Comment

https://doi.org/10.1038/s41467-025-65281-2

LDBT instead of DBTL: combining machine learning and rapid cell-free testing

Alia Clark-ElSayed, Isa Madrigal Harrison, Meagan L. Olsen, John T. Lazar, Michael C. Jewett & Andrew D. Ellington

Synthetic biology is defined by Design-Build-Test-Learn cycles. Recent advances in machine learning are changing the landscape; thus, we propose that "Learning" can precede "Design". Moreover, adopting cell-free platforms can further accelerate "Building" and "Testing" for megascale data generation and models.

Synthetic biology arose and has advanced by following the simple engineering mantra of Design-Build-Test-Learn (DBTL)[1]. In the first phase of this workflow, researchers define objectives for the desired biological function and then design the parts or system they want to use. This can include introducing novel components or redesigning existing parts for a novel application[2]. The Design phase relies on domain knowledge, expertise, and computational approaches for modeling[3]. In the Build phase, DNA constructs are synthesized, assembled into plasmids or other vectors, and then introduced into the characterization system. Systems include in vivo chassis such as bacteria, eukaryotic cells, mammalian cells, and plants, or in vitro cell-free systems and synthetic cells. The Test phase determines the Design and Build phases' efficacy by experimentally measuring the engineered biological constructs' performance. The Learning phase relies on analyzing data collected during testing and comparing it to objectives set during the Design stage. This enables researchers to inform the next Design round and iterate through additional rounds of the DBTL cycle until they have reached their desired function. These cycles streamline and simplify efforts to build biological systems by providing a systematic, iterative framework for engineering.

Machine learning provides a new opportunity for directly engineering proteins and pathways with desired functions but is challenging due to the complex relationship between a protein's sequence, structure, and thus, function. Although computational models have often yielded successes[4], there are still instances where models are unable to predict how sequence changes will affect protein folding[5], stability[6], or activity[7]. Additionally, protein function often depends on the environment in which the protein is expressed, which can be difficult to anticipate in silico, and characterizations often require painstaking transformation, expression, and purification. These roadblocks argue for a different approach to the overall synthetic biology workflow that places Learning to the fore, in the form of machine learning.

The DBTL paradigm described here is not unique to protein engineering or synthetic biology. This workflow closely resembles approaches used in established engineering disciplines such as mechanical engineering, where iteration involves first gathering information, processing it, identifying design revisions, and implementing those changes[8]. In mechanical engineering, physical laws are extensively employed to model parameters such as damping, friction, and stiffness[9]. Incorporating prior knowledge from machine learning models to refine and construct designs for testing can accelerate the path to functional solutions[10].

Unsurprisingly, machine learning has also become a driving force in the synthetic biology enterprise[11]. Machine learning approaches have become dominant not because they replace physics, but because current biophysical models are computationally expensive and limited in scope when applied to the complexity of biomolecules. Machine learning methods can economically leverage large biological datasets to detect patterns in high-dimensional spaces, enabling more efficient and scalable design. Protein language models that rely on attention mechanisms are useful for designing proteins as they can capture long-range evolutionary dependencies within amino acid sequences, enabling the prediction of structure-function relationships, albeit imperfectly today. Since these models are trained on large datasets consisting of millions of protein sequences or hundreds of thousands of structures, machine learning can precede and be directly incorporated into the Design phase, allowing researchers to increasingly be able to make zero-shot (without additional training[12]) predictions that improve the functionality of protein parts (Fig. 1a).

Sequence-based protein language models—such as ESM[13] and ProGen[14]—are trained on the evolutionary relationships between protein sequences embedded in all of phylogeny. These language models are thereby capable of tasks such as predicting beneficial mutations and inferring the function of protein sequences and have proven adept at zero-shot prediction of diverse antibody sequences[14] and predicting solvent exposed and charged amino acids[15]. Even in the absence of exact prediction, pre-trained protein language models have been used to design libraries for engineering biocatalysts that have yielded enantioselective bond formation[16].

Similarly, structural models learn from the ever-expanding databases of experimentally determined structures to enable powerful zero-shot design strategies. For example, MutCompute focuses on residue-level optimization by identifying probable mutations given the local environment. MutCompute uses a deep neural network trained on protein structures and can thereby associate an amino acid with its surrounding chemical environment, allowing for prediction of potentially stabilizing and functionally beneficial substitutions[17]. The success of this method is demonstrated in engineering a hydrolase for polyethylene terephthalate (PET) depolymerization, where proteins with mutations from MutCompute had increased stability and activity compared to wild-type[18]. In contrast, ProteinMPNN is a structure-based deep learning design tool that takes an entire protein structure as input and predicts new sequences that fold into that backbone[19]. ProteinMPNN has been used to design variants of TEV protease that

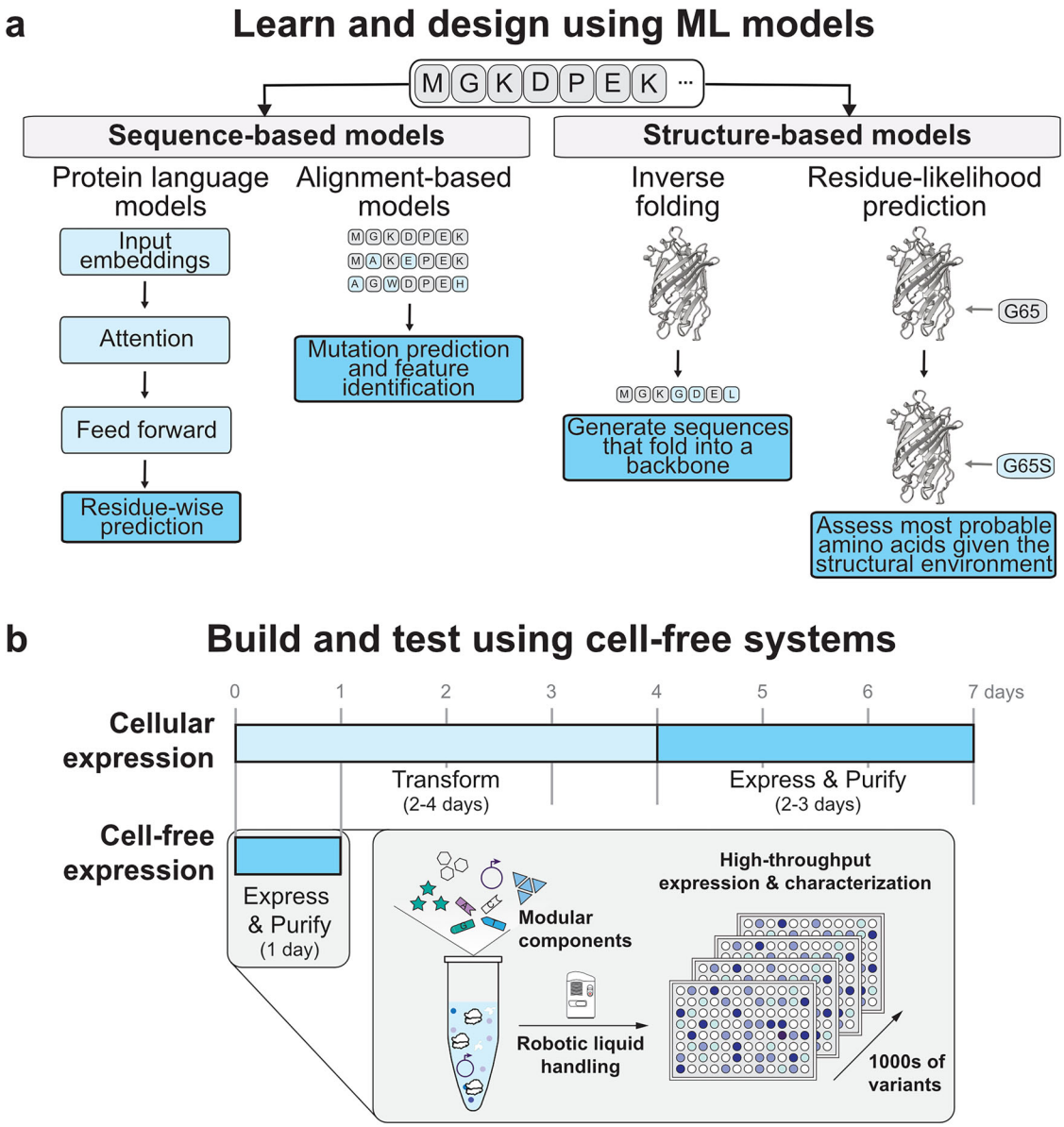

Fig. 1 | Proposed enhancements to current DBTL workflow. A Sequence- and structure-based machine learning (ML) models. Sequence-based models use amino acid sequences—they do not explicitly require knowledge of the protein structure. Structure-based models include tools to fold the protein sequence, to generate sequences that fold into the backbone, and optimize local structural regions of the protein. **B** Cell-free expression enables rapid, customizable protein synthesis and testing. Shifting from cell-based to cell-free platforms integrates with the DBTL pipeline, speeding up Build and Test steps.

improve catalytic activity compared to the parent sequence. Furthermore, it has been demonstrated that combining ProteinMPNN for sequence design with deep learning-based structure assessment (e.g., AlphaFold[20] and RoseTTAFold), leads to a nearly 10-fold increase in design success rates[21]. Hybrid approaches, such as physics-informed machine learning[22], also offer the potential to combine the predictive power of statistical models with the explanatory strength of physical principles.

In addition to purely sequence- and structure-based approaches, zero-shot methods have been augmented with additional evolutionary and biophysical information, illustrating how multiple layers of biological knowledge can enhance predictive power in protein engineering. In one example, researchers have improved upon the one-shot designed PET hydrolase by using a large language model trained on two datasets of PET hydrolase homologs and force-field-based algorithms, essentially exploring the evolutionary landscape[23]. Other examples include efforts to map sequence-fitness landscapes across multiple regions of chemical space to simultaneously engineer multiple distinct specialized enzymes[24].

Beyond models that generate designs based on a protein's sequence or structure, there are also machine learning-guided engineering models that focus on functional prediction. Two protein

properties that are frequently targeted for optimization are thermostability and solubility. The software Prethermut predicts the effects of single- or multi-site mutations using machine learning methods that were trained on a collection of experimentally measured thermodynamic stability changes of mutant proteins[25]. Similarly, Stability Oracle was trained on a collection of stability data and protein structures, using a graph-transformer architecture to learn pairwise representations of residues[26]. As an output, Stability Oracle predicts the $\Delta\Delta G$ of the protein. Both approaches can be used to eliminate potentially destabilizing mutations or to identify stabilizing ones. Finally, DeepSol is a deep learning-based tool for predicting protein solubility, relying on mapping the primary sequence (via sets of k-mers) to solubility[27]. These examples likely presage many future efforts to more finely predict functionality.

Classic synthetic biology methods play a large role in translating computational predications into the physical, biological systems, but the DBTL paradigm can be further accelerated by using cell-free methods for expression and testing of predictions (Fig. 1b)[28]. Cell-free gene expression leverages protein biosynthesis machinery obtained from either crude cell lysates or purified components[29] to activate in vitro transcription and translation. Synthesized DNA templates can be provided to cell-free systems for protein expression without intermediate, time-intensive cloning steps, and the expressed proteins can be used directly or can be further purified. Cell-free expression systems are rapid (>1 g/L protein in <4 h[30]), enable production of products that are otherwise toxic to a live cell[31], are readily scalable from the pL to kL scale[32], and can be coupled with colorimetric or fluorescent-based assays for high-throughput sequence to function mapping of protein variants[33]. The required cellular machinery can be obtained from organisms across the tree of life, and DNA and reagents can be readily exchanged due to the modular nature of cell-free expression platforms, enabling facile customization of the reaction environment. Incorporation of non-canonical amino acids and post-translational protein modifications like glycosylation and phosphorylation has also been achieved, positioning cell-free expression platforms as a highly productive and versatile strategy for high-throughput synthesis and testing of nearly any protein product or enzymatic pathway[34,35].

Cell-free systems can be readily combined with liquid handling robots and microfluidics to further scale the number of reactions and speed at which researchers can traverse the classic DBTL cycle[36]. For example, DropAI leveraged droplet microfluidics and multi-channel fluorescent imaging to screen upwards of 100,000 picoliter-scale reactions[32]. Biofoundries (e.g., ExFAB) are also increasingly leveraging cell-free platforms[37] alongside existing high-throughput workflows. Closed-loop design platforms that leverage AI agents[38] to cycle through experiments are further expanding capacity. These high-throughput capabilities of cell-free expression systems provide a powerful tool to build large datasets for training machine learning models and to test in silico predictions, including data for solving the protein expression problem[39].

Cell-free expression platforms have already been effectively paired with machine learning techniques to advance protein and pathway design. Ultra-high-throughput protein stability mapping has been achieved through coupling in vitro protein synthesis with cDNA display, allowing the ΔG calculations of 776,000 protein variants[40]. This vast dataset has been extensively utilized to benchmark various zero-shot predictors for model predictability[41]. Additional protein engineering efforts have incorporated machine learning directly into the engineering campaign through training linear supervised models on over 10,000 reactions from iterative rounds of site saturation mutagenesis data to accelerate the identification of enzyme candidates with favorable properties, which has been applied towards engineering amide synthetases[24]. Pairing deep-learning sequence generation with cell-free expression, researchers have been able to computationally survey over 500,000 antimicrobial peptides (AMP) and select 500 optimal variants to experimentally validate, leading to 6 promising AMP designs[42]. Cell-free pathway prototyping has also dramatically benefitted from incorporation of machine learning. In vitro prototyping and rapid optimization of biosynthetic enzymes (or iPROBE) uses a training set of pathway combinations and enzyme expression levels to then predict optimal pathway sets via a neural network, which has been leveraged to improve 3-HB in a *Clostridium* host by over 20-fold[43]. In summary, cell-free systems have proven to be a powerful platform towards large-scale data generation and seamlessly integrating machine learning into both protein and pathway engineering campaigns.

Overall, even with machine learning enhancements, the classic DBTL cycle requires multiple turns to gain knowledge, and the Build-Test portions of the cycle can be especially slow. The field continues to rely heavily on empirical iteration rather than predictive engineering. We propose a paradigm shift, wherein in many cases, the data that would be "learned" by Build-Test phases may already be inherent in machine learning algorithms (or alternatively new "ground truth" data sets will be generated that form the basis of foundational models). Given the increasing success of zero-shot predictions, it may be possible to reorder the cycle (and, indeed, do away with cycling altogether) via "LDBT", where Learn-Design (based on available or readily plumbed large data sets) allows an initial set of answers to be quickly built and tested, leading to a single cycle that can generate functional parts and circuits (Fig. 2). This process in turn brings synthetic biology closer to a Design-Build-Work model that relies on first principles, similar to that of disciplines like civil engineering. Such a shift would have transformative impacts on efforts to engineer biological systems and help reshape the bioeconomy.

To better enable the LDBT paradigm shift, additional (and preferably large, megascale) datasets linking sequence to structure and function must be assembled. Even with the use of machine learning-based design at the start of an LDBT cycle, it is likely that multiple iterations of designing, building, and testing biological systems will be required. There exist numerous machine learning strategies to efficiently search protein sequence space based on data generated during the Test stage. Traditional machine learning directed evolution (MLDE) utilizes sequence-function data, often with one-shot encoded mutations, to predict high-performing protein variants. MLDE has also been used with protein language models to more effectively capture long-range sequence dependencies and evolutionary information. For example, deep mutational scanning was used to train a machine learning model to predict membrane activities of antimicrobial peptides, resulting in the identification of a peptide with reduced toxicity but retained activity[44]. Bayesian optimization is another approach that allows protein engineering with few experimental measurements[45]. Usually, Gaussian processes are used to model both the predicted function and the uncertainty of protein variants. Such an approach was used to improve fatty alcohol production by two-fold with fewer than 100 experimental measurements[46]. Beyond single rounds of predictions, EVOLVEpro recently demonstrated success in engineering six different proteins with relatively few experimental data points by combining a protein language model with a regression model to learn the

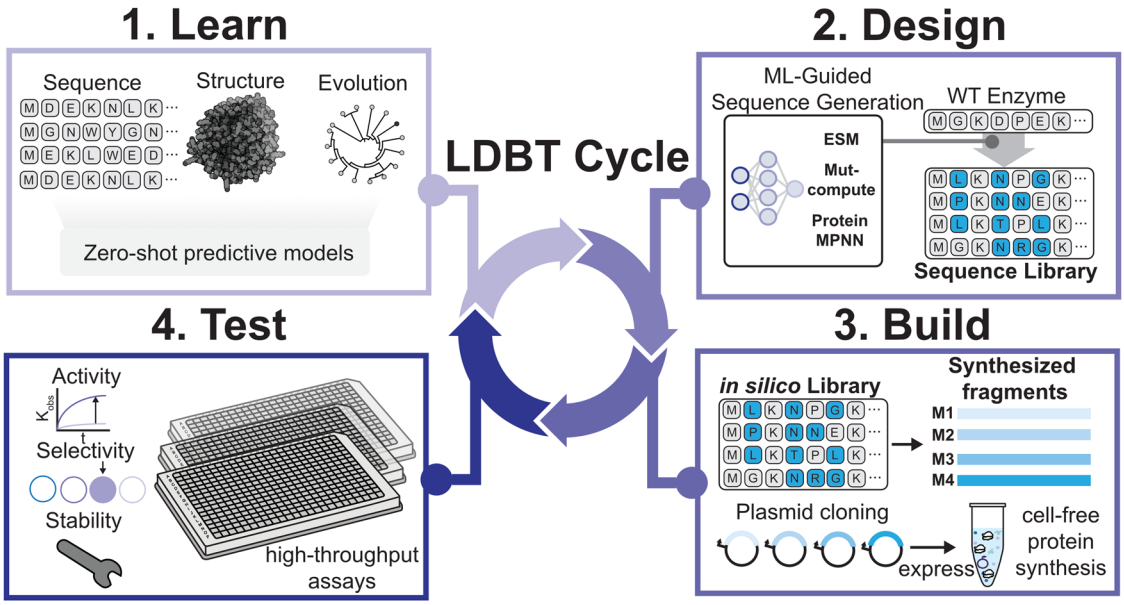

Fig. 2 | Learn-design-build-test instead of design-build-test-learn. Centering powerful new machine learning capabilities at the start of biotechnology development, complemented by high-throughput Build and Test assays, enables a shift towards LDBT instead of DBTL.

relationship between sequence embeddings and experimentally determined data[47]. By starting with small number of data points, a random forest regression model could be trained, and then after each round additional data points were added to the dataset to retrain the model, allowing the successive prediction of multi-mutant variants from single-variant data, a typically challenging task in engineering studies.

The development of predictive machine learning models depends on the availability of high-quality training data. Initiatives such as the Align Foundation facilitates the generation of open-access datasets to allow researchers to build on one another's work[48]. Community-driven design challenges play a key role, allowing researchers to evaluate and iteratively improve predictive models in protein engineering. However, the push for open-access data can be accompanied by tensions; for example, BaseData by Basecamp Research includes billions of protein sequences collected from diverse environments, and their public release raises questions regarding benefit sharing, legal frameworks, and data ownership[49]. Conversely, private companies are expansively developing proprietary datasets that may be inaccessible to the broader synthetic biology community, while new algorithms are also being held increasingly behind walls, at least upon initial release.

Ultimately, we envision enhanced machine learning approaches combined with cell-free protein synthesis as a facile way to express the necessary proteins (both homologs and mutants), wherein generalized assays can be used to quickly assess expression, function, and protein-protein interactions. Machine learning enhances the Learn phase by allowing zero-shot predictions of beneficial protein variants as well as enabling rapid analysis of experimental data. Cell-free systems (up to[50] and including synthetic cells[51]) accelerate the Design, Build, and Test phases through rapid evaluation of genetic constructs. Looking ahead, we anticipate that LDBT cycles may be limited primarily by the speed of DNA synthesis and data generation for models. It may be that bespoke, local DNA synthesis, rather than corporate delivery, will be the most viable option for the future to address this challenge, further revising where economies of scale may lie.

To extend these advances beyond protein engineering, further progress is required to expand modeling to additional biomolecules, pathways, and ultimately metabolism as a whole, and to continue to develop scalable methods to model even complex biological systems and functions. The greatest obstacles remain the scarcity of high-quality data and the difficulties inherent in its analysis[52]. Yet, the promise of rapidly going from "*desired function*" to "*designed sequence*" to "*working protein/function*" in a reimagined LDBT cycle holds promise to unlock the full design space of biology.

Alia Clark-ElSayed [1], Isa Madrigal Harrison [1], Meagan L. Olsen [2,3], John T. Lazar [4], Michael C. Jewett [2,3,4] ✉ & Andrew D. Ellington [1] ✉

[1]Department of Molecular Biosciences, University of Texas at Austin, Austin, TX, USA. [2]Department of Chemical and Biological Engineering, Northwestern University, Evanston, IL, USA. [3]Center for Synthetic Biology, Northwestern University, Evanston, IL, USA. [4]Department of Bioengineering, Stanford University, Stanford, CA, USA.
✉e-mail: mjewett@stanford.edu; ellingtonlab@gmail.com

Received: 8 August 2025; Accepted: 13 October 2025;
Published online: 05 November 2025

References

1. Liu, R., Bassalo, M. C., Zeitoun, R. I. & Gill, R. T. Genome-scale engineering techniques for metabolic engineering. *Metab. Eng.* **32**, 143–154 (2015).
2. Endy, D. Foundations for engineering biology. *Nature* **438**, 449–453 (2005).
3. Salis, H. M., Mirsky, E. A. & Voigt, C. A. Automated design of synthetic ribosome binding sites to control protein expression. *Nat. Biotechnol.* **27**, 946–950 (2009).
4. Kouba, P. et al. Machine Learning-Guided Protein Engineering. *ACS Catal.* **13**, 13863–13895 (2023).
5. Buel, G. R. & Walters, K. J. Can AlphaFold2 predict the impact of missense mutations on structure?. *Nat. Struct. Mol. Biol.* **29**, 1–2 (2022).
6. Fang, J. A critical review of five machine learning-based algorithms for predicting protein stability changes upon mutation. *Brief. Bioinform.* **21**, 1285–1292 (2020).

7. Livesey, B. J. & Marsh, J. A. Using deep mutational scanning to benchmark variant effect predictors and identify disease mutations. *Mol. Syst. Biol.* **16**, e9380 (2020).
8. Costa, R. & Sobek, D. K. Iteration in engineering design: inherent and unavoidable or product of choices made? In *Proc. 15th International Conference on Design Theory and Methodology* 669–674. https://doi.org/10.1115/DETC2003/DTM-48662 (ASMEDC, 2003).
9. Merino-Olagüe, M., Iriarte, X., Castellano-Aldave, C. & Plaza, A. Hybrid modelling and identification of mechanical systems using physics-enhanced machine learning. *Eng. Appl. Artif. Intell.* **159**, 111762 (2025).
10. Castle, S. D., Stock, M. & Gorochowski, T. E. Engineering is evolution: a perspective on design processes to engineer biology. *Nat. Commun.* **15**, 3640 (2024).
11. Gherman, I. M. et al. Bridging the gap between mechanistic biological models and machine learning surrogates. *PLOS Comput. Biol.* **19**, e1010988 (2023).
12. Meier, J. et al. Language models enable zero-shot prediction of the effects of mutations on protein function. Preprint at https://doi.org/10.1101/2021.07.09.450648 (2021).
13. Rives, A. et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl Acad. Sci.* **118**, e2016239118 (2021).
14. Nijkamp, E., Ruffolo, J. A., Weinstein, E. N., Naik, N. & Madani, A. ProGen2: Exploring the boundaries of protein language models. *Cell Syst.* **14**, 968–978.e3 (2023).
15. Kulikova, A. V. et al. Two sequence- and two structure-based ML models have learned different aspects of protein biochemistry. *Sci. Rep.* **13**, 13280 (2023).
16. Ding, K. et al. Machine learning-guided co-optimization of fitness and diversity facilitates combinatorial library design in enzyme engineering. *Nat. Commun.* **15**, 6392 (2024).
17. Shroff, R. et al. Discovery of Novel Gain-of-Function Mutations Guided by Structure-Based Deep Learning. *ACS Synth. Biol.* **9**, 2927–2935 (2020).
18. Lu, H. et al. Machine learning-aided engineering of hydrolases for PET depolymerization. *Nature* **604**, 662–667 (2022).
19. Dauparas, J. et al. Robust deep learning–based protein sequence design using ProteinMPNN. *Science* **378**, 49–56 (2022).
20. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
21. Bennett, N. R. et al. Improving de novo protein binder design with deep learning. *Nat. Commun.* **14**, 2625 (2023).
22. Omar, S. I., Keasar, C., Ben-Sasson, A. J. & Haber, E. Protein design using physics informed neural networks. *Biomolecules* **13**, 457 (2023).
23. Cui, Y. et al. Computational redesign of a hydrolase for nearly complete PET depolymerization at industrially relevant high-solids loading. *Nat. Commun.* **15**, 1417 (2024).
24. Landwehr, G. M. et al. Accelerated enzyme engineering by machine-learning guided cell-free expression. *Nat. Commun.* **16**, 865 (2025).
25. Tian, J., Wu, N., Chu, X. & Fan, Y. Predicting changes in protein thermostability brought about by single- or multi-site mutations. *BMC Bioinform.* **11**, 370 (2010).
26. Diaz, D. J. et al. Stability Oracle: a structure-based graph-transformer framework for identifying stabilizing mutations. *Nat. Commun.* **15**, 6170 (2024).
27. Khurana, S. et al. DeepSol: a deep learning framework for sequence-based protein solubility prediction. *Bioinformatics* **34**, 2605–2613 (2018).
28. Silverman, A. D., Karim, A. S. & Jewett, M. C. Cell-free gene expression: an expanded repertoire of applications. *Nat. Rev. Genet.* **21**, 151–170 (2020).
29. Shimizu, Y. et al. Cell-free translation reconstituted with purified components. *Nat. Biotechnol.* **19**, 751–755 (2001).
30. Jewett, M. C. & Swartz, J. R. Mimicking the Escherichia coli cytoplasmic environment activates long-lived and efficient cell-free protein synthesis. *Biotechnol. Bioeng.* **86**, 19–26 (2004).
31. Salehi, A. S. M. et al. Cell-free protein synthesis of a cytotoxic cancer therapeutic: onconase production and a just-add-water cell-free system. *Biotechnol. J.* **11**, 274–281 (2016).
32. Zhu, J. et al. AI-driven high-throughput droplet screening of cell-free gene expression. *Nat. Commun.* **16**, 2720 (2025).
33. GAN, R. et al. High-throughput regulatory part prototyping and analysis by cell-free protein synthesis and droplet microfluidics. *ACS Synth. Biol.* **11**, 2108–2120 (2022).
34. Hunt, A. C. et al. Cell-free gene expression: methods and applications. *Chem. Rev.* **125**, 91–149 (2025).
35. Garenne, D. et al. Cell-free gene expression. *Nat. Rev. Methods Prim.* **1**, 49 (2021).
36. Hunt, A. C. et al. A rapid cell-free expression and screening platform for antibody discovery. *Nat. Commun.* **14**, 3897 (2023).
37. Hérisson, J., Hoang, A. N., El-Sawah, A., Khalil, M. M. & Faulon, J.-L. Operate a cell-free biofoundry using large language models. Preprint at https://doi.org/10.1101/2024.10.28.619828 (2024).
38. Rapp, J. T., Bremer, B. J. & Romero, P. A. Self-driving laboratories to autonomously navigate the protein fitness landscape. *Nat. Chem. Eng.* **1**, 97–107 (2024).
39. Baranowski, C. et al. Can protein expression be 'solved'? Trends Biotechnol. 0, (2025).
40. Tsuboyama, K. et al. Mega-scale experimental analysis of protein folding stability in biology and design. *Nature* **620**, 434–444 (2023).
41. Notin, P. et al. ProteinGym: large-scale benchmarks for protein fitness prediction and design. *Adv. Neural Inf. Process. Syst.* **36**, 64331–64379 (2023).
42. Pandi, A. et al. Cell-free biosynthesis combined with deep learning accelerates de novo-development of antimicrobial peptides. *Nat. Commun.* **14**, 7197 (2023).
43. Karim, A. S. et al. In vitro prototyping and rapid optimization of biosynthetic enzymes for cell design. *Nat. Chem. Biol.* **16**, 912–919 (2020).
44. Randall, J. R., Vieira, L. C., Wilke, C. O. & Davies, B. W. Deep mutational scanning and machine learning for the analysis of antimicrobial-peptide features driving membrane selectivity. *Nat. Biomed. Eng.* **8**, 842–853 (2024).
45. Freschlin, C. R., Fahlberg, S. A. & Romero, P. A. Machine learning to navigate fitness landscapes for protein engineering. *Curr. Opin. Biotechnol.* **75**, 102713 (2022).
46. Greenhalgh, J. C., Fahlberg, S. A., Pfleger, B. F. & Romero, P. A. Machine learning-guided acyl-ACP reductase engineering for improved in vivo fatty alcohol production. *Nat. Commun.* **12**, 5825 (2021).
47. Jiang, K. et al. Rapid in silico directed evolution by a protein language model with EVOLVEpro. *Science* **387**, eadr6006 (2025).
48. The Align Foundation—Align to Innovate Public Research Data. https://alignbio.org/.
49. Vince, O. et al. Breaking through biology's data wall: expanding the known tree of life by over 10x using a global biodiscovery pipeline. Preprint at https://doi.org/10.1101/2025.06.11.658620 (2025).
50. Hodgman, C. E. & Jewett, M. C. Cell-free synthetic biology: thinking outside the cell. *Metab. Eng.* **14**, 261–269 (2012).
51. Adamala, K. P., Martin-Alarcon, D. A., Guthrie-Honea, K. R. & Boyden, E. S. Engineering genetic circuit interactions within and between synthetic minimal cells. *Nat. Chem.* **9**, 431–439 (2017).
52. Camacho, D. M., Collins, K. M., Powers, R. K., Costello, J. C. & Collins, J. J. Next-generation machine learning for biological networks. *Cell* **173**, 1581–1592 (2018).

Acknowledgements

This work was supported by the National Science Foundation (CBET – 2341123; TIPP – 2448653, 2449206, 2448820), the Department of Energy (DE-SC0023278, DE-NA0003525.), the National Research Foundation of Korea, and the Army Research Office (W911NF-22-2-0210). MLO acknowledges support from the National Science Foundation Graduate Research Fellowship under grant no. DGE-1842165. This work was supported by the Welch Foundation (F-1654) and the NIH (R01GM146093-01A1). Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number T32GM139796 to ACE.

Author contributions

All authors contributed to the conceptualization and writing of the manuscript.

Competing interests

M.C.J. has a financial interest in Pearl Bio, Inc., Synolo Therapeutics, Ridge Bio, and Stemloop Inc. M.C.J.'s interests are reviewed and managed by Stanford University and Northwestern University in accordance with their competing interest policies. All other authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Michael C. Jewett or Andrew D. Ellington.

Peer review information *Nature Communications* thanks Olivier Borkowski and Thomas Gorochowski for their contribution to the peer review of this work.