https://doi.org/10.1038/s41594-024-01244-3

Structure-guided drug discovery: back to the future

Cheryl H. Arrowsmith

Over the past 30 years, the field of structural biology and its associated biological insights have seen amazing progress. In this Comment, I recount several milestones in the field and how we can apply lessons from the past toward an exciting future, especially as it relates to drug discovery.

In 1994, when I published one of my first three-dimensional (3D) structures in *Nature Structural Biology* — the nuclear magnetic resonance structure of the tetramerization domain of p53 — I never imagined that today almost all independently folded proteins (or domains thereof) could be predicted with high accuracy, and that massive molecular machines and multiprotein complexes could be imaged in atomic detail by cryogenic electron microscopy (cryo-EM) and transmission microscopy. As I reflect on the trajectory of my research career from one-protein-at-a-time structural biology studies, to many-proteins-at-a-time studies (structural genomics), to aspirations toward high-throughput structure-guided drug discovery, I consider here some learnings and envision a bright future at the convergence of the fields of molecular, structural, computational and chemical biology.

The growth and value of open science, data sharing and transparency

In the pre-genomic era of the 1990s, structural biologists were often working on the same small subset of proteins that (1) were known and sequenced, (2) were small enough to be amenable to the early nuclear magnetic resonance and X-ray diffraction technologies, and (3) could be produced recombinantly. Layered onto this was the understandable need to work on 'hot' proteins (such as p53), as defined by the biological and medical studies of others. As high-resolution structures were so novel at the time, each new structure was an exciting revelation, and researchers were not expected to include related functional studies in these publications. The downside of the field at this time was that scientists were competing fiercely to be the first to solve the structure of each new 'hot' protein that was discovered by biologists. And although the Protein Data Bank (PDB) was a pioneer in the public data sharing of 3D structures, there was not a requirement to deposit structure coordinates or associated data into the PDB, and sharing of methods and data was not common, except for those developing methods and software for data collection and analysis.

Fast forward to the early genomic era circa 2000. The genomics community impacted our work profoundly, not only by unveiling a new universe of encoded proteins with unknown 3D structures and functions but also, and just as importantly, in my opinion, by establishing an ethos of data sharing and open science. Journals began to require deposition of structural and '-omics' data in relevant databases, which democratized access to this new knowledge. Structural biologists were no longer restricted to a very small proportion of the protein universe, and the concept of structural genomics came about. Many initiatives were started with various objectives, from solving new protein folds so that eventually all proteins could be computationally modeled, to populating the PDB with structures of drug targets to enable drug discovery, to providing structural annotation of cellular signaling and disease pathways. A common objective of all these programs was technology development in protein production, automation, and data collection and analysis to accelerate the rate of structure determination. The structural genomics community also embraced open science principles, often placing structures in the PDB well before their publication, and collaborated to establish standards and requirements for the deposition of associated data in the PDB and in publications.

Subsequent progress in the field throughout the 2000s enabled us to tackle increasingly difficult proteins and biological questions. The number of structures in the PDB was growing exponentially, and structural biology was becoming increasingly integral to understanding the molecular aspects of biological systems, as indicated by the expansion of the name of this journal. Interestingly, structural biologists, out of necessity, were becoming the most skilled scientists in producing and characterizing purified proteins. The impact of these advances was especially felt in the pharmaceutical and biotechnology field.

Protein science fuels chemical biology

By the 2010s, the PDB was full of structures of known and putative drug targets — some only recognized as such years later. No longer were structures of candidate drug targets mostly delivered too late to support a drug-discovery program, and structure-enabled drug discovery was increasingly common. It was also becoming clear that pharmacological tools, such as chemical probes[1,2], were among the most impactful reagents for investigating and understanding biological and disease pathways[3], and most of the wherewithal for pursuing a chemical probe or drug-discovery program (other than synthetic chemistry) was a by-product of a structural biology study: protein expression and purification protocols and reagents and, very often, biophysical, enzymatic and cellular assays for characterizing the protein system.

Recognizing the value of protein-based abilities, we at the Structural Genomics Consortium launched programs in chemical probe discovery for proteins involved in epigenetic, kinase and ubiquitin signaling, as well as probes donated from the internal programs of our pharmaceutical industry partners. We rigorously characterized these chemical tools for selectivity and potent cellular activity and made them available as unencumbered reagents for the global research community. Their use has catalyzed groundbreaking research that has led to more than 80 clinical trials and preclinical programs. The open science aspect of our medicinal chemistry public–private partnership was very unusual at the time, given the tendency for secrecy and patenting in the pharmaceutical field. But we all realized that science would advance more quickly through the sharing of these early drug-discovery tools, particularly for linking the inhibition of specific drug targets to the modulation of disease pathology. The value of chemical probes as tools for uncovering and understanding new drug targets inspired us to launch Target 2035 (ref. 4) – a global initiative to find selective pharmacological modulators for most human proteins by the year 2035.

Protein science meets AI

Concurrent with these developments, cryo-EM was also maturing to the stage at which it could contribute atomic-level models of spectacularly complicated proteins and macromolecular complexes, and today it is also a major tool in molecular biology and drug discovery. The other key advance over the past decade has been the progress in artificial intelligence (AI) and machine learning, which, combined with the wealth of public domain structures in the PDB, was able to achieve one of the holy grails of protein science: accurate prediction of 3D protein structure from sequence[5,6], as validated by the long-standing community benchmarking program CASP[7].

So, what is next? I believe the future is bright with ever-improving technologies for enabling and exploiting protein and molecular structure research in biology and medicine. Undoubtedly AI and machine learning will serve important roles in creating highly enabling tools when the requisite amount of well-curated training data and unbiased benchmarks and standards are available. A logical extension of the success of AlphaFold2 and analogous initiatives would be the ability to predict drug-like or endogenous small molecules that bind to a given protein. However, the paucity of robust, curated, public-domain protein–small molecule interaction data is currently a major impediment to this goal[8]. Such experimental data are still very challenging to generate, as they require substantial infrastructure, dedicated experimental efforts and massive datasets of positive and negative data, formatted in a uniform machine-readable format. Although public databases such as ChEMBL and PubChem are laudable resources, they often lack negative data and encompass data from heterogenous experimental protocols, which makes comparative analyses difficult. Moreover, community-developed data-quality standards and protocols for sharing massive datasets are also needed for both protein–small molecule interaction data and the rapidly increasing amount of cryo-EM data. I argue that a collaborative and open-science approach is again needed to address these issues, as was the case for genomics initiatives.

Structural biologists paved the way for AlphaFold2 with massive amounts of standardized, curated, open-access data coupled with the CASP benchmarking program. As a community, we can do the same for drug discovery. I am optimistic that this will happen, due to the growing trend in open science and multidisciplinary collaborations. For example, efforts are underway to generate the requisite data, to store and enable open access to machine-learning-ready data (https://aircheck.ai), and to benchmark progress in computational hit-finding (CACHE (Critical Assessment of Computational Hit-finding Experiments))[9]. As these and related initiatives grow, we can envision the eventual development of computational methods powerful enough to make protein–ligand discovery a largely computational exercise in the future. This, in turn, will greatly accelerate progress toward making Target 2035, and our understanding of the human proteome, a reality.

Cheryl H. Arrowsmith [1,2,3] ✉

[1]Princess Margaret Cancer Centre, Toronto, Ontario, Canada. [2]Structural Genomics Consortium, University of Toronto, Toronto, Ontario, Canada. [3]Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada.

✉e-mail: cheryl.arrowsmith@uhn.ca

Published online: 14 March 2024

References

1. Frye, S. *Nat. Chem. Biol.* **6**, 159–161 (2010).
2. Arrowsmith, C. et al. *Nat. Chem. Biol.* **11**, 536–541 (2015).
3. Edwards, A. et al. *Nature* **470**, 163–165 (2011).
4. Carter, A. J. et al. *Drug Discov. Today* **26**, 607–612 (2019).
5. Jumper, J. et al. *Nature* **596**, 583–589 (2021).
6. Baek, M. et al. *Science* **373**, 871–876 (2021).
7. Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K. & Moult, J. *Proteins* **91**, 1539–1549 (2020).
8. *Nature* **617**, 438 (2023).
9. Ackloo, S. et al. *Nat. Rev. Chem.* **6**, 287–295 (2022).

Acknowledgements

C.H.A. is funded by the Canadian Institutes for Health Research (FDN154328 and OGB190363), Canada foundation for Innovation, National Institutes of Health (U19 AI171292), Huntington Society of Canada, Princess Margaret Cancer Centre, The Princess Margaret Cancer Foundation, and Ontario Ministry of Health. The Structural Genomics Consortium is a registered charity (no. 1097737) that receives funds from Bayer AG, Boehringer Ingelheim, Bristol Myers Squibb, Genentech, Genome Canada through Ontario Genomics Institute (OGI-196), EU/EFPIA/OICR/McGill/KTH/Diamond Innovative Medicines Initiative 2 Joint Undertaking (EUbOPEN grant 875510), Janssen, Merck (known as EMD in Canada and the United States), Pfizer and Takeda.

Competing interests

The author declares no competing interests.