

being strictly dependent on the individual structure's resolution³. Optimistically, the cellular context could help to build a more complete understanding of intrinsically disordered proteins. For example, within the regular arrangement of nuclear pore rings are disordered Phe-Gly repeats. Although high-resolution information of these repeats remains elusive by cryo-ET, knowledge of their location within tomograms might be harnessed towards understanding the interplay between structured and unstructured cellular elements⁴. Another exciting prospect of cryo-ET is the description of hitherto unknown or elusive supercomplexes within the cell. A stirring recent example is the in-cell structure of the actively transcribing-translating expressome⁵.

Many outstanding questions regarding cellular structures will require integration of temporal and

compositional information. Towards this goal, advances in correlative light and electron microscopy expand the range of questions addressable by cryo-ET³. Moreover, visually unresolvable, molecular-level information can be integrated using cross-linking mass spectrometry^{5,6}. Finally, particular molecular architectures may be specific to organisms or tissues. It is thus inspiring that structural work within multicellular samples and visualization of molecular structures within clinical samples appear within reach^{7,8}.

Collating all these emerging opportunities, we believe that the exploration of cell and tissue architecture represents an exciting future avenue for structural biology. This endeavor requires a blend of skills and seamless collaborations between cell and structural biologists, towards their common goal of a mechanistic understanding of biological complexity. □

Michael R. Wozny¹ and Wanda Kukulski^{1,2} ✉

¹MRC Laboratory of Molecular Biology, Cambridge, UK. ²Institute of Biochemistry and Molecular Medicine, University of Bern, Bern, Switzerland.

✉e-mail: wanda.kukulski@ibmm.unibe.ch

Published online: 7 May 2021
<https://doi.org/10.1038/s41592-021-01131-5>

References

1. Kendrew, J. C. et al. *Nature* **181**, 662–666 (1958).
2. Guo, Q. et al. *Cell* **172**, 696–705.e12 (2018).
3. Ader, N.R. et al. *Elife* **8**, e40712 (2019).
4. Zila, V. et al. *Cell* **184**, 1032–1046.e18 (2021).
5. O'Reilly, F. J. et al. *Science* **369**, 554–557 (2020).
6. von Kügelgen, A. et al. *Cell* **180**, 348–358.e15 (2020).
7. Schaffer, M. et al. *Nat. Methods* **16**, 757–762 (2019).
8. Weiss, G.L. et al. *Science* **369**, 1005–1010 (2020).
9. Pettersen, E. F. et al. *Protein Sci.* **30**, 70–82 (2021).
10. Kremer, J. R., Mastronarde, D. N. & McIntosh, J. R. *J. Struct. Biol.* **116**, 71–76 (1996).

Competing interests

The authors declare no competing interests.



Computational protein modeling and the next viral pandemic

Computational protein modeling rapidly advances structural knowledge of viral proteins, but methods for modeling protein complexes still need improvement.

Oleksandr Narykov, Suhas Srinivasan and Dmitry Korin

It has been one year since the release of the first SARS-CoV-2 genome¹, which provided scientists with critical knowledge about its proteins. Thanks to the unprecedented experimental efforts by scientists worldwide, we have now obtained structural knowledge about most SARS-CoV-2 proteins, determining their three-dimensional (3D) shapes. Perhaps even more critical is the structural knowledge of the protein complexes that underlie the basics of viral functioning. Months before the experimental protein structures were solved, computational efforts by several groups provided researchers with accurate 3D models of the viral proteins and their physical interactions with each other and with host proteins. This 3D molecular information is instrumental in basic research, to understand mechanisms behind the viral entry and replication, as well as in structure-based drug design, to determine

new antiviral targets, or in vaccine development, to study effects of novel mutations on antigen–antibody binding. Given that it is not ‘if’, but ‘when’ a new viral pandemic will emerge², it is crucial to know whether computational modeling methods can facilitate structural characterization of viral proteins and their essential complexes. After one year of intensive research by the structural biology community, we have accumulated enough data to evaluate the impact of computational modeling efforts toward understanding the structural nature of the virus.

Structural genomics efforts to characterize the protein repertoire of a virus are usually carried out by comparative—or template-based—modeling³. A newer technique, de novo protein modeling⁴, does not require a template structure and may complement existing methods. Template-based models are often more

accurate than de novo ones; however, the former technique is dependent on previously solved structures of homologous proteins or protein complexes while the latter can be applied to novel proteins. The latest success in protein modeling has been primarily due to recent technological innovations in the development of novel protein structure prediction algorithms, which use deep learning and are empowered by advances in graphical processing unit (GPU)-accelerated computing. We surveyed accurate template-based and de novo models of SARS-CoV-2 proteins and protein complexes that were also experimentally solved to determine (i) model accuracy when compared with the experimental structure and (ii) how far ahead of the experimental structures they were obtained (Fig. 1). We considered comparative models generated by our group⁵ and de novo models reported by AlphaFold⁶ and C-I-TASSER⁷,

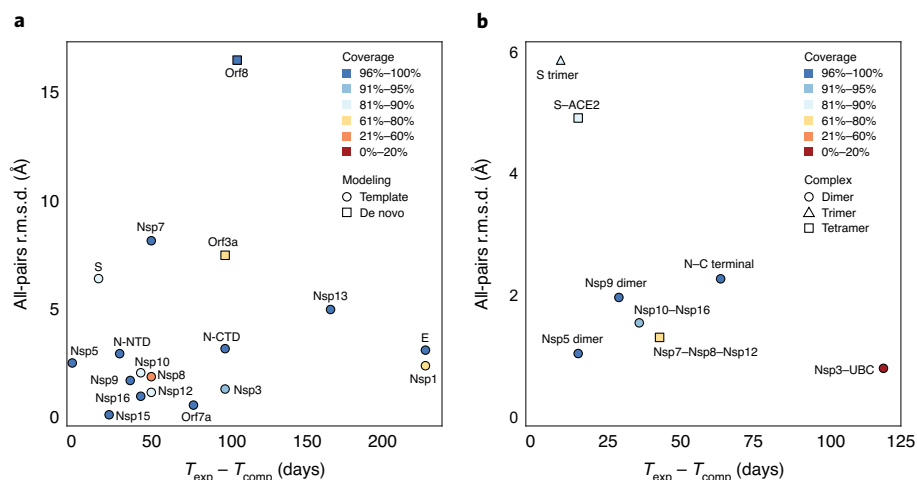


Fig. 1 | Evaluation of computational approaches for modeling 3D structures of SARS-CoV-2 proteins and related protein complexes. a, Analysis of 17 individual proteins that were both experimentally characterized and computationally modeled, using comparative (circles) and de novo (squares) methods. **b**, Analysis of 8 protein complexes; each complex consists of two (circle), three (triangle) or four (square) protein subunits. For each modeled protein or protein complex, its r.m.s.d. error between the model and experimental structure, the number of days between the releases of experimental and computational structures, and the model's coverage of the protein sequence (color) are calculated.

which have also contributed to structural characterization of SARS-CoV-2 proteins (Fig. 1a and Supplementary Table 1). Of the 29 putative proteins, 17 were at least partially experimentally and computationally resolved, while 5, including key structural protein M, were characterized only computationally. Six putative proteins have not been structurally characterized at all. The computational methods were fairly accurate, producing an average root mean squared deviation (r.m.s.d.) error of 4.1 Å for all 17 proteins (Supplementary Note). On average, computational models covered roughly 80% of the viral protein sequence, while experimental structures covered 82%. Most importantly, 3D models of viral proteins were released on average 86 days earlier than the corresponding experimental structures.

Even if we had structural knowledge of all SARS-CoV-2 proteins, our understanding of the virus's functional units would be far from complete: most, if not all, viral proteins carry out their functions by forming macromolecular complexes. Recent efforts to map all protein complexes formed by SARS-CoV-2 proteins have identified hundreds of putative interactions⁸. Unfortunately, only a small fraction of these complexes have been structurally characterized (Fig. 1b and Supplementary Table 2): 18 protein complexes have been

characterized experimentally and 16 computationally. Overall, for 13 protein complexes, the structure was both modeled and resolved experimentally. For 5 of these, an incorrect oligomer conformation was derived from homologous complexes; for the remaining 8, the computational models yielded accurate protein complexes in correct conformations, with an average r.m.s.d. of 2.6 Å over the entire multimeric structure (Supplementary Information). The models were available on average 53 days earlier than experimental structures, covering on average 77% of all protein sequences involved in the complex. Lastly, for 4 modeled complexes, no experimental structures have yet been obtained.

In the 2011 science fiction movie *Contagion*, which went viral [sic] in 2020, scientists were shown looking at a structure of a viral surface protein bound to the host receptor just a couple of days after the viral genome was sequenced. That speed is not yet possible experimentally, but can already be achieved using computational modeling. Modeling 3D shapes of the viral proteins and their key complexes brings structural knowledge of the virus several critical months earlier than experiments can. We expect that computational models will be increasingly helpful in designing experiments to test neutralizing antibodies,

studying the role of emerging mutations, and understanding the molecular mechanisms behind viral infections. Furthermore, we envision a new generation of artificial intelligence (AI)-driven protein modeling tools, such as AlphaFold 2 (ref. ⁹), providing even greater improvement in protein models for novel viruses. Still, de novo modeling should be used with caution and backed up by experiments when characterizing viral proteins because their remarkably diverse structural repertoire might not be captured during training of an AI method. Furthermore, structural characterization of the macromolecular complexes formed by viral proteins presents a major challenge. Thus, development of the new methods for accurate de novo characterization of protein complexes, akin to AI-driven protein structure prediction methods, is the next frontier. □

Oleksandr Narykov^{1,4}, Suhas Srinivasan^{1,2,4} and Dmitry Korkin^{1,2,3} ✉

¹Computer Science Department, Worcester Polytechnic Institute, Worcester, MA, USA.

²Data Science Program, Worcester Polytechnic Institute, Worcester, MA, USA. ³Bioinformatics and Computational Biology Program, Worcester Polytechnic Institute, Worcester, MA, USA.

⁴These authors contributed equally: Oleksandr Narykov, Suhas Srinivasan.

✉e-mail: korkin@korkinlab.org

Published online: 7 May 2021
<https://doi.org/10.1038/s41592-021-01144-0>

References

- Zhou, P. et al. *Nature* **579**, 270–273 (2020).
- Burton, D. R. & Topol, E. J. *Nature* **590**, 386–388 (2021).
- Martí-Renom, M. A. et al. *Annu. Rev. Biophys. Biomol. Struct.* **29**, 291–325 (2000).
- Kuhlman, B. & Bradley, P. *Nat. Rev. Mol. Cell Biol.* **20**, 681–697 (2019).
- Srinivasan, S. et al. *Viruses* **12**, 360 (2020).
- Senior, A. W. et al. *Nature* **577**, 706–710 (2020).
- Zheng, W. et al. *Proteins* **87**, 1149–1164 (2019).
- Gordon, D. E. et al. *Nature* **583**, 459–468 (2020).
- Callaway, E. *Nature* **588**, 203–204 (2020).

Acknowledgements

This work was supported by a National Institute of Health grant (1R01GM135919) to D.K.

Author contributions

D.K. designed and supervised the project. O.N. and S.S. collected the data. All authors analyzed the data, wrote the original draft, and contributed to review and editing.

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/s41592-021-01144-0>.