Modelling protein complexes with crosslinking mass spectrometry and deep learning

Kolja Stahl[1], Oliver Brock[2], Juri Rappsilber[1,3,4,§]

[1] Technische Universität Berlin, Chair of Bioanalytics, 10623 Berlin, Germany
[2] Technische Universität Berlin, Robotics and Biology Laboratory, 10623 Berlin, Germany
[3] Si-M/"Der Simulierte Mensch", a Science Framework of Technische Universität Berlin and Charité - Universitätsmedizin Berlin, 10623 Berlin, Germany.
[4] Wellcome Centre for Cell Biology, University of Edinburgh, Edinburgh EH9 3BF, UK

[§] corresponding author: juri.rappsilber@tu-berlin.de

Scarcity of structural and evolutionary information on protein complexes poses a challenge to deep learning-based structure modelling. We integrated experimental distance restraints obtained by crosslinking mass spectrometry (MS) into AlphaFold-Multimer, by extending AlphaLink to protein complexes. Integrating crosslinking MS data substantially improves modelling performance on challenging targets, by helping to identify interfaces, focusing sampling, and improving model selection. This extends to single crosslinks from whole-cell crosslinking MS, suggesting the possibility of whole-cell structural investigations driven by experimental data.

Solving the structure of protein complexes is key to understanding life at the molecular level. The advent of deep learning-based methods has significantly improved the reliability of single protein structure prediction[1]. However, the general lack of co-evolutionary information and the lower number of solved structures makes predicting the structure of protein complexes a more difficult task[2]. Experimental distance restraints, such as proximal residue pairs revealed by crosslinking MS, can identify interactions[3–5] and their interfaces[6]. The experimental information can supplement information derived from evolutionary relationships in guiding structure prediction[7].

In our previous work, we improved modelling of challenging single protein targets by leveraging distance restraints in AlphaLink[7], a deep learning-based method that integrates crosslinking mass spectrometry (MS) data directly into the pair representation of AlphaFold2[1] to bias the prediction. Here, we extend AlphaLink to protein complexes. This required changing the code base from OpenFold[8] to Uni-Fold[9], using a training set of protein complexes[10], and fine-tuning the network with simulated succinimidyl 4,4-azipentanoate (SDA) crosslinks. This soluble crosslinker provides restraints between residues receptive to its first reaction step, Lys, Ser, Thr and Tyr, and other amino acid on protein surfaces[11].

We evaluated AlphaLink on challenging heteromeric CASP15 (Critical Assessment of Structure Prediction[12]) targets with simulated crosslinks. Moreover, we validated it on in-cell crosslinking MS data from *Bacillus subtilis*[5] and a virally modified Cullin4-RING ubiquitin ligase (CRL4) complex[13,14]. We focused the evaluation on heteromeric assemblies, since homomers pose a different challenge. The inherent ambiguity of "self" crosslinks in homo-multimeric assemblies rarely permits distinguishing inter- from intra-chain restraints[6].

Integrating crosslinking MS data substantially improves prediction quality of challenging CASP15 targets

AlphaLink with distance restraints vastly outperforms AlphaFold-Multimer[2] (from here on referred to as AlphaFold). It achieves similar or better results on challenging heteromeric CASP15 targets than the best performing algorithms that participated in CASP15, which used up to 2400x more sampling to derive predictions. Incorporating simulated SDA crosslinks in the modelling of eight challenging heteromeric CASP15 targets (H1129, H1134, H1140, H1141, H1142, H1144, H1166, H1167) substantially improved the DockQ[15] score from 0.14 to 0.48 on average, compared to the AlphaFold baseline (**Fig. 1a**). The test set comprised both dimers and multimeric assemblies. Except for H1129, crosslinking MS data helped to produce at least acceptable solutions according to DockQ score (DockQ ≥ 0.23[15]). To ensure comparability, we used the same multiple sequence alignments (MSAs) as AlphaFold and fine-tuned AlphaLink on the v2 network weights of AlphaFold. The simulated crosslink-derived distance restraints had 10% sequence coverage and 20% false-discovery rate[16], resulting in 33 links in the median per protein-protein interaction (including 7 false links, in the median). For comparison purposes, we selected the best predictions based on the highest model confidence[2] (0.8 * interface predicted TM-score (ipTM) + 0.2 * predicted TM-score), which allows us to pick close to the best model (**Extended Data Fig. 1**). Indeed, ipTM and DockQ correlate (ipTM 0.6 equates roughly DockQ 0.4)[2]. However, it is worth noting that further improvements in model selection for AlphaLink can be achieved by considering crosslink satisfaction. Selecting first by crosslink satisfaction and then for model confidence, we succeed in selecting high-quality predictions even when the model confidence is not discriminative (**Extended Data Fig. 2**). The quality of the selected model increases substantially in certain cases, e.g., the H1141 DockQ score improves from 0.07 (incorrect[15]) to 0.28 (acceptable) (**Fig. 1b**). We compare the performance of AlphaLink with and without crosslinks to exclude the influence of other parameters, such as having trained AlphaLink on larger crops than AlphaFold v2.2. Indeed, we see that the observed improvements are the result of integrating crosslinks (**Extended Data Fig. 3**).

Crosslinking MS data improve modelling of antibody-antigen targets

The largest improvements were observed on nanobody-antigen (yellow shaded area in **Fig. 1a**) and antibody-antigen targets (red shaded area in **Fig. 1a**), where the co-evolutionary signal is lower[17]. In these cases, crosslinking MS data drastically aided prediction. AlphaFold models are incorrect for 5 out of 6 targets (DockQ < 0.23[15]) while AlphaLink generates at least medium quality models for 5 out of 6 targets (DockQ > 0.49[15]). Notably, for H1142, H1166, and H1167,

AlphaLink produced better median score predictions than the top-ranked CASP15 submissions (**Fig. 1a,c,d**). The DockQ score for H1142 improved from 0.01 for AlphaFold and 0.1 in CASP15 to 0.68 by using AlphaLink. All true links are satisfied, while 100% of the false links are rejected. This demonstrates a high resilience of AlphaLink towards noise when predicting protein complexes, as was already observed for single proteins[7]. Similarly, the DockQ score for H1166 improved from 0.22 (AlphaFold) to 0.65 (AlphaLink), again with crosslink satisfaction and noise rejection being 100%.

Current limitations

We identified a number of challenges that AlphaLink inherits from AlphaFold. Although the predicted backbones for H1142 and H1166 were well aligned to the native structure (**Fig. 1c,d**), side chain interactions were poorly predicted. In the interfaces of H1142 and H1166, nine out of 39 and 38 out of 138 native contacts, respectively, were missed due to wrong side chain orientations (highlighted by the red circle in **Fig. 1c**). The addition of side chains in the very last step of the prediction pipeline limits their influence on the AlphaFold/AlphaLink predictions. Side chains provide specific interactions across protein-protein interfaces and accordingly play a crucial role in docking, providing additional contacts to support interface prediction. Including side chains earlier might aid the prediction of targets such as H1134 that have only few contacts in the interface and are more flexible (**Extended Data Fig. 4**), plausibly explaining the large spread in the DockQ scores in CASP15 (**Extended Data Fig. 5**). Such flexible targets may benefit from increased sampling and recycling, or a modified loss function that incorporates physical terms. Indeed, most methods employed during CASP15 used substantially more sampling than AlphaLink here. For example, Wallner *et al.*, one of the top predictors, generated as many as 30,000 samples for some targets (6000 for H1134) and increased the number of recycling iterations[18]. Based on these observations, we increased the number of samples moderately for H1134 from 10 to 100 and the number of recycling iterations from 3 to 20 and see improved results (mean DockQ score increases from 0.45 to 0.53) (**Extended Data Fig. 6**). The results of H1129 and H1141 indicate remaining challenges of our approach. AlphaFold predicts chain B of H1129 separately better than within the complex (0.89 vs 0.79), suggesting a potential problem with the fold-and-dock approach of AlphaFold and thus AlphaLink. Regarding H1141, we observe two clusters that differ in the relative orientation of the subunits. As both satisfy all crosslinks (blue in **Extended Data Fig. 7**), SDA crosslinks were not restrictive enough to return a single cluster. However, the better cluster (model confidence 0.86 versus 0.38) corresponds to the crystal structure (DockQ score 0.72).

A single crosslink obtained in cells can dramatically improve model quality

Encouraged by the success of AlphaLink with simulated data, we modelled 135 dimeric protein-protein interactions (PPIs) with real data from *Bacillus subtilis* cells[5] crosslinked *in situ* with disuccinimidyl sulfoxide (DSSO)[19]. This soluble crosslinker provides restraints between Lys, Ser, Thr, and Tyr groups. The data are sparse (median 1 crosslink per PPI), however even one crosslink can drastically improve the results (**Fig. 2a**). For example, the model confidence of the

CodY-YppF interaction[5,20] improves from 0.25 to 0.81 based on a single crosslink (**Fig. 2b**). For RpoA-RpoC (PDB 6WVK), the DockQ improves from 0.003 to 0.69 based on four in-cell crosslinks (**Fig. 2c**). Each individual crosslink would have sufficed to improve model quality substantially (DockQ 0.69-0.7). In the case of the YlaN-Fur interaction[5,20], AlphaFold predicts two conformations. One of these does not agree with the crosslinking MS data and does not support an interaction (model confidence 0.41). AlphaLink only predicts a single conformation (model confidence 0.84) consistent with the crosslink data (**Extended Data Fig. 8**). AlphaLink is fine-tuned on model_1, in comparison, AlphaFold predicts the targets with 5 different networks. This improves the model confidence on average by 0.03 points (**Extended Data Fig. 9).** As seen with the CASP15 targets, crosslinks help to focus sampling on the interesting regions, reducing the amount of sampling required. Overall, integrating crosslinking MS data increases the median model confidence from 0.42 (AlphaFold) to 0.6 (AlphaLink). With AlphaLink, 12 additional interactions (total of 46, gain of 35%) reach a model confidence > 0.75. Remarkably, the AlphaLink network was not trained on DSSO data, which differ in linked residues, linker length, and density of data from simulated SDA data. Thus, our results suggest a broad applicability of AlphaLink to varying crosslinker chemistry and the typically low data densities achieved in large-scale and whole-cell crosslinking MS experiments.

Crosslinking MS-driven modelling of a multi-protein complex

Finally, we challenged AlphaLink to model, with the help of real sulfo-SDA data, the 6-subunit $CRL4^{DCAF1\text{-}CtD}/Vpr_{mus}$/SAMHD1 assembly[14] (360 kDa, 3118 AA) (**Fig. 3**), using a crystal structure and cryo-EM density[14] as ground truth. The accessory protein Vpr from certain simian immunodeficiency viruses targets the DCAF1 substrate receptor of host Cullin4-RING ubiquitin ligases (CRL4), to recruit and mark the restriction factor SAMHD1 for proteasomal degradation and thus to stimulate virus replication[14,21]. The v2.2 network weights of AlphaFold fail to return meaningful structures of this assembly. This is overcome by moving to v2.3 network weights, which have been trained on larger complexes. Leveraging crosslinks improves the model confidence from 0.56 (AlphaFold v2.3) to 0.64 (AlphaLink v2.3) (**Fig. 3a**). The CRL4 subunit CUL4A shows substantial movements, resulting in three conformational states according to cryo-EM data (**Fig. 3b**). AlphaFold predicts a contact between CUL4A and DCAF1 that is not supported by the cryo-EM density (**Fig. 3c**). By contrast, AlphaLink predicts no contact (**Fig. 3d**). In addition, crosslinks allow AlphaLink to position the viral Vpr protein inside the experimental density in agreement with the crystal structure (PDB 6ZX9) (**Fig. 3e, green, f**), while AlphaFold places Vpr incorrectly. Our predicted Vpr-DCAF1 interaction model achieved a DockQ score of 0.56 compared to 0.04 by AlphaFold (**Fig. 3f**).

Conclusion

Our findings demonstrate the successful extension of AlphaLink, an experiment-assisted AI approach, to predict protein complex structures. We can now start visualising at pseudo-atomic resolution protein-protein interactions inside cells, by leveraging the very scarce data of whole-cell crosslinking MS. With this breakthrough, protein-protein interaction studies move

from delivering links (who) to structures (how), inside cells and at scale. Importantly, in-cell crosslinking reveals protein interactions that are often lost upon cell lysis[5]. While lysis is required frequently for large-scale study of protein complexes, doing so without prior crosslinking challenges especially transient and fragile features of complexes. AlphaLink with whole-cell crosslinking will therefore accelerate discovering hitherto hidden aspects of biology, to advance our understanding of life and widen our therapeutic options in cases of disease.

References

1. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
2. Evans, R. *et al.* Protein complex prediction with AlphaFold-Multimer. *bioRxiv* 2021.10.04.463034 (2022) doi:10.1101/2021.10.04.463034.
3. Tang, X. & Bruce, J. E. Chemical Cross-Linking for Protein–Protein Interaction Studies. in *Mass Spectrometry of Proteins and Peptides: Methods and Protocols* (eds. Lipton, M. S. & Paša-Tolic, L.) 283–293 (Humana Press, 2009).
4. Lenz, S. *et al.* Reliable identification of protein-protein interactions by crosslinking mass spectrometry. *Nat. Commun.* **12**, 3564 (2021).
5. O'Reilly, F. J. *et al.* Protein complexes in cells by AI-assisted structural proteomics. *Mol. Syst. Biol.* **19**, e11544 (2023).
6. Maiolica, A. *et al.* Structural analysis of multiprotein complexes by cross-linking, mass spectrometry, and database searching. *Mol. Cell. Proteomics* **6**, 2200–2211 (2007).
7. Stahl, K., Graziadei, A., Dau, T., Brock, O. & Rappsilber, J. Protein structure prediction with in-cell photo-crosslinking mass spectrometry and deep learning. *Nat. Biotechnol.* (2023) doi:10.1038/s41587-023-01704-z.
8. Ahdritz, G. *et al.* OpenFold: Retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization. *bioRxiv* 2022.11.20.517210 (2022) doi:10.1101/2022.11.20.517210.
9. Li, Z. *et al.* Uni-Fold: An Open-Source Platform for Developing Protein Folding Models beyond AlphaFold. *bioRxiv* 2022.08.04.502811 (2022) doi:10.1101/2022.08.04.502811.
10. Townshend, R., Bedi, R., Suriana, P. & Dror, R. End-to-end learning on 3d protein structure for interface prediction. *Adv. Neural Inf. Process. Syst.* **32**, (2019).
11. Belsom, A., Schneider, M., Fischer, L., Brock, O. & Rappsilber, J. Serum Albumin Domain

Structures in Human Blood Serum by Mass Spectrometry and Computational Biology. *Mol. Cell. Proteomics* **15**, 1105–1116 (2016).

12. Moult, J., Pedersen, J. T., Judson, R. & Fidelis, K. A large-scale experiment to assess protein structure prediction methods. *Proteins* **23**, ii–v (1995).
13. Mahon, C., Krogan, N. J., Craik, C. S. & Pick, E. Cullin E3 ligases and their rewiring by viral factors. *Biomolecules* **4**, 897–930 (2014).
14. Banchenko, S. *et al.* Structural insights into Cullin4-RING ubiquitin ligase remodelling by Vpr from simian immunodeficiency viruses. *PLoS Pathog.* **17**, e1009775 (2021).
15. Basu, S. & Wallner, B. DockQ: A Quality Measure for Protein-Protein Docking Models. *PLoS One* **11**, e0161879 (2016).
16. Fischer, L. & Rappsilber, J. Quirks of Error Estimation in Cross-Linking/Mass Spectrometry. *Anal. Chem.* **89**, 3829–3833 (2017).
17. Yin, R., Feng, B. Y., Varshney, A. & Pierce, B. G. Benchmarking AlphaFold for protein complex modeling reveals accuracy determinants. *Protein Sci.* **31**, e4379 (2022).
18. Wallner, B. TS037-Wallner group AlphaFold with aggressive sampling. *predictioncenter.org* https://predictioncenter.org/casp15/doc/presentations/Day2/Predictor_BWallner_CASP15.pdf.
19. Kao, A. *et al.* Development of a novel cross-linking strategy for fast and accurate identification of cross-linked peptides of protein complexes. *Mol. Cell. Proteomics* **10**, M110.002212 (2011).
20. Pedreira, T., Elfmann, C. & Stülke, J. The current state of SubtiWiki, the database for the model organism Bacillus subtilis. *Nucleic Acids Res.* **50**, D875–D882 (2022).
21. Fregoso, O. I. *et al.* Evolutionary toggling of Vpx/Vpr specificity results in divergent recognition of the restriction factor SAMHD1. *PLoS Pathog.* **9**, e1003496 (2013).
22. Vallat, B., Webb, B., Westbrook, J. D., Sali, A. & Berman, H. M. Development of a Prototype System for Archiving Integrative/Hybrid Structure Models of Biological Macromolecules.

Structure **26**, 894–904.e2 (2018).

23. Kahraman, A., Malmström, L. & Aebersold, R. Xwalk: computing and visualizing distances in cross-linking experiments. *Bioinformatics* **27**, 2163–2164 (2011).
24. Méndez, R., Leplae, R., De Maria, L. & Wodak, S. J. Assessment of blind predictions of protein-protein interactions: current status of docking methods. *Proteins* **52**, 51–67 (2003).

Acknowledgements

We thank Andrea Graziadei and David Schwefel for their feedback on the manuscript. We are grateful to the Uni-Fold team for providing a fully open-source and trainable reimplementation of AlphaFold-Multimer (https://github.com/dptech-corp/Uni-Fold).

Funding

This research was supported by the Wellcome Trust [Grant number 227166] (JR, KS) and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany´s Excellence Strategy – EXC 2008 – 390540038 (JR) – UniSysCat (JR) and under Germany's Excellence Strategy – EXC 2002/1 "Science of Intelligence" – project number 390523135 (OB). The Wellcome Centre for Cell Biology is supported by core funding from the Wellcome Trust [203149]. For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Author contributions

Software development: KS

Data Analysis: KS

Supervision: JR, OB

Writing- initial draft: KS, JR, OB

Writing- editing and revision: KS, JR, OB

Competing interests

The other authors declare no competing interests.

Data and materials availability

AlphaLink models based on experimental crosslinks have been deposited as integrative/hybrid models in PDB-Dev[22] with accession code XXX.
The code for AlphaLink is deposited at https://github.com/Rappsilber-Laboratory/AlphaLink2.
Model weights were deposited at Zenodo: doi:10.5281/zenodo.8007238

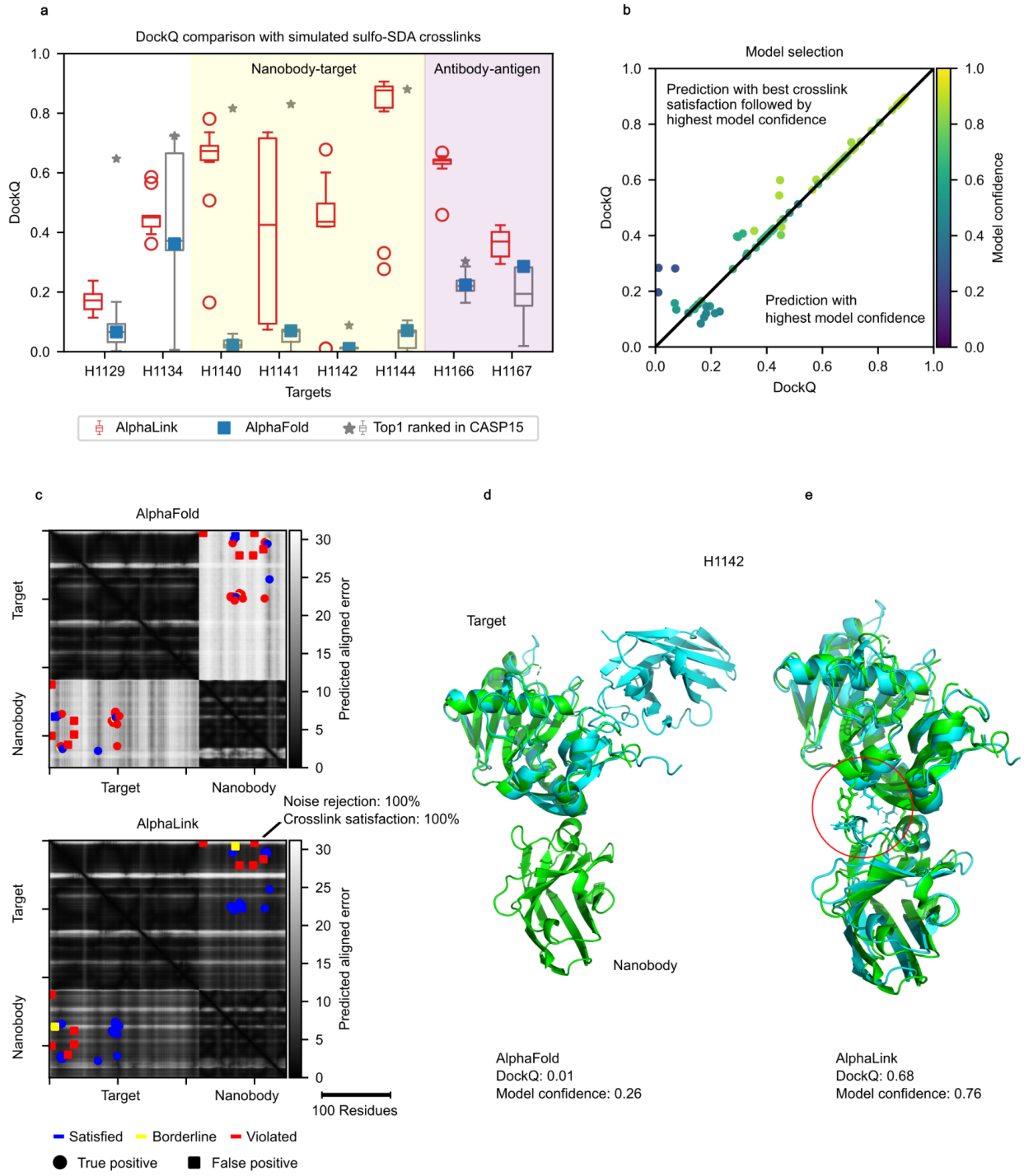

Fig. 1: AlphaLink performance on simulated data.

a- DockQ comparison on 8 heteromeric CASP15 targets with simulated SDA crosslinks. The AlphaLink boxplot shows the highest model confidence model (N = 10) for 10 randomly sampled crosslink sets. The top ranked predictions (grey) correspond to all submissions that were ranked

first by the participants (model_1), the best for each target is highlighted with an asterisk. AlphaFold is highlighted in blue (model_1). The yellow shaded area highlights nanobody, the red shaded area antibody targets. In all boxplots, the line shows the median and the whiskers represent the 1.5x interquartile range. **b-** AlphaLink's ranking performance (DockQ) with model confidence vs selecting first by crosslink satisfaction and second by model confidence. **c-** PAE map comparison of AlphaFold and AlphaLink for H1142. **d-** AlphaFold prediction for H1142 (cyan) aligned to the crystal structure (green). **e-** AlphaLink prediction for H1142 (cyan) aligned to the crystal structure (green). The red circle highlights differences in side chain orientations in the interface.

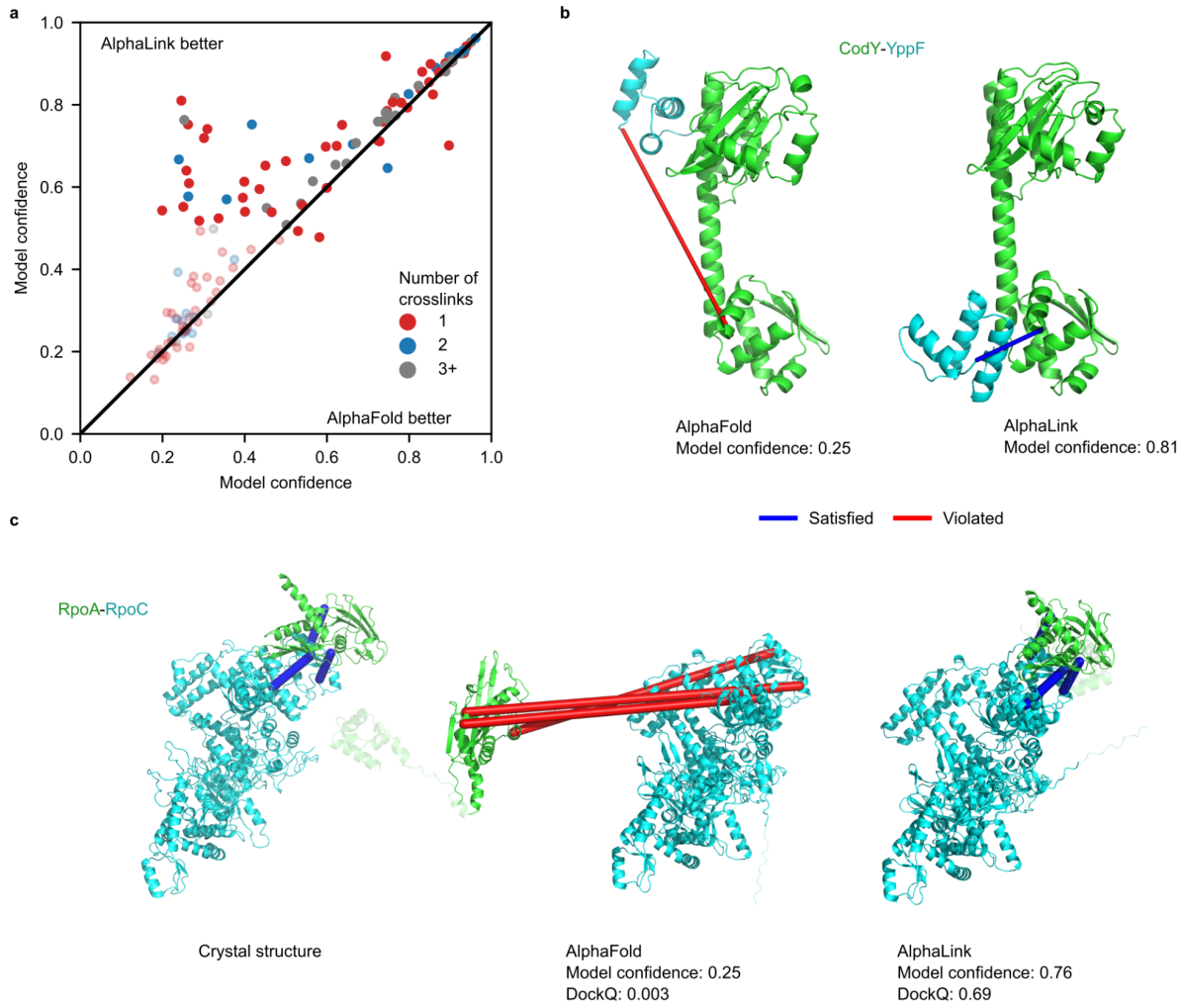

Fig. 2: AlphaLink performance on real data from *Bacillus subtilis*.
a- Highest model confidence of AlphaLink v2.2 (N = 5) vs AlphaFold v2.2 (M = 5) on the *Bacillus subtilis* data on K = 135 dimeric protein-protein interactions. Low confidence predictions for both methods (model confidence < 0.5) have higher transparency. **b-** AlphaFold (left) and AlphaLink (right) prediction for CodY-YppF with annotated crosslinks (blue - satisfied, red - violated). **c-** Comparison of crystal structure (left), AlphaFold prediction (middle) and AlphaLink prediction (right) for RpoA-RpoC with annotated crosslinks (blue - satisfied, red - violated). We omitted one crosslink that is not covered by the crystal structure.

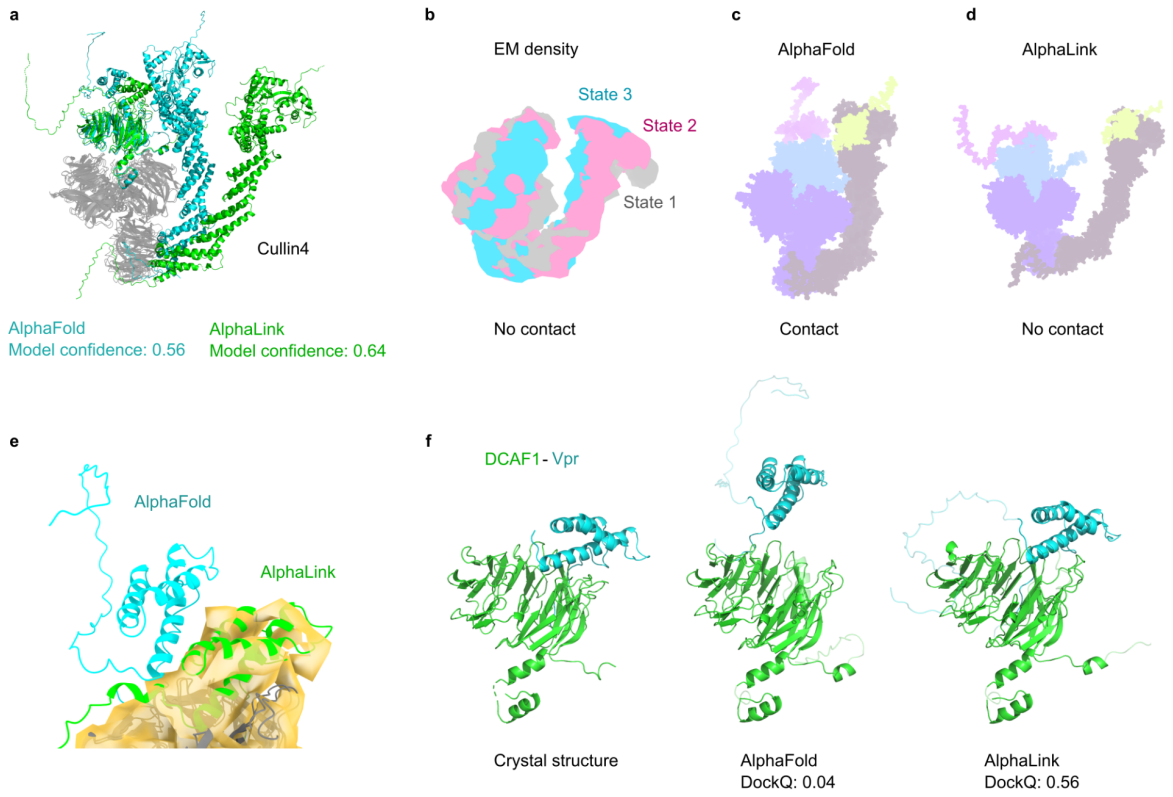

Fig. 3: AlphaLink performance on real data from CRL4 with v2.3 weights.
a- AlphaFold (v2.3) (cyan) and AlphaLink (v2.3) (green) prediction for a virally modified CRL4 assembly. SAMHD1 is not visualised for clarity. **b-** Schematic view of the three states of Cullin4 observed in the EM density (EMD-10612 to EMD-10614). **c-** Schematic view of the AlphaFold prediction, showing a contact not present in the density (**b**). **d-** Schematic view of the AlphaLink prediction. **e-** Placement of the viral Vpr protein for AlphaFold (cyan) and AlphaLink (green) in the EM density. **f-** Predictions of DCAF1-Vpr compared to the crystal structure (PDB 6ZX9).