The case for post-predictional modifications in the AlphaFold Protein Structure Database

To the editor — AlphaFold2 has arrived to change workflows in structural biology, for good. However, the algorithm does not account for essential modifications that affect protein structure and function, which gives us only part of the picture. Here we discuss how this omission can be addressed in a relatively straightforward manner, which leads to a complete structural prediction of complex biomolecular systems.

The recent release of the AlphaFold Protein Structure Database[1] by DeepMind and EMBL-EBI marks a major breakthrough in structural biology, as it makes available to the scientific community worldwide highly accurate structural predictions for 20,000 proteins from humans and proteins from 20 other biologically relevant organisms that include *Escherichia coli*. Like many scientists that work on macromolecular structure, we are genuinely excited about this development, yet we feel that there is a non-negligible potential for misinterpretation of its content in its current form. In particular, the protein-only predictions in the AlphaFold database means that cofactors and, most importantly, co- and post-translational modifications are understandably — owing to the scope of the technique — excluded. Among the most relevant co- and post-translational modifications is protein glycosylation — relevant and very visible, as recent studies of the dynamics of a fully glycosylated SARS-CoV-2 spike protein illustrate[2,3]. Indeed, between 50% and 70% of those 20,000 predicted human proteins are believed to be glycosylated[4], but none of this is yet visibly highlighted on the database. Detailed information on the likelihood of modifications is readily available through AlphaFolds's links to Uniprot (https://www.uniprot.org), and thus we strongly encourage the users of this fantastic new resource to check the information available on Uniprot before downloading a model.

Within this framework, we believe that the absence of cofactors and of co- or post-translational modifications in the models in the AlphaFold Protein Structure Database might be remediated through the use of sequence and structure-based comparative studies. Indeed, in the specific case of glycosylation, the algorithms that are implemented by DeepMind have digested inter-residue distances from the Protein Data Bank (PDB)[5], where glycosylated proteins often exhibit either full or partial glycan structures; therefore, the space where unmodeled modifications, such as protein glycosylation, should have appeared is somehow preserved in AlphaFold models, which allows for these structural features to be directly grafted onto a model. To demonstrate the potential of this approach, we have developed proof-of-concept functionality that grafts protein glycosylation from a library of structurally equilibrated glycan blocks, obtained from molecular dynamics[6], onto an AlphaFold model.
This task has been automated and integrated into the new Python interface of the carbohydrate-specific Privateer software[7] and is available to all on its GitHub repository (https://github.com/glycojones/privateer.git). Figure 1 shows AlphaFold model P29016 (depicted in magenta) of a human T cell surface glycoprotein Cd1b, superposed onto the protein's crystal structure PDB 5WL1. The latter was expressed in an insect cell line and it shows a characteristic double core-fucosylation of the N-glycans, which were omitted in Fig. 1 for clarity. The N-glycan our tool grafted onto the AlphaFold model is not just compatible with the available space, but it shows a high complementarity to the protein surface, where the Man6 core is involved with Trp 23 in a CH-π interaction[8], as seen in the crystal structure.

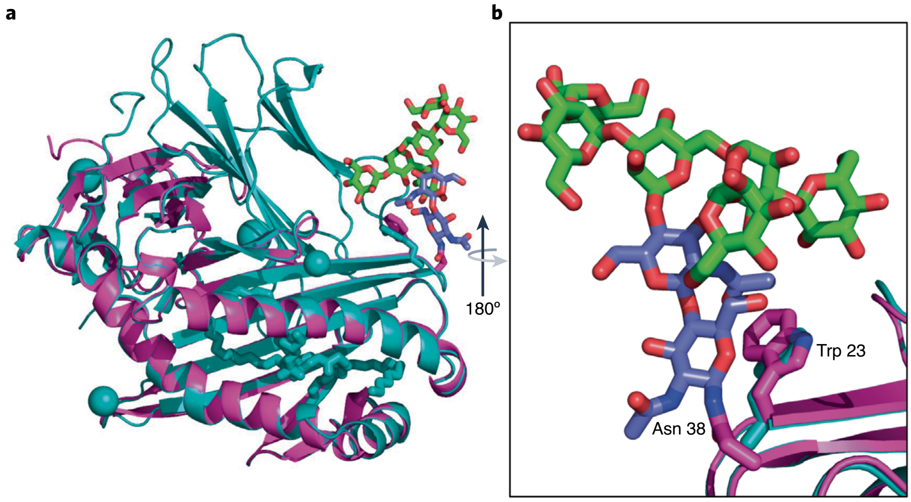

Fig. 1 | Grafting an N-glycan onto an AlphaFold model. a, Structural alignment of the crystal structure of human CD1b in complex with phosphatidylglycerol (PDB 5WL1), shown in cyan, onto the model predicted by AlphaFold (accession code P29016), shown in magenta. The N-glycosylation at position N38 was reconstructed with Privateer[7], where the linked Man6 structure was selected from a library of highly populated conformers at equilibrium, obtained from molecular dynamics simulations at 300 K[6]. **b**, Close-up view of the grafted Man6, with the structure rotated around the *z*-axis by 180°, represented in sticks with colouring compliant to the Symbol Nomenclature for Glycans scheme. The relative positions of the Trp 23 sidechain stacking the Man6 core are highlighted in sticks in both the crystal structure (cyan) and in the AlphaFold model (magenta).

We would like to emphasize that this approach may also be useful to complete the AlphaFold models in the database with other types of modifications. For example, the AlphaFold model P6887, a hemoglobin subunit beta, contains a heme binding site with just enough space for a heme cofactor. Certain structure completions will only be feasible via automated comparative analyses against available structural information — for example, co-translational modifications such as myristoylation[9], or O-GlcNAcylation[10] — while others such as N-glycosylation or tryptophan mannosylation, which rely on consensus sequences, will be more amenable to prediction. As comparative studies would have to rely on experimental structural information, positional uncertainty (for example, a pLDDT-like score[11]) may be estimated by comparing the placed coordinates to a superposition of the available structural information. However, in the particular case of protein glycosylation, we see more of a compositional problem; indeed, the biggest challenge would be to get a good estimation of what glycoform is linked to each sequon. Experimental structures offer only partial information owing to limiting factors such as mobility and micro-heterogeneity[12], so other sources of knowledge (for example, glycomics and molecular dynamics simulations) ought to be used, especially when attempting to model full-length glycans, which is something we are sure the glycobiology community will appreciate. We are expanding the Privateer software to address these cases, by harnessing the rich information available in glycomics databases[13].

To conclude, we think that these early results are highly encouraging to serve as a rallying point for the developers' community to complete and enrich the predicted protein models with likely modifications, to bring them to their fullest potential and to correctly inform the next generation of structural biology studies. ❒

Haroldas Bagdonas[1], Carl A. Fogarty[2], Elisa Fadda[2]✉ and Jon Agirre[1]✉

[1]*York Structural Biology Laboratory, Department of Chemistry, University of York, York, UK.*
[2]*Department of Chemistry and Hamilton Institute, Maynooth University, Maynooth, Ireland.*
✉*e-mail: elisa.fadda@mu.ie; jon.agirre@york.ac.uk*

Published online: 29 October 2021
https://doi.org/10.1038/s41594-021-00680-9

References

1. Tunyasuvunakool, K. et al. *Nature* **596**, 590–596 (2021).
2. Casalino, L. et al. *ACS Cent. Sci.* **6**, 1722–1734 (2020).
3. Turoňová, B. et al. *Science* **370**, 203–208 (2020).
4. An, H. J., Froehlich, J. W. & Lebrilla, C. B. *Curr. Opin. Chem. Biol.* **13**, 421–426 (2009).
5. Berman, H., Henrick, K. & Nakamura, H. *Nat. Struct. Biol.* **10**, 980 (2003).
6. Fogarty, C. A. & Fadda, E. *J. Phys. Chem. B* **125**, 2607–2616 (2021).
7. Agirre, J. et al. *Nat. Struct. Mol. Biol.* **22**, 833–834 (2015).
8. Hudson, K. L. et al. *J. Am. Chem. Soc.* **137**, 15152–15160 (2015).
9. Udenwobele, D. I. et al. *Front. Immunol.* **8**, 751 (2017).
10. Zhu, Y. et al. *Chem. Biol.* **11**, 319–325 (2015).
11. Jumper, J. et al. *Nature* **596**, 583–589 (2021).
12. Atanasova, M., Bagdonas, H. & Agirre, J. *Curr. Opin. Struct. Biol.* **62**, 70–78 (2020).
13. Bagdonas, H., Ungar, D. & Agirre, J. *Beilstein J. Org. Chem.* **16**, 2523–2533 (2020).

Acknowledgements

H.B. is funded by The Royal Society grant RGF/R1/181006. J.A. is the Royal Society Olga Kennard Research Fellow award ref. UF160039. C.A.F. is funded by the Irish Research Council (IRC) Government of Ireland Postgraduate Scholarship Programme. Data and methods are available at https://doi.org/10.5281/zenodo.5290624

Competing interests

The authors declare no competing interests.