with a biotinylated primer ordered from Integrated DNA Technologies (IDT). Resulting in two different proves, one with three ZnF268 binding sites "GCGTGGGCG" and one with random bases to assess for unspecific off-target binding **(table S3)**.

A.2.2 VARIANT ENRICHMENT AND SELECTION

Based on the CIS-Display DNA-protein pulldown Protocol (Qi, Bennett, and Isalan 2024), *in vitro* expression reactions were prepared for each DNA bait. 3-4μg of library DNA template was mixed into *E. coli* S30 extract system for linear templates (Promega) following the manufacturer's instructions supplemented with $ZnCl_2$ (50μM final concentration) to a total reaction volume of 50μL. The *in vitro* TnT (transcription and translation) reaction was incubated at 37°C and cooled at 4°C to stop the reaction. The synthesized protein library was diluted 1:10 and blocked before adding 25pM of target bait and control bait respectively. After 1 hour incubation with gentle shaking to increase the interactions, proves were pulled down using Dynabeads™ M-280 Streptavidin (Invitrogen) magnetic beads. The beads were washed 8-12 times to remove non-binding proteins, and remaining proteins were eluted in nuclease free water. Different incubation times, bait amounts and washing conditions were optimized to increase the enrichment of binding variants. To prepare the input template DNA for the next round, recovered sequences from selected variants were PCR amplified from the eluate using KAPA HiFi HotStart ReadyMix (Figure S3).

Sequencing libraries were prepared amplifying the 300bp ZnF coding sequence from the selected barcodes contained in the target and control eluates **(fig. S3)**. KAPA HiFi HotStart ReadyMix to add illumina adaptors.

A.3 SEQUENCING DATA ANALYSIS AND ENRICHMENT CALCULATION

The first variant library was sequenced using a NextSeq 2000 (pair-end, 300 cycles) kit at the CRG-Genomics facility. Given the lower number of variants, the second library was sequenced using a Miseq V3 (paired-end, 300 cycles) at UPF-Genomics facility. Both times accomplishing a coverage of 100x at least. Quality control of the raw reads was performed with FastQC (v0.11.9) (Andrews et al. (2012)). Pair-end reads were merged with PEAR (v0.9.6) (Zhang et al. (2013)) and then trimmed with Cutadapt (v3.7) (Martin (2011)). Variant calling was performed using the function vcountPDict() from Biostrings R package (v2.70.3) (Hervé Pagès (2024)), without allowing for any mismatch. Variant counts were then used to calculate enrichment scores for each protein throughout all rounds with limma (v3.58.1) (Ritchie et al. (2015)) and edgeR (4.0.16) (Chen et al. (2025)) R packages. Technical replicate counts were added and low count proteins were filtered out with the function filterByExpr(). For normalisation, the library sizes were taken into account and, after converting the raw counts to Counts Per Million (CPM), they were log-transformed. The log(CPM) counts were fitted into a linear model and for each protein we extracted the slope (LogFC), which if it was higher than 0, it was considered positive enrichment and if it was lower than 0, negative enrichment.

A.4 ZINC FINGER SYNEVO THROUGHPUT

The approximate rate of DNA input (genotype) to DNA output (phenotype) is $\sim$5 phenotypes/second in the current approach. Given a set of DNA instructions library ($\sim$500,000 sequences), this can be protein barcoded and phenotyped in an automatable experimental pipeline of about 8 hours, followed by barcode sequencing overnight.

A.5 DIRECT PREFERENCE OPTIMIZATION

We have applied the weighted and ranked form of DPO as previously described in (Stocco et al. (2024)).

A.6 *In vitro* DPO OPTIMIZATION OF ZYMCTRL

For ZymCTRL's alignment with experimental data, we decided to refine it with ProtRL for 20 epochs with the hyperparameters stated in **table S2**. Four different DPO models were trained combining the use of two ProtRL implementations (ranked and weighted) and two different enrichment scores. The sequences were paired with the slope extracted from the counts analysis and an additional measure was tested, the difference between the abundance in the last and the first round ($\Delta log(CPM)$), dubbed dynamic range. For each model, we generated sequences at epochs 5, 10, 15 and 20. The training and generation took place over the course of two days in 2 H100 GPUs. Once generated, the sequences were filtered following the same criteria as the 1st library (**Methods**).

A.7 *In silico* DPO OPTIMIZATION OF ZYMCTRL

The fine-tuned (FT) ZymCTRL model was used as a reference and iteratively updated through the reinforcement learning (RL) campaign. Contact probabilities were determined by folding the protein sequence with the DNA strand "AGCGTGGGCGT" using AlphaFold3 (AF3) (Abramson et al. (2024)). The contact probability score was computed as the difference between the contact probabilities of the target nucleotides and those of off-target nucleotides, ensuring a specificity-driven selection metric. Each sequence was assigned a weight based on a composite scoring function, incorporating: 1) Contact probability – computed as described above, 1) TM-score – determined using Foldseek and using WT as template (PDB:1ZAA), 3) Inverse of sequence identity (sequence dissimilarity) was computed with MMseqs2 (Steinegger & Söding (2017)) using WT sequence as template, to promote diversity in sequence generation, 4) Key residue ratio, calculated by aligning generated sequences with the WT sequence and computing the fraction of conserved key residues,

5) Length penalty, computed as a Gaussian function of the ratio between WT length and generated sequence length, and 6) a Gaussian function of sequence alignment length obtained with Foldseek. Finally we also considered the 7) gaussian of the ratio of the sequence fractional charge, defined as the sum of charged residues divided by sequence length, and the WT fractional charge (0.3). Training was conducted on a GPU H100 for each task, ranging from a few GPU hours to one day of training to reach the 30 iterations for DPO. The hyperparameters for DPO are stated in **table S2**.

A.8 Simulation of Directed Evolution

The mutation rate acquired for the experiment and the biases for the type of mutations to apply were obtained from the Mutazyme II DNA polymerase enzyme, from the Agilent GeneMorph II Random Mutagenesis Kit. As the mutation rate is 3-16 mutations per Kb (per PCR), a mutation rate of 9,5 per 1 Kb was approximated, meaning 1 mutation in each protein of nearly 90 amino acids.

To perform Multiple Sequence Alignments, Clustal Omega (v1.2.4) (Sievers & Higgins (2014)) was used with default parameters. Distance matrices were calculated using the function dist.alignment() from seqinr R package (v.4.2-36) (Charif & Lobry (2007)). Multidimensional Scaling representations were built with scikit-learn (v1.6.1) (Pedregosa et al. (2011)) MDS implementation.

B Supplementary data

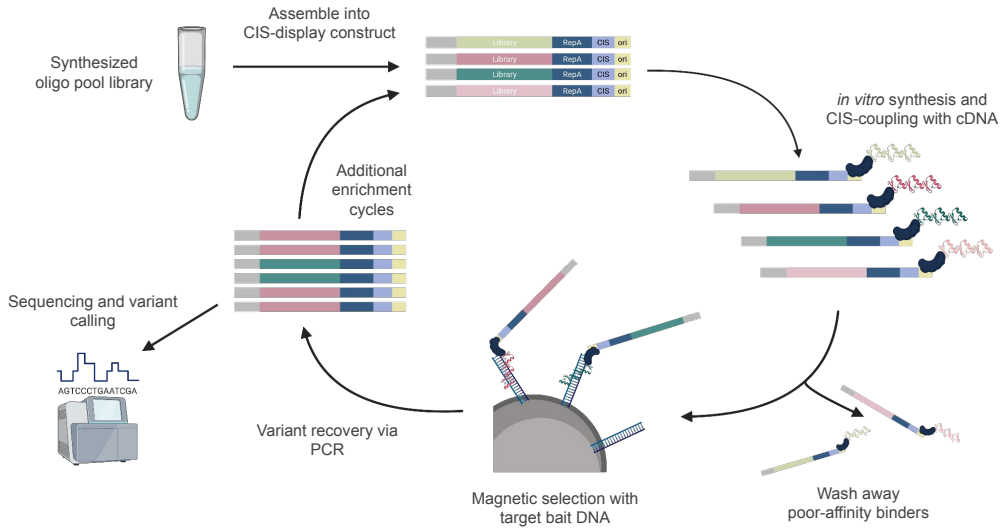

Figure S1: CIS-display protocol diagram as showcased in Qi et al. (2024). Created in `https://BioRender.com`.

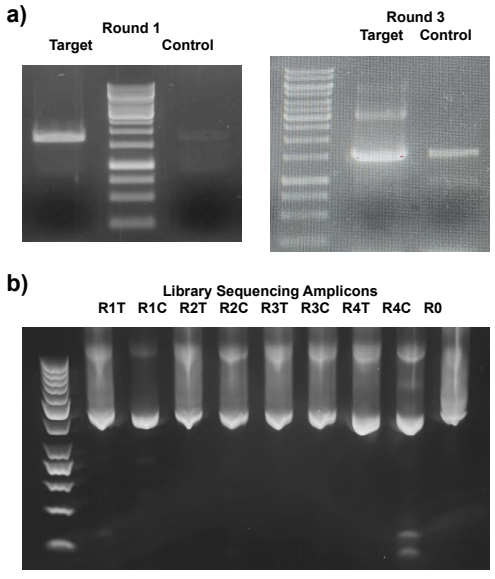

Figure S2: **(a)** PCR products from amplification under the same conditions of pulldown eluates from round 1 and 3 of the first library. **(b)** Barcode recovery amplification product of every sample for library preparation (RnT - sample pulldown with Zif268 target, RnC - control).

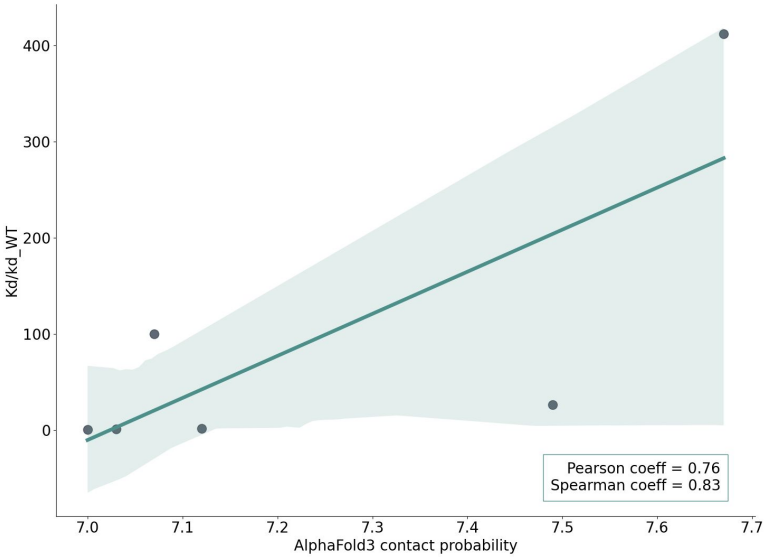

Figure S3: Correlation between AF3 contact probabilities and affinity measures for 6 published Zif268 variants (Elrod-Erickson & Pabo (1999))

a)

Design method	Library variants	Recovery at 4th round
ZymCTRL	574	**51%**
ESM3	1 441	**20%**
ProGen2	274	**29%**
ESM-IF	676	**53%**
DMS	1 740	**99%**
Foldseek	477	**96%**
Total	**5 184**	**62%**

b)

Trained epochs	Library variants	Recovery at 2nd round
Epoch 5	190	**69%**
Epoch 10	249	**68%**
Epoch 15	1	**100%**
Epoch 20	509	**66%**
Total	**949**	**67%**

Table S1: **(a)** Number of variants of the first library divided by method of generation and percentage of recovery at the last experimental round. **(b)** Number of variants of the second library divided by epoch of generation and percentage of recovery at the last experimental round.

Hyperparameters (DPO)	
β	0.01
Seed number	1998
Learning rate	1×10-7
Batch size	5
epochs	5
Train/Test split	0.2
Adams	(0.9, 0.98)
ε	1× 10-8
Adam decay	0.1

Table S2: Hyperparemeters used for the training of ProtRL unless otherwise specified in the text.

Oligo	Sequence
Target DNA bait oligo	NNNNNNGCGTGGGCGNNNNNNCGCCCACGCNNNNNNGCGTGGGCGNNNNNNTCACAGTCAGTCCACACGTC
Control DNA bait oligo	NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNTCACAGTCAGTCCACACGTC
Biotinylated amplification primer	Biotin-GACGTGTGGACTGACTGTGA

Table S3: Sequences of DNA bait probes ordered to IDT. ZnF268 target sequence underlined.