Cell Systems

CellPress

Perspective

Illuminating the universe of enzyme catalysis in the era of artificial intelligence

Jason Yang,[1] Francesca-Zhoufan Li,[2] Yueming Long,[1] and Frances H. Arnold[1,2,*]
[1]Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, CA 91125, USA
[2]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA 91125, USA
*Correspondence: frances@cheme.caltech.edu
https://doi.org/10.1016/j.cels.2025.101372

SUMMARY

Scientific research has revealed only a minuscule fraction of the enzymes that evolution has generated to power life's essential chemical reactions—and an even tinier fraction of the vast universe of *possible* enzymes. Beyond the enzymes already annotated lie an astronomical number of biocatalysts that could enable sustainable chemical production, degrade toxic pollutants, and advance disease diagnosis and treatment. For the past few decades, directed evolution has been a powerful strategy for reshaping enzymes to access new chemical transformations: by harnessing nature's existing diversity as a starting point and taking inspiration from nature's most powerful design process, evolution, to modify enzymes incrementally. Recently, artificial intelligence (AI) methods have started revolutionizing how we understand and compose the language of life. In this perspective, we discuss a vision for AI-driven enzyme discovery to unveil a world of enzymes that transcends biological evolution and perhaps offers a route to genetically encoding almost any chemistry.

DIRECTED ENZYME EVOLUTION

Nearly all of biology's chemistry is catalyzed by protein catalysts, the enzymes. Enzymes drive cell growth and development, repair wounds, and generate wonderfully complex and functional materials such as wood or shells—all while deriving energy and starting materials from their environments. Enzymes have an unrivaled ability to catalyze chemical reactions cleanly, selectively, and under mild, physiological conditions. These brilliant natural inventions offer powerful and sustainable chemical routes to make food ingredients, materials, fuels, and specialty chemicals; to degrade pollutants into recyclable or biodegradable forms; and to diagnose and treat disease.[1–4]

Imagine that we could expand biology's DNA-encoded chemistry to encompass even more useful chemical transformations, including the best of those invented by human chemists. Since the beginning of life, nature has explored only the tiniest fraction of the universe of life's *possible* chemistry. Although many natural reactions have proven valuable to humans, from cellulose hydrolysis for biofuel production[5] to antibiotic synthesis via polyketide synthases,[6] many applications will require enzymes that are not likely to be found in nature. Thus, the challenge is to identify protein sequences that encode desirable new functions, where overall performance in an application is quantified as "fitness." The different modalities relevant to protein engineering can be thought of as falling on a spectrum ranging from specific design goals (fitness/function) at one end to general design outputs (structures/sequences) at the other end (Figure 1).

However, the vast majority of the staggering number of *possible* proteins—orders of magnitude greater than the number of particles in the universe—are either nonfunctional or unsuitable (Figure 2A).[7,8] There are unfortunately many, many more ways to be non-functional in the library of possible genomes, just as the vast majority of Babel's books contain pure gibberish. The collection of all possible proteins is also vast, and the density of "meaningful" sequences is miniscule.

But, unlike the librarians of Borges' random book collection, we, the librarians of the protein collection, have an easier time finding the meaningful sequences because we have evolution to guide us. John Maynard Smith described a conceptual space such that each protein sequence is surrounded by all its single-mutant neighbors.[9] For evolution to occur, he argued, a functional protein must be surrounded by at least one other functional protein. Evolution could then pass through the network of functional protein sequences to explore new possibilities, one mutation at a time. This process has given rise to the billions upon billions of proteins found in nature today—rare meaningful sequences that we can literally scrape from the bottom of our shoes. And it can give rise to many more. Such an exploration, one (or a few) mutations at a time, can discover and improve enzymes, including ones that would be particularly useful to humans.[1] These directed evolution "greedy uphill walks" (Figure 2A) have led to the thousands of improved enzymes used in products from laundry detergents to disease therapies.[2]

Directed evolution tackles the problem of optimizing extant enzymes for specific fitness goals such as enhanced stability or enzymatic activity, but it can also be used to discover and explore whole new functional spaces. By taking advantage of the ability of enzymes to catalyze reactions that are not their native ones, a property called catalytic promiscuity,[10] one can obtain enzymes that operate on whole new substrates[11,12] and even perform chemistry that is unknown in nature, such as the

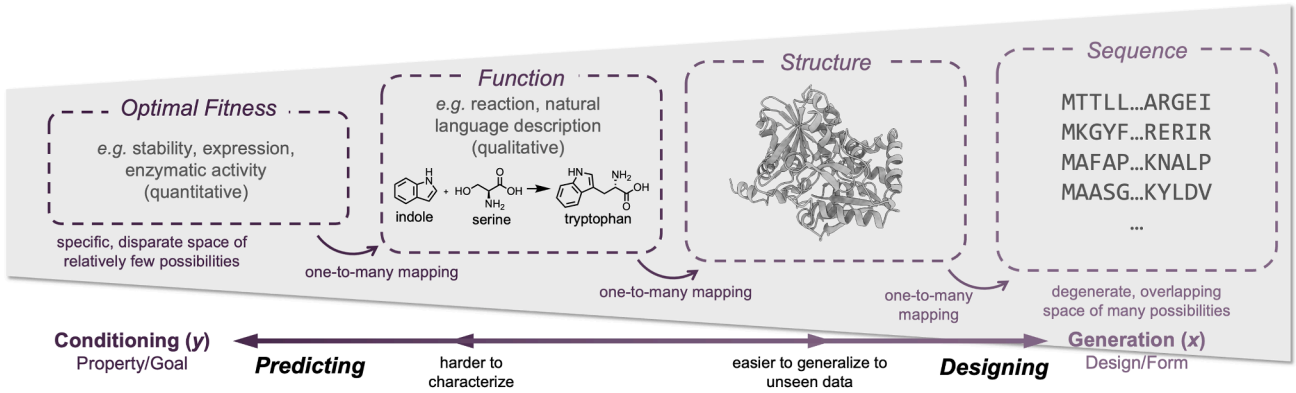

Figure 1. Spectrum of modalities relevant to protein engineering

Sequence design is often the ultimate downstream task, with other sources of information providing useful forms of conditioning (design goals). There are many solutions that satisfy each form of conditioning, as one moves from left to right on the spectrum.

formation or cleavage of carbon-silicon bonds.[2,13–15] One day, perhaps, all useful chemistry could be encoded in DNA and performed cleanly, as nature does. Of course, we do not know whether all relevant chemistry can be accomplished by enzymes—but with so many sequence possibilities, and with the assistance of cofactors like metals, it is likely that many, if not all, useful chemistries could be encoded in DNA and catalyzed by enzymes (Figure 2B).

AI IS REVOLUTIONIZING THE WAY WE ENGINEER ENZYMES

Despite its many successes, directed evolution is not the most efficient way to explore the vast space of possible enzymes. Directed evolution has been effective because many protein-fitness landscapes are smooth in at least some of their many dimensions, but it faces severe limitations when this is not the case.[16–18] It is also slow and resource intensive, and, as a local search in sequence space, it will ignore other parts of the space that may hold even greater functional potential (Figure 2A).[16] Furthermore, directed evolution requires a starting point with some level of the desired function or at least a strategy for acquiring that function through incremental steps. These are major bottlenecks in the quest to unveil new regions of enzyme functional space.

Alternatively, scientists have leveraged computational modeling to rationally design proteins.[19] Foundational work like ancestral sequence reconstruction[20] and structure-guided chimeragenesis[21,22] provided valuable strategies for diversifying protein sequences by allowing greater jumps in sequence space. Over time, knowledge of structure, intermolecular interactions, physical force fields, and enzyme mechanisms have been used to design enzymes entirely *de novo*.[23] Yet, enzyme design remains an extraordinary challenge because the sequence-function relationships are highly complex.[24] Protein binder design is often framed as a structure design problem, as binding interactions can be modeled relatively accurately with physics, given a target structure.[25,26] In contrast, enzyme engineering requires precisely positioned catalytic residues and an active site capable of accommodating reactant(s) and intermediates(s), often containing metals and cofactors that can be challenging to model computationally.[27] Transition state energies can be useful predictors of enzymatic activity,[28] but many enzyme reaction mechanisms remain poorly understood. Further complicating matters, enzymes are tasked

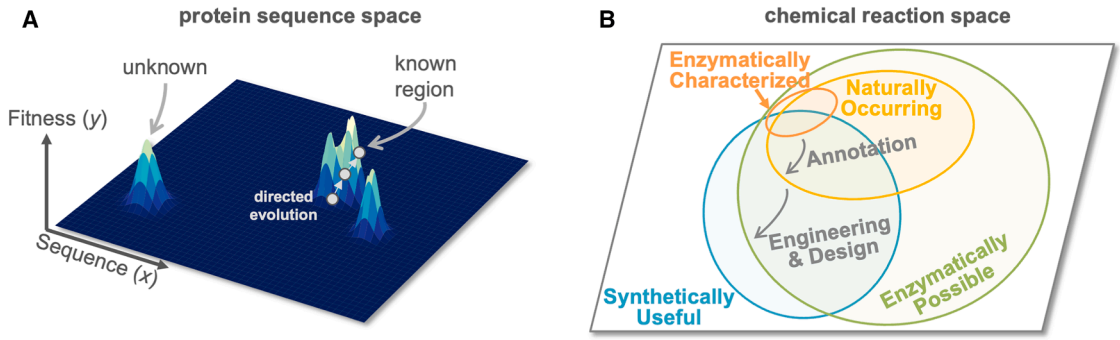

Figure 2. The vast design space of protein sequences and functions

(A) Directed evolution has been a powerful tool for expanding the functional space of enzymes and achieving specific fitness goals in a vast design space where most sequences are not functional. This search has been largely constrained to local search from existing functions, leaving many useful regions of sequence space underexplored and distant regions completely unknown.

(B) Only a minuscule fraction of enzymatic chemical reaction space (function space) that would be useful for human purposes has been characterized.

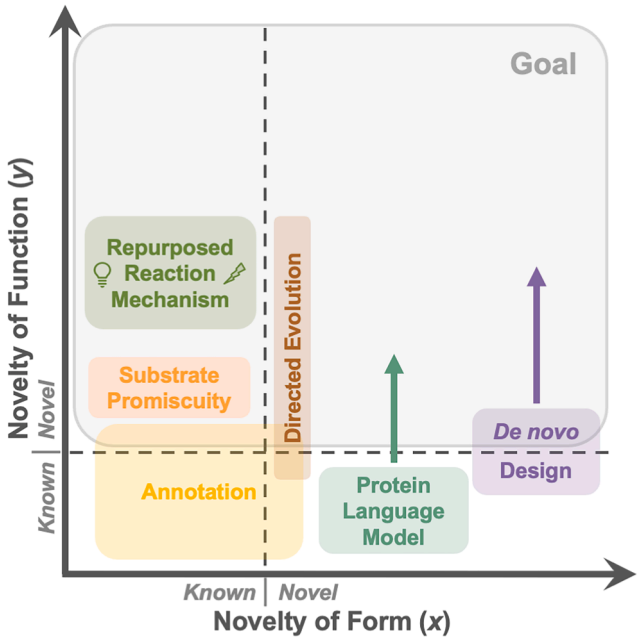

Figure 3. Most existing ML methods can generate proteins with novel sequences/structures (forms), but designing proteins with novel enzymatic functions is more difficult

Novel enzyme functions can be achieved by discovering sequences via functional annotation, taking advantage of promiscuity with respect to other substrates, exploiting novel chemical mechanisms of action, or modifying enzyme sequences through directed evolution. Developing better ML methods to unlock novel functions that are inaccessible from existing enzymes would be highly impactful.

with simultaneously balancing overall protein stability with dynamic conformation flexibility—an interplay that is poorly understood.[29–31] A single mutation, even one distant from the active site, can affect activity in ways that are difficult to explain, much less to predict.[1] Thus, before the advent of deep learning, impressive examples of rationally designed enzymes were relatively rare.[23,32,33]

In the past couple of years, artificial intelligence (AI) powered by machine-learning (ML) methods has almost completely overtaken previous computational approaches, enabling us to design proteins by learning primarily from data.[19,34–40] Generative ML can learn the distribution of proteins that nature has already made and humans have deposited in rapidly growing databases: it learns to compose new proteins using the "language" of nature's proteins to infer functional sequences that nature has not explored.[34,41–43] Such generative protein language models, from Potts models[44] to deep-learning models,[45–51] have generated sequences, for example, of functional lysozymes,[45] carbonic anhydrases,[49] and even Cas9 nucleases[52] with low sequence identity and activities comparable to naturally occurring enzymes (Figure 3). Although these generated proteins largely adopt existing structural folds, some efforts have shown that it is possible to push generations into novel structural space.[53,54]

Generative ML methods can also integrate structure and physics for *de novo* design (Figure 3) of novel folds[55]—mostly binders but also enzymes, including luciferases,[56] serine hydrolases,[57] and retro-aldolases.[58] A prevailing approach is to use a generative model[59,60] to design protein backbone structures that will accommodate a desired active-site conformation (motif scaffolding)[61–63] or a certain reaction transition state.[64,65] These structures are then mapped to sequences using inverse folding models based on deep learning,[66,67] as sequence design is generally the ultimate goal. However, this two-step process may be a relic of protein design in the pre-deep-learning era, where difficult problems were broken into more manageable steps. Recently, there have been efforts to design structures jointly with sequences[68–75] to directly scaffold active-site residues in enzymes (not just the backbone coordinates) and leverage all available data for more reasonable designs.[76] Structure prediction has quickly become an indispensable tool for validating these designs,[77–79] along with models that can predict enzymatic function from sequence.[80–83] Besides conditioning on structure, an interesting direction of research involves using descriptions of function as keywords or natural language descriptions to generate enzymes with the desired function.[47,71,84–86]

Utilizing AI-driven methods, researchers are achieving success rates orders of magnitude higher than those of design methods before the deep-learning era, and they are generating real, functional proteins.[58,87] Still, success rates are often low, and computational filters are necessary to improve hit rates.[88] Existing *de novo* structure design models also seem to consistently under-sample catalytically important motifs and over-sample alpha helices,[89] and there are many other potential areas for improvement.[90] Ultimately, AI enables bigger jumps in protein space, which could lead to more novelty in functional space (Figure 3) or more useful combinations of properties such as high thermostability and catalytic promiscuity. In the next section, we motivate and contextualize methods to achieve functional novelty.

ILLUMINATING NOVEL ENZYMATIC FUNCTIONS

Discovering enzymes with previously unknown functions is particularly valuable for applications ranging from chemical synthesis[91] to environmental remediation and gene editing. For example, degrading emerging environmental pollutants will require enzymes with new activities, and many therapeutics require Cas nucleases with modified specificities.[92] Although some chemistries can be accessed by modifying existing enzymes through directed evolution or by using existing enzyme scaffolds on new substrates or with repurposed mechanisms (Figure 3),[93–96] the overall discovery of functional novelty in enzymes could be greatly facilitated and automated by AI.

One of the most exciting aspects of engineering the biological world is that the wealth of biological data is still largely untapped, offering tremendous potential to uncover genes that encode new, useful enzymes. In fact, less than 1% of all known protein sequences are functionally annotated, and many extant proteins have never even been sequenced.[97] To complement an increase in unlabeled sequences from next-generation metagenomic sequencing studies that are being deposited in public databases,[98–100] ML models are advancing in their ability to predict and annotate protein functions, helping to decode the "dark matter" of protein sequences (those without functional associations) and showing higher sensitivity than sequence similarity approaches such as BLAST (Figure 3).[101–104] Different ML strategies are used to annotate genes,[105–107] such as semantic mining by incorporating genomic context, where related genes are

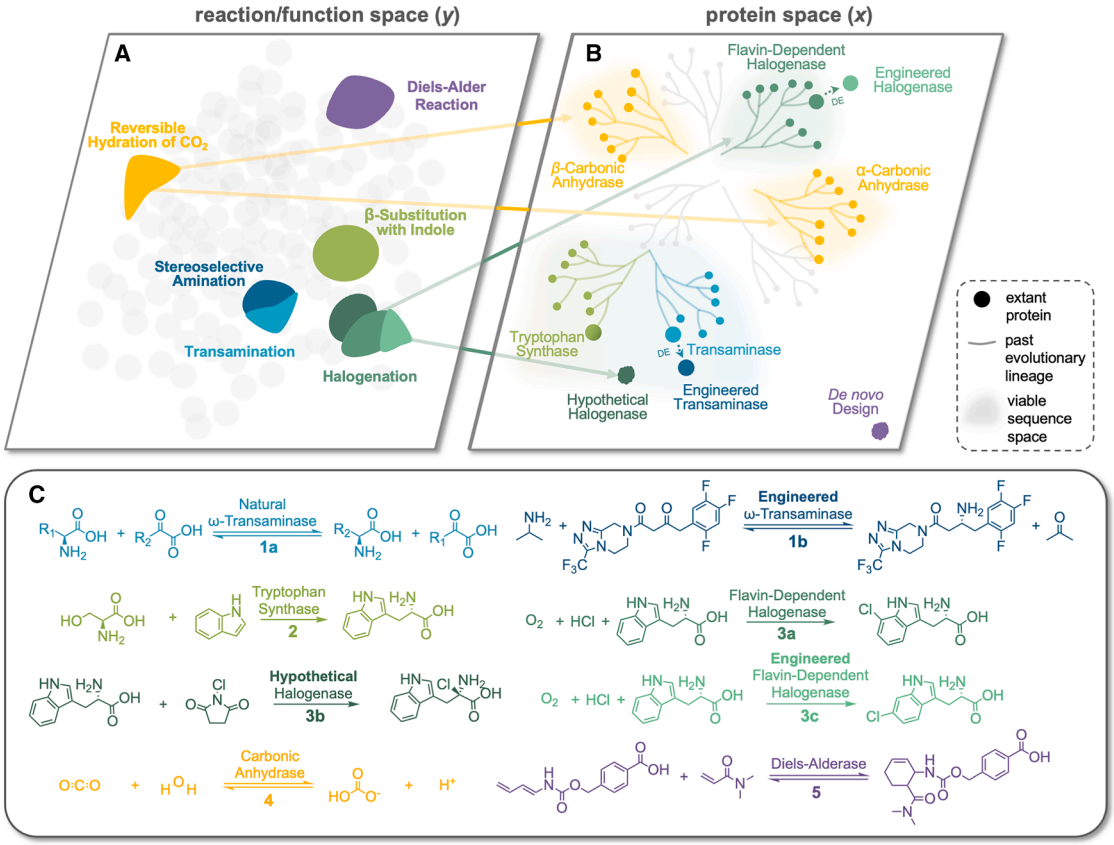

Figure 4. Conceptual framework for illuminating the universe of enzyme function

(A and B) Learning a mapping between (A) function space (*y*) and (B) protein space (*x*) using ML. Although directed evolution enables incremental, local expansions in function and sequence space, ML methods could be used to make larger jumps. Example of a moderate jump: inferring a hypothetical PLP-dependent halogenase that employs a similar catalytic mechanism to existing PLP-dependent transaminases and tryptophan synthases and may share some sequence/structural similarity. Example of a large jump: designing a *de novo* Diels-Alderase that shares no sequence similarity to existing proteins. There potentially exist multiple design solutions to achieve the same desired function, e.g., convergent evolution of natural carbonic anhydrases and different mechanistic routes to achieve halogenation. Solid circles refer to extant proteins, with translucent branches indicating their extinct ancestors.

(C) Example of a chemical reaction associated with each function, color coded.

often clustered.[108] Structure-similarity-based searching has also demonstrated success in the real world to discover new gene-editing enzymes that are related to existing ones but have different properties—such as smaller size and altered selectivity.[109–111] Similarly, ML methods have led to the discovery of synthetically valuable terpene synthases in previously uncharacterized taxa.[112] Other approaches frame the problem as enzyme retrieval, given target reactions.[81,113,114]

Still, there are several limitations to existing approaches. To start, many desirable enzymatic functions, such as reactions used in natural product synthesis or gene editing, fall outside the existing enzyme categorization scheme of enzyme commission (EC) numbers. Better analysis methods (i.e., for mass spectrometry data[115]) and databases[116,117] are needed to detect and catalog specific enzyme reactions/products to train ML models, respectively. Finally, most computational functional annotation strategies are only reliable for known protein families with known functions, and existing models have difficulty retrieving proteins for previously unannotated reactivities.[81]

Likewise, existing approaches to *generate* proteins, such as ancestral sequence reconstruction, chimeragenesis, and protein language models, primarily explore known functions within known enzyme families (within the translucent cloud of a single color in Figure 4).[118,119] Although *de novo* structure design can explore more novelty in sequence/structure space (Figure 4), it is also largely limited to designing enzymes for known reactions with known active-site configurations.[73] Design becomes much more difficult for novel reactions where such a "blueprint" is not available.

To fully harness the potential of generative ML models for protein engineering, the models need to generalize to new functions with real-world applications (Figure 3).[81,84] For example, future models could understand the catalytic mechanisms of natural pyridoxal phosphate (PLP)-dependent transaminases and tryptophan synthases to infer new mechanistically similar enzymes that could act as halogenases (Figure 4). *De novo* structure design for entirely new functions (e.g., the Diels-Alder reaction) is also possible.[120] The most successful models will likely integrate multiple modalities, learning from protein sequences, structures, reactions, mechanisms, and natural-language descriptions to capture enzyme function and provide a more holistic approach to enzyme design.[119] Unlike directed evolution,

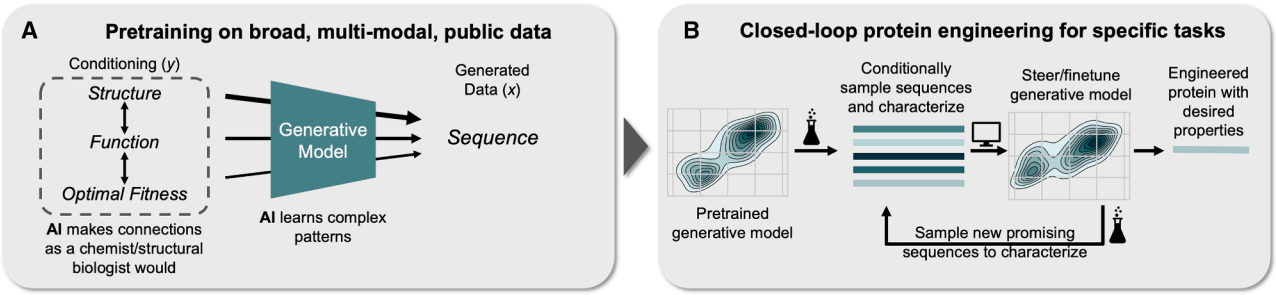

Figure 5. Vision for all-in-one enzyme (protein) sequence engineering with a generative model that learns the distribution of viable proteins, conditioned on properties such as fitness, function, and/or structure

Such a model would (A) be trained on publicly available data on natural proteins and (B) be updated/steered with data from experimental measurements. For more specific tasks, weak connections between modalities would benefit the most from steering. The thickness of the arrow shows the relative strength of the connection given existing data and/or tools. Many existing tasks in protein discovery and design fall under this framework, e.g., protein fitness optimization, inverse folding, natural language-guided functional protein generation, etc.

which is a local search, ML approaches would allow us to jump to new "islands" in functional and sequence/structure space (Figure 4). We elaborate on how to achieve this vision in the penultimate section.

THE IMPORTANCE OF ASSAY-LABELED DATA

Although we have been able to learn valuable patterns from nature's proteins, many useful reactions likely lie outside the distribution of known natural enzymatic functions, making it difficult for ML models trained solely on natural proteins to extrapolate effectively.[121] Currently, computational methods such as physics-based simulations and calculations are not yet accurate predictors for most enzymatic functions, so high-throughput wet-lab assay data collection and learning pipelines will continue to be vital for creating new biocatalysts.[122–124]

ML has increasingly been used to suggest combinations of beneficial mutations and reduce the burden of having to perform many generations of mutagenesis and screening for protein fitness optimization.[39,125] Sampled sequence-function data can be used to build a statistical model of the fitness landscape, enabling predictions about combinations of mutations with improved functions, i.e., semi-local expansions in functional and sequence space. We have used such supervised ML methods to make directed evolution more efficient and effective in the real world.[124,126–128] These models can also be updated in each generation (such as in active learning-assisted directed evolution),[129] and future exploration can be balanced with exploitation of previous knowledge.[130] Besides learning directly from labeled data collected through experiments, the models constructed from the data can additionally include what we already know of natural proteins, e.g., in the form of zero-shot scores based on evolutionary conservation[131–135] or protein structures and physics.[136,137] Overall, there are many useful supervised ML approaches combining the utility of unlabeled data on natural sequences with labeled assay data.[128,132,134–136,138,139]

A VISION FOR GENERAL, AUTOMATED PROTEIN DESIGN

Until now, designing new proteins has largely been treated as a separate problem from protein fitness optimization. Yet, generative models[34,41,42,140] are particularly well suited for both tasks: directed evolution struggles to find protein solutions that are distant in sequence space, and even supervised ML requires enumerating fitness predictions over the vast design space, which is intractable. Thus, *tackling protein design/optimization with a unified generative model that understands the rules of natural proteins while also integrating assay-labeled data will be needed to unlock new design possibilities* (Figure 5). Whereas directed evolution is limited to slow, local searches, the proposed model could leverage multiple data sources to travel to semi-distant regions of sequence space—regions unexplored by evolution but still viable for maximizing a desired function (Figure 4). Such a model would encompass many existing tasks in protein design, such as function-guided design and inverse folding, and lay the foundation for many more.

Crucially, such a generative model would need to be conditioned on relevant information to enable controllable generation.[43,141] Two aspects of the conditions are essential: (1) function, which qualitatively describes the reaction(s) that an enzyme can perform, and (2) fitness, which quantitatively measures performance for a particular function and can be a combination of properties such as catalytic rate, stability, and protein expression level (Figure 1). Some preliminary work has explored function-based conditional generation using language models[84,86] and diffusion models.[85,142] Function could be conveyed as reaction representations, natural language descriptions, or protein structural features—particularly for binder design, where current methods largely focus on structure generation and inverse folding. Nevertheless, these efforts remain largely fragmented, highlighting the critical need for a unified multimodal framework that integrates diverse forms of knowledge, enabling transfer learning across tasks and generalization to unseen functions (Figure 5A)—just as a chemist, structural biologist, and bioinformatician would collaborate to understand or predict a protein's function.

Efforts in this direction have included conditioning protein language models[40] on protein structures[47,74,143] or functional descriptors,[47,84,86] discrete diffusion models for protein sequence generation,[142] and the integration of natural language, protein structure, and protein sequence into unified representation learning frameworks.[81,82,85,114] However, current AI models do not have a generalized understanding of enzyme functions and reaction mechanisms and thus cannot design enzymes with

novel functions. To solve this, new ML architectures are needed to enable conditional sampling from probabilistic distributions and bridge disparate modalities spanning natural language, sequences, graphs, 3D coordinates, reaction mechanisms, tabular data, and more.[144] This will be a challenging task because most strategies used in computer vision and natural language processing are not perfectly suited for biological data, and different modalities are annotated for different proteins. Utilizing inductive biases from physics, such as information about electronics and transition states, may be crucial to improving the generalization ability of such models.

More difficult applications, such as optimizing an enzyme for a new-to-nature reaction that is not captured in the natural distribution of enzyme function, will require guiding or fine-tuning the generative model based on expensive (low-throughput) real-world fitness measurements (Figure 5B). Developing new ML techniques for steering generative models as part of active learning, or iterative cycles of experimentation and computational modeling, will be essential. For example, this could involve guidance with diffusion models[50,51,68,145–149] or reinforcement learning with language models.[138,150–152] In essence, these methods would shift the distribution of the generative model to *conditionally* sample proteins with desired properties (Figure 5B).[153,154] Even with the help of inductive biases from physics, some new-to-nature reactions will require such a large extrapolative leap that activity may be impossible to attain without first conditioning the model on related but more readily attainable protein-reaction data, analogous to the "substrate walking" approach often used in directed evolution.[155]

In the not-too-distant future, fully automated "cloud" labs, or autonomous laboratories, could perform experiments continuously and reproducibly, and this would be accessible to diverse users regardless of their discipline or expertise. An AI agent could design experiments, execute, analyze, and improve its model, and repeat this process in iterative cycles of active learning. This kind of self-improving AI-driven experimentation is already beginning to emerge, with increasing capabilities for autonomous hypothesis generation and testing.[156–164]

As part of long-term active learning, the development of a standardized enzyme sequence-fitness database will be indispensable. Although the pretraining step can utilize public data sources such as the Protein Data Bank (sequence-structure), reactions in BRENDA (sequence-function), and natural language descriptions from SwissProt and literature articles (sequence-function), there is a critical need for a sequence-fitness database comprising more specific fitness measurements such as data from enzymatic activity, stability, and expression assays, including environmental conditions, kinetic behaviors, substrate specificities, and reaction mechanisms. For this to become a reality, an open-source, collaborative initiative could be instrumental, where research institutions and industry players contribute to and benefit from the growth of the database. Standardized benchmarks and/or competitions could further accelerate the pace of advancement and foster a more integrated research community.[165] With these datasets, automated enzyme engineering will enable us to create new genetically encoded chemistry much more efficiently than human-driven processes alone. As ML models continue to advance, they will progressively improve by integrating knowledge across different modalities.

CLOSING THOUGHTS

Nature has been exploring the universe of enzymes long before any of us were born, albeit slowly and with its own objective: survival and reproduction. With directed evolution, we can steer enzymes toward functions useful to humans, but we are still limited to a slow and rather naive optimization process. But now we are armed with AI, a powerful tool that can help us traverse sequence space to uncover useful new enzymes. The outlook for protein design with AI is particularly promising for several reasons. (1) Proteins have a track record of reliable optimization for many functions by nature and by humans using nature's original optimization strategy, evolution. By contrast, there is no guarantee that other forms of matter—chemicals, materials, etc.—can be so readily improved. (2) Proteins have been explored by nature long before we came along, and a wealth of data is accumulating on these natural designs that can inform future designs. These literal life lessons are being collected in databases to inform future engineering. In comparison, we have limited data on other materials, most of which have only been constructed by humans since the industrial revolution. (3) Finally, and critically, we have the machinery to build nearly any desired protein sequence, encoded in DNA and produced by a cell factory. For other materials, synthesis remains a major bottleneck.[166] With advancements in biological sequencing,[167,168] synthesis,[169–171] and assay technology,[172] our knowledge of and ability to manipulate protein space will only become more powerful.

In the era of AI, we may wonder how effective ML approaches will be. To what extent can learning from natural proteins extrapolate to new sequences or even new structures? Is there enough density of natural functional enzymes to infer sequences that might lead to novel functions? AI workflows for new-to-nature enzyme activity will surely need assay-labeled data in addition to knowledge of natural proteins, but how many experiments will be necessary? John Maynard Smith articulated that evolution is possible because there is a high enough density of functional proteins in single-mutant neighbors, a realization that made it possible to access new functional proteins by directed evolution. Unfortunately, we have no precedent to predict how useful existing data on natural proteins will be for creating functional novelty. Nonetheless, exploiting all of the data we have will undoubtedly get us to our goals faster.

Overall, AI is learning to compose across multiple modalities, accelerating engineering toward functional proteins, and filling in the enzymatic universe. Although we have focused here on enzyme engineering, many of the ideas and approaches outlined here generalize to other protein engineering applications where unlabeled and labeled data can be utilized together. With regard to biosecurity, the generation of new toxins, resistance mechanisms, or pathogens does not lie far from our brave new world of enzymatic chemistry, so there will need to be monitoring of such tasks.[173–175] One day, we may replace inefficient, wasteful human chemistry with the highly selective, efficient chemistry performed by enzymes. Let us use AI to safely propel us toward that dream.

ACKNOWLEDGMENTS

This work was supported by a US Army Research Office cooperative agreement (W911NF-19-2-0026 to F.H.A.) and an Amgen Chem-Bio Engineering

award. J.Y. and F.-Z.L. are partially supported by National Science Foundation Graduate Research Fellowships. J.Y. is also supported by a Google PhD Fellowship. The authors would also like to thank Ariane Mora for reviewing the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Arnold, F.H. (2018). Directed Evolution: Bringing New Chemistry to Life. Angew. Chem. Int. Ed. Engl. *57*, 4143–4148. https://doi.org/10.1002/anie.201708408.

2. Reisenbauer, J.C., Sicinski, K.M., and Arnold, F.H. (2024). Catalyzing the future: recent advances in chemical synthesis using enzymes. Curr. Opin. Chem. Biol. *83*, 102536. https://doi.org/10.1016/j.cbpa.2024.102536.

3. Buller, R., Lutz, S., Kazlauskas, R.J., Snajdrova, R., Moore, J.C., and Bornscheuer, U.T. (2023). From nature to industry: Harnessing enzymes for biocatalysis. Science *382*, eadh8615. https://doi.org/10.1126/science.adh8615.

4. Pyser, J.B., Chakrabarty, S., Romero, E.O., and Narayan, A.R.H. (2021). State-of-the-Art Biocatalysis. ACS Cent. Sci. *7*, 1105–1116. https://doi.org/10.1021/acscentsci.1c00273.

5. Wilson, D.B. (2009). Cellulases and biofuels. Curr. Opin. Biotechnol. *20*, 295–299. https://doi.org/10.1016/j.copbio.2009.05.007.

6. Gomes, E.S., Schuch, V., and de Macedo Lemos, E.G.D.M. (2013). Biotechnology of polyketides: new breath of life for the novel antibiotic genetic pathways discovery through metagenomics. Braz. J. Microbiol. *44*, 1007–1034. https://doi.org/10.1590/S1517-83822013000400002.

7. Arnold, F.H. (2011). The Library of Maynard-Smith: My Search for Meaning in the Protein Universe. Microbe Mag. *6*, 316–318. https://doi.org/10.1128/microbe.6.316.1.

8. Dennett, D.C. (1995). Darwin's Dangerous Idea: Evolution and the Meanings of Life (Simon & Schuster).

9. Smith, J.M. (1970). Natural selection and the concept of a protein space. Nature *225*, 563–564. https://doi.org/10.1038/225563a0.

10. Khersonsky, O., and Tawfik, D.S. (2010). Enzyme Promiscuity: A Mechanistic and Evolutionary Perspective. Annu. Rev. Biochem. *79*, 471–505. https://doi.org/10.1146/annurev-biochem-030409-143718.

11. Bayer, T., Wu, S., Snajdrova, R., Baldenius, K., and Bornscheuer, U.T. (2025). An Update: Enzymatic Synthesis for Industrial Applications. Angew. Chem. Int. Ed. Engl. *64*, e202505976. https://doi.org/10.1002/anie.202505976.

12. Bell, E.L., Finnigan, W., France, S.P., Green, A.P., Hayes, M.A., Hepworth, L.J., Lovelock, S.L., Niikura, H., Osuna, S., Romero, E., et al. (2021). Biocatalysis. Nat. Rev. Methods Primers *1*, 46. https://doi.org/10.1038/s43586-021-00044-z.

13. Kan, S.B.J., Lewis, R.D., Chen, K., and Arnold, F.H. (2016). Directed evolution of cytochrome c for carbon–silicon bond formation: Bringing silicon to life. Science *354*, 1048–1051. https://doi.org/10.1126/science.aah6219.

14. Sarai, N.S., Fulton, T.J., O'Meara, R.L., Johnston, K.E., Brinkmann-Chen, S., Maar, R.R., Tecklenburg, R.E., Roberts, J.M., Reddel, J.C.T., Katsoulis, D.E., et al. (2024). Directed evolution of enzymatic silicon-carbon bond cleavage in siloxanes. Science *383*, 438–443. https://doi.org/10.1126/science.adi5554.

15. Chen, K., and Arnold, F.H. (2020). Engineering new catalytic activities in enzymes. Nat. Catal. *3*, 203–213. https://doi.org/10.1038/s41929-019-0385-5.

16. Romero, P.A., and Arnold, F.H. (2009). Exploring protein fitness landscapes by directed evolution. Nat. Rev. Mol. Cell Biol. *10*, 866–876. https://doi.org/10.1038/nrm2805.

17. Fröhlich, C., Bunzel, H.A., Buda, K., Mulholland, A.J., Van Der Kamp, M. W., Johnsen, P.J., Leiros, H.S., and Tokuriki, N. (2024). Epistasis arises from shifting the rate-limiting step during enzyme evolution of a β--lactamase. Nat. Catal. *7*, 499–509. https://doi.org/10.1038/s41929-024-01117-4.

18. Miton, C.M., Buda, K., and Tokuriki, N. (2021). Epistasis and intramolecular networks in protein evolution. Curr. Opin. Struct. Biol. *69*, 160–168. https://doi.org/10.1016/j.sbi.2021.04.007.

19. Albanese, K.I., Barbe, S., Tagami, S., Woolfson, D.N., and Schiex, T. (2025). Computational protein design. Nat. Rev. Methods Primers *5*, 1–28. https://doi.org/10.1038/s43586-025-00383-1.

20. Joy, J.B., Liang, R.H., McCloskey, R.M., Nguyen, T., and Poon, A.F.Y. (2016). Ancestral Reconstruction. PLoS Comput. Biol. *12*, e1004763. https://doi.org/10.1371/journal.pcbi.1004763.

21. Silberg, J.J., Endelman, J.B., and Arnold, F.H. (2004). SCHEMA-Guided Protein Recombination. Methods Enzymol. *388*, 35–42. https://doi.org/10.1016/S0076-6879(04)88004-2.

22. Otey, C.R., Landwehr, M., Endelman, J.B., Hiraga, K., Bloom, J.D., and Arnold, F.H. (2006). Structure-Guided Recombination Creates an Artificial Family of Cytochromes P450. PLoS Biol. *4*, e112. https://doi.org/10.1371/journal.pbio.0040112.

23. Huang, P.-S., Boyken, S.E., and Baker, D. (2016). The coming of age of de novo protein design. Nature *537*, 320–327. https://doi.org/10.1038/nature19946.

24. Reardon, S. (2024). Five Tasks that Still Challenge Protein Designers. Nature *635*, 246–248. https://doi.org/10.1038/d41586-024-03595-9.

25. Cao, L., Coventry, B., Goreshnik, I., Huang, B., Sheffler, W., Park, J.S., Jude, K.M., Marković, I., Kadam, R.U., Verschueren, K.H.G., et al. (2022). Design of protein binding proteins from target structure alone. Nature *605*, 551–560. https://doi.org/10.1038/s41586-022-04654-9.

26. Pacesa, M., Nickel, L., Schmidt, J., Pyatova, E., Schellhaas, C., Kissling, L., Barendse, P., Choudhury, J., Kapoor, S., Alcaraz-Serna, A., et al. (2024). BindCraft: one-shot design of functional protein binders. Preprint at bioRxiv. https://doi.org/10.1101/2024.09.30.615802.

27. Siddiqui, S.A., Stuyver, T., Shaik, S., and Dubey, K.D. (2023). Designed Local Electric Fields—Promising Tools for Enzyme Engineering. JACS Au. JACS Au *3*, 3259–3269. https://doi.org/10.1021/jacsau.3c00536.

28. Fersht, A. (1999). Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding (W. H. Freeman and Company).

29. Tokuriki, N., and Tawfik, D.S. (2009). Protein Dynamism and Evolvability. Science *324*, 203–207. https://doi.org/10.1126/science.1169375.

30. Tokuriki, N., Stricher, F., Serrano, L., and Tawfik, D.S. (2008). How Protein Stability and New Functions Trade Off. PLoS Comput. Biol. *4*, e1000002. https://doi.org/10.1371/journal.pcbi.1000002.

31. Olsson, M.H.M., Parson, W.W., and Warshel, A. (2006). Dynamical Contributions to Enzyme Catalysis: Critical Tests of A Popular Hypothesis. Chem. Rev. *106*, 1737–1756. https://doi.org/10.1021/cr040427e.

32. Röthlisberger, D., Khersonsky, O., Wollacott, A.M., Jiang, L., DeChancie, J., Betker, J., Gallaher, J.L., Althoff, E.A., Zanghellini, A., Dym, O., et al. (2008). Kemp elimination catalysts by computational enzyme design. Nature *453*, 190–195. https://doi.org/10.1038/nature06879.

33. Chu, A.E., Lu, T., and Huang, P.S. (2024). Sparks of function by de novo protein design. Nat. Biotechnol. *42*, 203–215. https://doi.org/10.1038/s41587-024-02133-2.

34. Hsu, C., Fannjiang, C., and Listgarten, J. (2024). Generative models for protein structures and sequences. Nat. Biotechnol. *42*, 196–199. https://doi.org/10.1038/s41587-023-02115-w.

35. Wen, S., Zheng, W., Bornscheuer, U.T., and Wu, S. (2025). Generative artificial intelligence for enzyme design: Recent advances in models and applications. Curr. Opin. Green Sustain. Chem. *52*, 101010. https://doi.org/10.1016/j.cogsc.2025.101010.

36. Wu, W., Wang, T., Zhang, Y., Deng, N., Song, X., Bi, Z., Yao, Z., Chen, K., Li, M., Niu, Q., et al. (2025). Advanced Deep Learning Methods for Protein

Structure Prediction and Design. Preprint at arXiv. https://doi.org/10.48550/arXiv.2503.13522.

37. Kyro, G.W., Qiu, T., and Batista, V.S. (2025). A Model-Centric Review of Deep Learning for Protein Design. Preprint at arXiv. https://doi.org/10.48550/arXiv.2502.19173.

38. Winnifrith, A., Outeiral, C., and Hie, B.L. (2024). Generative artificial intelligence for de novo protein design. Curr. Opin. Struct. Biol. *86*, 102794. https://doi.org/10.1016/j.sbi.2024.102794.

39. Yang, J., Li, F.-Z., and Arnold, F.H. (2024). Opportunities and Challenges for Machine Learning-Assisted Enzyme Engineering. ACS Cent. Sci. *10*, 226–241. https://doi.org/10.1021/acscentsci.3c01275.

40. Fan, W., Zhou, Y., Wang, S., Yan, Y., Liu, H., Zhao, Q., Song, L., and Li, Q. (2025). Computational Protein Science in the Era of Large Language Models (LLMs). Preprint at arXiv. https://doi.org/10.48550/arXiv.2501.10282.

41. Wu, Z., Johnston, K.E., Arnold, F.H., and Yang, K.K. (2021). Protein sequence design with deep generative models. Curr. Opin. Chem. Biol. *65*, 18–27. https://doi.org/10.1016/j.cbpa.2021.04.004.

42. Ruffolo, J.A., and Madani, A. (2024). Designing proteins with language models. Nat. Biotechnol. *42*, 200–202. https://doi.org/10.1038/s41587-024-02123-4.

43. Ferruz, N., and Höcker, B. (2022). Controllable protein design with language models. Nat. Mach. Intell. *4*, 521–532. https://doi.org/10.1038/s42256-022-00499-z.

44. Russ, W.P., Figliuzzi, M., Stocker, C., Barrat-Charlaix, P., Socolich, M., Kast, P., Hilvert, D., Monasson, R., Cocco, S., Weigt, M., et al. (2020). An evolution-based model for designing chorismate mutase enzymes. Science *369*, 440–445. https://doi.org/10.1126/science.aba3304.

45. Madani, A., Krause, B., Greene, E.R., Subramanian, S., Mohr, B.P., Holton, J.M., Olmos, J.L., Xiong, C., Sun, Z.Z., Socher, R., et al. (2023). Large language models generate functional protein sequences across diverse families. Nat. Biotechnol. *41*, 1099–1106. https://doi.org/10.1038/s41587-022-01618-2.

46. Ferruz, N., Schmidt, S., and Höcker, B. (2022). ProtGPT2 is a deep unsupervised language model for protein design. Nat. Commun. *13*, 4348. https://doi.org/10.1038/s41467-022-32007-7.

47. Hayes, T., Rao, R., Akin, H., Sofroniew, N.J., Oktay, D., Lin, Z., Verkuil, R., Tran, V.Q., Deaton, J., Wiggert, M., et al. (2025). Simulating 500 million years of evolution with a language model. Science *387*, 850–858. https://doi.org/10.1126/science.ads0018.

48. Hesslow, D., Zanichelli, N., Notin, P., Poli, I., and Marks, D. (2022). RITA: a Study on Scaling Up Generative Protein Sequence Models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2205.05789.

49. Munsamy, G., Illanes-Vicioso, R., Funcillo, S., Nakou, I.T., Lindner, S., Ayres, G., Sheehan, L.S., Moss, S., Eckhard, U., Lorenz, P., et al. (2024). Conditional language models enable the efficient design of proficient enzymes. Preprint at bioRxiv. https://doi.org/10.1101/2024.05.03.592223.

50. Alamdari, S., Thakkar, N., van den Berg, R., Tenenholtz, N., Strome, R., Moses, A.M., Lu, A.X., Fusi, N., Amini, A.P., and Yang, K.K. (2024). Protein generation with evolutionary diffusion: sequence is all you need. Preprint at bioRxiv. https://doi.org/10.1101/2023.09.11.556673.

51. Wang, X., Zheng, Z., Ye, F., Xue, D., Huang, S., and Gu, Q. (2024). Diffusion Language Models Are Versatile Protein Learners. In Proceedings of the 41st International Conference on Machine Learning, pp. 52309–52333.

52. Ruffolo, J.A., Nayfach, S., Gallagher, J., Bhatnagar, A., Beazer, J., Hussain, R., Russ, J., Yip, J., Hill, E., Pacesa, M., et al. (2024). Design of highly functional genome editors by modeling the universe of CRISPR-Cas sequences. Published online July 30, 2025. Nature. https://doi.org/10.1038/s41586-025-09298-z.

53. Verkuil, R., Kabeli, O., Du, Y., Wicky, B.I.M., Milles, L.F., Dauparas, J., Baker, D., Ovchinnikov, S., Sercu, T., and Rives, A. (2022). Language models generalize beyond natural proteins. Preprint at bioRxiv. https://doi.org/10.1101/2022.12.21.521521.

54. Subramanian, A.M., Martinez, Z.A., Lourenço, A.L., Liu, S., and Thomson, M. (2025). Unexplored regions of the protein sequence-structure map revealed at scale by a library of foldtuned language models. Preprint at bioRxiv. https://doi.org/10.1101/2023.12.22.573145.

55. Buller, R., Damborsky, J., Hilvert, D., and Bornscheuer, U.T. (2025). Structure Prediction and Computational Protein Design for Efficient Biocatalysts and Bioactive Proteins. Angew. Chem. Int. Ed. Engl. *64*, e202421686. https://doi.org/10.1002/anie.202421686.

56. Yeh, A.H.-W., Norn, C., Kipnis, Y., Tischer, D., Pellock, S.J., Evans, D., Ma, P., Lee, G.R., Zhang, J.Z., Anishchenko, I., et al. (2023). De novo design of luciferases using deep learning. Nature *614*, 774–780. https://doi.org/10.1038/s41586-023-05696-3.

57. Lauko, A., Pellock, S.J., Sumida, K.H., Anishchenko, I., Juergens, D., Ahern, W., Jeung, J., Shida, A.F., Hunt, A., Kalvet, I., et al. (2025). Computational design of serine hydrolases. Science *388*, eadu2454. https://doi.org/10.1126/science.adu2454.

58. Anishchenko, I., Kipnis, Y., Kalvet, I., Zhou, G., Krishna, R., Pellock, S.J., Lauko, A., Lee, G.R., An, L., Dauparas, J., et al. (2024). Modeling protein-small molecule conformational ensembles with ChemNet. Preprint at bioRxiv. https://doi.org/10.1101/2024.09.25.614868.

59. Watson, J.L., Juergens, D., Bennett, N.R., Trippe, B.L., Yim, J., Eisenach, H.E., Ahern, W., Borst, A.J., Ragotte, R.J., Milles, L.F., et al. (2023). De novo design of protein structure and function with RFdiffusion. Nature *620*, 1089–1100. https://doi.org/10.1038/s41586-023-06415-8.

60. Ingraham, J.B., Baranov, M., Costello, Z., Barber, K.W., Wang, W., Ismail, A., Frappier, V., Lord, D.M., Ng-Thow-Hing, C., Van Vlack, E.R., et al. (2023). Illuminating protein space with a programmable generative model. Nature *623*, 1070–1078. https://doi.org/10.1038/s41586-023-06728-8.

61. Zheng, Z., Zhang, B., Didi, K., Yang, K.K., Yim, J., Watson, J.L., Chen, H.-F., and Trippe, B.L. (2025). MotifBench: A standardized protein design benchmark for motif-scaffolding problems. Preprint at arXiv. https://doi.org/10.48550/arXiv.2502.12479.

62. Wang, Q., Liu, X., Zhang, H., Chu, H., Shi, C., Zhang, L., Bai, J., Liu, P., Li, J., Zhu, X., et al. (2024). Cytochrome P450 Enzyme Design by Constraining the Catalytic Pocket in a Diffusion Model. Research (Wash D. C) *7*, 0413. https://doi.org/10.34133/research.0413.

63. Lovelock, S.L., Crawshaw, R., Basler, S., Levy, C., Baker, D., Hilvert, D., and Green, A.P. (2022). The road to fully programmable protein catalysis. Nature *606*, 49–58. https://doi.org/10.1038/s41586-022-04456-z.

64. Hua, C., Liu, Y., Zhang, D., Zhang, O., Luan, S., Yang, K.K., Wolf, G., Precup, D., and Zheng, S. (2024). EnzymeFlow: Generating Reaction-specific Enzyme Catalytic Pockets through Flow Matching and Co-Evolutionary Dynamics. Preprint at arXiv. https://doi.org/10.48550/arXiv.2410.00327.

65. Hua, C., Lu, J., Liu, Y., Zhang, O., Tang, J., Ying, R., Jin, W., Wolf, G., Precup, D., and Zheng, S. (2024). Reaction-conditioned De Novo Enzyme Design with GENzyme. Preprint at arXiv. https://doi.org/10.48550/arXiv.2411.16694.

66. Dauparas, J., Anishchenko, I., Bennett, N., Bai, H., Ragotte, R.J., Milles, L.F., Wicky, B.I.M., Courbet, A., de Haas, R.J., Bethel, N., et al. (2022). Robust deep learning-based protein sequence design using ProteinMPNN. Science *378*, 49–56. https://doi.org/10.1126/science.add2187.

67. Hsu, C., Verkuil, R., Liu, J., Lin, Z., Hie, B., Sercu, T., Lerer, A., and Rives, A. (2022). Learning inverse folding from millions of predicted structures. In Proceedings of the 39th International Conference on Machine Learning, *162*, pp. 8946–8970.

68. Lisanza, S.L., Gershon, J.M., Tipps, S.W.K., Sims, J.N., Arnoldt, L., Hendel, S.J., Simma, M.K., Liu, G., Yase, M., Wu, H., et al. (2024). Multistate and functional protein design using RoseTTAFold sequence space diffusion. Published online October 7, 2024. Nat. Biotechnol. https://doi.org/10.1038/s41587-024-02395-w.

69. Huguet, G., Vuckovic, J., Fatras, K., Thibodeau-Laufer, E., Lemos, P., Islam, R., Liu, C.-H., Rector-Brooks, J., Akhound-Sadegh, T., Bronstein, M., et al. (2024). Sequence-Augmented SE(3)-Flow Matching For Conditional Protein Backbone Generation. Adv. Neural. Inf. Process. Syst. *37*, 33007–33036.

70. Campbell, A., Yim, J., Barzilay, R., Rainforth, T., and Jaakkola, T. (2024). Generative Flows on Discrete State-Spaces: Enabling Multimodal Flows with Applications to Protein Co-Design. In Proceedings of the 41st International Conference on Machine Learning, pp. 5453–5512.

71. Lu, A.X., Yan, W., Robinson, S.A., Kelow, S., Yang, K.K., Gligorijevic, V., Cho, K., Bonneau, R., Abbeel, P., and Frey, N.C. (2025). All-Atom Protein Generation with Latent Diffusion. Preprint at bioRxiv. https://doi.org/10.1101/2024.12.02.626353.

72. Chu, A.E., Kim, J., Cheng, L., El Nesr, G., Xu, M., Shuai, R.W., and Huang, P.-S. (2024). An all-atom protein generative model. Proc. Natl. Acad. Sci. USA *121*, e2311500121. https://doi.org/10.1073/pnas.2311500121.

73. Ahern, W., Yim, J., Tischer, D., Salike, S., Woodbury, S.M., Kim, D., Kalvet, I., Kipnis, Y., Coventry, B., Altae-Tran, H.R., et al. (2025). Atom level enzyme active site scaffolding using RFdiffusion2. Preprint at bioRxiv. https://doi.org/10.1101/2025.04.09.648075.

74. Wang, X., Zheng, Z., Ye, F., Xue, D., Huang, S., and Gu, Q. (2025). DPLM-2: A Multimodal Diffusion Protein Language Model. In 13th International Conference on Learning Representations.

75. Geffner, T., Didi, K., Cao, Z., Reidenbach, D., Zhang, Z., Dallago, C., Kucukbenli, E., Kreis, K., and Vahdat, A. (2025). La-Proteina: Atomistic Protein Generation via Partially Latent Flow Matching. Preprint at arXiv. https://doi.org/10.48550/arXiv.2507.09466.

76. Wang, C., Alamdari, S., Domingo-Enrich, C., Amini, A.P., and Yang, K.K. (2025). Toward deep learning sequence-structure co-generation for protein design. Curr. Opin. Struct. Biol. *91*, 103018. https://doi.org/10.1016/j.sbi.2025.103018.

77. Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A.J., Bambrick, J., et al. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature *630*, 493–500. https://doi.org/10.1038/s41586-024-07487-w.

78. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589. https://doi.org/10.1038/s41586-021-03819-2.

79. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. Science *379*, 1123–1130. https://doi.org/10.1126/science.ade2574.

80. Yu, T., Cui, H., Li, J.C., Luo, Y., Jiang, G., and Zhao, H. (2023). Enzyme function prediction using contrastive learning. Science *379*, 1358–1363. https://doi.org/10.1126/science.adf2465.

81. Yang, J., Mora, A., Liu, S., Wittmann, B.J., Anandkumar, A., Arnold, F.H., and Yue, Y. (2024). CARE: a Benchmark Suite for the Classification and Retrieval of Enzymes. In Advances in Neural Information Processing Systems.

82. Su, J., Zhou, X., Zhang, X., and Yuan, F. (2024). ProTrek: Navigating the Protein Universe through Tri-Modal Contrastive Learning. Preprint at bioRxiv. https://doi.org/10.1101/2024.05.30.596740.

83. Zhou, X., Han, C., Zhang, Y., Su, J., Zhuang, K., Jiang, S., Yuan, Z., Zheng, W., Dai, F., Zhou, Y., et al. (2025). Decoding the Molecular Language of Proteins with Evola. Preprint at bioRxiv. https://doi.org/10.1101/2025.01.05.630192.

84. Yang, J., Bhatnagar, A., Ruffolo, J.A., and Madani, A. (2025). Function-Guided Conditional Generation Using Protein Language Models with Adapters. Preprint at arXiv. https://doi.org/10.48550/arXiv.2410.03634.

85. Liu, S., Li, Y., Li, Z., Gitter, A., Zhu, Y., Lu, J., Xu, Z., Nie, W., Ramanathan, A., Xiao, C., et al. (2025). A Text-guided Protein Design Framework. Nat. Mach. Intell. *7*, 580–591. https://doi.org/10.1038/s42256-025-01011-z.

86. Dai, F., Fan, Y., Su, J., Wang, C., Han, C., Zhou, X., Liu, J., Qian, H., Wang, S., Zeng, A., et al. (2025). Toward De Novo Protein Design from Natural Language. Preprint at bioRxiv. https://doi.org/10.1101/2024.08.01.606258.

87. Notin, P., Rollins, N., Gal, Y., Sander, C., and Marks, D. (2024). Machine learning for functional protein design. Nat. Biotechnol. *42*, 216–228. https://doi.org/10.1038/s41587-024-02127-0.

88. Johnson, S.R., Fu, X., Viknander, S., Goldin, C., Monaco, S., Zelezniak, A., and Yang, K.K. (2025). Computational scoring and experimental evaluation of enzymes generated by neural networks. Nat. Biotechnol. *43*, 396–405. https://doi.org/10.1038/s41587-024-02214-2.

89. Lu, T., Liu, M.H., Chen, Y., Kim, J., and Huang, P.-S. (2025). Assessing Generative Model Coverage of Protein Structures with SHAPES. Cell Syst. *23*, 101347. https://doi.org/10.1016/j.cels.2025.101347.

90. Listov, D., Goverde, C.A., Correia, B.E., and Fleishman, S.J. (2024). Opportunities and challenges in design and optimization of protein function. Nat. Rev. Mol. Cell Biol. *25*, 639–653. https://doi.org/10.1038/s41580-024-00718-y.

91. Yu, T., Boob, A.G., Volk, M.J., Liu, X., Cui, H., and Zhao, H. (2023). Machine learning-enabled retrobiosynthesis of molecules. Nat. Catal. *6*, 137–151. https://doi.org/10.1038/s41929-022-00909-w.

92. Chen, P.J., and Liu, D.R. (2023). Prime editing for precise and highly versatile genome manipulation. Nat. Rev. Genet. *24*, 161–177. https://doi.org/10.1038/s41576-022-00541-1.

93. Schwizer, F., Okamoto, Y., Heinisch, T., Gu, Y., Pellizzoni, M.M., Lebrun, V., Reuter, R., Köhler, V., Lewis, J.C., and Ward, T.R. (2018). Artificial Metalloenzymes: Reaction Scope and Optimization Strategies. Chem. Rev. *118*, 142–231. https://doi.org/10.1021/acs.chemrev.7b00014.

94. Cheng, L., Li, D., Mai, B.K., Bo, Z., Cheng, L., Liu, P., and Yang, Y. (2023). Stereoselective amino acid synthesis by synergistic photoredox-pyridoxal radical biocatalysis. Science *381*, 444–451. https://doi.org/10.1126/science.adg2420.

95. Jain, S., Ospina, F., and Hammer, S.C. (2024). A New Age of Biocatalysis Enabled by Generic Activation Modes. JACS Au *4*, 2068–2080. https://doi.org/10.1021/jacsau.4c00247.

96. Raps, F.C., and Hyster, T.K. (2025). Emergent Mechanisms in Biocatalysis. ACS Cent. Sci. *11*, 1029–1040. https://doi.org/10.1021/acscentsci.5c00245.

97. The UniProt Consortium (2023). UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res. *51*, D523–D531. https://doi.org/10.1093/nar/gkac1052.

98. Cornman, A., West-Roberts, J., Camargo, A.P., Roux, S., Beracochea, M., Mirdita, M., Ovchinnikov, S., and Hwang, Y. (2025). The OMG dataset: An Open MetaGenomic corpus for mixed-modality genomic language modeling. In 13th International Conference on Learning Representations.

99. Vince, O., Oldach, P., Pereno, V., Leung, M.H.Y., Greco, C., Minto-Cowcher, G., Ur-Rehman, S., Kam, K.Y.K., Chow, W., Bolton, E., et al. (2025). Breaking Through Biology's Data Wall: Expanding the Known Tree of Life by Over 10x using a Global Biodiscovery Pipeline. Preprint at bioRxiv. https://doi.org/10.1101/2025.06.11.658620.

100. Yang, K.K., Alamdari, S., Lee, A.J., Kaymak-Loveless, K., Char, S., Brixi, G., Domingo-Enrich, C., Wang, C., Lyu, S., Fusi, N., et al. (2025). The Dayhoff Atlas: scaling sequence diversity for improved protein generation. Preprint at bioRxiv. https://doi.org/10.1101/2025.07.21.665991.

101. Pavlopoulos, G.A., Baltoumas, F.A., Liu, S., Selvitopi, O., Camargo, A.P., Nayfach, S., Azad, A., Roux, S., Call, L., Ivanova, N.N., et al. (2023). Unraveling the functional dark matter through global metagenomics. Nature *622*, 594–602. https://doi.org/10.1038/s41586-023-06583-7.

102. Kim, H.R., Ji, H., Kim, G.B., and Lee, S.Y. (2025). Enzyme functional classification using artificial intelligence. Published online March 27, 2025. Trends Biotechnol. https://doi.org/10.1016/j.tibtech.2025.03.003.

103. Yeo, J., Han, Y., Bordin, N., Lau, A.M., Kandathil, S.M., Kim, H., Karin, E. L., Mirdita, M., Jones, D.T., Orengo, C., et al. (2025). Metagenomic-scale analysis of the predicted protein structure universe. Preprint at bioRxiv. https://doi.org/10.1101/2025.04.23.650224.

104. Zhang, Y., Moorhoff, F., Qiu, S., Dong, W., Medina-Ortiz, D., and Davari, M.D. (2025). Machine Learning-Driven Enzyme Mining: Opportunities, Challenges, and Future Perspectives. Preprint at arXiv. https://doi.org/10.48550/arXiv.2507.07666.

105. Jha, N., Kravitz, J., West-Roberts, J., Lu, C., Camargo, A.P., Roux, S., Cornman, A., and Hwang, Y. (2025). Gaia: An AI-enabled genomic

context–aware platform for protein sequence annotation. Sci. Adv. *11*, eadv5109. https://doi.org/10.1126/sciadv.adv5109.

106. Derry, A., Tartici, A., and Altman, R.B. (2025). Protein functional site annotation using local structure embeddings. Preprint at bioRxiv. https://doi.org/10.1101/2023.10.13.562298.

107. Levy Karin, E.L., and Steinegger, M. (2025). Cutting-edge deep-learning based tools for metagenomic research. Natl. Sci. Rev. *12*, nwaf056. https://doi.org/10.1093/nsr/nwaf056.

108. Merchant, A.T., King, S.H., Nguyen, E., and Hie, B.L. (2024). Semantic mining of functional de novo genes from a genomic language model. Preprint at bioRxiv. https://doi.org/10.1101/2024.12.17.628962.

109. Yoon, P.H., Zhang, Z., Loi, K.J., Adler, B.A., Lahiri, A., Vohra, K., Shi, H., Rabelo, D.B., Trinidad, M., Boger, R.S., et al. (2024). Structure-guided discovery of ancestral CRISPR-Cas13 ribonucleases. Science *385*, 538–543. https://doi.org/10.1126/science.adq0553.

110. Van Kempen, M., Kim, S.S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist, C.L.M., Söding, J., and Steinegger, M. (2023). Fast and accurate protein structure search with Foldseek. Nat. Biotechnol. *42*, 243–246. https://doi.org/10.1101/2022.02.07.479398.

111. Gilchrist, C.L.M., Mirdita, M., and Steinegger, M. (2024). Multiple Protein Structure Alignment at Scale with FoldMason. Preprint at bioRxiv. https://doi.org/10.1101/2024.08.01.606130.

112. Samusevich, R., Hebra, T., Bushuiev, R., Engst, M., Kulhánek, J., Bushuiev, A., Smith, J.D., Čalounová, T., Smrčková, H., Molineris, M., et al. (2025). Structure-enabled enzyme function prediction unveils elusive terpenoid biosynthesis in archaea. Preprint at bioRxiv. https://doi.org/10.1101/2024.01.29.577750.

113. Pate, S.C., Wang, E.H., Broadbelt, L.J., and Tyo, K.E.J. (2025). RC-GNN: A predictive model of enzyme-reaction pairs. Preprint at bioRxiv. https://doi.org/10.1101/2025.06.22.660952.

114. Mikhael, P.G., Chinn, I., and Barzilay, R. (2024). CLIPZyme: Reaction-Conditioned Virtual Screening of Enzymes. In Proceedings of the 41st International Conference on Machine Learning, pp. 35647–35663. https://doi.org/10.5555/3692070.3693523.

115. Bushuiev, R., Bushuiev, A., Samusevich, R., Brungs, C., Sivic, J., and Pluskal, T. (2025). Self-supervised learning of molecular representations from millions of tandem mass spectra using DreaMS. Published online May 23, 2025. Nat. Biotechnol. https://doi.org/10.1038/s41587-025-02663-3.

116. Heid, E., Probst, D., Green, W.H., and Madsen, G.K.H. (2023). EnzymeMap: curation, validation and data-driven prediction of enzymatic reactions. Chem. Sci. *14*, 14229–14242. https://doi.org/10.1039/D3SC02048G.

117. Probst, D., Manica, M., Nana Teukam, Y.G., Castrogiovanni, A., Paratore, F., and Laino, T. (2022). Biocatalysed synthesis planning using data-driven learning. Nat. Commun. *13*, 964. https://doi.org/10.1038/s41467-022-28536-w.

118. Martin-Alonso, C., Alamdari, S., Samad, T.S., Yang, K.K., Bhatia, S.N., and Amini, A.P. (2025). Deep learning guided design of protease substrates. Preprint at bioRxiv. https://doi.org/10.1101/2025.02.27.640681.

119. Paton, A., Boiko, D., Perkins, J., Cemalovic, N., Reschützegger, T., Gomes, G., and Narayan, A. (2024). Generation of connections between protein sequence space and chemical space to enable a predictive model for biocatalysis. Preprint at ChemRxiv. https://doi.org/10.26434/chemrxiv-2024-w4dtr.

120. Siegel, J.B., Zanghellini, A., Lovick, H.M., Kiss, G., Lambert, A.R., St Clair, J.L., Gallaher, J.L., Hilvert, D., Gelb, M.H., Stoddard, B.L., et al. (2010). Computational Design of an Enzyme Catalyst for a Stereoselective Bimolecular Diels-Alder Reaction. Science *329*, 309–313. https://doi.org/10.1126/science.1190239.

121. Fannjiang, C., and Listgarten, J. (2024). Is Novelty Predictable? Cold Spring Harb. Perspect. Biol. *16*, a041469. https://doi.org/10.1101/cshperspect.a041469.

122. Gantz, M., Mathis, S.V., Nintzel, F.E.H., Zurek, P.J., Knaus, T., Patel, E., Boros, D., Weberling, F.-M., Kenneth, M.R.A., Klein, O.J., et al. (2024). Microdroplet screening rapidly profiles a biocatalyst to enable its AI-assisted engineering. Preprint at bioRxiv. https://doi.org/10.1101/2024.04.08.588565.

123. Gantz, M., Mathis, S.V., Nintzel, F.E.H., Lio, P., and Hollfelder, F. (2024). On synergy between ultrahigh throughput screening and machine learning in biocatalyst engineering. Faraday Discuss. *252*, 89–114. https://doi.org/10.1039/D4FD00065J.

124. Thomas, N., Belanger, D., Xu, C., Lee, H., Hirano, K., Iwai, K., Polic, V., Nyberg, K.D., Hoff, K.G., Frenz, L., et al. (2025). Engineering highly active nuclease enzymes with machine learning and high-throughput screening. Cell Syst. *16*, 101236. https://doi.org/10.1016/j.cels.2025.101236.

125. Yang, K.K., Wu, Z., and Arnold, F.H. (2019). Machine-learning-guided directed evolution for protein engineering. Nat. Methods *16*, 687–694. https://doi.org/10.1038/s41592-019-0496-6.

126. Romero, P.A., Krause, A., and Arnold, F.H. (2013). Navigating the protein fitness landscape with Gaussian processes. Proc. Natl. Acad. Sci. USA *110*, E193–E201. https://doi.org/10.1073/pnas.1215251110.

127. Wu, Z., Kan, S.B.J., Lewis, R.D., Wittmann, B.J., and Arnold, F.H. (2019). Machine learning-assisted directed protein evolution with combinatorial libraries. Proc. Natl. Acad. Sci. USA *116*, 8852–8858. https://doi.org/10.1073/pnas.1901979116.

128. Wittmann, B.J., Yue, Y., and Arnold, F.H. (2021). Informed training set design enables efficient machine learning-assisted directed protein evolution. Cell Syst. *12*, 1026–1045.e7. https://doi.org/10.1016/j.cels.2021.07.008.

129. Yang, J., Lal, R.G., Bowden, J.C., Astudillo, R., Hameedi, M.A., Kaur, S., Hill, M., Yue, Y., and Arnold, F.H. (2025). Active learning-assisted directed evolution. Nat. Commun. *16*, 714. https://doi.org/10.1038/s41467-025-55987-8.

130. Vornholt, T., Mutný, M., Schmidt, G.W., Schellhaas, C., Tachibana, R., Panke, S., Ward, T.R., Krause, A., and Jeschek, M. (2024). Enhanced Sequence-Activity Mapping and Evolution of Artificial Metalloenzymes by Active Learning. ACS Cent. Sci. *10*, 1357–1370. https://doi.org/10.1021/acscentsci.4c00258.

131. Fram, B., Su, Y., Truebridge, I., Riesselman, A.J., Ingraham, J.B., Passera, A., Napier, E., Thadani, N.N., Lim, S., Roberts, K., et al. (2024). Simultaneous enhancement of multiple functional properties using evolution-informed protein design. Nat. Commun. *15*, 5141. https://doi.org/10.1038/s41467-024-49119-x.

132. Hsu, C., Nisonoff, H., Fannjiang, C., and Listgarten, J. (2022). Learning protein fitness models from evolutionary and assay-labeled data. Nat. Biotechnol. *40*, 1114–1122. https://doi.org/10.1038/s41587-021-01146-5.

133. Hie, B.L., Shanker, V.R., Xu, D., Bruun, T.U.J., Weidenbacher, P.A., Tang, S., Wu, W., Pak, J.E., and Kim, P.S. (2024). Efficient evolution of human antibodies from general protein language models. Nat. Biotechnol. *42*, 275–283. https://doi.org/10.1038/s41587-023-01763-2.

134. Jiang, K., Yan, Z., Di Bernardo, M., Sgrizzi, S.R., Villiger, L., Kayabolen, A., Kim, B.J., Carscadden, J.K., Hiraizumi, M., Nishimasu, H., et al. (2024). Rapid in silico directed evolution by a protein language model with EVOLVEpro. Science *387*, eadr6006. https://doi.org/10.1126/science.adr6006.

135. Ding, K., Chin, M., Zhao, Y., Huang, W., Mai, B.K., Wang, H., Liu, P., Yang, Y., and Luo, Y. (2024). Machine learning-guided co-optimization of fitness and diversity facilitates combinatorial library design in enzyme engineering. Nat. Commun. *15*, 6392. https://doi.org/10.1038/s41467-024-50698-y.

136. Li, F.-Z., Yang, J., Johnston, K.E., Gürsoy, E., Yue, Y., and Arnold, F.H. (2024). Evaluation of Machine Learning-Assisted Directed Evolution Across Diverse Combinatorial Landscapes. Preprint at bioRxiv. https://doi.org/10.1101/2024.10.24.619774.

137. Ding, D., Shaw, A.Y., Sinai, S., Rollins, N., Prywes, N., Savage, D.F., Laub, M.T., and Marks, D.S. (2024). Protein design using structure-based residue preferences. Nat. Commun. *15*, 1639. https://doi.org/10.1038/s41467-024-45621-4.

138. Sun, H., He, L., Deng, P., Liu, G., Zhao, Z., Jiang, Y., Cao, C., Ju, F., Wu, L., Liu, H., et al. (2025). Accelerating protein engineering with fitness

landscape modeling and reinforcement learning. Preprint at bioRxiv. https://doi.org/10.1101/2023.11.16.565910.

139. Landwehr, G.M., Bogart, J.W., Magalhaes, C., Hammarlund, E.G., Karim, A.S., and Jewett, M.C. (2025). Accelerated enzyme engineering by machine-learning guided cell-free expression. Nat. Commun. *16*, 865. https://doi.org/10.1038/s41467-024-55399-0.

140. Brookes, D., Park, H., and Listgarten, J. (2019). Conditioning by adaptive sampling for robust design. In Proceedings of the 36th International Conference on Machine Learning, *97*, pp. 773–782.

141. Ferruz, N., Heinzinger, M., Akdel, M., Goncearenco, A., Naef, L., and Dallago, C. (2023). From sequence to function through structure: Deep learning for protein design. Comput. Struct. Biotechnol. J. *21*, 238–250. https://doi.org/10.1016/j.csbj.2022.11.014.

142. Praljak, N., Yeh, H., Moore, M., Socolich, M., Ranganathan, R., and Ferguson, A.L. (2024). Natural Language Prompts Guide the Design of Novel Functional Protein Sequences. Preprint at bioRxiv. https://doi.org/10.1101/2024.11.11.622734.

143. Ruffolo, J.A., Bhatnagar, A., Beazer, J., Nayfach, S., Russ, J., Hill, E., Hussain, R., Gallagher, J., and Madani, A. (2024). Adapting protein language models for structure-conditioned design. Preprint at bioRxiv. https://doi.org/10.1101/2024.08.03.606485.

144. Guloglu, B., Bragança, M., Graves, A., Cameron, S., Atkinson, T., Copoiu, L., Laterre, A., and Barrett, T.D. (2025). AbBFN2: A flexible antibody foundation model based on Bayesian Flow Networks. Preprint at bioRxiv. https://doi.org/10.1101/2025.04.29.651170.

145. Frey, N.C., Hötzel, I., Stanton, S.D., Kelly, R.L., Alberstein, R.G., Makowski, E.K., Martinkus, K., Berenberg, D., Bevers, J., Bryson, T., et al. (2025). Lab-in-the-loop therapeutic antibody design with deep learning. Preprint at bioRxiv. https://doi.org/10.1101/2025.02.19.639050.

146. Stanton, S., Maddox, W., Gruver, N., Maffettone, P., Delaney, E., Greenside, P., and Wilson, A.G. (2022). Accelerating Bayesian Optimization for Biological Sequence Design with Denoising Autoencoders. In Proceedings of the 39th International Conference on Machine Learning, pp. 20459–20478.

147. Gruver, N., Stanton, S., Frey, N.C., Rudner, T.G.J., Hotzel, I., Lafrance-Vanasse, J., Rajpal, A., Cho, K., and Wilson, A.G. (2023). Protein Design with Guided Discrete Diffusion. Adv. Neural. Inf. Process. Syst. *36*.

148. Nisonoff, H., Xiong, J., Allenspach, S., and Listgarten, J. (2025). Unlocking Guidance for Discrete State-Space Diffusion and Flow Models. In 13th International Conference on Learning Representations.

149. Yang, J., Chu, W., Khalil, D., Astudillo, R., Wittmann, B.J., Arnold, F.H., and Yue, Y. (2025). Steering Generative Models with Experimental Data for Protein Fitness Optimization. Preprint at arXiv. https://doi.org/10.48550/arXiv.2505.15093.

150. Stocco, F., Artigues-Lleixà, M., Hunklinger, A., Widatalla, T., Güell, M., and Ferruz, N. (2025). Guiding Generative Protein Language Models with Reinforcement Learning. Preprint at arXiv. https://doi.org/10.48550/arXiv.2412.12979.

151. Widatalla, T., Rafailov, R., and Hie, B. (2024). Aligning protein generative models with experimental fitness via Direct Preference Optimization. Preprint at bioRxiv. https://doi.org/10.1101/2024.05.20.595026.

152. Blalock, N., Seshadri, S., Babbar, A., Fahlberg, S.A., Kulkarni, A., and Romero, P.A. (2025). Functional Alignment of Protein Language Models via Reinforcement Learning. Preprint at bioRxiv. https://doi.org/10.1101/2025.05.02.651993.

153. Hie, B.L., and Yang, K.K. (2022). Adaptive machine learning for protein engineering. Curr. Opin. Struct. Biol. *72*, 145–152. https://doi.org/10.1016/j.sbi.2021.11.002.

154. Freschlin, C.R., Fahlberg, S.A., and Romero, P.A. (2022). Machine learning to navigate fitness landscapes for protein engineering. Curr. Opin. Biotechnol. *75*, 102713. https://doi.org/10.1016/j.copbio.2022.102713.

155. Almhjell, P.J., Johnston, K.E., Porter, N.J., Kennemur, J.L., Bhethanabotla, V.C., Ducharme, J., and Arnold, F.H. (2024). The β-subunit of tryptophan synthase is a latent tyrosine synthase. Nat. Chem. Biol. *20*, 1086–1093. https://doi.org/10.1038/s41589-024-01619-z.

156. Boiko, D.A., MacKnight, R., Kline, B., and Gomes, G. (2023). Autonomous chemical research with large language models. Nature *624*, 570–578. https://doi.org/10.1038/s41586-023-06792-0.

157. Gao, S., Fang, A., Huang, Y., Giunchiglia, V., Noori, A., Schwarz, J.R., Ektefaie, Y., Kondic, J., and Zitnik, M. (2024). Empowering biomedical discovery with AI agents. Cell *187*, 6125–6151. https://doi.org/10.1016/j.cell.2024.09.022.

158. Rapp, J.T., Bremer, B.J., and Romero, P.A. (2024). Self-driving laboratories to autonomously navigate the protein fitness landscape. Nat. Chem. Eng. *1*, 97–107. https://doi.org/10.1038/s44286-023-00002-4.

159. Singh, N., Lane, S., Yu, T., Lu, J., Ramos, A., Cui, H., and Zhao, H. (2025). A Generalized Platform for Artificial Intelligence-powered Autonomous Enzyme Engineering. Nat. Commun. *16*, 5648. https://doi.org/10.1038/s41467-025-61209-y.

160. Gottweis, J., Weng, W.-H., Daryin, A., Tu, T., Palepu, A., Sirkovic, P., Myaskovsky, A., Weissenberger, F., Rong, K., Tanno, R., et al. (2025). Towards an AI co-scientist. Preprint at arXiv. https://doi.org/10.48550/arXiv.2502.18864.

161. Swanson, K., Wu, W., Bulaong, N.L., Pak, J.E., and Zou, J. (2024). The Virtual Lab: AI Agents Design New SARS-CoV-2 Nanobodies with Experimental Validation. Preprint at bioRxiv. https://doi.org/10.1101/2024.11.11.623004.

162. Roohani, Y., Lee, A., Huang, Q., Vora, J., Steinhart, Z., Huang, K., Marson, A., Liang, P., and Leskovec, J. (2025). BioDiscoveryAgent: An AI Agent for Designing Genetic Perturbation Experiments. In 13th International Conference on Learning Representations.

163. Simon, E., Swanson, K., and Zou, J. (2024). Language models for biological research: a primer. Nat. Methods *21*, 1422–1429. https://doi.org/10.1038/s41592-024-02354-y.

164. Huang, K., Zhang, S., Wang, H., Qu, Y., Lu, Y., Roohani, Y., Li, R., Qiu, L., Li, G., Zhang, J., et al. (2025). Biomni: A General-Purpose Biomedical AI Agent. Preprint at bioRxiv. https://doi.org/10.1101/2025.05.30.656746.

165. Armer, C., Kane, H., Cortade, D.L., Redestig, H., Estell, D.A., Yusuf, A., Rollins, N., Spinner, A., Marks, D., Brunette, T.J., et al. (2025). Results of the Protein Engineering Tournament: An Open Science Benchmark for Protein Modeling and Design. Proteins. https://doi.org/10.1002/prot.70008.

166. Gao, W., and Coley, C.W. (2020). The Synthesizability of Molecules Proposed by Generative Models. J. Chem. Inf. Model. *60*, 5714–5723. https://doi.org/10.1021/acs.jcim.0c00174.

167. Wittmann, B.J., Johnston, K.E., Almhjell, P.J., and Arnold, F.H. (2022). evSeq: Cost-Effective Amplicon Sequencing of Every Variant in a Protein Library. ACS Synth. Biol. *11*, 1313–1324. https://doi.org/10.1021/acssynbio.1c00592.

168. Long, Y., Mora, A., Li, F.-Z., Gürsoy, E., Johnston, K.E., and Arnold, F.H. (2025). LevSeq: Rapid Generation of Sequence-Function Data for Directed Evolution and Machine Learning. ACS Synth. Biol. *14*, 230–238. https://doi.org/10.1021/acssynbio.4c00625.

169. Yang, J., Ducharme, J., Johnston, K.E., Li, F.-Z., Yue, Y., and Arnold, F.H. (2023). DeCOIL: Optimization of Degenerate Codon Libraries for Machine Learning-Assisted Protein Engineering. ACS Synth. Biol. *12*, 2444–2454. https://doi.org/10.1021/acssynbio.3c00301.

170. Freschlin, C.R., Yang, K.K., and Romero, P.A. (2025). Scalable and cost-efficient custom gene library assembly from oligopools. Preprint at bioRxiv. https://doi.org/10.1101/2025.03.22.644747.

171. Zhu, D., Brookes, D.H., Busia, A., Carneiro, A., Fannjiang, C., Popova, G., Shin, D., Donohue, K.C., Lin, L.F., Miller, Z.M., et al. (2024). Optimal trade-off control in machine learning–based library design, with application to adeno-associated virus (AAV) for gene therapy. Sci. Adv. *10*, eadj3786. https://doi.org/10.1126/sciadv.adj3786.

172. Markin, C.J., Mokhtari, D.A., Sunden, F., Appel, M.J., Akiva, E., Longwell, S.A., Sabatti, C., Herschlag, D., and Fordyce, P.M. (2021). Revealing enzyme functional architecture via high-throughput microfluidic enzyme

kinetics. Science *373*, eabf8761. https://doi.org/10.1126/science.abf8761.

173. Baker, D., and Church, G. (2024). Protein design meets biosecurity. Science *383*, 349. https://doi.org/10.1126/science.ado1671.

174. Wittmann, B.J., Alexanian, T., Bartling, C., Beal, J., Clore, A., Diggans, J., Flyangolts, K., Gemler, B.T., Mitchell, T., Murphy, S.T., et al. (2024). Toward AI-Resilient Screening of Nucleic Acid Synthesis Orders: Process, Results, and Recommendations. Preprint at bioRxiv. https://doi.org/10.1101/2024.12.02.626439.

175. Ikonomova, S.P., Wittmann, B.J., Piorino, F., Ross, D.J., Schaffter, S.W., Vasilyeva, O., Horvitz, E., Diggans, J., Strychalski, E.A., Lin-Gibson, S., et al. (2025). Experimental Evaluation of AI-Driven Protein Design Risks Using Safe Biological Proxies. Preprint at bioRxiv. https://doi.org/10.1101/2025.05.15.654077.