**Article:**
Joosten, Robbie P and Agirre, Jon orcid.org/0000-0002-1086-0253 (2022) Whole-proteome structures shed new light on posttranslational modifications. PLoS Biology. e3001673. ISSN 1544-9173

https://doi.org/10.1371/journal.pbio.3001673

PLOS BIOLOGY

PRIMER

# Whole-proteome structures shed new light on posttranslational modifications

**Robbie P. Joosten**[1], **Jon Agirre**[2]*

**1** Biochemistry Department, Netherlands Cancer Institute, the Netherlands and Oncode Institute, Amsterdam, the Netherlands, **2** York Structural Biology Laboratory, Department of Chemistry, University of York, York, United Kingdom

* jon.agirre@york.ac.uk

*Accurate but protein-only AlphaFold models may show structural fingerprints of likely posttranslational modifications (PTMs). In this issue of* PLOS Biology, *Bludau and colleagues add a functional context to models by combining them with readily available proteomics results.*





**Citation:** Joosten RP, Agirre J (2022) Whole-proteome structures shed new light on posttranslational modifications. PLoS Biol 20(5): e3001673. https://doi.org/10.1371/journal.pbio.3001673

**Published:** May 27, 2022



**Funding:** This work was supported by The Royal Society fellowship code UF160039 (J.A.) and Horizon 2020 Project ID 871037 – iNext-Discovery (R.P.J.). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing interests:** The authors have declared that no competing interests exist.

**Abbreviations:** AFDB, AlphaFold Protein Structure Database; PAE, positional alignment error; PDB, Protein Data Bank; PTM, posttranslational modification.

The recent artificial intelligence revolution in protein structure prediction, spearheaded by DeepMind's AlphaFold [1] and swiftly seized upon by RoseTTAFold [2], is allowing scientists to arrive at an accurate structural model of a protein, or at least parts of it, in a matter of hours. This already diminutive lead time can be further compressed to mere seconds if the protein of interest is found in the complete set of proteins expressed by an organism (proteome) in the list of the ever expanding set of organisms covered by the AlphaFold Protein Structure Database (AFDB). The AFDB, released in 2021 and subsequently updated [3], is expected to cover the 100 million set of sequences in the proteomes available at UniRef90 [4]. It offers immediate access to predicted models of human proteins, alongside reliable estimates of their accuracy in the form of 2 metrics: pLDDT (per-residue confidence) and PAE (positional alignment error of each residue with respect to the rest). Structures with a consistently high pLDDT and very low PAE are expected to show an accuracy on par with experimentally determined protein models.

Human proteins are obvious targets for therapeutics; however, their function and structure are, more often than not, modulated or regulated by co- and posttranslational (covalent) modifications, plus ligands and cofactors (noncovalent). Those important moieties, not currently targeted by the AlphaFold algorithm, are conspicuously absent from predicted structures [5]: As an example, many more than half of all human proteins are expected to include either protein glycosylation [6], phosphorylation [7], or both. Thus, the analysis of AlphaFold structures of modified proteins can produce misleading results [5].

Recent studies have suggested that most predicted models are accurate enough to include space for the absent modifications, ligands, and cofactors to be added postprediction [5,8]. Importantly, these endeavours can only be as successful as our ability to pinpoint their occurrence and location on a protein's structure. In a slightly different case, transplanting likely ligands (e.g., a heme group onto hemoglobin or a polysaccharide onto a glycoside hydrolase) onto AlphaFold models by homology with experimental structure models becomes

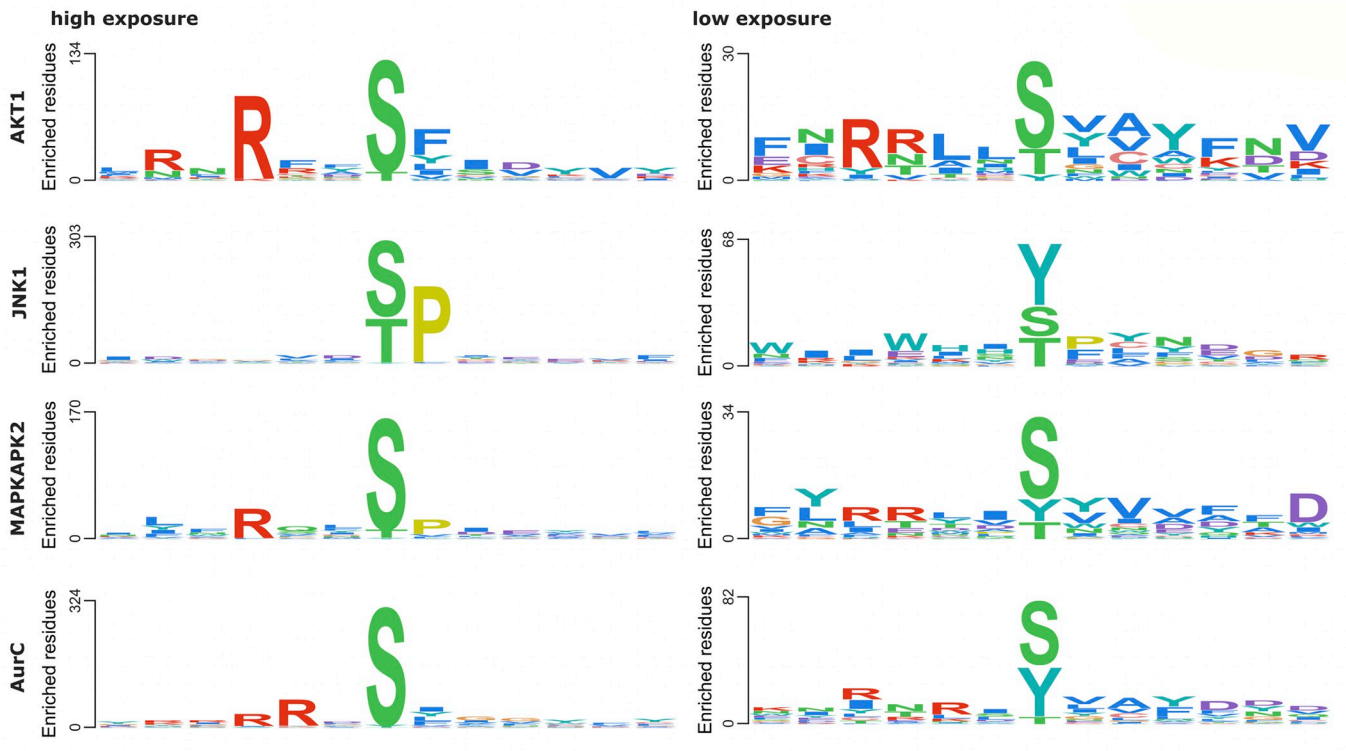


**Fig 1. Different sequence profiles in exposed and solvent-excluded phosphosites (STY) as recognised by different kinases.** The availability of accurate 3D models now allows for this direct mapping of sequence profiles onto structures and allows estimating their solvent accessibility. Extracted from Fig 3d of Bludau and colleagues [10].

https://doi.org/10.1371/journal.pbio.3001673.g001

increasingly error-prone when the homology becomes more distant. In the case that homology is absent altogether, transferable knowledge from experimental structure models is absent as well, and this process becomes a speculative docking experiment.

In the absence of experimental structural models, the extensive proteomics datasets available today can provide information on co- and posttranslational modifications (PTMs) on the respective target proteins [9]. Furthermore, the covalent transference of modifications onto protein often follows a consensus sequence—e.g., N-glycosylation on Asn-X-Ser/Thr where X is any amino acid other than proline; these consensus sequences are variably well studied across modifications. Crucially, mapping proteomics and bioinformatics information onto AlphaFold 3D models may allow us to not just complete models, but to learn more about the structural fingerprints left by modifications: the structure of their protein scaffold and their environment. In this issue, Bludau and colleagues [10] discuss the first results from the implementation of such an approach, targeting different modification types including phosphorylation, ubiquitination, and more.

Not all PTMs are made equal: They may play different roles depending on whether they are buried or exposed to solvent (Fig 1), added to a correctly folded region, a misfolded region, or to an intrinsically disordered one. On that last point, the synergy with AlphaFold brings another important contribution to the table: Because AlphaFold has been trained on data from the structured parts of ordered proteins—a precondition for atomic positions to be well resolved in both X-ray crystallography and electron cryo-microscopy, the 2 main techniques contributing structures to the Protein Data Bank (PDB)—there is a good correlation between intrinsic disorder and low prediction confidence as measured by AlphaFold;s pLDDT [1].

Bludau and colleagues [10] use this knowledge to select PTMs that are enriched for having regulatory functions. These regulatory modification sites show a preference for short intrinsically disordered regions such as the activation loops in protein kinases. In addition, the authors use AlphaFold models to show that different regulatory modification sites have a strong tendency to flock together in 3D and not just in sequence space, hinting at coregulation or even cross talk between different types of PTMs [10].

The work, as one of the first systematic analyses of the functional importance of PTMs, lays an important foundation for new experimental studies targeting PTMs in specific proteins. The authors provide software tools to shortlist the modification sites of regulatory importance, thereby allowing more focused experimental studies. Importantly, the software—for which source code is available from the "structuremap" and "alphamap" repositories at https://github.com/MannLabs—will also enable richer annotation of PTMs on AlphaFold entries. To this end, we think the results from Bludau and colleagues [10] would make a worthy contribution to the recently introduced 3D-Beacons database, which aims to become a reference point for structural knowledge (https://www.ebi.ac.uk/pdbe/pdbe-kb/3dbeacons).

## References


1. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021; 596:583–9. https://doi.org/10.1038/s41586-021-03819-2 PMID: 34265844
2. Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science. 2021; 373:871–6. https://doi.org/10.1126/science.abj8754 PMID: 34282049
3. Tunyasuvunakool K, Adler J, Wu Z, Green T, Zielinski M, Žídek A, et al. Highly accurate protein structure prediction for the human proteome. Nature. 2021; 596:590–6. https://doi.org/10.1038/s41586-021-03828-1 PMID: 34293799
4. Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH, UniProt Consortium. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics. 2015; 31:926–32. https://doi.org/10.1093/bioinformatics/btu739 PMID: 25398609
5. Bagdonas H, Fogarty CE, Fadda E, Agirre J. The case for post-predictional modification in the AlphaFold Protein Structure Database. Nat Struct Mol Biol. 2021. https://doi.org/10.1038/s41594-021-00680-9 PMID: 34716446
6. An HJ, Froehlich JW, Lebrilla CB. Determination of Glycosylation Sites and Site-specific Heterogeneity in Glycoproteins. Curr Opin Chem Biol. 2009; 13:421. https://doi.org/10.1016/j.cbpa.2009.07.022 PMID: 19700364
7. Sharma K, D'souza RCJ, Tyanova S, Schaab C, Wiśniewski JR, Cox J, et al. Ultradeep human phosphoproteome reveals a distinct regulatory nature of Tyr and Ser/Thr-based signaling. Cell Rep. 2014; 8:1583–94. https://doi.org/10.1016/j.celrep.2014.07.036 PMID: 25159151
8. Hekkelman ML, de Vries I, Joosten RP, Perrakis A. AlphaFill: enriching the AlphaFold models with ligands and co-factors. https://doi.org/10.1101/2021.11.26.470110
9. Ochoa D, Jarnuczak AF, Viéitez C, Gehre M, Soucheray M, Mateus A, et al. The functional landscape of the human phosphoproteome. Nat Biotechnol. 2020; 38:365–73. https://doi.org/10.1038/s41587-019-0344-3 PMID: 31819260
10. Bludau I, Willems S, Zeng W-F, Strauss MT, Hansen FM, Tanzer MC, et al. The structural context of posttranslational modifications at a proteome-wide scale. PLoS Biol. 2022; 20(5): e3001636. https://doi.org/10.1371/journal.pbio.3001636 PMID: 35576205