https://doi.org/10.1038/s41580-023-00659-y

Challenges and opportunities in macromolecular structure determination

Xiao-chen Bai, Tamir Gonen, Angela M. Gronenborn, Anastassis Perrakis, Andrea Thorn & Jianyi Yang

Embedded within the complexity of biological systems lies a formidable task: deciphering the intricate architecture of macromolecules. In this Viewpoint, a panel of experts discuss the key challenges and opportunities of macromolecular structure determination, highlighting the crucial synergy between empirical experimentation and artificial intelligence-based techniques in unravelling these complexities.

What are the main challenges of protein structure determination today and how does your research aim at addressing these?

Angela Gronenborn: Structural biologists are keenly aware of the fact that to determine structures by any of the three major experimental methodologies, X-ray crystallography, NMR and cryo-EM, purified, homogeneous materials are a prerequisite. This often involves lengthy biochemical explorations such as optimization of expression constructs, isotope labelling strategies, buffer conditions, temperature and pH. However, even if suitable samples are available, using one methodology alone will only provide a partial picture of the system at hand. Therefore, my research uses, as much as possible, integrated structural biology approaches whereby each technique addresses specific features of the 'functional protein'. As I stated previously: "The current reality of all scientific disciplines engaged in elucidating protein structure from sequence to structure to function is one of multiple methods, models, and representations, investigating different features of a phenomenon in a variety of contexts"[1]. Given that each method can only provide a partial representation of an object under study, different theoretical and experimental approaches, models and descriptions have to be integrated to yield a holistic picture of the phenomenon at hand. This has been termed 'integrative pluralism'[2].

Xiao-chen Bai: Since 2013, single-particle cryo-electron microscopy (cryo-EM) has emerged as a formidable competitor to X-ray crystallography. It has been widely used to determine the structures of large protein complexes or membrane proteins that are difficult to crystallize. However, a significant challenge arises from the inherent dynamic nature of most macromolecular protein complexes during cellular processes. When protein samples are extracted from cells, they often contain various conformational or compositional states. This poses a major obstacle to the structural determination of these complexes using single-particle cryo-EM. To overcome this challenge, researchers have developed various 3D classification methods to separate particles with distinct conformational states. Unfortunately, the current 3D classification approaches cannot handle high heterogeneity or continuous conformational variability. Consequently, regions with substantial structural flexibility can remain undetectable in the cryo-EM map, resulting in incomplete structures. This limitation hampers our understanding of protein dynamics, which has a crucial role in biological processes.

For example, receptor tyrosine kinases (RTKs) consist of an extracellular domain (ECD), a single-pass transmembrane domain (TM) and an intracellular kinase domain (ICD). However, in previously published cryo-EM structures of full-length RTKs, the TM and ICD were either poorly resolved or not resolved at all because of continuous motions between these domains. Consequently, these structures fail to provide insights into how the ECD, TM and ICD of RTKs work cooperatively. To address this issue, we are currently using biochemical approaches to trap the RTKs in a relatively static state.

Furthermore, capturing the structures of weakly associated complexes using cryo-EM presents additional challenges. This is primarily because of the low sample concentrations (in the low micromolar range) required for cryo-EM, which can cause the dissociation of the unstable protein complexes. By contrast, in X-ray crystallography, crystallization requires very high protein concentrations, promoting complex formation. Moreover, weakly associated complexes can become even more unstable in the thin layer of solution on cryo-EM grids owing to frequent contact with the hydrophobic air–water interface.

Tamir Gonen: Despite technological progress, establishing structures of membrane proteins is a key challenge. Most drugs, small molecules, natural products and so on bind to and act via membrane proteins; hence, knowledge of these structures is of great interest for drug discovery. Yet, most of these proteins are small and cannot be easily resolved using methods such as single-particle cryo-EM. In addition, they are difficult to make, purify and analyse using crystallographic techniques. To overcome these challenges, we developed a method called microcrystal electron diffraction, or MicroED[3]. Instead of using X-rays, MicroED uses electrons to examine the atomic structure of molecules in tiny crystals. The main difference between MicroED and X-ray crystallography is that MicroED requires crystals that are much smaller, about one-billionth of the size needed for X-ray crystallography. This means that MicroED solves the problem of working with membrane proteins, making them accessible for study. We use MicroED to investigate protein–drug interactions. We showed that with this technique, we can determine the structure not only of the protein but also of drugs bound or the drugs in isolation. The latter is done rapidly without the need to purify or crystallize the substance and typically atomic resolution structures can be obtained with MicroED within minutes. The process can be automated so that very large volumes of samples are analysed and feed precision of drug binding and discovery pipelines is optimized.

Jianyi Yang: There has been a lot of excitement about the use of computational methods based on machine learning, such as AlphaFold (now AlphaFold2)[4] and RoseTTAFold[5] for protein structure prediction. However, at present, the usefulness of these techniques is mainly

for static structure, with no or limited consideration of biological function. Given the tremendous progress in the prediction of static structure, the main challenges are in the prediction of biological function-oriented structure (ref. 6). This includes structures with multiple functional states, folding pathways, intrinsically disordered proteins, complex structures (protein–protein, protein–nucleic acids and protein–small molecules) and mutation effects. New computational algorithms are being developed to address these challenges, based on the success of machine learning-based static protein structure prediction. A new frontier is integration of experimental data (for example, cross-linking, NMR, SAXS and cryo-EM) into computational algorithms.

Anastassis Perrakis: The main challenge is to understand protein structure and interactions in the context of conformational states and dynamics, both for 'structured' proteins, but also for their 'unstructured' regions. We have a glimpse of catalytic processes for quite a few enzymes, numerous examples of small and large conformational changes in structured regions of proteins, and many disorder-to-order (or order-to-disorder) transitions upon forming complexes between proteins and other macromolecules. But these are only in vitro glimpses of what happens in a cell. Basic chemistry is unlikely to change between a crystal or a solution and a cell, but the dynamics are likely to be affected by a myriad of things that are different. Concurrently, we hardly scratch the surface in understanding macromolecular interactions: deciphering how exactly drugs interact with their (membrane) protein targets can allow designing small-molecule drugs to improve human health or enzymes to lead the transition green industry and economy; understanding how proteins recognize each other can allow us to design antibodies or biological polymers.

My team is trying to combine the more 'classical' structure determination techniques — X-ray crystallography and cryo-EM — with biophysical methods to study the interactions of specific proteins with other biological macromolecules. For example, we are interested in understanding how the two tubulin detyrosinating enzymes we recently described recognize different microtubule species, what is the role of their large unstructured regions in this process, if their activity is processive (performing consecutive reactions without releasing the microtubules) or distributive — and all these ideally inside cells. At the same time, we are exploring the potential of machine learning methods in predicting the binding of biological ligands to proteins, extending on our AlphaFill approach (ref. 7), which adds co-factors and ligands to protein structure models predicted by AlphaFold based on sequence and structure similarity with experimentally determined structures.

Andrea Thorn: Many of the most exciting questions currently are linked to dynamics — the motion of molecular machines, the signal transduction in membrane proteins or pathogen–host interactions. However, crystallography and cryo-EM mainly give us models that represent single low-energy states. In fact, we try to avoid 'disorder' and 'heterogeneity' as much as we can in sample preparation and omit related data in processing, be it in map or phase reconstruction. And as a consequence of this, machine learning-based fold prediction — which is becoming widely used (ref. 4) — is being trained on rigid molecular models, and thus, cannot give us a full picture of flexibility, dynamics or folding. My group addresses this huge challenge — we aim to understand the solvation of macromolecules, which drives flexibility, and the molecular movement itself. It would be great to make these aspects part of our interpretation of experimental data. We also try to combine information from different sources — NMR, electron microscopy, crystallography, biochemistry, molecular dynamics calculations and so forth — in order to see beyond individual structures and exploit their information content fully.

What, in your opinion, is a key advancement that is needed to tackle these challenges?

A.G.: If considering NMR, the methodology with which I am most intimately familiar, the main challenges are sensitivity and the length of time it takes to collect and analyse the data. This applies to both solution and solid-state NMR. For sensitivity enhancement, hyperpolarization approaches will need to be vigorously developed further, combined with pushing ultra-high field magnet technology. To speed up data collection and analysis, automation is absolutely essential. X-ray crystallography and cryo-EM are way ahead of NMR in that regard. Automation of NMR resonance assignments is an outright must and needs to be implemented as soon as possible. The same holds for computerized determination of protein folds (models), based on NMR-extracted distances and angles. NMR needs to be democratized and passed on from experts to the broader community. Every competent chemist, biochemist or biologist should be able to learn and apply NMR via user-friendly software tools and data analysis platforms. Artificial intelligence (AI) needs to be part of this process and help speed it along (see also next question).

X.B.: There is a clear and pressing need to develop more advanced image-sorting

The contributors

Xiaochen Bai Xiaochen Bai is an associate professor at UT Southwestern Medical Center. He has been working on cryo-EM method development and structural determination for more than a decade. His lab currently focuses on the structural and functional studies of receptor tyrosine kinases.

Tamir Gonen Tamir Gonen is a membrane biophysicist and an expert in crystallography and cryo-EM. He is a professor of biological chemistry and physiology at the David Geffen School of Medicine at UCLA and an investigator of the Howard Hughes Medical Institute and a member of the Royal Society of New Zealand. His group develops methodologies for studying medically important membrane protein structure and dynamics.

Angela Gronenborn Angela M. Gronenborn is a structural biologist who has developed and applied NMR methods since 1980. She had led research groups at NIMR in London, at the Max Planck Institute in Martinsried, the intramural research programme of NIDDK, NIH, and finally at the School of Medicine of the University of Pittsburgh from 2005 to the present. She also serves as director of the Pittsburgh Center for HIV Protein Interactions.

Anastassis Perrakis Anastassis (Tassos) Perrakis was trained as an X-ray crystallographer and biochemist at the EMBL. Throughout his career, he provides structural understanding into key questions about the relationship of structure and function in cell biology and biochemistry, and develops software and resources for deciphering and understanding protein structure and function.

Andrea Thorn Andrea Thorn works both experimentally, solving molecular structures from viruses and fungi by crystallography and cryo-EM, and computationally, developing methods for experimental data in structural biology. She started and led the international Coronavirus Structural Task Force and her group develops AI-based tools for diffraction diagnostics (AUSPEX) and for reconstruction map annotation (HARUSPEX).

Jianyi Yang Jianyi Yang is a distinguished professor of mathematics and interdisciplinary sciences at Shandong University. He has made notable contributions to the field of protein structure and function prediction through his co-development of several widely used algorithms, including trRosetta, I-TASSER, COACH and BioLiP. His research group, known as Yang-Server, ranked at the top in the prediction of protein tertiary structure in the CASP15 experiment.

algorithms, potentially utilizing machine learning, for single-particle cryo-EM, which can effectively handle the vast structural heterogeneity and continuous motions in protein complexes. Such technological advancements would allow us to capture structural snapshots of dynamic protein complexes and transient protein–protein interactions. Moreover, the development of improved electron-detecting cameras and next-generation phase plates is critical to further enhance the quality of cryo-EM micrographs, which would enable the identification of even smaller conformational variabilities.

Additionally, there is an urgent need to devise better methods for cryo-EM grid preparation. The hydrophobic air–water interface of cryo-EM grids often leads to the dissociation of weak protein complexes. One possible solution involves coating continuous carbon or graphene on top of cryo-EM grids to absorb proteins and keep them away from the air–water interface. However, these supporting films introduce additional background noise that reduces the sharpness of particle images, making them less than ideal for high-resolution cryo-EM structural determination. Consequently, the development of improved supporting films, preferably utilizing two-dimensional protein crystals, becomes crucial. In theory, the background noise contributed by such a two-dimensional crystal film could be eliminated computationally, ensuring that the image quality of the particles remains uncompromised. Furthermore, improving the vitrification method (that is, eliminating the use of blotting paper) would minimize protein damage caused by the air–water interface.

Nanodiscs or liposomes have been developed as membrane mimetics to stabilize membrane proteins for structural studies. However, these methods come with certain limitations. Firstly, nanodiscs are often too small to be applied to study large membrane protein assemblies. Secondly, liposomes possess highly curved lipid bilayers that fail to fully represent the native membrane environment. Therefore, there is a clear demand for improved membrane mimetics to facilitate future structural studies of membrane proteins.

T.G.: In many cases, predicting the structure of a protein is easier than actually determining its structure through experiments. However, when it comes to membrane proteins, predicting their structures is very challenging. This is because there are not many experimental structures available for membrane proteins, and this limits the information we can use in machine learning pipelines. In addition, using the sequence data of membrane proteins is tough because certain sections that seem hydrophilic may actually be hidden inside the membrane, protected from the inner hydrophobic core of the membrane by protein structures. This makes it difficult to predict their structures.

Similarly, because we have limited information about the structure of the membrane itself, it is hard to predict how these proteins might fold and exist within the cell membrane. To overcome these difficulties, we need methods that can determine the structures of membrane proteins while they are embedded in membranes, using very small amounts of material. MicroED and electron crystallography, in particular, are excellent methods for determining the structures of membrane proteins because they allow us to study them in their natural environment, the lipid bilayer. When the resolution is high enough, these methods can even reveal the structure of the lipid membrane itself.

J.Y.: As mentioned before, an important challenge is understanding the biological functions–structure relationships of macromolecules. One of the major barriers to developing machine learning-based algorithms for structure–function prediction is the lack of a significant amount of diverse training data. The success of statistic structure prediction has been closely tied to the over 50 years of efforts to deposit high-quality structures in the Protein Data Bank. However, it is less likely to obtain sufficient data through pure experiments in the near future. A close collaboration between computational scientists and experimental scientists is needed to speed up the procedure of experimental determination of structure–function relationships. I hope that there would be some international collaborations in this direction. Nevertheless, at this stage, with limited data on structure–function relationships, it may be helpful to combine physics-based models with machine learning algorithms.

A.P.: There is no single key. Understanding conformational states calls for procedures to better sample the full spectrum of conformational states in single particle cryo-electron microscopy; we now interpret a small fraction of the information available in the 'grids'. Macromolecular structure in the context of a cell can benefit from better methods in sample preparation and faster, more accurate, data acquisition for cryo-electron tomography. More efficient methods for X-ray, NMR and cryo-EM methods for determining the structure of ligands bound to protein (or DNA and RNA) will be key to create large datasets that we will need to correlate with functional data in biochemical and cellular assays. Multimodal imaging mass spectrometry with spatial resolution and single-molecule biophysical methods with time resolution are still in their infancy and have the potential to contribute towards an integrated approach to understand macromolecular structure.

A.T.: We can already measure many effects of the solvation, dynamics and mobility of large macromolecular complexes, not only in NMR, SAXS, serial crystallography or cross-linking mass spectrometry but also as part of normal diffraction patterns in macromolecular crystallography, and micrographs in single-particle cryo-EM. However, we lack tools to interpret these data fully. We need a better fundamental understanding of these processes so that we can interpret experiments better and utilize it for computation. Luckily, the structural biology community rises to the challenge with more sensitive measurements, AI-guided interpretation and integrative structural biology, where different methods inform each other. Key to this will also be developing an understanding of the underlying nature and structure of errors in crystallographic and cryo-EM data, or more precisely, understanding the information content of our measurements.

Techniques based on AI have undoubtedly revolutionized the field. But can computational approaches ever fully substitute experimental protein determination? How can both approaches be used for maximum benefit?

A.G.: AI is here to stay, and I hope and anticipate that AI-based algorithms will take the tedium out of NMR structure determination. It is already clear that for lots of proteins, folds can be determined by machine learning-based programmes[4,5]. The way to take advantage of predicted models is to devise clever experiments that can verify or falsify the predictions. For example, if a few 'NMR-active' atoms are judiciously introduced into a protein at structurally informative positions, they can serve to provide long-range distances for testing the validity of a model predicted with machine learning. Fluorine atoms and paramagnetic

tags with readout by ^{19}F NMR constitute such an approach[8]. In a similar vein, one can focus the NMR experiments on atomic details: if a single amino acid change in a protein is thought to be associated with a disease mutant, one can introduce NMR-active isotopes (such as ^{15}N and ^{13}C) into this particular region of the protein and compare the conformations of the mutant variant to that of the wild-type protein.

For examining interactions between small molecules and proteins in solution (as pursued in structure-guided drug design efforts), computational methods can provide initial suggestions, which can be refined by experiments. AI will help to automate high-throughput NMR screening approaches by designing and deconvoluting pools of molecules, identifying hits, assisting in the analysis of large datasets and predicting ligand binding affinities.

And, although not implemented yet, machine learning may also help to predict protein motions one day, when enough experimental data on protein dynamics are available in databases that can be used for training the networks.

X.B.: It has been demonstrated that machine learning-based techniques can accurately predict the structures of individual proteins or simple protein–protein complexes. However, challenges arise when using machine learning to predict the structures of large macromolecular machines or higher-order protein assemblies, such as large oligomers, long filaments or liquid-like condensates. The intricate nature of these complex structures makes precise structural prediction with machine learning difficult. In addition, while computational methods excel at identifying the lowest-energy conformation of a protein, they are not suitable for depicting the conformational ensembles of intrinsically flexible proteins.

Nevertheless, machine learning-based approaches can complement experimental structural determination methods effectively. For instance, many cryo-EM maps of dynamic protein complexes, as well as these maps reconstructed using cryo-electron tomography (cryo-ET), are determined at medium resolution (5–10 Å), and these low-quality maps are challenging to interpret. In such cases, machine learning can aid in building accurate models of individual protein components or domains within the complex computationally. These machine learning-predicted protein segments can then be fitted into the low-quality cryo-EM maps or tomograms using rigid-body docking, resulting in a complete model of the entire protein assembly.

Machine learning can also play a significant part in data processing for single-particle cryo-EM. Various machine learning-based approaches have been employed in tasks such as particle picking, particle dynamics analysis, resolution estimation and map sharpening. Furthermore, for high-resolution cryo-EM maps with a resolution better than 3.5 Å, machine learning approaches can automatically perform precise modelling, significantly reducing potential modelling errors introduced by human.

T.G.: AI cannot replace experiments; its role is to complement them and provide valuable information to assist in experimental design. However, experimental validation is crucial for confirming predictions. The accuracy of predictions relies on the quality of the input data, and there are instances where predictions can be incorrect, leading scientists astray. In our own laboratory experiments, machine learning successfully generated a reasonable model to interpret MicroED data for a new protein, but it failed to identify the structural changes occurring at the active site of the protein. This crucial information, obtained through MicroED, was necessary to understand the mechanism of action of the enzyme.

Therefore, I would argue that machine learning-powered approaches can offer important insights to guide experimental design, but these predictions must be followed by actual experiments to validate and expend the findings.

J.Y.: It depends on the specific real-world applications. For qualitative studies that focus on cellular function, such as in transcriptomics and metabolomics, theoretically predicted structures by machine learning approaches may be sufficient to understand many biological processes. However, for quantitative studies, such as in structure-based drug design, it is necessary to validate the computational predictions through experiments. In this case, computational predictions can be used to accelerate or assist experimental design, whereas experimental solutions can provide feedback to further improve computational algorithms. By combining the strengths of both approaches, we can achieve more accurate predictions and accelerate structure–function relationship determinations.

A.P.: AI and experiment will go hand in hand. The only reason AI can now understand and predict protein structures is because we used experimental methods to determine hundreds of thousands of structures and made them publicly available in the Protein Data Bank. By now AI is really good in determining structures of single proteins and is getting pretty good in determining structures of protein complexes. We already use that to our advantage to design more complex and more challenging experiments to reject or validate hypotheses. In these experiments, we often experimentally determine more complex structures: proteins with other proteins, with DNA and RNA, with all kinds of bioactive lipids and with a variety of polysaccharides, and small molecules, natural or synthetic. More advanced AI will need to be designed to learn from these new data. Then, we will challenge ourselves by ever more complicated structural questions. Eventually, more complex AI will get a glimpse to states, dynamics and interactions and the next challenge might be to first measure and then understand thermodynamics in the cellular context. In all cases, there is a wealth of experiments ahead, for a few decades to come.

A.T.: Of course not! AI (or, more exactly, machine learning or neural networks) cannot fully replace experiments, since these tools are trained not only on experimental data but typically also on our interpretation of them, which is limited. For example, single-state models of 3D coordinates for the atoms and typically a single value for the positional uncertainty of each atom cannot fully describe the ensemble of similar structures we typically observe in crystallography, let alone single-particle methods. Interpretation would also be better if it included, for example, radiation damage as a process, charges (in cryo-EM), better descriptions of detergent for membrane structures, metal ion coordination, RNA conformational space and solvation. However, machine learning is an extremely useful technique for hypothesis generation and to find underlying patterns in data. Once the pattern is found, for example, the folded low-energy state of a main chain, machine learning can inform and maybe replace routine experiments. Even more exciting, by using explainable AI techniques[9], a trained neural network can be the starting point in order to understand the underlying concepts. One approach is to 'ask' which parts of a given input were the most important contributors in the determination of the output, for example, by layer-wise relevance propagation[10,11]. Related methods allow to analyse which neurons contributed most

strongly, so we can not only follow the path of information through the neural network but also get parametrization for say a molecule or a map, with neural network training. This can be enhanced through human-readable outputs: Alphafold1 predicted a distance matrix and torsion angle distributions between networks, which were interpretable by humans. In AlphaFold2, all information is fed directly into the next network, so we unfortunately lost that transparency in the process. I think that explainable AI will reconcile us more with these methods: as scientists, we really like to understand how something works, not only use the result. All in all, experiments will always be necessary, but AI opens a lot of cool synergies for method developers to explore!

In your opinion, what are the new frontiers in macromolecular structure determination? What can we expect in the near future?

A.G.: Macromolecular structure determination has to move beyond the "divide and conquer" approach, away from analysing macromolecules in isolation. In future, fewer and fewer biochemically reconstituted systems will be looked at, and complexes and assemblies that are isolated from native cellular or tissue environments will be studied. Cryo-EM and cryo-ET are making big strides in this arena already. NMR spectroscopic approaches under near physiological conditions, such as in-cell NMR, provide new opportunities for studying cellular processes and visualizing the 'functioning' molecule, both with respect to structure and dynamics. Functioning often requires an ensemble of different conformers, only one (or few) of which may be functionally competent. NMR can characterize such ensembles, both structurally and with respect to their motions at atomic resolution.

The complexity of biological phenomena, linked to the inherent incompleteness of any representation, requires the use of multiple methodologies, experimental, computational and theoretical. As is universally appreciated, individual types of structural data are limited in scope, accuracy and generality, and using complementary information in an integrative fashion helps to overcome any shortcomings. Such integration will need to become the norm and can be advanced through machine learning.

As I have argued before, whereas no single theoretical or empirical approach describes all the features that are relevant to the structure of a 'functional protein', being aware of the incompleteness of representations should serve as a constant reminder of the need to integrate multiple models, methods and representations. This is what integrative pluralism in the sciences is about and why we need to practice it[2].

X.B.: The structures and functions of most proteins are regulated by surrounding molecules in the cellular environment, such as the cytoskeleton, other proteins, membranes and small metabolites. When studying purified proteins, the structural information obtained may lack these essential cellular factors. However, recent advancements in cryo-ET combined with sub-tomogram averaging have opened up an exciting new frontier: the direct visualization of macromolecular machines or large membrane proteins within their cellular context. This approach does not require isolating the proteins and allows for the structural determination of protein complexes in their native environment. The powerful combination of cryo-correlative light and electron microscopy (cryo-CLEM) along with cryo-focused ion-beam (cryo-FIB) milling (thinning) of cellular samples enables precise localization of protein complexes within cells. Sub-tomogram averaging, akin to single-particle analysis, is an emerging technique that improves the resolution of cryo-ET reconstructions. Cryo-ET can also be used to investigate the structures of large protein clusters and liquid-like condensates, expanding our understanding of these complex cellular structures.

In traditional structural approaches, proteins are often overexpressed and purified from recombinant sources. However, these recombinant proteins lack critical regulatory factors found in native environments, such as post-translational modifications, low-abundance binding partners and endogenous small molecules. Therefore, another exciting frontier is the structural study of protein complexes directly purified from animal tissues using cryo-EM. Cryo-EM requires sample amounts smaller than those required by X-ray crystallography and can tolerate a certain degree of impurity, making it feasible to analyse protein complexes from native sources. By combining cryo-EM analysis with mass spectrometry, we can identify obscure components that are crucial for complex formation and function.

T.G.: I believe that in the next 10 years, we will gain thorough understanding of key membrane proteins and I am confident that MicroED will become valuable for determining the structures that were beyond the reach of other methods. However, it is not just about understanding individual protein structures. We also need to learn more about how the membrane is organized and how these proteins work in the wider context of the membrane environment. This is something that AI currently struggles with. Experimental structure determination provides us with a snapshot, but we need to gather many snapshots to get a better idea of how these proteins function. With this knowledge, we can potentially develop new medications and treatments.

J.Y.: The new frontiers in macromolecular structure determination are changing from statistic structure to biological function-oriented structure. In the near future, we can expect significant progress in combining experimental data (for example, cryo-EM density map) and computational algorithms, which will greatly accelerate the determination or prediction of biological function-oriented structure with reduced cost. We hope that proteome structure determination would become as easy as genome sequencing in future, paving the way towards a better understanding of the biological function of macromolecules.

A.P.: We are already taking tiny glimpses of a bright future: looking inside cells, we see different conformational states of ribosomes, we identify tens of proteins inside microtubules, we enumerate post-translational modifications of the proteins and the RNA of ribosomes, and we are making snapshots of the life cycle of viruses. In the near future, we will achieve better spatial and time resolution, seeing smaller and more complex macromolecular machines: detailed pictures of 'transcription bubbles', DNA replication and replication initiation complexes, molecular structures of DNA repair foci, or kinetochores attached to microtubules during mitosis; and we might understand the membraneless compartments inside the cell and the role of phase separation and of intrinsically disordered regions of proteins in their formation. Better glimpses to the 'macromolecular interactions' puzzle will allow to design biologics as therapeutics, and amassing millions of protein–ligand complexes in public repositories – as the community did for hundreds of thousands of protein structures – will allow us to design new drugs in record times. New generative AI will enable addressing the Sustainable Development Goals by providing platforms for the design of new enzymes that degrade plastics or make environment-friendly biopolymers.

A.T: We ultimately aim to understand the function and structure of molecules in the context of a living cell. This will require bridging across atomic resolution methods and also integrating insights across different scales of analysis and combining all of this information in a meaningful way computationally. To this end, we need to overcome the reproducibility crisis in science and leverage all the advances in techniques, automation and machine learning in the coming years: it is a truly exciting time.

Xiao-chen Bai [1] ✉, **Tamir Gonen** [2,3,4] ✉, **Angela M. Gronenborn** [5] ✉, **Anastassis Perrakis** [6] ✉, **Andrea Thorn** [7] ✉ **& Jianyi Yang** [8] ✉

[1]Department of Biophysics, University of Texas Southwestern Medical Center, Dallas, TX, USA. [2]Department of Biological Chemistry, University of California Los Angeles, Los Angeles, CA, USA. [3]Howard Hughes Medical Institute, University of California Los Angeles, Los Angeles, CA, USA. [4]Department of Physiology, University of California Los Angeles, Los Angeles, CA, USA. [5]Structural Biology, University of Pittsburgh, Pittsburgh, PA, USA. [6]Oncode Institute, Division of Biochemistry, Netherlands Cancer Institute, Amsterdam, Netherlands. [7]Institute for Nanostructure and Solid State Physics, University of Hamburg, Hamburg, Germany. [8]MOE Frontiers Science Center for Nonlinear Expectations, Research Center for Mathematics and Interdisciplinary Sciences, Shandong University, Qingdao, China.
✉e-mail: Xiaochen.Bai@UTSouthwestern.edu; tgonen@g.ucla.edu; amg100@pitt.edu; a.perrakis@nki.nl; andrea.thorn@uni-hamburg.de; yangjy@sdu.edu.cn

Published online: 17 October 2023

References

1. Mitchell, S. & Gronenborn, A. M. After 50 years, why are protein X-ray crystallographers still in business? *Br. J. Philos. Sci.* **68**, 703–723 (2017).
2. Gronenborn, A. M. Integrated multidisciplinarity in the natural sciences. *J. Biol. Chem.* **294**, 18162–18167 (2019).
3. Nannenga, B. L. & Gonen, T. The cryo-EM method microcrystal electron diffraction (MicroED). *Nat. Methods* **16**, 369–379 (2019).
4. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
5. Baek, M. et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
6. Peng, Z. et al. Protein structure prediction in the deep learning era. *Curr. Opin. Struct. Biol.* **77**, 102495 (2022).
7. Hekkelman, M. L. et al. AlphaFill: enriching AlphaFold models with ligands and cofactors. *Nat. Methods* **20**, 205–213 (2023).
8. Gronenborn, A. M. Small, but powerful and attractive: ^{19}F in biomolecular NMR. *Structure* **30**, 6–14 (2022).
9. Thorn, A. Artificial intelligence in the experimental determination and prediction of macromolecular structures. *Curr. Opin. Struct. Biol.* **74**, 102368 (2022).
10. Montavon G, et al. in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (eds Samek, W. et al.) pp. 193–209. (Springer International, 2019).
11. Bach, S. et al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE* **10**, e0130140 (2015).

Author contributions

The authors contributed equally to all aspects of the article.

Competing interests

The authors declare no competing interests.