Computational Design of Monomeric Variants of Multimeric Enzymes

Jakub Lála[*]
Department of Materials
Imperial College London
London, United Kingdom
jakublala@gmail.com

Arnav Cheruku[*]
Department of Life Sciences
Imperial College London
London, United Kingdom
archeruku@gmail.com

Stefano Angioletti-Uberti
Department of Materials
Imperial College London
AminoAnalytica Ltd.
Nanograb Ltd.
London, United Kingdom
sangiole@imperial.ac.uk

Abstract

Multimeric enzymes often place catalytic residues at subunit interfaces, coupling activity to correct assembly and concentration. The need for correct assembly from multiple units complicates bacterial expression, scale-up, and implementation into industrial bioprocesses. By construction, a single-chain monomer folding into the exact same active-site geometry could, in principle, alleviate these challenges, while still retaining its catalytic activity. Following this reasoning, we present a general computational design protocol to monomerize multimeric enzymes while preserving their active-site geometry. Our method combines Monte Carlo optimization with an energy-based formalism, specifying design constraints as energy terms defined through the outputs of a protein-folding algorithm. Specifically, by driving the sequence search to active-site geometries close to that of the multimeric enzyme, we generate monomeric variants with low RMSD to the multimeric active site, in line with commonly used thresholds to filter potential candidates for lab testing. As a case study, we redesign the homotetrameric formolase enzyme, a critical synthetic enzyme for conversion of carbon dioxide into valuable C_3 chemicals. Given the protocol's generality, we believe these results represent an important foundational step toward speeding up the search for industrially practical synthetic enzyme mimics.

1 Introduction

Enzymes are ubiquitous in nature, and much of biotechnology aims to replicate their defining property: high selectivity and high-turnover catalysis under process conditions. Nature, guided by hundreds of millions of years of evolution, provides an enormous repertoire. Many enzymes are used directly, while others serve as starting points for optimization. The latter is often needed because natural enzymes evolve under constraints unlike those of industrial use. For example, natural enzymes evolve to work at moderate temperature, near-neutral pH, and in dilute conditions while catalyzing multiple reactions, whereas industrial processes often require stability in harsher environments and strict selectivity for a single transformation.

For multimeric enzymes – whose catalytic sites often form at subunit interfaces – the mismatch between natural selection and process needs is especially pronounced. First, subunit dissociation – driven by dilution, shear, pH shifts, or surface attachment – distorts interfacial geometry, reducing activity [Fernandez-Lafuente, 2009, Mateo et al., 2020]. Second, yields and lot-to-lot consistency suffer because multiple polypeptides must be co-expressed at defined stoichiometries. Imbalances create orphan subunits that degrade or aggregate, making expression tuning another process variable.

[*] Equal contribution.

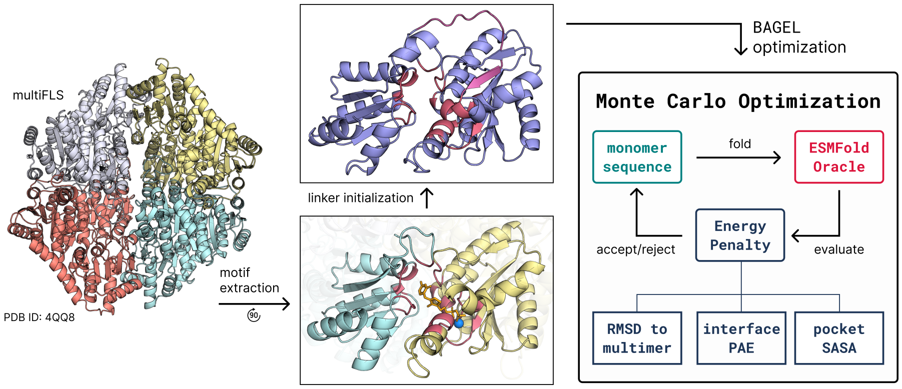

Figure 1: **Overview of the Monomerization Protocol.** First, we retrieve the multimeric formolase from the Protein Data Bank (PDB), extract a functional motif encompassing the active site, and initialize a random linker (magenta) between contiguous parts of the two chains forming the interface (blue) of the active site (red). Second, we use this initial sequence to run an energy-driven search through sequence design space with BAGEL [Lála et al., 2025].

Third, immobilization and reactor integration are harder for bulky multimers: they pack less efficiently into meso-/macroporous carriers (e.g., metal-organic frameworks), face orientation and diffusional constraints, and often require interfacial cross-linking to prevent dissociation, typically at the expense of specific activity [Sheldon and van Pelt, 2013, Bayne et al., 2013, Wang et al., 2020].

To at least partially alleviate these problems, one could try to replace a complex multimeric enzyme with a single-chain mimic. In order to reach this goal, a central question is: *is there a monomeric amino acid sequence that preserves the catalytic geometry of a multimeric enzyme's interfacial active site?* In a broader context, this question is known as the *motif scaffolding* problem, recently addressed using a multitude of machine-learning–based algorithms Wang et al. [2022], Hansen et al. [2024], Ahern et al. [2025]. Nevertheless, monomerization remains largely handcrafted, relying on manual interface redesign and residue swaps, either with the 3D structure [Liu et al., 2020] or without it [Tong et al., 2005]. With progress in deep learning-based folding, a general, practical, on-demand *in silico* protocol is now within reach.

To that end, and to the best of our knowledge, we propose a general approach that enables the monomerization of any multimeric enzyme, whose active site lies at the interface of its protein chains, depicted in Figure 1. As a test case, we design a monomeric variant of the tetrameric formolase (FLS) enzyme (PDB ID 4QQ8, UniProt Q9F4L3). FLS is a key component of an enzymatic cascade that converts carbon dioxide into valuable C_3 chemicals. For details, refer to Appendix A.

2 Method

General Protocol. Naively trying to find the right monomeric sequence that will fold into the exact geometry of a multimeric, interfacial active site is a prohibitively difficult task. Therefore, we build on the previously developed energy landscape framework BAGEL [Lála et al., 2025], formalizing the design task as a minimization over a set of design constraints.

In practice, we minimize an energy $E_\Omega = \sum_i w_i E_i$, which is a weighted combination of energy terms E_i, each of which is a function of folded structure X and confidence metrics Φ, retrieved from a protein-folding oracle. To minimize E_Ω, we iteratively mutate the associated protein sequence s to s', employing Monte Carlo (MC) optimization with the usual Metropolis acceptance criterion:

$$\alpha(s \to s'; T) = \min \left[1, \exp\left(-\frac{E_\Omega(X', \Phi') - E_\Omega(X, \Phi)}{T} \right) \right] \tag{1}$$

T is the effective temperature balancing exploration and exploitation of the search space. $(X, \Phi) = \mathcal{F}(s)$ and $(X', \Phi') = \mathcal{F}(s')$ are the folded structures and folding confidence metrics corresponding to sequences s and s' respectively. We use ESMFold [Lin et al., 2023] to model the folding oracle $\mathcal{F}$. We alternate between low and high temperatures, effectively performing simulated tempering [Marinari and Parisi, 1992] to improve the convergence of the minimization.

On top of common energy terms usually employed to achieve high-confidence designs, fully provided in Table 1 in Appendix B, we highlight three other essential energy terms for driving the sequence search. First, we use a template-matching energy that penalizes deviations in the geometry of active-site heavy atoms between the designed monomer (monoFLS) and the original multimeric enzyme (multiFLS). This deviation is quantified as the root mean squared deviation (RMSD) of the active site, also referred to as a *motif*. Second, to improve the stability of the (former) interface, we minimize the interface Predicted Aligned Error (iPAE). This ensures the two linked domains behave as a single rigid body, i.e., do not move relative to one another. Lastly, we ensure that the monomerization does not obstruct the active site by maximizing its solvent-accessible surface area (SASA), thus preserving substrate accessibility.

Monomeric Formolase. Given the vast space of protein sequences, we need to wisely choose the starting point of our minimization. For FLS, we initialize the search with a functional motif derived from the crystal structure of FLS (PDB ID 4QQ8). We extract residues within 5 Å of a reaction intermediate, similar to Wang et al. [2022]. The intermediate dihydroxyacetone-thiamine-pyrophosphate (DHA-TPP) and Mg^{2+} were retrieved and superimposed into the crystal structure from Siegel et al. [2015]. This yields a functional motif of 12 residue islands (39 residues in total), which we consider essential for catalytic activity and thus do not mutate during the optimization. Scaffolding these many islands in such a way as to provide a catalytically active interfacial geometry is non-trivial, hence we continue the motif extraction by finding the two contiguous sequences from each multimeric domain that connect all the islands from each chain, resulting in two *effective* domains (blue and yellow in the middle of Figure 1). We fuse the sequences of these domains with a randomly initialized mutable 15-residue linker, yielding a single monomeric chain that can be used as a starting point for our optimization procedure. In this work, we run 10 independent optimization runs, each initialized with a different linker, with the entire sequence apart from the immutable residue islands free to evolve throughout.

Orthogonal Folding and Molecular Simulations. After generating candidates with MC, we further filter our designs with *orthogonal* protein-folding algorithms, namely AlphaFold2 [Jumper et al., 2021], Boltz-2 [Passaro et al., 2025], and Chai-1 [Boitreaud et al., 2024] – both within the single-sequence (ESM-2) and multiple sequence alignment (MSA) modes. Lastly, to establish the stability of the active site pocket, we run vanilla molecular dynamics for 500 ns for both multiFLS and the designed monoFLS variants, with simulation details provided in Appendix D.

3 Results

Optimization Results. Starting from high values of motif RMSD, iPAE and active-site SASA, the optimization yields low-energy designs, shown on an example in Figure 4 in Appendix B. From each run, we select up to 10 sequences with the lowest energy E_Ω that also meet the following criteria: a motif RMSD of less than 1.0 Å, a pLDDT score higher than 0.75, and a pTM score above 0.7 (all retrieved from ESMFold-derived structures). We choose this motif RMSD threshold as previously described by [Lauko et al., 2025]. This results in 40 monoFLS design variants.

Orthogonal Folding. To assess the robustness of our designs beyond ESMFold, we evaluated them with multiple orthogonal folding algorithms. Motif RMSD distributions across the 40 monoFLS variants (Figure 5 in Appendix C) show close agreement between ESMFold and Chai-1, while AlphaFold2 and Boltz-2 predict higher RMSD values, suggesting poorer active-site conservation. Interestingly, rank correlations (Figure 6) are stronger between ESMFold and AlphaFold2/Boltz-2 than with Chai-1. Based on the average motif RMSD across all models, we identified four top *successful* candidates (Figure 2; Table 2 in Appendix C). These are also the only designs supported by at least one orthogonal model – specifically Chai-1 (MSA) – with a motif RMSD below 1.5 Å, a threshold previously used to identify successful designs [Ahern et al., 2025]. This convergence suggests these monoFLS variants are the most promising for experimental validation.

These variants all come from the same optimization trajectory, differing by at most ten point mutations with effectively near-identical ESMFold-derived structures, shown in Figure 2A. Nevertheless, once folded with other oracles, the same variant exhibits a different geometry near the active site, shown in Figure 2B. This then extends to the other variants as well (see Figure 7 in Appendix C). We do

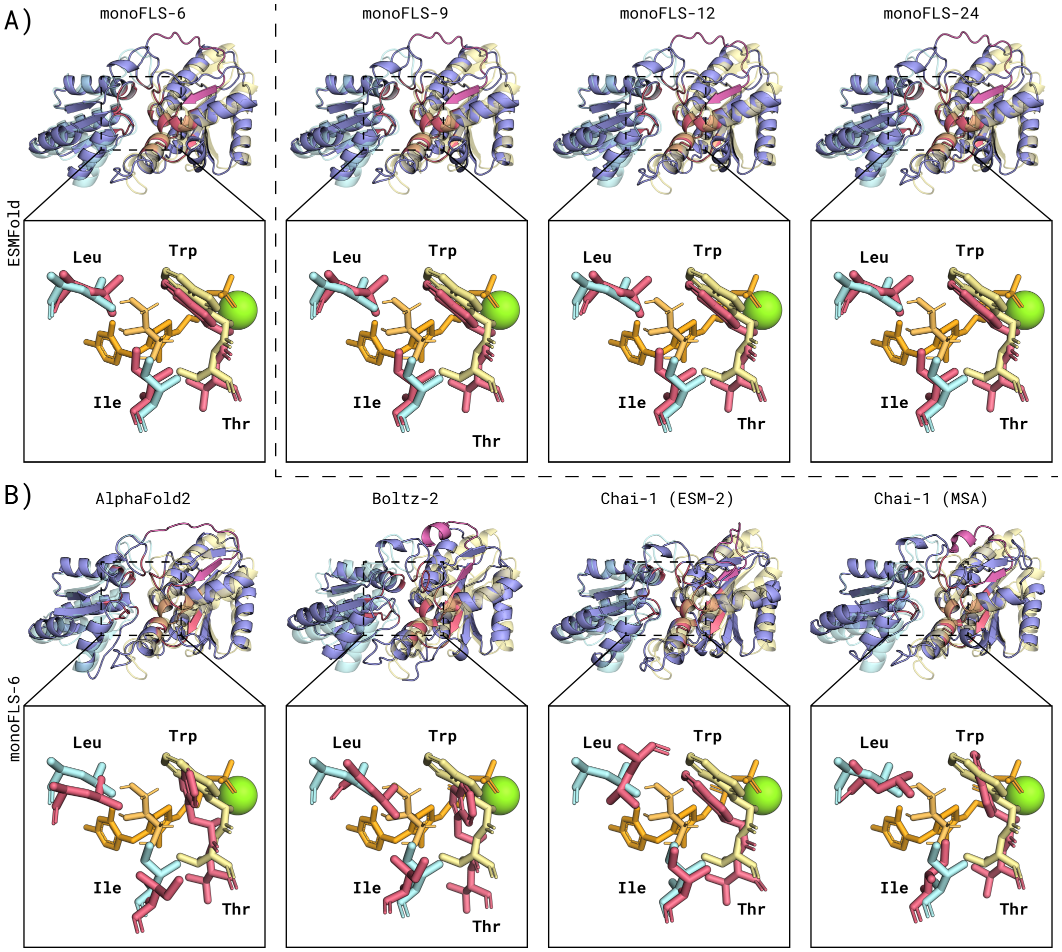

Figure 2: **Folded Structures of Final Designs.** A) ESMFold-predicted structures for the four final monoFLS variants show near-identical global folds and entry-site geometries. B) For monoFLS-6, different folding algorithms predict distinct folds at the entry. In all panels, we show the full monoFLS with an inset of the entry to the active site. We compare multiFLS (cyan, yellow) to the monoFLS scaffold (blue) and motif (red). We include the DHA-TPP (orange) intermediate and magnesium ion (green). For clarity, only residues at the active-site entry are shown, highlighting our focus on maintaining pocket accessibility.

not further investigate whether this comes from the inaccuracy of the algorithms, or from genuine misfolding.

Molecular Simulations. Figure 3 indicates monoFLS-9 is similarly stable to the native multiFLS after 500 ns of simulation, both in terms of domain–domain separation and motif RMSD. If we were to choose a single sequence for experimental validation, these results suggest monoFLS-9 should be brought forward. Detailed interpretation of the simulations is given in Appendix D.

4 Discussion

Optimization Results. We show that our protocol is able to monomerize multiFLS, with a relatively low motif RMSD and iPAE, and a reasonable active-site SASA. Despite the SASA decreasing from the initialized sequence (before minimization) in Figure 4, we consider the strategy of SASA-maximization as successful in keeping the active-site pocket accessible, when relying on the geometry of the entry residues produced by the folding algorithms. Nevertheless, we notice the poor sequence diversity of the monoFLS designs, and only a single run yielding results passing our metric filters. We label this as a limitation of the current MC protocol, which cannot produce diverse sequences within

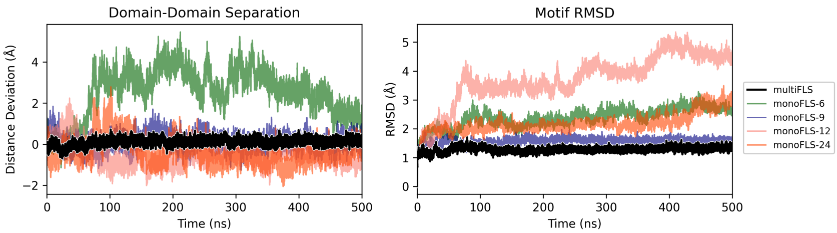

Figure 3: **Molecular Dynamics Trajectories.** (left) Time evolution of the domain–domain separation deviation. (right) Time evolution of the motif RMSD to the multiFLS crystal structure.

a single optimization. We could achieve greater diversity by initializing many more independent runs, increasing the temperature during the tempering regime, or with a more sophisticated framework, such as GFlowNets [Jain et al., 2023].

Folding Models. Our protocol is inherently limited by the quality of the predictions of the protein-folding algorithms. We used multiple protein-folding algorithms to filter out adversarial designs by ESMFold. Nevertheless, given the algorithms' similar architectures and training data, these oracles do not constitute truly orthogonal predictions. Still, despite observing effectively identical geometries in the ESMFold-derived structures, other models fold the active site quite differently even though the sequences are very similar (see Figure 7 in Appendix C). We hypothesize that the structural differences from minor mutations across orthogonal models could be leveraged for more effective filtering, helping to avoid adversarial ESMFold sequences as lab candidates. Furthermore, when compared to popular frameworks such as RFDiffusion Ahern et al. [2025], given that we design directly in sequence space, and such frameworks *always* rely a final validation re-folding, in the limit of sufficient algorithmic optimization and diverse enzyme classes, we expect to have these methods behave similarly.

Stability and Pocket Collapse. First, molecular simulations show the two domains in monoFLS are less stable than in multiFLS. To tackle this, we plan to employ additional energy terms that would further improve the formed interface, on top of the already employed iPAE energy. Moreover, the optimization could potentially be run for longer, as the iPAE does not seem to converge yet (Figure 4 in Appendix B). Importantly, simulations let us pinpoint dynamics-driven issues without synthesis, so we will continue to use them as oracles for design filtering. Second, despite seeing the entry residues retain the open geometry in the folded structures, after the initial equilibration, we observe the active-site pocket collapsing. We do not necessarily attribute this to a failure of our designs, as we observe the same collapse in multiFLS as well (Figure 8). We hypothesize this is because the co-factor is not present, or due to a general inaccuracy of the force field used. Further work will explore other force fields, including ones that can model co-factors [Takaba et al., 2024].

Experimental Validation. It has been shown that monomeric variants can have some key properties degraded. Upon monomerization, aminopimelate decarboxylase has attenuated catalytic activity [Peverelli et al., 2016], anthranilate phosphoribosyltransferase has worsened thermostability [Schwab et al., 2008], or an insulin-degrading enzyme loses its regulatory properties [Song et al., 2010]. We hypothesize that our monomerization procedure might lead to similar worsening of such properties, at the benefit of creating a more compact version. Therefore, we view these designs as the first batch of sequences for subsequent optimization through directed evolution [Arnold, 2017, Yang et al., 2019].

5 Conclusion

We proposed a general protocol to monomerize multimeric enzymes with an interfacial active site, useful in many bioindustrial processes. On the specific example of homotetrameric formolase, we found four candidate monomeric sequences that we consider the most promising for subsequent experimental validation. We view our method as a foundational step toward the general simplification and miniaturization of multimeric proteins. We recognize the limitation of showcasing the method on a single enzyme only, and thus invite the community to use our protocol on other multimeric enzymes, testing the monomerized variants in their respective experimental assays.

Acknowledgements

J.L. acknowledges the President's PhD scholarship at Imperial College London for funding. We acknowledge computational resources and support provided by the Imperial College Research Computing Service (`http://doi.org/10.14469/hpc/2232`). This research was also supported by grants from NVIDIA and utilized NVIDIA's GPU access through the Academic Grant for Simulation and Modeling.

Code and data availability

The scripts and data can be provided on the request by the reviewers, and will be fully released upon the acceptance of the manuscript. All generations were run with `biobagel` v0.1.4 package readily available through PyPI.

Use of Large Language Models

We used a large language model (LLM) for copy-editing only (grammar and phrasing). No ideas, methods, results, figures, code, or references were generated by the LLM. All text was reviewed and verified by the authors.

References

Roberto Fernandez-Lafuente. Stabilization of multimeric enzymes: Strategies to prevent subunit dissociation. *Enzyme and Microbial Technology*, 45(6–7):405–418, December 2009. ISSN 0141-0229. doi: 10.1016/j.enzmictec.2009.08.009. URL `http://dx.doi.org/10.1016/j.enzmictec.2009.08.009`.

Cesar Mateo, Benevides C. C. Pessela, Manuel Fuentes, Rodrigo Torres, Lorena Betancor, Aurelio Hidalgo, Gloria Fernandez-Lorente, Roberto Fernandez-Lafuente, and Jose M. Guisan. *Stabilization of Multimeric Enzymes via Immobilization and Further Cross-Linking with Aldehyde-Dextran*, page 175–187. Springer US, 2020. ISBN 9781071602157. doi: 10.1007/978-1-0716-0215-7_11. URL `http://dx.doi.org/10.1007/978-1-0716-0215-7_11`.

Roger A. Sheldon and Sander van Pelt. Enzyme immobilisation in biocatalysis: why, what and how. *Chem. Soc. Rev.*, 42(15):6223–6235, March 2013. ISSN 1460-4744. doi: 10.1039/c3cs60075k. URL `http://dx.doi.org/10.1039/C3CS60075K`.

Lauren Bayne, Rein V. Ulijn, and Peter J. Halling. Effect of pore size on the performance of immobilised enzymes. *Chemical Society Reviews*, 42(23):9000, 2013. ISSN 1460-4744. doi: 10.1039/c3cs60270b. URL `http://dx.doi.org/10.1039/c3cs60270b`.

Xiaoliang Wang, Pui Ching Lan, and Shengqian Ma. Metal–organic frameworks for enzyme immobilization: Beyond host matrix materials. *ACS Central Science*, 6(9):1497–1506, August 2020. ISSN 2374-7951. doi: 10.1021/acscentsci.0c00687. URL `http://dx.doi.org/10.1021/acscentsci.0c00687`.

Jue Wang, Sidney Lisanza, David Juergens, Doug Tischer, Joseph L. Watson, Karla M. Castro, Robert Ragotte, Amijai Saragovi, Lukas F. Milles, Minkyung Baek, Ivan Anishchenko, Wei Yang, Derrick R. Hicks, Marc Expòsit, Thomas Schlichthaerle, Jung-Ho Chun, Justas Dauparas, Nathaniel Bennett, Basile I. M. Wicky, Andrew Muenks, Frank DiMaio, Bruno Correia, Sergey Ovchinnikov, and David Baker. Scaffolding protein functional sites using deep learning. *Science*, 377(6604):387–394, July 2022. ISSN 1095-9203. doi: 10.1126/science.abn2100. URL `http://dx.doi.org/10.1126/science.abn2100`.

Anders Lønstrup Hansen, Frederik Friis Theisen, Ramon Crehuet, Enrique Marcos, Nushin Aghajari, and Martin Willemoës. Carving out a glycoside hydrolase active site for incorporation into a new protein scaffold using deep network hallucination. *ACS Synthetic Biology*, 13(3):862–875, February 2024. ISSN 2161-5063. doi: 10.1021/acssynbio.3c00674. URL `http://dx.doi.org/10.1021/acssynbio.3c00674`.

Woody Ahern, Jason Yim, Doug Tischer, Saman Salike, Seth M. Woodbury, Donghyo Kim, Indrek Kalvet, Yakov Kipnis, Brian Coventry, Han Raut Altae-Tran, Magnus Bauer, Regina Barzilay, Tommi S. Jaakkola, Rohith Krishna, and David Baker. Atom level enzyme active site scaffolding using rfdiffusion2. April 2025. doi: 10.1101/2025.04.09.648075. URL http://dx.doi.org/10.1101/2025.04.09.648075.

Hu Liu, Mingming Cao, Ying Wang, Bo Lv, and Chun Li. Bioengineering oligomerization and monomerization of enzymes: learning from natural evolution to matching the demands for industrial applications. *Critical Reviews in Biotechnology*, 40(2):231–246, January 2020. ISSN 1549-7801. doi: 10.1080/07388551.2019.1711014. URL http://dx.doi.org/10.1080/07388551.2019.1711014.

Yufeng Tong, David Hughes, Lisa Placanica, and Matthias Buck. When monomers are preferred: A strategy for the identification and disruption of weakly oligomerized proteins. *Structure*, 13(1): 7–15, January 2005. ISSN 0969-2126. doi: 10.1016/j.str.2004.10.018. URL http://dx.doi.org/10.1016/j.str.2004.10.018.

Jakub Lála, Ayham Al-Saffar, and Stefano Angioletti-Uberti. Bagel: Protein engineering via exploration of an energy landscape. July 2025. doi: 10.1101/2025.07.05.663138. URL http://dx.doi.org/10.1101/2025.07.05.663138.

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

E Marinari and G Parisi. Simulated tempering: A new monte carlo scheme. *Europhysics Letters (EPL)*, 19(6):451–458, July 1992. ISSN 1286-4854. doi: 10.1209/0295-5075/19/6/002. URL http://dx.doi.org/10.1209/0295-5075/19/6/002.

Justin B. Siegel, Amanda Lee Smith, Sean Poust, Adam J. Wargacki, Arren Bar-Even, Catherine Louw, Betty W. Shen, Christopher B. Eiben, Huu M. Tran, Elad Noor, Jasmine L. Gallaher, Jacob Bale, Yasuo Yoshikuni, Michael H. Gelb, Jay D. Keasling, Barry L. Stoddard, Mary E. Lidstrom, and David Baker. Computational protein design enables a novel one-carbon assimilation pathway. *Proceedings of the National Academy of Sciences*, 112(12):3704–3709, March 2015. ISSN 1091-6490. doi: 10.1073/pnas.1500545112. URL http://dx.doi.org/10.1073/pnas.1500545112.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, July 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03819-2. URL http://dx.doi.org/10.1038/s41586-021-03819-2.

Saro Passaro, Gabriele Corso, Jeremy Wohlwend, Mateo Reveiz, Stephan Thaler, Vignesh Ram Somnath, Noah Getz, Tally Portnoi, Julien Roy, Hannes Stark, David Kwabi-Addo, Dominique Beaini, Tommi Jaakkola, and Regina Barzilay. Boltz-2: Towards accurate and efficient binding affinity prediction. June 2025. doi: 10.1101/2025.06.14.659707. URL http://dx.doi.org/10.1101/2025.06.14.659707.

Jacques Boitreaud, Jack Dent, Matthew McPartlon, Joshua Meier, Vinicius Reis, Alex Rogozhnikov, and Kevin Wu. Chai-1: Decoding the molecular interactions of life. October 2024. doi: 10.1101/2024.10.10.615955. URL http://dx.doi.org/10.1101/2024.10.10.615955.

Anna Lauko, Samuel J. Pellock, Kiera H. Sumida, Ivan Anishchenko, David Juergens, Woody Ahern, Jihun Jeung, Alexander F. Shida, Andrew Hunt, Indrek Kalvet, Christoffer Norn, Ian R. Humphreys, Cooper Jamieson, Rohith Krishna, Yakov Kipnis, Alex Kang, Evans Brackenbrough, Asim K. Bera, Banumathi Sankaran, K. N. Houk, and David Baker. Computational design of serine hydrolases. *Science*, 388(6744), April 2025. ISSN 1095-9203. doi: 10.1126/science.adu2454. URL http://dx.doi.org/10.1126/science.adu2454.

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, 2023. URL https://arxiv.org/abs/2203.04115.

Kenichiro Takaba, Anika J. Friedman, Chapin E. Cavender, Pavan Kumar Behara, Iván Pulido, Michael M. Henry, Hugo MacDermott-Opeskin, Christopher R. Iacovella, Arnav M. Nagle, Alexander Matthew Payne, Michael R. Shirts, David L. Mobley, John D. Chodera, and Yuanqing Wang. Machine-learned molecular mechanics force fields from large-scale quantum chemical data. *Chemical Science*, 15(32):12861–12878, 2024. ISSN 2041-6539. doi: 10.1039/d4sc00690a. URL http://dx.doi.org/10.1039/D4SC00690A.

Martin G. Peverelli, Tatiana P. Soares da Costa, Nigel Kirby, and Matthew A. Perugini. Dimerization of bacterial diaminopimelate decarboxylase is essential for catalysis. *Journal of Biological Chemistry*, 291(18):9785–9795, April 2016. ISSN 0021-9258. doi: 10.1074/jbc.m115.696591. URL http://dx.doi.org/10.1074/jbc.M115.696591.

Thomas Schwab, Darko Skegro, Olga Mayans, and Reinhard Sterner. A rationally designed monomeric variant of anthranilate phosphoribosyltransferase from sulfolobus solfataricus is as active as the dimeric wild-type enzyme but less thermostable. *Journal of Molecular Biology*, 376(2):506–516, February 2008. ISSN 0022-2836. doi: 10.1016/j.jmb.2007.11.078. URL http://dx.doi.org/10.1016/j.jmb.2007.11.078.

Eun Suk Song, David W. Rodgers, and Louis B. Hersh. A monomeric variant of insulin degrading enzyme (ide) loses its regulatory properties. *PLoS ONE*, 5(3):e9719, March 2010. ISSN 1932-6203. doi: 10.1371/journal.pone.0009719. URL http://dx.doi.org/10.1371/journal.pone.0009719.

Frances H. Arnold. Directed evolution: Bringing new chemistry to life. *Angewandte Chemie International Edition*, 57(16):4143–4148, November 2017. ISSN 1521-3773. doi: 10.1002/anie.201708408. URL http://dx.doi.org/10.1002/anie.201708408.

Kevin K. Yang, Zachary Wu, and Frances H. Arnold. Machine-learning-guided directed evolution for protein engineering. *Nature Methods*, 16(8):687–694, July 2019. ISSN 1548-7105. doi: 10.1038/s41592-019-0496-6. URL http://dx.doi.org/10.1038/s41592-019-0496-6.

Tanja G. Mosbacher, Michael Mueller, and Georg E. Schulz. Structure and mechanism of the thdp-dependent benzaldehyde lyase from pseudomonas fluorescens. *The FEBS Journal*, 272(23): 6067–6076, November 2005. ISSN 1742-4658. doi: 10.1111/j.1742-4658.2005.04998.x. URL http://dx.doi.org/10.1111/j.1742-4658.2005.04998.x.

Sean Poust, James Piety, Arren Bar-Even, Catherine Louw, David Baker, Jay D. Keasling, and Justin B. Siegel. Mechanistic analysis of an engineered enzyme that catalyzes the formose reaction. *ChemBioChem*, 16(13):1950–1954, July 2015. ISSN 1439-7633. doi: 10.1002/cbic.201500228. URL http://dx.doi.org/10.1002/cbic.201500228.

Nathan H. Chen, Karrera Y. Djoko, Frédéric J. Veyrier, and Alastair G. McEwan. Formaldehyde stress responses in bacterial pathogens. *Frontiers in Microbiology*, 7, March 2016. ISSN 1664-302X. doi: 10.3389/fmicb.2016.00257. URL http://dx.doi.org/10.3389/fmicb.2016.00257.

Tao Cai, Hongbing Sun, Jing Qiao, Leilei Zhu, Fan Zhang, Jie Zhang, Zijing Tang, Xinlei Wei, Jiangang Yang, Qianqian Yuan, Wangyin Wang, Xue Yang, Huanyu Chu, Qian Wang, Chun You, Hongwu Ma, Yuanxia Sun, Yin Li, Can Li, Huifeng Jiang, Qinhong Wang, and Yanhe Ma. Cell-free chemoenzymatic starch synthesis from carbon dioxide. *Science*, 373(6562):1523–1527, September 2021. ISSN 1095-9203. doi: 10.1126/science.abh4049. URL http://dx.doi.org/10.1126/science.abh4049.

Ahmad Abolpour Homaei, Reyhaneh Sariri, Fabio Vianello, and Roberto Stevanato. Enzyme immobilization: an update. *Journal of Chemical Biology*, 6(4):185–205, August 2013. ISSN 1864-6166. doi: 10.1007/s12154-013-0102-9. URL http://dx.doi.org/10.1007/s12154-013-0102-9.

J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, A. Courbet, R. J. de Haas, N. Bethel, P. J. Y. Leung, T. F. Huddy, S. Pellock, D. Tischer, F. Chan, B. Koepnick, H. Nguyen, A. Kang, B. Sankaran, A. K. Bera, N. P. King, and D. Baker. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378(6615):49–56, October 2022. ISSN 1095-9203. doi: 10.1126/science.add2187. URL `http://dx.doi.org/10.1126/science.add2187`.

Daniella Pretorius, Georgi I. Nikov, Kono Washio, Steve-William Florent, Henry N. Taunt, Sergey Ovchinnikov, and James W. Murray. Designing novel solenoid proteins with in silico evolution. April 2025. doi: 10.1101/2025.04.23.646631. URL `http://dx.doi.org/10.1101/2025.04.23.646631`.

Peter Eastman, Raimondas Galvelis, Raúl P. Peláez, Charlles R. A. Abreu, Stephen E. Farr, Emilio Gallicchio, Anton Gorenko, Michael M. Henry, Frank Hu, Jing Huang, Andreas Krämer, Julien Michel, Joshua A. Mitchell, Vijay S. Pande, João PGLM Rodrigues, Jaime Rodriguez-Guerra, Andrew C. Simmonett, Sukrit Singh, Jason Swails, Philip Turner, Yuanqing Wang, Ivy Zhang, John D. Chodera, Gianni De Fabritiis, and Thomas E. Markland. Openmm 8: Molecular dynamics simulation with machine learning potentials. *The Journal of Physical Chemistry B*, 128(1): 109–116, December 2023. ISSN 1520-5207. doi: 10.1021/acs.jpcb.3c06662. URL `http://dx.doi.org/10.1021/acs.jpcb.3c06662`.

James A. Maier, Carmenza Martinez, Koushik Kasavajhala, Lauren Wickstrom, Kevin E. Hauser, and Carlos Simmerling. ff14sb: Improving the accuracy of protein side chain and backbone parameters from ff99sb. *Journal of Chemical Theory and Computation*, 11(8):3696–3713, July 2015. ISSN 1549-9626. doi: 10.1021/acs.jctc.5b00255. URL `http://dx.doi.org/10.1021/acs.jctc.5b00255`.

Arkajyoti Sengupta, Zhen Li, Lin Frank Song, Pengfei Li, and Kenneth M. Merz. Parameterization of monovalent ions for the opc3, opc, tip3p-fb, and tip4p-fb water models. *Journal of Chemical Information and Modeling*, 61(2):869–880, February 2021. ISSN 1549-960X. doi: 10.1021/acs.jcim.0c01390. URL `http://dx.doi.org/10.1021/acs.jcim.0c01390`.

A Formolase Details and Significance

Formolase (FLS) is a synthetic, thiamine-pyrophosphate (TPP)-dependent carboligase engineered from benzaldehyde lyase (BAL) to condense formaldehyde into dihydroxyacetone (DHA), enabling a one-carbon (C1) assimilation route that funnels CO_2-derived formate/methanol into dihydroxyacetone phosphate (DHAP) via the "FLS pathway" [Siegel et al., 2015, Mosbacher et al., 2005]. FLS's catalytic efficiency toward formaldehyde is comparatively low ($k_{\text{cat}}/K_M \approx 4.7\ \mathrm{M}^{-1}\,\mathrm{s}^{-1}$), and at low formaldehyde concentration it also accumulates glycolaldehyde, reflecting the first carboligation step [Siegel et al., 2015, Poust et al., 2015]. Formaldehyde's inherent cytotoxicity further complicates *in vivo* deployment of this route [Chen et al., 2016]. Nevertheless, FLS is central to state-of-the-art cell-free carbon-utilization cascades (e.g., the artificial starch anabolic pathway, ASAP), where it can dominate enzyme mass requirements (reported at around 86% of total protein required in ASAP), highlighting FLS as a practical bottleneck for scale-up [Cai et al., 2021]. Monomerizing – and thus miniaturizing – FLS would reduce expression and immobilization burdens, particularly for porous carriers such as metal-organic frameworks, where multimer dissociation and pore-size constraints hinder large enzymes, thereby improving enzyme packing density, stability, and process economics [Fernandez-Lafuente, 2009, Homaei et al., 2013, Wang et al., 2020].

B Optimization Details

For a full description of the energy terms and the energy optimization procedure, refer to the original technical report on BAGEL [Lála et al., 2025]. Table 1 summarizes the philosophy behind choosing these terms and lists the corresponding weights; the optimization parameters are given in the caption. Note, however, given the stochastic nature of MC optimization, using these parameters will yield only qualitatively similar results across repeated runs.

Table 1: Energy terms used in the minimization within BAGEL with their descriptions and weights. Here, we use ESMFold as the folding oracle. Optimization parameters: $T_{\text{low}} = 0.1$, $T_{\text{high}} = 1.2$, $n_{\text{steps, high}} = 100$, $n_{\text{steps, low}} = 400$, $n_{\text{cycles}} = 100$, and `preserve_best_system_every_n_steps`= 500. The last parameter specifies how many MC steps are taken before reverting to the best sequence so far, which improves optimization convergence.

Energy Term	**Description**	**Weight**
Folded-structure terms		
`SurfaceAreaEnergy`	Reduces the amount of exposed surface area, by approximating the solvent-accessible surface area. With a negative weight, this term incentivizes maximizing exposed surface area.	-1.0
`HydrophobicEnergy`	Penalizes hydrophobic residues exposed on the solvent-accessible surface.	2.0
`GlobularEnergy`	Promotes compactness of the designed chain (penalizing overly extended conformations) to encourage a globular, well-packed fold, further aiding in miniaturization.	1.0
`TemplateMatchEnergy`	Limits deviation of the RMSD of a set of residues to a pre-defined template. We compute the RMSD on all the heavy atoms including the side chains.	5.0
Confidence-metric terms		
`PTMEnergy`	Encourages high predicted TM-score (pTM), biasing toward correct global topology of the folded model.	1.0
`OverallPLDDTEnergy`	Maximizes overall predicted Local Distance Difference Test (pLDDT) confidence across the model, promoting locally well-determined structure.	2.0
`PAEEnergy`	Minimizes interface Predicted Aligned Error (iPAE) across the specified residue groups (here, the two domains fused by a linker), encouraging them to behave as a rigid body, with a well-defined interface.	5.0

Figure 4 shows the evolution of the energy terms of the best system found so far throughout the optimization. This single run produced all of the four monoFLS variants. We clearly start off with a large motif RMSD (template-matching energy), showing that the randomly fused domains do not recover an active-site geometry close to that of the multiFLS. Moreover, as the optimization evolves, iPAE energy decreases as well, suggesting the domains start to form a well-packed interface. More noticeably, although we try to maximize SASA, the evolution shows a reduction of the SASA energy (unweighted). Nevertheless, we do not consider this a problem, as the initial structure was likely unstructured relative to our design objective, with a high motif RMSD, and thus likely exposed more of the surface in the pocket. Only later, once the motif RMSD drives the search to the correct geometry, does the SASA energy start to play the role of ensuring the pocket accessibility.

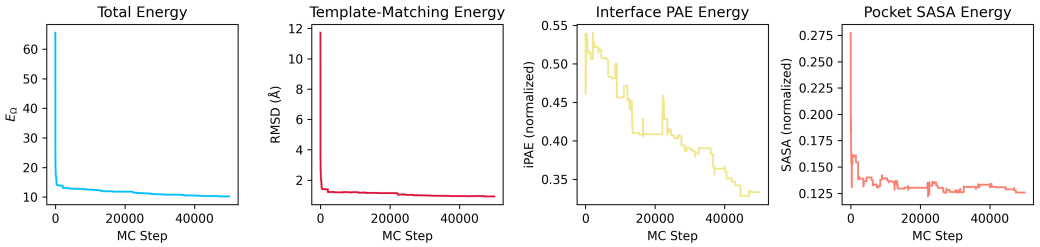

Figure 4: **Optimization Evolution.** System energy E_Ω, template-matching energy (i.e., motif RMSD), iPAE energy, and SASA of the active site pocket, as a function of the Monte Carlo steps of one of the best optimization runs, i.e., the one which produced all the successful monoFLS designs. All energies are provided unweighted.

C Further Results of Orthogonal Folding

Below we provide results of the orthogonal folding. Table 2 lists the four final monoFLS designs, reporting motif RMSD from each folding model and the full-protein RMSD from Chai-1 (MSA). The *full-protein RMSD* is the deviation of the monoFLS backbone between the orthogonally folded structures and the original ESMFold one. While common in RFDiffusion-based enzyme design [Dauparas et al., 2022], we do not use the full-protein RMSD for filtering. Prior work has also shown it can inversely correlate with success for solenoid proteins [Pretorius et al., 2025].

We report the full distributions of motif RMSD and full-protein RMSD in Figure 5, where Chai-1 generally yields lower RMSD values. Then, in Figure 6, we show the correlations between motif RMSD from ESMFold and each orthogonal folding algorithm respectively. We observe that if only using the previously discussed set of 40 designs that pass the filter of motif RMSD below 1.0 Å, we obtain almost no correlation. Only upon including more variants with larger ESMFold-derived motif RMSD do we see a better correlation, between ESMFold and the other models, with AlphaFold2 and Boltz-2 correlating more strongly than Chai-1. This is despite the behavior in Figure 5, which might suggest otherwise, given the very low RMSD values for Chai-1 structures. We do not investigate this further here, but note that filtering can obscure correlations between different folding algorithms, which could otherwise be used to increase confidence in specific designs. The decision of whether, when, and how to apply such filtering should therefore be made carefully and without bias.

Table 2: Final monomeric designs. All are within 39-41% sequence identity to the starting monomeric sequence with the randomly initialized linker.

	Motif RMSD					Backbone RMSD
Design	ESMFold	AlphaFold2	Boltz-2	Chai-1 (ESM-2)	Chai-1 (MSA)	Chai-1 (MSA)
monoFLS-6	0.932	1.822	2.405	2.271	1.459	1.951
monoFLS-9	0.931	2.196	2.336	2.350	1.413	2.584
monoFLS-12	0.933	2.136	2.330	2.289	1.495	1.631
monoFLS-24	0.929	1.910	2.283	2.525	1.419	1.778

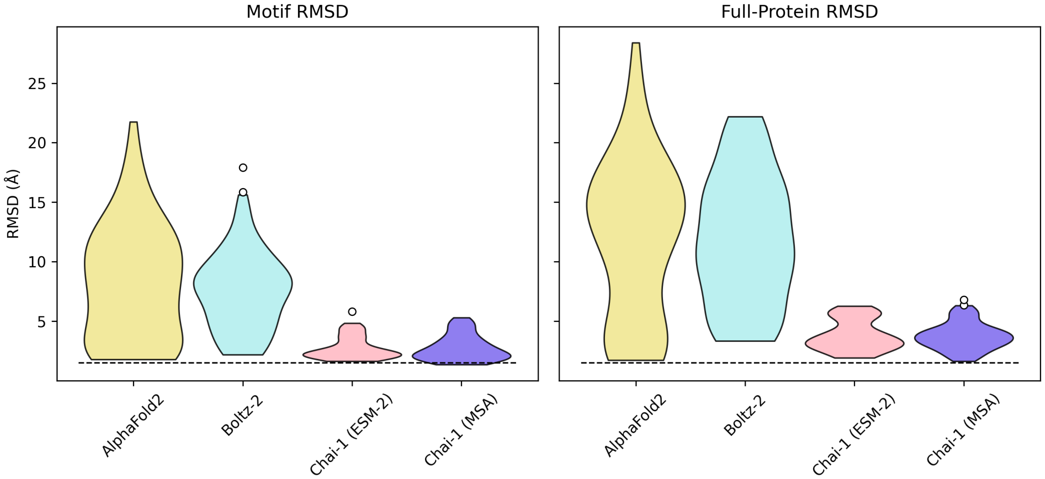

Figure 5: **Orthogonal Folding Distributions.** Motif and full-protein RMSD distributions from orthogonal validation across four folding algorithms. Motif RMSD is against the crystal structure of multiFLS; full-protein RMSD is against the ESMFold-derived monoFLS designs. The dashed line marks the 1.5 Å threshold. Across both metrics, ESMFold agrees most closely with Chai-1, whereas AlphaFold2 and Boltz-2 distributions are shifted toward higher RMSD values.

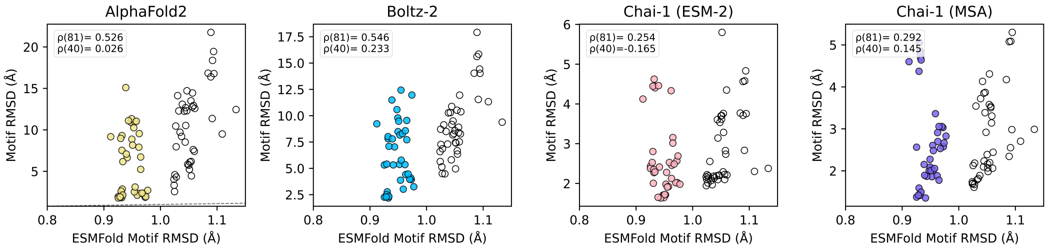

Figure 6: **Orthogonal Motif RMSD Correlations.** Comparison of motif RMSD predicted by ESMFold and four orthogonal folding models across 40 monoFLS designs (colored circles). We also include 41 additional variants that pass the initial filter of pLDDT score above 0.75, pTM above 0.7, but have a motif RMSD between 1.0 Å and 1.2 Å (uncolored circles). Spearman rank correlation coefficients (ρ) are computed for both the original set (40) and the more lenient set (81).

D Molecular Simulation Details

D.1 Simulation Protocol

We perform vanilla molecular dynamics with OpenMM [Eastman et al., 2023]. For the multimeric FLS enzyme, we removed TPP and Mg^{2+} from the crystal structure (4QQ8) and used it as the starting point, while for the designed monomers, we use the ESMFold-predicted structures as the starting points. Prior to simulation, all structures were pre-processed with `pdbfixer` to repair missing atoms or residues and to add hydrogen atoms. The Amber ff14SB force field [Maier et al., 2015] was used for the protein, which was solvated in an explicit solvent box using the TIP3P-FB water model [Sengupta et al., 2021], with a minimum padding of 1 nm between the solute and the box edges. Note that the simulation includes only the protein and the solvent, and thus should be considered as non-catalytic in the sense of attempting to reconstruct the catalytic transition state. The goal is rather to establish a sense of stability between the two effective domains upon their fusion.

Energy minimization was performed in two stages. In the first stage, all heavy atoms were harmonically restrained to allow hydrogen atoms to relax. In the second stage, the harmonic restraints were released and full system minimization was carried out. Following minimization, equilibration was conducted for 200 ps in the NPT ensemble at 300 K and 1 atm. Temperature control was maintained

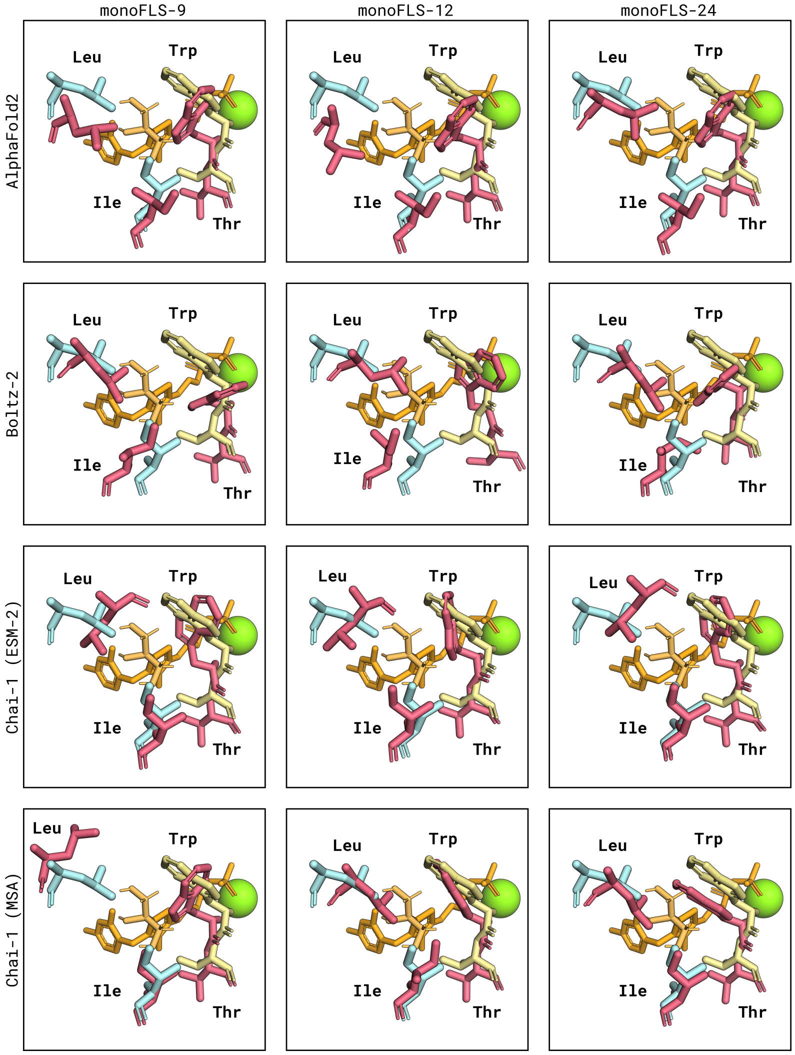

Figure 7: **Orthogonally Folded Active Sites.** We show the remaining three monoFLS variants, specifically the geometries of four key residues in the entry of the active site for all orthogonal folding algorithms.

with a Langevin thermostat (friction coefficient of 1 ps^{-1}), while pressure was regulated with a Monte Carlo barostat. Nonbonded interactions were treated using the particle-mesh Ewald (PME) method, employing a real-space cutoff of 1 nm (10 Å) and an Ewald error tolerance of 10^{-4}. Bond lengths involving hydrogen were constrained, and water molecules were kept rigid, permitting the use of a 2 fs integration timestep. Production simulations were subsequently carried out under these conditions.

D.2 Results

First, Figure 3 showed the domain–domain separation distance (between the two domains before introducing the linker) and the time-evolution of the motif RMSD (to multiFLS). We specifically show the deviation between the two centers of mass between the respective domains, when compared to the first frame of the trajectory, i.e., after the initial equilibration. The deviation is measured with respect to the initial snapshot of the production run, hence evolution into negative values show the domains coming closer together during the simulation. Three of the designs (monoFLS-9, monoFLS-12, monoFLS-24) show similar deviation to the native multiFLS, which, however, seems to be the most stable in terms of fluctuations of this metric throughout the 500 ns. monoFLS-6 shows large deviations, but then comes closer in terms of domain–domain separation. In terms of motif RMSD relative to multiFLS, monoFLS-9 shows the closest agreement throughout the trajectory, while the others gradually drift away.

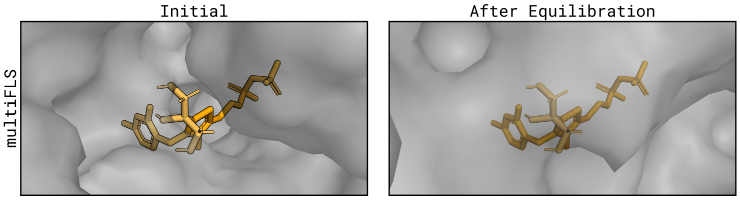

Figure 8: **Pocket Collapse of multiFLS.** The original multimeric enzyme appears to have its pocket collapse after initial equilibration, when compared to its crystal structure (Initial). We show the intermediate DHA-TPP in orange to clearly indicate where the effective pocket is located.

Second, we notice that during the simulation of the native multiFLS, the pocket collapses right after equilibration, shown in Figure 8. This might suggest that the force field cannot fully capture the relevant biochemical dynamics. The pocket is visualized using PyMOL's surface rendering, which uses a triangulated solvent-excluded surface with a default 1.4 Å probe. This observation is qualitative, as static solvent-excluded surface images are sensitive to probe radius and smoothing, and we thus do not over-interpret them, but provide them for completeness. Nevertheless, we believe such simulations can still serve as a lower-throughput validation step for filtering designs before committing to experimental synthesis. A more reliable quantitative assessment, such as trajectory-based solvent flux or water-occupancy analysis along the entry residues, would be more appropriate but is beyond the scope of this paper.