https://doi.org/10.1038/s41587-025-02607-x

The Protein Design Archive (PDA): insights from 40 years of protein design

While natural proteins are incredibly diverse biomacromolecules, they present only a subset of the sequence and structural space that is chemically possible[1]. The field of protein design has the potential to illuminate the unexplored regions of this space, to deepen our understanding of sequence–structure–function relationships, and to enable the development of novel proteins that could be used to address societal challenges.

Protein design has evolved dramatically over the past 40 years from rational design to contemporary data-driven approaches. The field has now reached milestones such as the design of small-molecule[2] and protein[3] binders, entirely novel folds[4–6], and has been recognized with a Nobel Prize. However, it can be difficult to identify current challenges and opportunities within the field as no easily searchable resource provides an overview of proteins that have been designed to date. Such a resource would improve our ability to learn from previous efforts and guide the development of future design methods.

Here, we present the Protein Design Archive (PDA), a website (https://pragmaticprotein-design.bio.ed.ac.uk/pda/) and database that records de novo protein designs that have been structurally characterized. As of the beginning of 2025, the PDA database contains over 1,500 designed protein structures, including historic designs that are not described in any publication. Data in the PDA were collected through automated querying of the Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB)[7] for polypeptides of synthetic origin (National Center for Biotechnology Information taxonomy identifier 32630). The results were then hand curated, on the basis of knowledge of the field and literature research, to ensure only high-quality data were included. A detailed methodology is provided in the Supplementary Information and regularly updated lists of manually included and excluded entries are provided on the website and on the GitHub repository for the project (https://github.com/wells-wood-research/protein-design-archive).

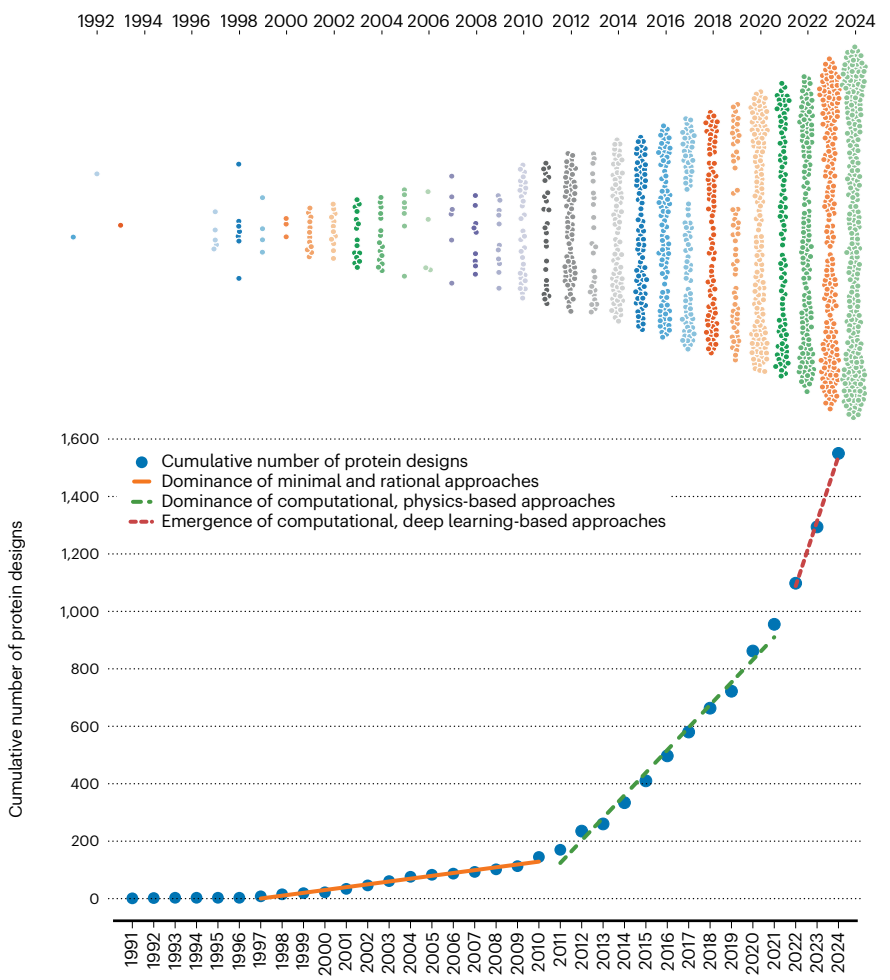

Fig. 1 | Rapid increase in the number designed proteins that have been structurally characterized. **a**, Growth of the number of protein designs contained in the PDA database over time. This timeline is also the focus of the Home page view of the PDA website. The designs are arranged as points on a timeline plot from 1991 to 2024, showing the rapid growth of designs across the years. **b**, Growth of designed protein structures. Three regions have been distinguished on the basis of the dominant approaches to protein design over these time periods: minimal and rational (1997–2010), computational physics-based (2010–2021), and computational deep learning-based (from 2021). Statistics of the fitted curves: minimal and rational: slope = 9.88, R^2 = 97%; computational physics-based: slope = 78.61, R^2 = 99%; computational deep learning-based: slope = 226.00, R^2 = 99%.

The PDA's interface enables browsing, data export, and filtering by criteria such as free text search, release date and novelty, which is measured by maximum structure- or sequence-based similarity to natural proteins. The website is organized around the Home page, structured as a timeline (Fig. 1a) followed by an alphabetical list displaying all protein designs. Each design has a dedicated Details page, summarizing information such

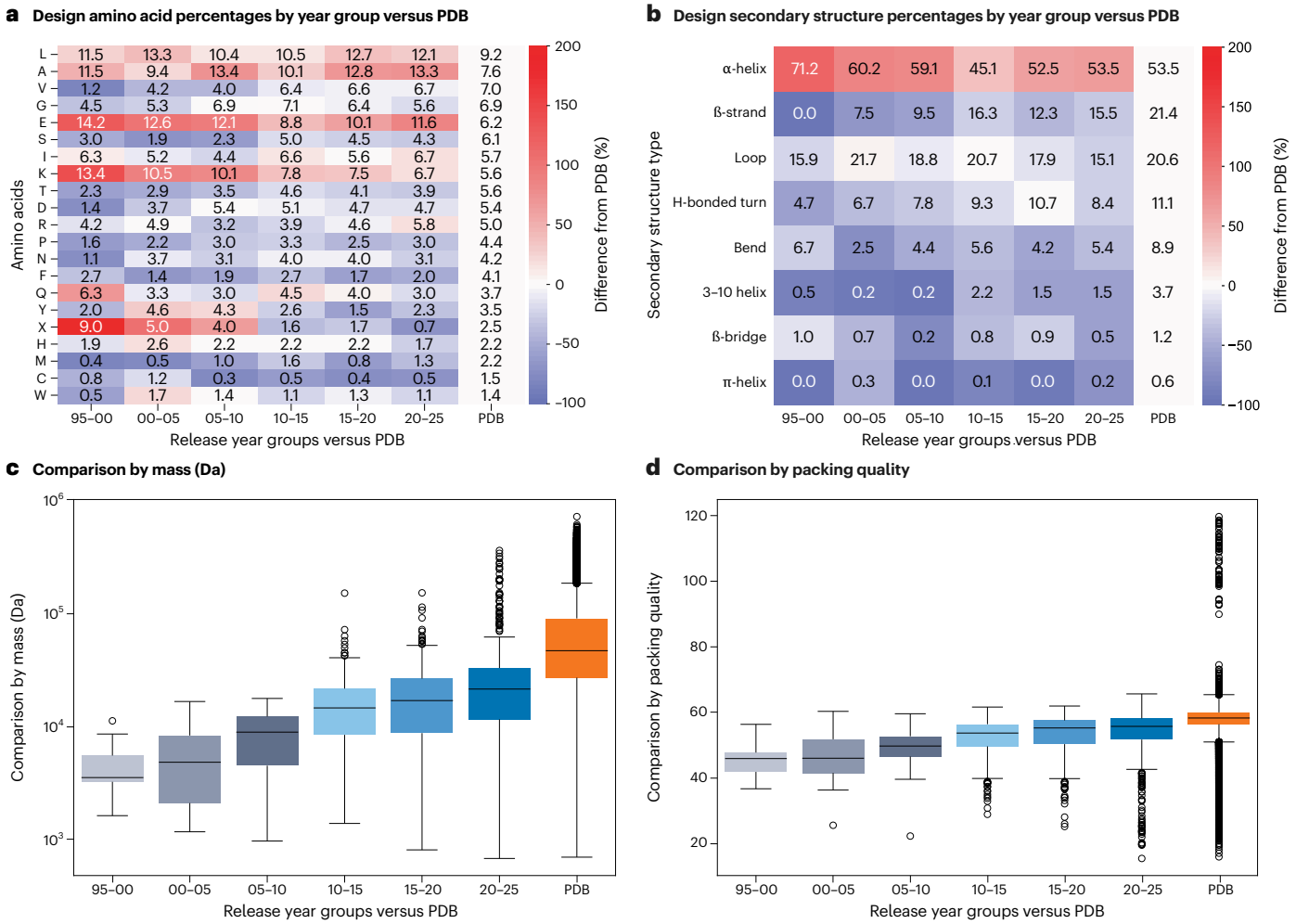

Fig. 2 | Designed protein complexity compared to the PDB over time. a, A heat map showing amino acid proportions for designs versus the PDB by release year group. The heat map is colored by the percentage increase (red) or decrease (blue) against the PDB. X denotes non-standard amino acids. **b**, A heat map showing secondary structure proportions for designs vs the PDB by release year group. The heat map is colored by the percentage increase (red) or decrease (blue) against the PDB. **c**, Box plots of mass in daltons for the designed proteins by release year group against the PDB. The *y* axis is on a logarithmic scale. **d**, Box plots of packing density of the designed proteins by release year group against the PDB.

as source publication, three-dimensional structure, sequence, and similarity scores to other designed and natural proteins.

Systematic analysis of the PDA reveals a rapid increase in the number and complexity of designs over time (Fig. 2). A growth curve (Fig. 1b and Supplementary Tables 1 and 2) reveals three distinct regions distinguished by their dominant approach to protein design: minimal and rational (1997–2010), computational, physics-based (2010–2021) and computational, deep learning-based (from 2021). The annual rate of release of new designs increased almost tenfold between the regions characterized by minimal and rational methods and computational, physics-based. It is likely that multiple factors catalyzed this shift, but worthy of particular mention is the release of more accessible computational methods and software for protein design, such as Rosetta3[8] and PyRosetta[9] in 2009. In recent years (2021–2024) the number of designs released annually has almost tripled compared with the previous decade, potentially marking the transition into a new era of protein design — one driven by artificial intelligence.

Analysis of designed proteins using DE-STRESS[10] uncovers biases toward specific amino acid usage and secondary structure types, such as a larger proportion of α-helical secondary structure, when compared to the PDB (Fig. 2a,b). Despite these differences, there is evidence that designed proteins are becoming more complex in recent years (Fig. 2c,d). Sequence-based similarity bit scores, calculated using MMseqs2[11], and structure-based similarity LDDT (local distance difference test) scores, calculated using Foldseek[12], are routinely computed and published as part of our data preparation and monthly website updates. These data give a valuable insight into similar natural proteins and related designs. A comprehensive analysis of the field, encompassing all designs, is presented in Supplementary Figs. 1–4 and Supplementary Tables 3 and 4.

Building the PDA posed some challenges, including inconsistent metadata and a lack of universally accepted definition of 'de novo design'. We have attempted to be as inclusive as possible with our definition of designed

protein while providing the user with tools to filter the proteins in the PDA according to their own criteria. The principles for selection for inclusion in the PDA can be summarized as follows: (1) The entity is a polymer of amino acids. (2) The entity is not found in nature. (3) The entity was created with a clear design intent. As such, the PDA does not include mutants of natural proteins. Many designed peptides are included, although we use a minimum sequence length threshold of 14 amino acids, which is aligned with the limits of the similarity analysis tool that we have used (see above). Exceptions are made for polypeptides that have played a pivotal role in protein design because of either their novelty at the time of publication or their impact on our understanding of sequence-to-structure relationships that guide design principles.

One current limitation of the PDA is that it includes only structural data and, while this is critical for validating designs, it accounts for only a fraction of information collected regarding designed proteins and protein design methodologies. Furthermore, these data represent only 'successful' designs. To drive further progress, information regarding all aspects of the design process must be collected. In the future, we intend to expand the PDA to include information on design protocols, sequence information and results of biochemical and biophysical analysis.

There have been many attempts to capture progress in the ever-evolving field of protein design. While many excellent reviews exist[13–15], rapid progress ensures that no single review can comprehensively cover the breadth of the field, and these efforts will become increasingly difficult with the accelerated rise in the number of designed proteins. The PDA offers a solution by providing a regularly updated, open access repository with methods described in detail to ensure transparency and reproducibility. Serving as a central resource for the community, the PDA will enable trend analysis, benchmarking of novel design tools, and the identification of rational design principles for functional design. Ultimately, we believe that the PDA will be invaluable in guiding development of protein design, thereby accelerating advancements, such as biologic development and enzyme design, and helping the field reach its full potential.

Data availability

The PDA is available without registration at https://pragmaticproteindesign.bio.ed.ac.uk/pda/. Source data are provided with this paper.

Code availability

Code supporting PDA data collection and processing is available on GitHub: https://github.com/wells-wood-research/chronowska-stam-wood-2024-protein-design-archive. This includes files detailing manual curation of the PDA dataset, such as csv files listing proteins that were manually included and excluded, and how information was manually corrected. Code used to build the PDA website is available on GitHub: https://github.com/wells-wood-research/protein-design-archive.

Marta Chronowska[1], Michael J. Stam[1], Derek N. Woolfson[2], Luigi F. Di Costanzo[3] & Christopher W. Wood[1]

[1]University of Edinburgh, School of Biological Sciences, Institute of Quantitative Biology, Biochemistry and Biotechnology, Edinburgh, UK. [2]University of Bristol, Schools of Chemistry and of Biochemistry, Bristol BioDesign Institute, Max Planck-Bristol Centre, Bristol, UK. [3]Department of Agriculture, University of Napoli Federico II, Portici, Italy.

e-mail: luigi.dicostanzo4@unina.it; chris.wood@ed.ac.uk

Published online: 21 March 2025

References

1. Baker, D. *Protein Sci.* **28**, 678–683 (2019).
2. Lu, L. et al. *Science* **384**, 106–112 (2024).
3. Gainza, P. et al. *Nature* **617**, 176–184 (2023).
4. Kuhlman, B. et al. *Science* **302**, 1364–1368 (2003).
5. Thomson, A. R. et al. *Science* **346**, 485–488 (2014).
6. Anishchenko, I. *Nature* **600**, 547–552 (2021).
7. Burley, S. K. et al. *Nucleic Acids Res.* **49**, D437–D451 (2021). (D1).
8. Leaver-Fay, A. et al. *Methods Enzymol.* **487**, 545–574 (2011).
9. Chaudhury, S., Lyskov, S. & Gray, J. J. *Bioinformatics* **26**, 689–691 (2010).
10. Stam, M. J. & Wood, C. W. *Protein Eng. Des. Sel.* **34**, gzab029 (2021).
11. Steinegger, M. & Söding, J. *Nat. Biotechnol.* **35**, 1026–1028 (2017).
12. van Kempen, M. *Nat. Biotechnol.* **42**, 243–246 (2024).
13. Huang, P.-S., Boyken, S. E. & Baker, D. *Nature* **537**, 320–327 (2016).
14. Woolfson, D. N. *J. Mol. Biol.* **433**, 167160 (2021).
15. Kortemme, T. *Cell* **187**, 526–544 (2024).

Acknowledgements

We thank the members of the Wells Wood Research Group for testing and feedback on the PDA website. M.C. is supported by a PhD studentship from the UK Research and Innovation-funded EastBio Doctoral Training Partnership programme. M.J.S., C.W.W. and D.N.W. are supported by a UKRI Biotechnology and Biological Sciences Research Council Strategic Longer and Larger award (BB/X003027/1).

Author contributions

M.C. and C.W.W. created the website, with support from M.J.S. M.C. and L.F.D.C. collected the data. M.C. and M.J.S. processed and analyzed the data. C.W.W. and D.N.W. conceived the project and secured the funding. M.C. with support of M.J.S. and C.W.W. prepared the manuscript; D.N.W. and L.F.D.C. edited the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-025-02607-x.