

Supplementary Materials: MetaEnzyme

Anonymous Authors

1 GENERAL PROTEINS FOR SEQDESIGN TASK

1.1 Baselines

The histogram depicted in Figure 5 within the main body, along with the comprehensive information provided in Table 1, showcases a selection of noteworthy works in the field. These works, predominantly open-source and easily reproducible, offer valuable insights into computational enzyme design. Notably, studies falling into Group 1 and Group 3 of Table 1 have been trained on a relatively modest-scale CATH training set, comprising approximately 18k training pairs. Despite the modest training data, earlier benchmarks such as Structured Transformer [5] and GVP-GNN [6] remain influential, boasting competitive performances and lightweight architectures. Additionally, ProteinMPNN [1] by Baker’s group has garnered attention for its advancements in performance and speed, backed by impressive biological validation experiments. PiFold [3] follows suit with further efficiency and effectiveness enhancements. Noteworthy among the selections in Group 2 is ESM-IF [4], distinguished by its robust open-source nature and robust data augmentation strategies, trained on a vast dataset of approximately 12M training pairs sourced from AlphaFoldDB [12]. It’s worth mentioning that our MetaEnzyme architectures share similarities with ESM-IF, hence we choose ESM-IF as a primary baseline for comparison in our study.

1.2 Detailed Comparison for General Protein Sequence Design

Comparison of the CATH, Ts50, and Ts500 datasets using perplexity and AAR metrics, as shown in Table 1.

2 DETAILS OF IN VITRO WET EXPERIMENTS

A schematic representation of the structure of the reversible glycine cleavage system (rGCS)[8, 10, 13] is shown in Figure 1.

2.1 Enzyme Preparation

The genes coding for H, P, and T proteins were amplified from *E. coli* K12 genomic DNA, and then cloned into the expression vector pET28a (NdeI and XhoI). Then synthesized the genes coding for the mutants of P-protein. *E. coli* BL21 (DE3) harboring the resulting constructs was cultured in LB medium supplemented with 50 mg/L of kanamycin or 100mg/L of ampicillin at 37 °C until the OD600 of the culture reached 0.6-0.8. Isopropyl-beta-D-thiogalactopyranoside (IPTG) was added to a final concentration of 0.2mM to induce protein expression for 12 h at 16 °C. The H-protein was expressed with 150 μ M lipoic acid added exogenously to obtain lipoylated H-protein directly.

All the proteins were purified using His-tag affinity chromatography (AKTA, GE Healthcare, USA) equipped with a nickel column (HisTrapTM HP, 5 mL). Buffer A containing 500mM NaCl, 50mM Tris-HCl, and 20mM imidazole (pH = 7.4) was used to elute non-target proteins, and buffer B containing 500mM NaCl, 50mM Tris-HCl, and 500mM imidazole (pH = 7.4) was used to elute the target

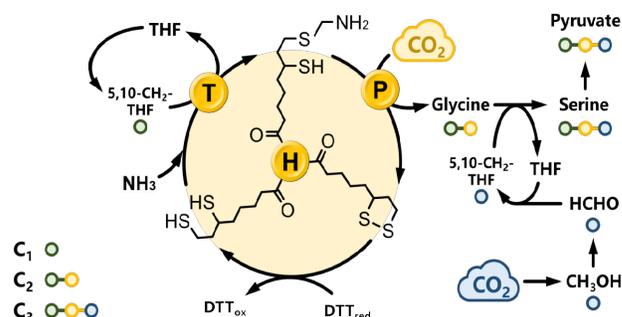


Figure 1: Conceptual representation of the structure of the reversible glycine cleavage system (rGCS). The glycine cleavage system comprises four proteins: T, P, L, and H proteins. These proteins do not form stable complexes, with the inherent carbon-fixing role of the P protein being the primary focus of our investigation. Specifically, T-protein is responsible for fixing one carbon (formaldehyde), P-protein fixes another carbon (carbon dioxide), and H-protein shuttles between the two enzymes to facilitate the transfer of amino methyl groups.

protein adsorbed in the nickel column. The purified enzymes were checked by SDS-PAGE.

2.2 Measurement of Glycine Production Rate

The rGCS reaction mixture contained 50 mM Tris-HCl (pH 7.5), 0.5 mM THF, 20 mM formaldehyde, 20 mM DTT, 0.5 μ M PLP, 50 mM NH₄Cl, 50mM NaHCO₃, 5 μ M P protein or mutant, 3 μ M T protein, and 40 μ M H protein. Glycine concentration in the reaction mixture was determined by HPLC after pre-column dansyl chloride derivatization. To this end, 40 μ L of a reaction mixture was mixed with 160 μ L of 0.2M NaHCO₃ and 200 μ L of 20 mM dansyl chloride in acetonitrile. Derivatization was carried out at 30°C for 30 minutes. The samples were analyzed on an Agilent InfinityLab Poroshell HPH-C18 4.6×100 mm, 2.7-Micron column (Agilent, USA) and Agilent 1260 Infinity HPLC system at DAD 338 nm after OPA derivatization (Agilent, USA) were used to measure the concentration of glycine. The mobile phase is composed of acetonitrile and 20mM potassium phosphate buffer pH6.0 (25:75 v/v) at a flow rate of 0.8 mL/min.

REFERENCES

- [1] Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J Ragotte, Lukas F Milles, Basile IM Wicky, Alexis Courbet, Rob J de Haas, Neville Bethel, et al. 2022. Robust deep learning-based protein sequence design using ProteinMPNN. *Science* 378, 6615 (2022), 49–56.
- [2] Zhangyang Gao, Cheng Tan, Stan Li, et al. 2022. AlphaDesign: A graph protein design method and benchmark on AlphaFoldDB. *arXiv preprint arXiv:2202.01079* (2022).
- [3] Zhangyang Gao, Cheng Tan, and Stan Z Li. 2022. PiFold: Toward effective and efficient protein inverse folding. *arXiv preprint arXiv:2209.12643* (2022).

#	Models	Perplexity					AAR(%)				
		All	Short	Single-chain	Ts50	Ts500	All	Short	Single-chain	Ts50	Ts500
	*Natural frequencies [4]	17.97	18.12	18.03	-	-	9.5	9.6	9.0	-	-
	SPIN [7]	-	-	-	-	-	-	-	-	30.3	30.3
	SPIN2 [9]	-	12.11	12.61	-	-	-	-	-	33.6	36.6
	StrucTransformer [5]	6.85	8.54	9.03	5.60	5.16	36.4	28.3	27.6	42.40	44.66
	Structured GNN [6]	6.55	8.31	8.88	5.40	4.98	37.3	28.4	28.1	43.89	45.69
	*ESM-IF [4]	6.44	8.18	6.33	-	-	38.3	31.3	38.5	-	-
	GCV [11]	6.05	7.09	7.49	5.09	4.72	37.64	32.62	31.10	47.02	47.74
	*GVP-GNN-large [6]	6.17	7.68	6.12	-	-	39.2	32.6	<u>39.4</u>	-	-
	GVP-GNN [6]	5.29	7.10	7.44	4.71	4.20	40.2	32.1	32.0	44.14	49.14
	AlphaDesign [2]	6.30	7.32	7.63	5.25	4.93	41.31	34.16	32.66	48.36	49.23
	ProteinMPNN [1]	4.61	6.21	6.68	3.93	3.53	45.96	36.35	34.43	54.43	58.08
	PiFold [3]	4.55	<u>6.04</u>	6.31	3.86	3.44	51.66	<u>39.84</u>	38.53	<u>58.72</u>	60.42
2	ESM-IF [4]	<u>3.99</u>	6.30	6.29	<u>3.43</u>	<u>3.34</u>	<u>52.51</u>	34.74	34.25	56.66	<u>60.85</u>
3	Ours (MetaEnzyme)	3.88	5.24	5.36	2.84	3.21	54.94	40.92	39.50	62.68	60.97

Table 1: Comparison on the CATH, Ts50, and Ts500 datasets. The best results are bolded, and the second best underlined. Group 1 & 3 are trained on a tiny CATH training set only, while Group 2 is trained on a large CATH+AlphaDB training set. The *Short* and *Single-chain* are subsets of CATH test set *ALL*. *: evaluated on CATH 4.3; Others: evaluated on CATH 4.2.

- [4] Chloe Hsu, Robert Verkuil, Jason Liu, Zeming Lin, Brian Hie, Tom Sercu, Adam Lerer, and Alexander Rives. 2022. Learning inverse folding from millions of predicted structures. *bioRxiv* (2022).
- [5] John Ingraham, Vikas Garg, Regina Barzilay, and Tommi Jaakkola. 2019. Generative models for graph-based protein design. *Advances in neural information processing systems* 32 (2019).
- [6] Bowen Jing, Stephan Eismann, Patricia Suriana, Raphael JL Townshend, and Ron Dror. 2020. Learning from protein structure with geometric vector perceptrons. *arXiv preprint arXiv:2009.01411* (2020).
- [7] Zhixiu Li, Yuedong Yang, Eshel Faraggi, Jian Zhan, and Yaoqi Zhou. 2014. Direct prediction of profiles of sequences compatible with a protein structure by neural networks with fragment-based local and energy-based nonlocal profiles. *Proteins: Structure, Function, and Bioinformatics* 82, 10 (2014), 2565–2573.
- [8] Jianming Liu, Han Zhang, Yingying Xu, Hao Meng, An Ping Zeng, Nathalie Le Bot, Enda Bergin, and Fiona Gillespie. 2023. Turn air-captured CO₂ with methanol into amino acid and pyruvate in an ATP/NAD(P)H-free chemoenzymatic system. *Nature Communications* (2023).
- [9] James O’Connell, Zhixiu Li, Jack Hanson, Rhys Heffernan, James Lyons, Kuldip Paliwal, Abdollah Dehzangi, Yuedong Yang, and Yaoqi Zhou. 2018. SPIN2: Predicting sequence profiles from protein structures using deep neural networks. *Proteins: Structure, Function, and Bioinformatics* 86, 6 (2018), 629–633.
- [10] Kazuko Okamura-Ikeda, Harumi Hosaka, Nobuo Maita, Kazuko Fujiwara, Akiyasu C. Yoshizawa, Atsushi Nakagawa, and Hisaaki Taniguchi. 2010. Crystal Structure of Aminomethyltransferase in Complex with Dihydropolyl-H-Protein of the Glycine Cleavage System. *Journal of Biological Chemistry* 285, 24 (2010), 18684–18692.
- [11] Cheng Tan, Zhangyang Gao, Jun Xia, and Stan Z Li. 2022. Generative de novo protein design with global context. *arXiv preprint arXiv:2204.10673* (2022).
- [12] Mihaly Varadi, Stephen Anyango, Mandar Deshpande, Sreenath Nair, Cindy Natassia, Galabina Yordanova, David Yuan, Oana Stroe, Gemma Wood, Agata Laydon, et al. 2022. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic acids research* 50, D1 (2022), D439–D444.
- [13] Hao Yu, Xueqing Hu, Yingru Zhang, Jiajia Wang, Zhongya Ni, Yan Wang, and Huirong Zhu. 2023. GLDC promotes colorectal cancer metastasis through epithelial-mesenchymal transition mediated by Hippo signaling pathway. (2023).