Protein structure prediction with in-cell photo-crosslinking mass spectrometry and deep learning

nature biotechnology

Article

https://doi.org/10.1038/s41587-023-01704-z

Protein structure prediction with in-cell photo-crosslinking mass spectrometry and deep learning

Received: 26 September 2022

Accepted: 6 February 2023

Published online: 20 March 2023

Check for updates

Kolja Stahl[1,7], Andrea Graziadei[2,7], Therese Dau[2,6], Oliver Brock[1,3] ✉ & Juri Rappsilber[2,4,5] ✉

While AlphaFold2 can predict accurate protein structures from the primary sequence, challenges remain for proteins that undergo conformational changes or for which few homologous sequences are known. Here we introduce AlphaLink, a modified version of the AlphaFold2 algorithm that incorporates experimental distance restraint information into its network architecture. By employing sparse experimental contacts as anchor points, AlphaLink improves on the performance of AlphaFold2 in predicting challenging targets. We confirm this experimentally by using the noncanonical amino acid photo-leucine to obtain information on residue–residue contacts inside cells by crosslinking mass spectrometry. The program can predict distinct conformations of proteins on the basis of the distance restraints provided, demonstrating the value of experimental data in driving protein structure prediction. The noise-tolerant framework for integrating data in protein structure prediction presented here opens a path to accurate characterization of protein structures from in-cell data.

AlphaFold2 has shown unprecedented performance in CASP14, the Critical Assessment of protein Structure Prediction[1–3], predicting two-thirds of the CASP targets with an approximately 1 Å root-mean-square deviation (r.m.s.d.) from the native backbone path[4]. This success, together with the reliable metrics provided by AlphaFold2 regarding the predicted accuracy of its models, is a tremendous achievement whose impact on life sciences is still unfolding.

AlphaFold2 predicts static models based on static input data. AlphaFold2 was trained on two information sources, the protein structures in the Protein Data Bank (PDB) and multiple sequence alignments (MSAs). This approach is challenged by targets that have insufficient evolutionary information, generating less confident or erroneous predictions[3]. For some classes of proteins, such as viral proteins, proteins from understudied organisms, antibodies[5] and synthetic proteins, but also clinically relevant mutations[6], evolutionary information may be misleading. Moreover, the x-ray structures underlying the model poorly reflect structural flexibility, multiple conformations and dynamic interactions. Structural restraints observed on proteins in solution, ideally in the cell, could help resolve these problems. Adding such restraints to the AlphaFold2 framework may then steer the prediction towards structural states occurring in situ under specific conditions.

Crosslinking mass spectrometry (MS) is capable of providing distance restraints that can be used in protein structure prediction[7–9]. In particular, photo amino acids (photo-AA) are readily incorporated by both prokaryotic and eukaryotic cells[10–12], which opens up the possibility of probing the in situ conformation of proteins. Unlike most soluble crosslinkers, where data can be polluted by rare protein

[1]Robotics and Biology Laboratory, Technische Universität Berlin, Berlin, Germany. [2]Technische Universität Berlin, Chair of Bioanalytics, Berlin, Germany. [3]Science of Intelligence, Research Cluster of Excellence, Berlin, Germany. [4]Si-M/'Der Simulierte Mensch', a Science Framework of Technische Universität Berlin and Charité - Universitätsmedizin Berlin, Berlin, Germany. [5]Wellcome Centre for Cell Biology, University of Edinburgh, Edinburgh, UK. [6]Present address: Fritz Lipmann Institute, Leibniz Institute on Aging, Jena, Germany. [7]These authors contributed equally: Kolja Stahl, Andrea Graziadei. ✉e-mail: oliver.brock@tu-berlin.de; juri.rappsilber@tu-berlin.de

https://doi.org/10.1038/s41587-023-01704-z

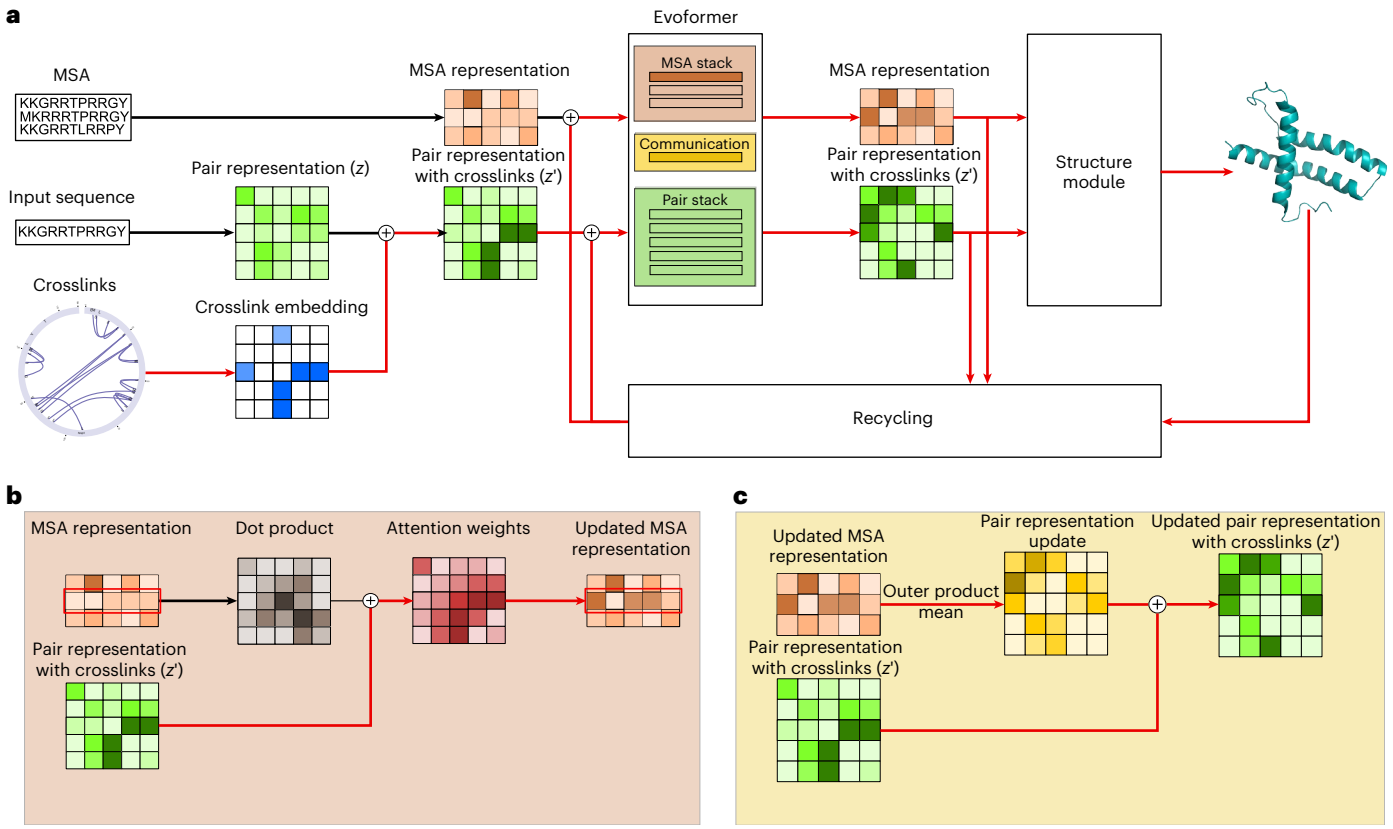


Fig. 1 | Information flow in AlphaLink. a, Overview of the information flow in AlphaLink. Crosslinks (blue) are embedded and added onto the pair representation (green). Impact of crosslinks shown in red. **b**, Crosslinks influence the retrieval of co-evolutionary information. They are used as a bias in the MSATransformer. **c**, The pair representation is updated with information from the MSAs that have been biased with the crosslinks.

states, photochemistry accurately represents in-solution ensembles[13,14]. Furthermore, photo-AA crosslinks yield comparably tight distance restraints that align well with co-evolutionary contacts, which are the basis of most protein structure prediction methods, including AlphaFold2. They are in theory capable of 'zero length' crosslinking from the side chain to any heavy atom via a reactive carbene[10] or alkyl diazo[15] intermediate. Photo-leucine (photo-L) was used in mapping conformations and binders in purified systems[11,12] but has not been used so far for in situ structure analyses. In general, the incorporation of amino acid analogs into the proteome is advantageous for crosslinking studies, because they allow the introduction of genetically encoded chemical entities that can react chemo-selectively at known locations in proteins[16].

In this Article, we introduce AlphaLink, a structure prediction method that integrates experimental data from photo-AA crosslinking directly into the AlphaFold2 architecture. AlphaLink uses deep learning to merge co-evolutionary relationships and crosslinking data in distance space, exploiting the complementary nature of the data. We demonstrate that AlphaLink can leverage noisy experimental contacts to improve predictions of challenging targets on both simulated and real experimental data, steering predictions towards the in situ conformation of proteins. To test AlphaLink, we perform a large-scale crosslinking MS study with photo-L, identifying 615 in situ residue–residue contacts in *Escherichia coli* membrane fractions, unlocking the power of photo-AA in mapping proximal residues directly in cells. We show that even sparse crosslinking MS data can anchor predictions to particular conformational states, opening up the possibility of probing dynamics by hybrid experimental/deep learning approaches. We further extend AlphaLink to arbitrary distance restraints by introducing a second representation that encodes distance restraints as distograms[3].

Results

AlphaLink: integrating crosslinks into AlphaFold2 via OpenFold

Crosslinking MS data have been used to guide candidate selection for AlphaFold-multimer in protein–protein interaction studies and validate models[17,18]. To fully leverage the potential of crosslinking MS data in protein structure prediction, we develop AlphaLink, a framework incorporating crosslinks directly into OpenFold[19]. OpenFold is a trainable reproduction of AlphaFold2. The creators of OpenFold verified that the implementation produces identical results. OpenFold primarily exploits co-evolutionary relationships. The main difficulty in merging multiple information sources is to find a suitable representation that facilitates integration and at the same time avoids information loss. OpenFold operates both in distance space (Evoformer) and in 3D space (Structure Module). Photo-AA crosslinking MS data provide distance restraints that naturally fit into the distance space of OpenFold, since they yield similar distances to co-evolutionary contacts by directly linking amino acids via diazirine chemistry. Co-evolutionary relationships and photo-AA crosslinks provide complementary and corroborating information. The sparsity of crosslinks can be compensated with co-evolutionary information. Accurate crosslinking data can act as an anchor in these cases. AlphaLink exploits this relationship by merging crosslinking MS and co-evolutionary data via the Evoformer, injecting crosslinks into the pair representation (z), yielding a consistent and unified constraint set (Fig. 1).

We introduce two representations to encode crosslinking information. The experimental data are represented as either soft labels or distance distributions (distograms). In the case of soft labels, each contact is weighted by the link-level false discovery rate (FDR) of the dataset (1-FDR) or, if present, the per-restraint FDR to indicate confidence in

crosslink assignment. Distograms allow us to generalize to arbitrary distance restraints. A particular crosslinker (or distance restraint in general) is represented by a distance distribution. Contact-like restraints can be represented by uniformly distributed distograms for the given cutoff. We model uncertainty directly in the representation by adjusting the probability mass according to the FDR. The distogram is designed to match the distogram that is predicted by the Evoformer from the pair representation that consists of 64 bins. We use the same binning for the first 64 bins and extend the distogram further to 128 bins, spanning from 2.3125 Å to 42 Å.

We embed the restraints and add them to the pair representation of OpenFold, which is later mapped into 3D space (Fig. 1a). The embedding is similar to the recycling embedding in AlphaFold2. The Evoformer jointly updates the MSA and the pair representation. The MSA transformer (Fig. 1b) retrieves co-evolutionary information and updates the MSA representation. The retrieval is biased with the pair representation that includes the experimental crosslinking information supplied by the user. The outer product mean (Fig. 1c) in turn updates the pair representation. This coupling maximizes synergy between MSA and experimental information and allows the network to perform noise rejection, that is, the rejection of misassigned experimental or co-evolutionary relationships or of contacts that do not support other strands of information leading to a consensus model.

We initialized OpenFold with the original weights of AlphaFold2 and fine-tuned the network with the newly added crosslinking bias. We followed the refinement training regime outlined in the AlphaFold2 paper, except that we subsampled the number of effective sequences (N_{eff}) to simulate challenging targets. In light of the limited availability of experimental crosslink data for training, we simulated photo-crosslinking MS data (Methods) that included simulated experimental noise in the form of false residue–residue contacts at the given FDR.

Integrating photo-AA crosslinks enables noise-tolerant prediction of challenging targets

We tested AlphaLink on 49 challenging CAMEO targets ($N_{\mathrm{eff}} \leq 25$, no MSA subsampling, Supplementary Data 1) (Fig. 2a). AlphaLink outperforms AlphaFold2, substantially improving the performance on targets with more than 20 crosslinks. Integrating simulated photo-L data improves the TM score on average by 19.2 ± 16.3% (95% confidence interval) (Fig. 2a). Encoding the crosslinks as distograms instead performs virtually the same (Extended Data Fig. 1a).

We further curated a second benchmark dataset consisting of 60 CASP14 targets and 45 CAMEO targets (Supplementary Data 1). To simulate challenging targets and to control for the MSA influence, we subsampled the MSAs to $N_{\mathrm{eff}} = 10$ and ignored structural templates. Here AlphaLink improves the TM score on average by 15.2% (Extended Data Fig. 1b). For particularly challenging targets ($N = 28$), where AlphaFold2 fails to predict the correct fold (TM score ≤0.5), the TM score improves on average by 50.6% (Fig. 2b). AlphaLink predicts the correct fold (TM score >0.5) of 14 of these. We tested the noise rejection capabilities of AlphaLink on 60 CASP14 targets by adding false links to simulate multiple noise levels. The performance is roughly constant with 10%, 20% or 50% false links (Fig. 2c) and still outperforms AlphaFold2, demonstrating AlphaLinks' robustness to different noise levels. Overall, the method achieves a crosslink satisfaction (<10 Å Cα–Cα) on average of 85 ± 1.2% (95% confidence interval) after three recycling iterations, and 88.3 ± 1.2% (95% confidence interval) of the simulated crosslinks with <10 Å Cα–Cα in the crystal structure are satisfied.

The sparse crosslink data act as anchor points that serve to pull the entire prediction towards the right solution (Fig. 2d). For CASP target T1064 ($N_{\mathrm{eff}} = 10$), four crosslinking restraints are sufficient to both drive the prediction to the native state (TM score improves from 0.28 to 0.86) and to decrease the predicted aligned error across the whole protein, including areas not covered by the crosslinking data. The crosslinking information has a wide-ranging impact due to its combination with the co-evolutionary and structural information embedded in the pair representation, which is used as a bias to retrieve contacts consistent with experimental data. Effectively, this improves the efficiency of using co-evolutionary information in AlphaFold2. Extended Data Fig. 1c shows the effect of using different distograms to encode a restraint between residues 11 and 103 in T1064. The Evoformer predicts a narrower distogram when using the expected distance distribution of photo-AA crosslinks as a prior, when compared with the uniform prior of an upper bound distance restraint. This representation slightly improves the prediction (TM score 0.68 to 0.7). The performance as a function of the number of crosslinks per residue is shown in Extended Data Fig. 1d. The performance generally increases with an increase in the number of crosslinks per residue. The main advantage of the distogram representation is enabling the user to inject distance restraints from different crosslinkers or even different experimental approaches into AlphaLink.

We test the performance of AlphaLink at different N_{eff} levels to investigate the effect of crosslinks on targets with varying difficulty (Fig. 2e). The performance of both AlphaFold2 and AlphaLink deteriorates in absence of sufficiently large MSAs (Fig. 2e). Crosslinks can compensate for smaller MSA sizes. In fact, photo-AA crosslinks alone without any MSA information allow us to predict the correct fold (TM score >0.5) of 43/105 benchmark targets, compared with 13/105 for AF2 without MSA information. The mean improvement in TM score increases to 75 ± 13.5% (95% confidence interval) over all targets (Fig. 2f). The benefit of crosslinks slowly disappears with a $N_{\mathrm{eff}} > 50$. This is at least partly due to the fact that most crosslinks are already satisfied when predicting with full MSAs (Fig. 2e). Rather than finding any solution that fits the crosslinks, our network appropriately weighs crosslinking MS information against the MSAs and uses it to guide the prediction to a more accurate solution. Note that as MSA size increases, the network will rely more on MSA information than on crosslinks—hence, we implement settings with different MSA subsamplings in the AlphaLink software package.

In summary, AlphaLink enables users to use sparse distance restraints to bias AlphaFold2 predictions, robustly handling noise, directly at the inference stage, due to their synergistic implementation in the network design.

Photo-L as an in situ structural probe

To generate a large-scale experimental photo-AA dataset required for testing such an application, we derived in situ structural restraints on the *E. coli* membrane fraction by crosslinking MS of cells grown on photo-L-containing medium. We optimized the growth protocol to maximize incorporation while maintaining a low level of cytotoxicity (750 µM photo-L in the medium, Extended Data Fig. 2a), ultraviolet (UV) illuminated the cells for crosslinking and then enriched the cell membrane of the crosslinked cells. The proteins were digested, and the resulting peptides subjected to two-dimensional fractionation, combining strong cation exchange and size exclusion chromatography (Extended Data Fig. 2b). Mass spectrometric analysis then led to the identification of 615 residue pairs involving 112 proteins at 5% link-level FDR (Fig. 3a, Extended Data Fig. 2c and Supplementary Data 2 and 3). Several crosslinks are detected among β-barrel proteins and proteins in the intermembrane space, including porins and known membrane complexes (Fig. 3a and Extended Data Fig. 2a). When visualized on known protein structures, the experimental crosslinks provide a median distance of 11.1 ± 8.1 Å Cα–Cα (mean ± standard deviation) (Fig. 3b), indicating the contact-like nature of these crosslinks in line with their implementation in AlphaLink. This is further supported by the fact that we exclude crosslinks within the same tryptic peptide, and between consecutive peptides in our analysis.

https://doi.org/10.1038/s41587-023-01704-z

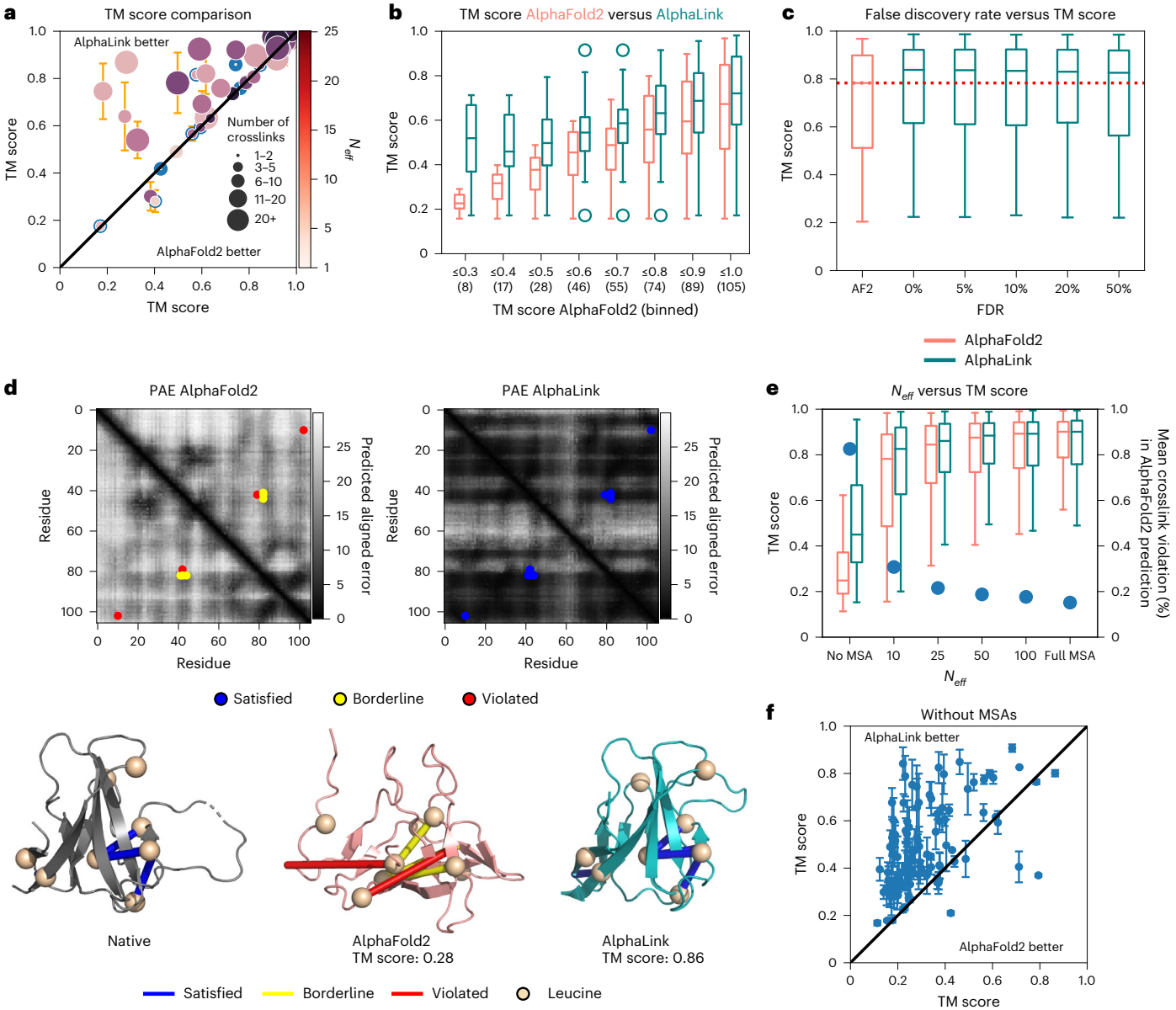


Fig. 2 | AlphaLink performance comparison against AlphaFold2. a, TM score comparison on 49 CAMEO targets with $N_{eff} \leq 25$. Error bars represent the 95% confidence intervals ($N = 10$). Points show the mean. TM score improves on average by 19.2%. **b**, Performance on 60 CASP14 and 45 CAMEO targets broken down by TM score ($N_{eff} = 10$). AlphaLink improves on average by 15.2%. Number of targets in each range bin in brackets below. **c**, Performance on 60 CASP14 targets ($N_{eff} = 10$) with different noise levels (FDR 0%, 5%, 10%, 20% and 50%). AlphaLink improves in the median for all noise levels. Performance shows robust noise rejection. Dotted line shows median performance of AlphaFold2. **d**, Predicted aligned error of AlphaFold2 (left) and AlphaLink (right) on T1064 with $N_{eff} = 10$ (top) and predicted structures (bottom). Light regions signify high uncertainty. Sparse restraints decrease uncertainty across the whole protein. Satisfied crosslinks <10 Å Cα–Cα highlighted in blue, borderline crosslinks (10–15 Å Cα–Cα) in yellow, and violated crosslinks >15 Å Cα–Cα in red. Possible crosslinking sites (leucines) are shown as spheres. Regions with violated crosslinks in the AlphaFold2 prediction (left) increase in certainty (darker regions). TM score improves from 0.28 to 0.86. **e**, Performance on 60 CASP14 targets ($N_{eff} = 10$) as a function of MSA size ($N = 100$, 10 MSAs and 10 crosslink sets). Dots represent the mean percentage of nonsatisfied crosslinks (>10 Å Cα–Cα) in the AlphaFold2 prediction. Improvement on average for all but full MSA size. Crosslink violation decreases and crosslink utility diminishes with increasing MSA size. Largest utility for $N_{eff} < 25$. **f**, Performance without MSAs on 60 CASP14 and 45 CAMEO targets. AlphaLink predicts the correct fold (TM score >0.5) for 43/105 (13/105 for AlphaFold2). Error bars represent the 95% confidence interval ($N = 10$). Points show the mean. In all box plots, the line shows the median and the whiskers represent the 1.5× interquartile range.

Photo-L provides validation for the in situ conformation of multiprotein complexes such as the AcrAB-TolC multidrug efflux pump, ribosome and ATP synthase (Fig. 3a). The crosslinks are consistent with previously characterized conformations of the bacterial outer membrane barrel assembly machinery (Bam). However, a link between the P2 and P3 domains highlights the flexibility of these modules (Fig. 3c), which are known to undergo large structural rearrangements in outer membrane protein folding and insertion. A total of 153 crosslinks are detected for the highly abundant protein OmpA. OmpA is made up of a β-barrel connected via a 20-residue linker to a C-terminal domain. It is also known to oligomerize in vivo, and this interaction is thought to be mediated by the C-terminal domain. The crosslinks between the β-barrel, linker and C-terminal domain highlight the relative flexibility of these modules (Fig. 3c) and point to potential contacts

https://doi.org/10.1038/s41587-023-01704-z

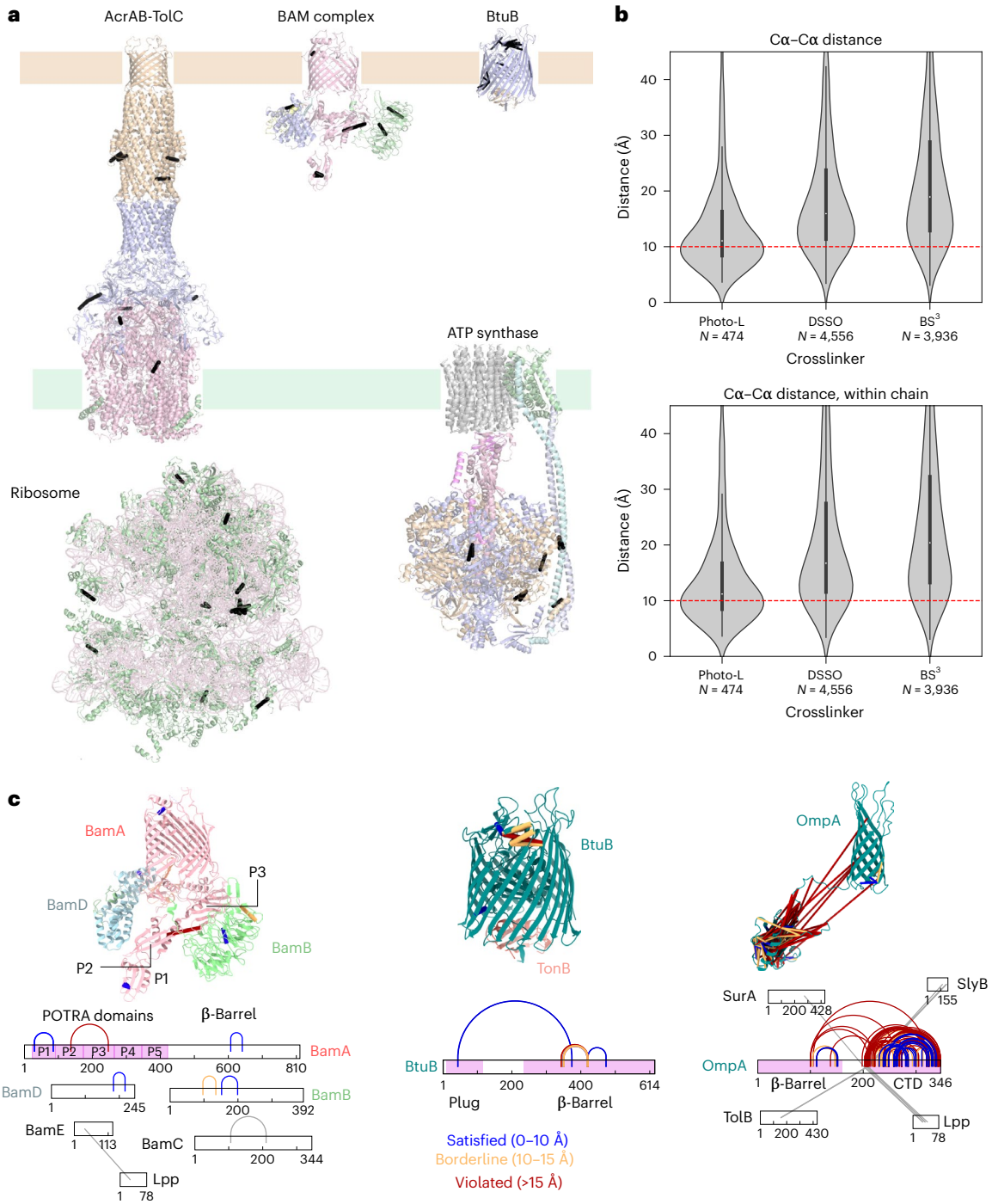


Fig. 3 | In situ photo-L crosslinking MS in *E. coli*. **a**, Distance restraints from in-cell photo-L crosslinking MS mapped onto cellular complexes. **b**, Distance distributions for photo-L crosslinks mapped onto known structures, taking only a single conformer per protein. Bissulfosuccinimidyl suberate (BS^3) and disuccinimidyl sulfoxide (DSSO) distograms are obtained by mapping the crosslinks of Lenz et al.[33]. The distograms are derived by accounting for homo-multimers (top) or mapping to only within-chain distances (bottom). **c**, Distance restraint analysis of outer membrane proteins crosslinked with photo-L. The dot represents the median, and the whiskers represent the 1.5× interquartile range.

made between multiple copies of the C-terminal domain. In several plug-containing β-barrel proteins, such as FhuA and BtuB, photo-L links the position of the central plug with the membrane barrel in a way that is consistent with previous structures (Fig. 3c), validating the arrangement of these two modules in the functional cycle of the proteins. These crosslinks highlight the potential of photo-L to provide in situ residue–residue contacts regardless of solvent accessibility, providing insight into function for critical domain contacts.

Structure prediction with in situ photo-L data

To test AlphaLink on experimental data, we predicted the proteins in the crosslinking MS dataset of the *E. coli* membrane fraction. We focused our evaluation on the 31 targets with high-resolution structures that had a median of five crosslinks (Fig. 4). Each target was predicted with ten randomly subsampled MSAs at $N_{eff} = 10$, yielding 310 predictions (Supplementary Data 4). We subsampled the MSAs to counter overfitting, because the targets were probably part of the AlphaFold2

https://doi.org/10.1038/s41587-023-01704-z

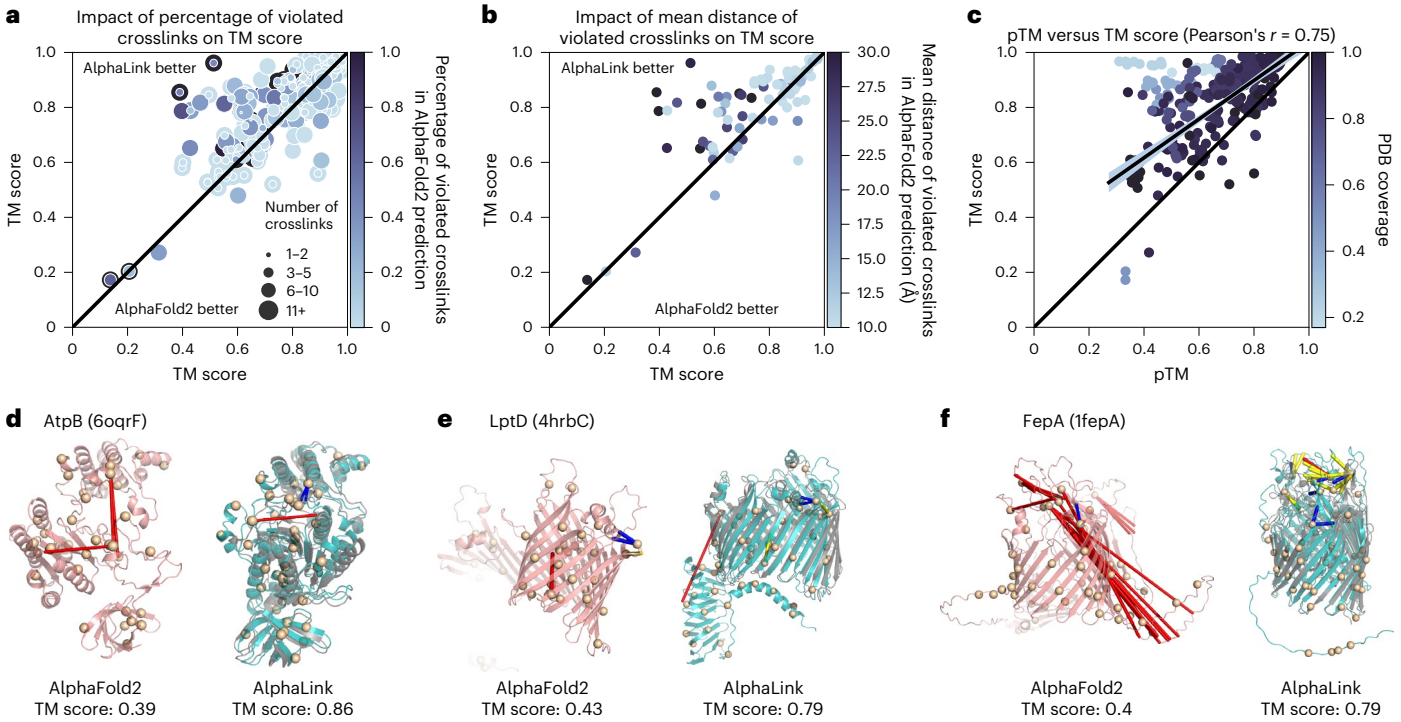


Fig. 4 | Structure prediction with in-cell photo-L crosslinking MS data of the *E. coli* membrane fraction. a, Comparison of TM score with annotated number of links (marker sizes) and percentage of nonsatisfied (>10 Å) crosslinks (color gradient) in the AlphaFold2 prediction. Performance improvement is bigger for targets with a higher percentage of nonsatisfied crosslinks in the base prediction (darker circles). Each target is predicted ten times with different MSA subsamples at $N_{eff} = 10$. AlphaLink outperforms AlphaFold2 on average. **b**, Comparison of TM score with annotated mean distance of nonsatisfied crosslinks in the base AlphaFold2 prediction (color gradient). Prediction quality improves with stronger crosslink violations (darker circles). **c**, We show the calibration of the pTM. On predictions that are at least 80% covered by the crystal structure, the correlation is 0.75. The true TM score is generally underestimated, meaning that the pTM score of AlphaLink is a conservative estimate. The shaded area corresponds to the 95% confidence interval. Line shows the linear fit. **d**, Prediction of the ATP synthase subunit AtpB by AlphaFold2 and AlphaLink using in-cell photo-L crosslinks at $N_{eff} = 10$. **e**, Prediction of the outer membrane lipopolysaccharide assembly protein. **f**, Prediction of the ferrienterobactin receptor. In all three cases, the in-cell crosslinking data helps AlphaLink position different regions of the protein relative to each other, yielding a performance improvement over AlphaFold2. The crystal structure of the target protein is shown in gray, overlaid with the AlphaLink prediction.

training set. Even with $N_{eff} = 10$, 65% of the AlphaFold2 predictions exceed a TM score of 0.8. AlphaLink improves performance measured by TM score on average by 5.2 ± 1.9% (95% confidence interval) across all proteins relative to AlphaFold2.

On targets where AlphaFold2 does not provide accurate models (TM score <0.8), AlphaLink with experimental data improves the TM score on average by 15.9 ± 4.6% (95% confidence interval). The improvement increases to 47.8 ± 24.8% (95% confidence interval) for AlphaFold2 predictions below a TM score of 0.5. We predict the correct fold (TM score >0.5) for ten additional proteins. This shows that simulated crosslinking MS data successfully model the features of experimental photo-AA restraints. For the 204 AlphaFold2 predictions with a TM score of 0.8 or higher, the performance is unaffected. At high TM scores, side-chain conformations begin to play a role, and crosslinking MS data do not have the resolution necessary to improve side-chain predictions.

To better judge the utility of the crosslinks for a given target, we include the percentage of nonsatisfied crosslinks in the baseline AlphaFold2 prediction (Fig. 4a) and also consider the mean distance of the nonsatisfied crosslinks in the AlphaFold2 prediction (Fig. 4b). We set the cutoff for violated crosslinking restraints to 10 Å Cα–Cα in the crystal structure. Many targets are not completely covered by the crystal structure. Therefore, we can analyze only a subset of the crosslinks. Crosslinks that are already satisfied in the AlphaFold2 predictions do not contribute novel information. On average, there are 0.5 violated crosslinking restraints per prediction at a cutoff of 10 Å Cα–Cα. Indeed, the TM score improvement of AlphaLink generally increases wherever AlphaFold2 makes a prediction containing unsatisfied crosslinks. We further show that the predictions that improved the most have unsatisfied crosslinks with large distances in the baseline prediction (Fig. 4b). Here crosslinks add the most value, and for some predictions a single crosslink is enough to improve the quality considerably (TM score 0.39 to 0.86 for target AtpB). Extended Data Fig. 2d shows two examples where adding crosslinks negatively impacts the prediction quality. In the case of OmpF there are multiple overlength crosslinks (highlighted in red in the native structure) that might stem from crosslinking different subunits, since OmpF is a homo-multimer. For the ATP synthase α subunit there is one overlength crosslink that is probably a false positive. Here, although the link is rejected in the end, it still induces a domain movement that leads to a worse prediction.

To investigate the correlation between predicted and true TM score for the predictions of the membrane fraction, we compute the fit on the predictions where the crystal structure covers at least 80% of the protein (Fig. 4c). The Pearson correlation coefficient is 0.75. We generally underestimate the true TM score. The correlation is in line with the baseline AlphaFold2 model (Extended Data Fig. 3), indicating that model confidence estimates of AlphaLink are comparable to AlphaFold2, allowing for users to reliably interpret predictions.

Extended Data Fig. 4 shows the predicted TM score (pTM) on a total of 96 targets, which include proteins where no structure is available. Each protein was predicted with one randomly subsampled MSA ($N_{eff} = 10$). The pTM indicates possible improvements over AlphaFold2 on these structures as well.

https://doi.org/10.1038/s41587-023-01704-z

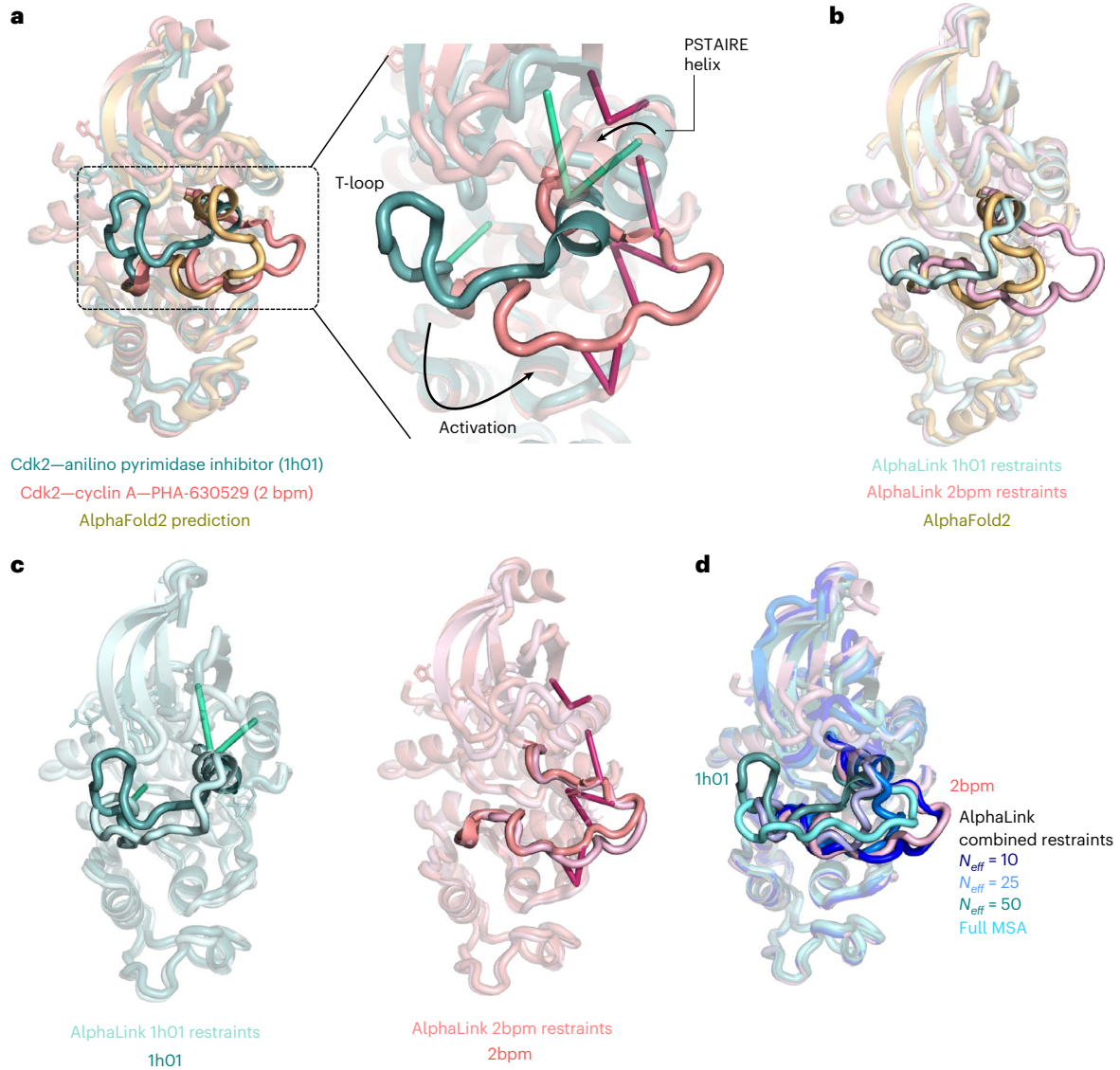


Fig. 5 | Photo-AA data guiding prediction of specific conformational states. **a**, Left: structures of the monomeric, inhibited conformation of Cdk2 (teal)[34] and the cyclin A-activated conformation (salmon)[35] overlaid with the AlphaFold2 prediction of Cdk2 performed at N_{eff} = 10. Right: focus on the T-loop and PSTAIRE helix involved in protein activation, with the two photo-AA restraint sets fed to AlphaLink colored according to the corresponding protein state. **b**, Comparison of the AlphaFold2 prediction with the two predictions of AlphaLink made with restraint sets corresponding to the active or inactive conformation of Cdk2, showing that the photo-AA data drive the prediction to either the active or inhibited conformation. **c**, Middle: overlay of the AlphaLink prediction with the crystal structure for the inhibited state. Right: overlay of the AlphaLink prediction with the structure for the cyclin A-bound state, showing the entire conformation of the loop is correctly predicted despite only sparse restraints being present. **d**, Outcome of predicting with a combined set of restraints. At low N_{eff} values, the crosslinks drive the prediction towards the cyclin E-bound state. As the MSA information increases, the prediction is steered more towards the inhibited state and closer to the AlphaFold2 prediction.

Probing conformational dynamics in situ

To probe whether experimental distance restraints can act as anchors to drive predictions towards different energy minima in multistate proteins, we simulate a proof-of-concept experiment on the human cyclin-dependent protein kinase Cdk2, a drug target in cancer therapy[20]. Activation of Cdk2 in the S phase proceeds via a conformational change in the T-loop (residues 145–165) and the PSTAIRE helix (residues 45–55) triggered by binding of cyclin A[21]. There are several structures of Cdk2 in various states of activation[22,23]. If Cdk2 is predicted without structural templates with AlphaFold2 (N_{eff} = 10), the T-loop is predicted in an intermediate conformation between the apo, auto-inhibited state and the cyclin A-bound conformation (Fig. 5a). Presumably, the intermediate conformation of this loop in the AlphaFold2 prediction is a consequence of co-evolutionary information driving it towards both the open and the closed state. When run with full MSA information, all AlphaFold2 predictions converge to the cyclin A-bound state (Extended Data Fig. 5a), failing to predict the inactive conformation.

We simulate two photo-crosslinking MS experiments in which the protein was acquired in either its inhibited or in its cyclin A-bound states, generating two sets of sparse restraints for the T-loop (Supplementary Table 1). Such experiments may be carried out on the purified protein or in cells before protein purification. We then predict the Cdk2 structure using AlphaLink with these restraints, showing that the loop structure is driven towards the appropriate conformation (Fig. 5b). The crosslinks act as anchor points positioning the whole T-loop in the appropriate configuration for the cyclin A-bound state, with a Cα r.m.s.d. of 1.24 Å on residues 145–165 to PDB 2bpm (Fig. 5c).

The same is true for the inactive state of the loop. In this case however, lack of leucine and lysine residues around T160 in the structure leads to a lack of sufficient restraints to capture the fully closed loop conformation, leading to a slightly higher Cα r.m.s.d. to the target structure (3.19 Å to 1h01), while still outperforming the AlphaFold2 prediction (Cα r.m.s.d. 6.29 Å to 1h01). This higher r.m.s.d. is also consistent with the fact that the T-loop is not rigid in its inhibited, dephosphorylated state, as highlighted by multiple crystal structures and molecular dynamics simulations[24]. AlphaLink successfully folds the short helix within the T-loop (residues 147–153) in the inactive state, and unfolds them into an extended conformation when given restraints for the cyclin A-bound state. It also correctly predicts the position and rotation of the PSTAIRE helix, despite having only two restraints in this region in the inactive conformation dataset, and three in the active dataset. In the case of a mixture of restraints, the prediction converges on the cyclin-A bound state at $N_{eff} = 10$ (Fig. 5d). This conformer is not produced by AlphaFold2. Increasing the MSA steers the prediction towards a middle ground that is more similar to the AlphaFold2 prediction. We interpret this as the algorithm performing noise rejection on a subset of crosslinks in the mixture and using the rest as anchor points to drive a prediction towards a particular solution.

To further show the influence of the MSA, we predict the conformation of the fold-switching protein KaiB (Extended Data Fig. 5b) with photo-L crosslinks simulated for the ground state, the fold-switched state or a mixture added on top of random sets of simulated photo-L crosslinks. At low N_{eff}, AlphaLink predicts both conformers accurately when given unique sets of crosslinks, but as MSA evidence gets larger, the prediction converges to one state for both sets. This result reproduces the outcome of running AlphaFold on KaiB with different, clustered subsamples of the MSA[25]. Predictions with mixed crosslinks lead to different outcomes at different N_{eff} values, as observed in the case of Cdk2, pointing to the fact that crosslinking is weighted against the MSA depending on the information content and size of both strands of information. In multiple simulated crosslinking datasets for the protein selecase (Extended Data Fig. 5b), even without MSAs, most predictions end up in the conformation observed in the monomeric state of the protein state, although some predictions corresponding to the bound state are observed when given unique crosslinks in the absence of MSA information.

These results demonstrate that AlphaLink can be used to obtain high-quality predictions of particular conformations of proteins given sets of restraints obtained under different conditions, enabling direct monitoring of conformational states in solution and in situ.

Discussion

We presented AlphaLink, a method for integrating crosslinking MS restraints derived from photo-AA-labeled cells into AlphaFold2, via OpenFold. Merging photo-crosslinking MS and MSA information in a deep learning framework allows us to leverage their respective strengths and compensate for their weaknesses. Our approach uses the experimental data to bias the retrieval of evolutionary relationships by the Evoformer updating the pair representation. The iterative nature of the AlphaLink architecture leads to noise rejection and robustness to experimental error. Our implementation of experimental restraints also translates to other methods with similar architectures, such as OmegaFold[26], which replace MSAs with protein language models (Extended Data Fig. 6).

The results in this study were achieved by refining the AlphaFold2 model parameters with simulated photo-AA data, as we were not able to fully retrain the OpenFold network to derive model parameters due to computational resource limitations. Nevertheless, the results demonstrate an improvement in prediction quality for challenging targets as a result of incorporating photo-AA restraints. The prediction times increase 1.4× compared with AlphaFold2 (Extended Data Fig. 7).

The information sources have different characteristics that match well. Crosslinking MS provides concrete distance information that can corroborate or refute amino acid associations picked up by co-evolution[7,27,28]. As such, crosslinking MS information has already been used to independently validate models from AlphaFold2 (refs. [17,18]). Moreover, genetic code engineering allows the use of amino acid analogs to substitute for encoded amino acids in protein translation. Thus, leucine positions in the proteome can be replaced to varying extents by photo-L. This amino acid has been linked with the evolutionary development of tertiary folds[29] and is usually found in the hydrophobic cores of proteins. Leucine crosslinking may therefore provide critical information that can guide fold prediction effectively.

In AlphaLink, crosslinks can act as anchors in the prediction itself, since the sparsity of crosslinks is compensated with co-evolutionary information that fills in and extrapolates the missing information. This also enables the software to use co-evolutionary information to perform noise rejection on experimental data. AlphaLink provides a framework for training AlphaFold2-style predictors with a number of data sources providing contacts and/or distance restraints, such as mutagenesis, nuclear magnetic resonance restraints, fluorescence resonance energy transfer and crosslinking MS performed with different crosslinkers. As a test, we fine-tuned the network with simulated sulfo-SDA crosslinks[9] and could successfully predict our test set (Extended Data Fig. 8).

We validate AlphaLink against CASP14/CAMEO targets that were not part of AlphaLink or AlphaFold2 training using synthetic data, and *E. coli* membrane proteins using in-cell photo-L crosslinking MS data. The crosslinking MS data enabled the systematic testing of predictions of 31 proteins against crystal structures with experimental information. While the gains observed on these targets are more modest than with the CASP14 and CAMEO set, these proteins have known structures that were part of the training set of AlphaFold2. This makes them inherently easier for AlphaFold2 to predict, as a result of data leakage. We show that AlphaLink accurately estimates model confidence with various metrics (predicted lDDT-CA score (pLDDT), predicted TM-score (pTM) and predicted aligned error (PAE)), providing the user with valuable information on what conclusions may be drawn from a particular structure prediction, in a manner comparable to AlphaFold2 (Extended Data Figs. 3, 4 and 9). This is a considerable improvement over the performance of crosslinks in CASP13, where crosslinks were included as information in the data-assisted category and led to a decrease in prediction quality[30].

Beyond improving predictions on challenging targets, simulated here by low N_{eff}, AlphaLink opens up the investigation of multiple conformational states by a combination of protein structure prediction and experimental information. This enables the structural characterization of cellular processes in defined biological conditions and may eventually be used to design binders and inhibitors to target specific protein states. Unlike other methods that rely on manipulation of the MSA[25,31,32], AlphaLink uses experimental information to drive the prediction of multiple conformational states. Because the algorithm weighs experimental evidence against evolutionary information, the nature and size of the MSA plays a role in driving the prediction. Thus, a high N_{eff} can 'overpower' experimental evidence. In this regard, subsampling the MSA is a way to tune down the weight of the MSA. In the analyses of KaiB and selecase, AlphaLink can be run with multiple MSA subsamplings or even combined with sequence clustering[25] to characterize the full range of conformations for given combinations of experimental and MSA evidence. Intriguingly, for both KaiB and Cdk2, running AlphaLink with crosslinks from mixtures of conformers led to predictions coinciding with one state at a low N_{eff}, then predictions in between and finally another state at high N_{eff}. In the case of selecase,

sequence clustering did not produce the alternative conformation at all[25], while AlphaLink could produce the alternative conformation in the absence of MSA information.

Taken together, our results show that AlphaLink successfully leverages experimental restraints via deep learning to improve protein structure prediction. We present a workflow based on photo-AA crosslinking MS, which provides contact-like distance information, and obtain the first large-scale photo-AA crosslinking MS dataset inside cells. We then implement photo-AA-based protein structure prediction in AlphaLink. Our method leverages a list of generic contacts, represented as explicit distance restraints or as distograms, to guide the OpenFold pipeline towards structures consistent with the experimental data. The workflow outlined here thus provides a general framework for the hybrid experiment-assisted AI prediction of protein structure, investigating the structure–function relationship of proteins directly in situ without any genetic manipulation.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-023-01704-z.

References

1. Pereira, J. et al. High-accuracy protein structure prediction in CASP14. *Proteins* **89**, 1687–1699 (2021).
2. Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K. & Moult, J. Critical assessment of methods of protein structure prediction (CASP)-Round XIII. *Proteins* **87**, 1011–1020 (2019).
3. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
4. Jumper, J. & Hassabis, D. Protein structure predictions to atomic accuracy with AlphaFold. *Nat. Methods* **19**, 11–12 (2022).
5. Abanades, B., Georges, G., Bujotzek, A. & Deane, C. M. ABlooper: fast accurate antibody CDR loop structure prediction with accuracy estimation. *Bioinformatics* **38**, 1877–1880 (2022).
6. Buel, G. R. & Walters, K. J. Can AlphaFold2 predict the impact of missense mutations on structure? *Nat. Struct. Mol. Biol.* **29**, 1–2 (2022).
7. Graziadei, A. & Rappsilber, J. Leveraging crosslinking mass spectrometry in structural and cell biology. *Structure* **30**, 37–54 (2022).
8. Leitner, A., Faini, M., Stengel, F. & Aebersold, R. Crosslinking and mass spectrometry: an integrated technology to understand the structure and function of molecular machines. *Trends Biochem. Sci.* **41**, 20–32 (2016).
9. 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).
10. Suchanek, M., Radzikowska, A. & Thiele, C. Photo-leucine and photo-methionine allow identification of protein-protein interactions in living cells. *Nat. Methods* **2**, 261–267 (2005).
11. Häupl, B., Ihling, C. H. & Sinz, A. Combining affinity enrichment, cross-linking with photo amino acids, and mass spectrometry for probing protein kinase D2 interactions. *Proteomics* **17**, e1600459 (2017).
12. Kohl, B., Brüderlin, M., Ritz, D., Schmidt, A. & Hiller, S. Protocol for high-yield production of photo-leucine-labeled proteins in *Escherichia coli*. *J. Proteome Res.* **19**, 3100–3108 (2020).
13. Belsom, A. & Rappsilber, J. Anatomy of a crosslinker. *Curr. Opin. Chem. Biol.* **60**, 39–46 (2021).
14. Ziemianowicz, D. S., Ng, D., Schryvers, A. B. & Schriemer, D. C. Photo-cross-linking mass spectrometry and integrative modeling enables rapid screening of antigen interactions involving bacterial transferrin receptors. *J. Proteome Res.* **18**, 934–946 (2019).
15. West, A. V. et al. Labeling preferences of diazirines with protein biomolecules. *J. Am. Chem. Soc.* **143**, 6691–6700 (2021).
16. Agostini, F. et al. Biocatalysis with unnatural amino acids: enzymology meets xenobiology. *Angew. Chem. Int. Ed. Engl.* **56**, 9680–9703 (2017).
17. O'Reilly, F. J. et al. Protein complexes in *Bacillus subtilis* by AI-assisted structural proteomics. Preprint at *bioRxiv* https://doi.org/10.1101/2022.07.26.501605 (2022).
18. Burke, D. F. et al. Towards a structurally resolved human protein interaction network. *Nat. Struct. Mol. Biol.* **30**, 216–225 (2023).
19. Ahdritz, G. et al. OpenFold: retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization. Preprint at *bioRxiv* https://doi.org/10.1101/2022.11.20.517210 (2022).
20. Tadesse, S., Caldon, E. C., Tilley, W. & Wang, S. Cyclin-dependent kinase 2 inhibitors in cancer therapy: an update. *J. Med. Chem.* **62**, 4233–4251 (2019).
21. Stevenson, L. M., Deal, M. S., Hagopian, J. C. & Lew, J. Activation mechanism of CDK2: role of cyclin binding versus phosphorylation. *Biochemistry* **41**, 8528–8534 (2002).
22. De Bondt, H. L. et al. Crystal structure of cyclin-dependent kinase 2. *Nature* **363**, 595–602 (1993).
23. van Montfort, R. L. M., Workman, P., Martin, M. P., Endicott, J. A. & Noble, M. E. M. Structure-based discovery of cyclin-dependent protein kinase inhibitors. *Essays Biochem.* **61**, 439–452 (2017).
24. Barrett, C. P. & Noble, M. E. M. Molecular motions of human cyclin-dependent kinase 2. *J. Biol. Chem.* **280**, 13993–14005 (2005).
25. Wayment-Steele, H. K., Ovchinnikov, S., Colwell, L. & Kern, D. Prediction of multiple conformational states by combining sequence clustering with AlphaFold2. Preprint at *bioRxiv* https://doi.org/10.1101/2022.10.17.512570 (2022).
26. Wu, R. et al. High-resolution de novo structure prediction from primary sequence. Preprint at *bioRxiv* https://doi.org/10.1101/2022.07.21.500999 (2022).
27. Burger, L. & van Nimwegen, E. Disentangling direct from indirect co-evolution of residues in protein alignments. *PLoS Comput. Biol.* **6**, e1000633 (2010).
28. Morcos, F. et al. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl Acad. Sci. USA* **108**, E1293–E1301 (2011).
29. Kubyshkin, V. & Budisa, N. The alanine world model for the development of the amino acid repertoire in protein biosynthesis. *Int. J. Mol. Sci.* **20**, 5507 (2019).
30. Fajardo, J. E. et al. Assessment of chemical-crosslink-assisted protein structure modeling in CASP13. *Proteins* **87**, 1283–1297 (2019).
31. Heo, L. & Feig, M. Multi-state modeling of G-protein coupled receptors at experimental accuracy. *Proteins* **90**, 1873–1885 (2022).
32. Del Alamo, D., Sala, D., Mchaourab, H. S. & Meiler, J. Sampling alternative conformational states of transporters and receptors with AlphaFold2. *eLife* **11**, e75751 (2022).
33. Lenz, S. et al. Reliable identification of protein–protein interactions by crosslinking mass spectrometry. *Nat. Commun.* **12**, 3564 (2021).
34. Beattie, J. F. et al. Cyclin-dependent kinase 4 inhibitors as a treatment for cancer. Part 1: identification and optimisation of substituted 4,6-bis anilino pyrimidines. *Bioorg. Med. Chem. Lett.* **13**, 2955–2960 (2003).
35. Pevarello, P. et al. 3-Aminopyrazole inhibitors of CDK2/cyclin A as antitumor agents. 2. Lead optimization. *J. Med. Chem.* **48**, 2944–2956 (2005).

Methods

Crosslink simulation

We considered several representations for photo-crosslinking MS-derived contacts (Supplementary Table 2), and ultimately decided to train with 10 Å Cα–Cα contacts, since it agrees well with the mean distances observed in experimental data (Fig. 3b), closely resembles previous definitions of contact restraints[36,37] and represents experimentally observed distance distributions more accurately than simulating crosslinks from leucine Cδ1 to the nearest nonhydrogen atom.

We simulate crosslinks by taking all residue pairs where one residue is a leucine or lysine and the other residue has an atom that is within 10 Å of the Cα atom of leucine/lysine. We consider only crosslinks that are not within the same or consecutive tryptic peptides. The links are randomly subsampled to 10% to match the expected coverage of the real data. We further add 5% of stochastic noise to match the expected FDR. During training, we always simulate at least one incorrect crosslink. The FDR can therefore be much higher. The link-level FDR is simulated by shuffling the crosslinks and counting the number of incorrect links observed so far. The minimum FDR is 5%. Crosslink statistics for the CASP14/CAMEO set can be found in Supplementary Tables 3–5.

Distogram sampling

We sample distograms on-the-fly during training to condition the network. There are three different types of distogram: first, a uniformly distributed distogram that represents contact information; second, a distogram based on the expected distances for a specific distance bin; third, a distogram based on the expected distances for photo-L and photo-K crosslinks (10 Å). For all distograms the probabilities sum to 1-FDR below the chosen bin and to FDR beyond the chosen bin.

Integration of crosslinks

We add a crosslink embedding layer to OpenFold by specifying an additional linear layer that maps the soft-label contact map, or in case of the distogram network, the distograms into the 128-dimensional *z*-space present in the AlphaFold2/OpenFold architecture. The projection is added to the pair representation (*z*). In addition, we learn a group embedding layer to indicate groups of possibly ambiguous crosslinks, which enables us to deal with restraints with positional ambiguity. The group embedding is also added to the pair representation.

MSA subsampling

For refining the network, we subsample the MSAs to a specific N_{eff}. The N_{eff} corresponds to the number of nonredundant sequences in a MSA below a specific sequence identity. We set the sequence identity to 80%. We subsample the MSAs in training in each epoch to a random N_{eff} between 1 and 25, generating a uniform distribution of N_{eff} values across all targets. In subsampling, the MSA is shuffled and sequences are added until the desired N_{eff} is reached.

Conformer experiment

For KaiB (Q79V61) and selecase (Q58610) we produced a total of 200 predictions that contained 100 simulated crosslinks (FDR 5%) and a set of 4–7 crosslinks unique to each conformation. The mixture crosslinking set is subsampled from both sets to ensure similar coverage. The unique crosslinks for each conformation are added on top.

Fine-tuning of AlphaFold2

To avoid training OpenFold from scratch, we start with the AlphaFold2 2.0 (https://github.com/deepmind/alphafold/releases/tag/v2.0.0) weights provided by Deepmind and refine the network on 13,000 proteins from the trRosetta[38] training set with simulated photo-AA crosslinking data. We use OpenFold v0.1.0 (based on GitHub from January 2022: commit 894905b9da941ed10e797c5ba15af75692cee1b4). To encourage the network to use the crosslinking data, we subsample the MSA to a N_{eff} between 1 and 25 (uniformly). MSAs were generated with the reduced database setting. We train and test with model_5_ptm, which does not use any template information. Fine-tuning specifically on low N_{eff} targets does not substantially change the performance of AlphaLink. We predicted the same structures without crosslinks in AlphaLink to verify that fine-tuning the network on low N_{eff} targets is not the reason for our improvements (Extended Data Fig. 10).

We follow the refinement training regime outlined in the AlphaFold2 paper, except that, due to memory constraints, we do not expand the MSA cluster size. Since our method is specifically targeting proteins with few MSAs, this is not a problem. We train for five epochs on five GPUs, which takes roughly 5 days.

Evaluation set up

For the baseline, we use OpenFold with the original AlphaFold2 weights provided by Deepmind. The creators of OpenFold verified that the implementation produces identical results. To assess AlphaLink and Openfold performance, we perform predictions with the model_5_ptm setting, which does not include templates and predicts the TM score as an auxiliary loss.

To ensure comparability, we make predictions deterministic. We disable masking out parts of the MSA input. Especially on targets with small MSAs, masking out parts of the input leads to big variances between runs, since it affects the N_{eff}. Here reconstruction can increase the N_{eff}. We use a fixed set of ten subsampled MSAs and a fixed set of ten subsampled crosslinks. Normally, MSAs will be subsampled on-the-fly to 128 sequences. The rest is aggregated with the ExtraMSAStack. We cap our MSA size at 128 for the subsampled MSAs to remove variance. Since we mostly evaluate on MSAs with $N_{\text{eff}} = 10$, where the MSA size is far below 128, we seldom reach this limit in practice.

If not denoted otherwise, the results we will show use the soft-label representation, which is trained for the particular crosslinker type and performs slightly better.

Our main comparison metric is the template modeling score (TM score), which measures the similarity between two protein structures. The TM score is calibrated in a way that structures with a TM score above 0.5 generally assume the same fold. A TM score of 1.0 signifies a perfect match. TM scores <0.2 correspond to random structures.

We use the same databases as Deepmind to recreate the CASP14 settings: UniRef90: v2020_01, MGnify: v2018_12, Uniclust30: v2018_08, BFD: only version available, PDB: downloaded 14 May 2020, PDB70: 13 May 2020. The MSAs are generated with the reduced database setting. We evaluate the CASP14 targets on the full sequence, not splitting them into domains. We evaluate on 45 CAMEO targets that were released after AlphaFold2. We consider only CAMEO targets where AlphaFold2 does not exceed a TM score of 0.8. We used SMART with pFam annotation[39] to divide the CAMEO targets into single/multidomain, ignoring low-complexity regions.

For predictions of conformational states of Cdk2, potential photo-L and photo-K crosslink sites were derived from structures of inhibited (1h01) (ref. [34]) and cyclin-A bound states (2bpm) (ref. [35]). Separate AlphaLink predictions were submitted with either dataset at $N_{\text{eff}} = 10$. AlphaFold2 predictions were performed at $N_{\text{eff}} = 10$ with the model_5_ptm setting. For Extended Data Fig. 5a, AlphaFold2 predictions were carried out with full MSA size and default model settings (five random seeds per model, five models predicted).

Photo-L crosslinking of *E. coli* cells

For optimization of photo-L concentration in the medium, *E. coli* K12 were grown in LB medium overnight at 37 °C. The cultures were diluted (1:100) into M9 minimal medium containing 0.2% glucose and varying concentrations of photo-L (0, 5, 25, 250, 500, 750, 1,000 and 2,000 µM photo-L) in 96-well plates in a Microplate reader Infinite M200 Tecan. Three colonies were used per condition. Cell growth was monitored via OD_{600}.

For crosslinking MS experiments, *E. coli* were grown with 0.75 mM photo-L for 22 h at 37 °C in 100 ml of M9 minimal medium. Thirty million cells were then pelleted for 15 min at 4,000*g*. The pellet was resuspended in fresh M9 minimal medium to a concentration of 1 million cells ml^{-1}. UV crosslinking was then performed by exposing the cell suspension for 20 min on ice in a CL-1000L crosslinking device (UVP). Cells were then pelleted again, washed with PBS buffer and snap frozen.

Membrane enrichment

Cells were resuspended in 20 mM Tris, pH 7.4 and lysed by four freeze–thaw cycles followed by sonication. Larger cell debris was removed through centrifugation at 2,000*g* for 20 min. The supernatant was then cleared centrifuged at 16,000*g* for 20 min at 4 °C. The pellet was washed with 20 mM Tris pH 7.6, 1 M NaCl.

Proteome digestion and peptide fractionation

The pellets were solubilized in PBS buffer and subsequently mixed with NuPage LDS sample buffer (Life Technologies) and run into a 4–12% Bis-Tris SDS–PAGE gel (Life Technologies). Gels were stained using Imperial Protein Stain (Thermo Scientific), and the whole proteome was cut out and prepared for in-gel digestion. Proteins were reduced with 10 mM dithiothreitol (Sigma Aldrich) for 30 min at 37 °C, followed by alkylation with 55 mM iodoacetamide (Sigma Aldrich) for 20 min at room temperature in the dark. Gel pieces were incubated with 13 ng μl^{-1} trypsin (Pierce, Thermo Fisher Scientific) at 37 °C for 16 h in 10 mM ammonium bicarbonate, 10% acetonitrile (ACN). The samples were cleaned up using Sep-Pak C18 cartridges (Waters) before strong cation exchange chromatography.

The peptides were resuspended in SCX loading buffer (10 mM KH_2PO_4 and 30% ACN) and separated on a polysulfoethyl A column (PolyLC, PolySulfoethyl A 100 × 2.1 mm^2, 3 µm) using SCX elution buffer (10 mM KH_2PO_4, 30% ACN and 1 M KCl). Separation of peptides was accomplished using a nonlinear gradient with running buffer B (30% ACN, 1 M KCl and 10 mM KH_2PO_4, pH 3.0), as described[40]. Fractions of 200 µl each were collected over the elution window (approximately 18 column volumes). Collected fractions of interest from five runs were pooled, desalted using Stage-Tips and stored at −20 °C.

Crosslinked peptides in each SCX fraction (labeled fractions 16–22 in the JPOST deposition) were subsequently enriched by size-exclusion chromatography using a Superdex Peptide 3.2/300 column (GE Healthcare). The mobile phase consisted of 30% (v/v) ACN and 0.1% trifluoroacetic acid, running at a flow rate of 10 µl min^{-1}. The earliest five peptide-containing fractions (50 µl each, labeled SEC6–10) were collected and dried in a vacuum concentrator. Whenever amounts were insufficient for liquid chromatography (LC)–MS analysis, adjacent fractions were pooled.

LC–MS acquisition of photo-L crosslinked *E. coli* membranes

Acquisition of crosslinked peptide spectra was performed on a Fusion Lumos Tribrid Mass Spectrometer (Thermo Fisher Scientific) connected to an Ultimate 3000 UHPLC system (Dionex) operating with XCalibur 4.4 and Tune 3.4. Chromatography was performed with mobile 0.1% (v/v) formic acid as mobile phase A, and 80% (v/v) ACN, 0.1% (v/v) formic acid as mobile phase B. The samples were dissolved in 1.6% ACN (Honeywell Fluka), 0.1% formic acid (Honeywell Fluka) and separated on an Easy-Spray column (C-18, 50 cm, 75 µm internal diameter, 2 µm particle size, 100 Å pore size) running with 300 nl min^{-1} flow rate using optimized gradients for each offline fraction (ranging from 2% B to 55% B over 62.5, 92.5 or 152.5 min, then to 55% in 2.5 min and to 95% in 2.5 min).

The MS data were acquired in data-dependent mode using the top-speed setting with a 3 s cycle time. For every cycle, the full-scan mass spectrum was recorded in the Orbitrap at a resolution of 120,000 in the range of 400 to 1,450 m/z. Ions with a precursor charge state between +3 and +7 were isolated and fragmented with a decision tree strategy[41]. Higher-energy collisional dissociation energies optimized for mass and charge of a precursor species were applied[41]. The fragmentation spectra were then recorded in the Orbitrap with a resolution of 50,000. Dynamic exclusion was enabled with single repeat count and 60 s exclusion duration.

Identification and statistical validation of crosslinked peptides

LC-MS raw data were searched against the *E. coli* K12 proteome (download from UniProt February 2020) using MaxQuant 1.6.17 (ref. [42]) (Supplementary Data 3). The top 1,200 proteins by iBAQ were used as the database for the crosslink search. For the crosslink search, raw data were processed using MSconvert 3.0.22 (ref. [43]) to recalibrate precursor masses and convert to mgf format. An open modification search with MSfragger 3.4 (ref. [44]) was used to quantify modifications in the sample. The peak files were then searched with xiSEARCH 1.7.6.4 (ref. [45]) with the following settings: MS1 tolerance: 3 ppm; MS2 tolerance: 5 ppm, allowing up to two missing monoisotopic peaks and three missed tryptic cleavages. Cysteine carbamidomethylation was defined as a fixed modification. Oxidation of methionine, deamidation of asparagine and methylation of glutamic acid were defined as variable modifications. $-CH_3SOH/-H_2O/-NH_3$ were defined as losses. In the crosslink search, the photo-L crosslinker was defined as follows: the linkage mass was set to −16.0313 Da and the specificity set to leucine to any amino acid. Variable modifications on leucine to account for photo-L reactions with water (1.97926 Da) or within a peptide (−16.0313) were also defined. Noncovalent gas-phase associations were included in the search[46].

The spectral matches were filtered before FDR estimation to crosslinked peptide matches with a minimum of three fragments matched per peptide. Results were then filtered to 5% crosslink-level FDR in xiFDR 2.1.5.5 (ref. [45]) with the boosting feature for error thresholding at lower levels enabled. The minimum peptide length was set to 6. Consecutive peptides were removed from the analysis. The resulting residue pairs are in Supplementary Data 2.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Crosslinking MS data are deposited in ProteomeXChange JPOST[47] with accession code JPST001851. Models and corresponding MSA and simulated crosslinking data have been deposited on ModelArchive[48] with accession code ma-rap-alink. AlphaLink models based on experimental crosslinks have been deposited as integrative/hybrid models in PDB-Dev[49] with accession codes PDBDEV_00000165 to PDBDEV_00000198 (group ID PDBDEV_G_1000001). Source data are provided with this paper.

Code availability

The code for AlphaLink is deposited at https://github.com/lhatsk/AlphaLink.

References

36. Di Lena, P., Nagata, K. & Baldi, P. Deep architectures for protein contact map prediction. *Bioinformatics* **28**, 2449–2457 (2012).
37. Wang, S., Li, W., Zhang, R., Liu, S. & Xu, J. CoinFold: a web server for protein contact prediction and contact-assisted protein folding. *Nucleic Acids Res.* **44**, W361–W366 (2016).
38. Du, Z. et al. The trRosetta server for fast and accurate protein structure prediction. *Nat. Protoc.* **16**, 5634–5651 (2021).
39. Letunic, I. & Bork, P. 20 years of the SMART protein domain annotation resource. *Nucleic Acids Res.* **46**, D493–D496 (2018).

40. Chen, Z. A. et al. Architecture of the RNA polymerase II-TFIIF complex revealed by cross-linking and mass spectrometry. *EMBO J.* **29**, 717–726 (2010).
41. Kolbowski, L., Mendes, M. L. & Rappsilber, J. Optimizing the parameters governing the fragmentation of cross-linked peptides in a tribrid mass spectrometer. *Anal. Chem.* **89**, 5311–5318 (2017).
42. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).
43. Chambers, M. C. et al. A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* **30**, 918–920 (2012).
44. Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D. & Nesvizhskii, A. I. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat. Methods* **14**, 513–520 (2017).
45. Mendes, M. L. et al. An integrated workflow for crosslinking mass spectrometry. *Mol. Syst. Biol.* **15**, e8994 (2019).
46. Giese, S. H., Belsom, A., Sinn, L., Fischer, L. & Rappsilber, J. Noncovalently associated peptides observed during liquid chromatography–mass spectrometry and their effect on cross-link analyses. *Anal. Chem.* **91**, 2678–2685 (2019).
47. Okuda, S. et al. jPOSTrepo: an international standard data repository for proteomes. *Nucleic Acids Res.* **45**, D1107–D1111 (2017).
48. Schwede, T. et al. Outcome of a workshop on applications of protein models in biomedical research. *Structure* **17**, 151–159 (2009).
49. 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).

Acknowledgements

We thank F. O'Reilly for support in acquiring the experimental data and F. Schildhauer for growth curves of *E. coli* in the presence and absence of photo-L. We are grateful to the OpenFold team for providing a fully open-source and trainable reimplementation of AlphaFold2. OpenFold may be accessed at https://github.com/aqlaboratory/openfold. This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—EXC 2008—390540038 (J.R.)—UniSysCat and project 329673113 (J.R.) and under Germany's Excellence Strategy—EXC 2002/1 'Science of Intelligence'—project number 390523135 (O.B.). This research was funded, in part, by the Wellcome Trust (grant number 203149) (J.R.). The Wellcome Centre for Cell Biology is supported by core funding from the Wellcome Trust (203149) (J.R.). 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: K.S. Data analysis: K.S. and A.G. Crosslinking MS: T.D., A.G. and J.R. Supervision: J.R. and O.B. Writing—initial draft: A.G., K.S., J.R. and O.B. Writing—editing and revision: A.G., K.S., T.D., J.R. and O.B.

Competing interests

The authors declare no competing interests.

Additional information

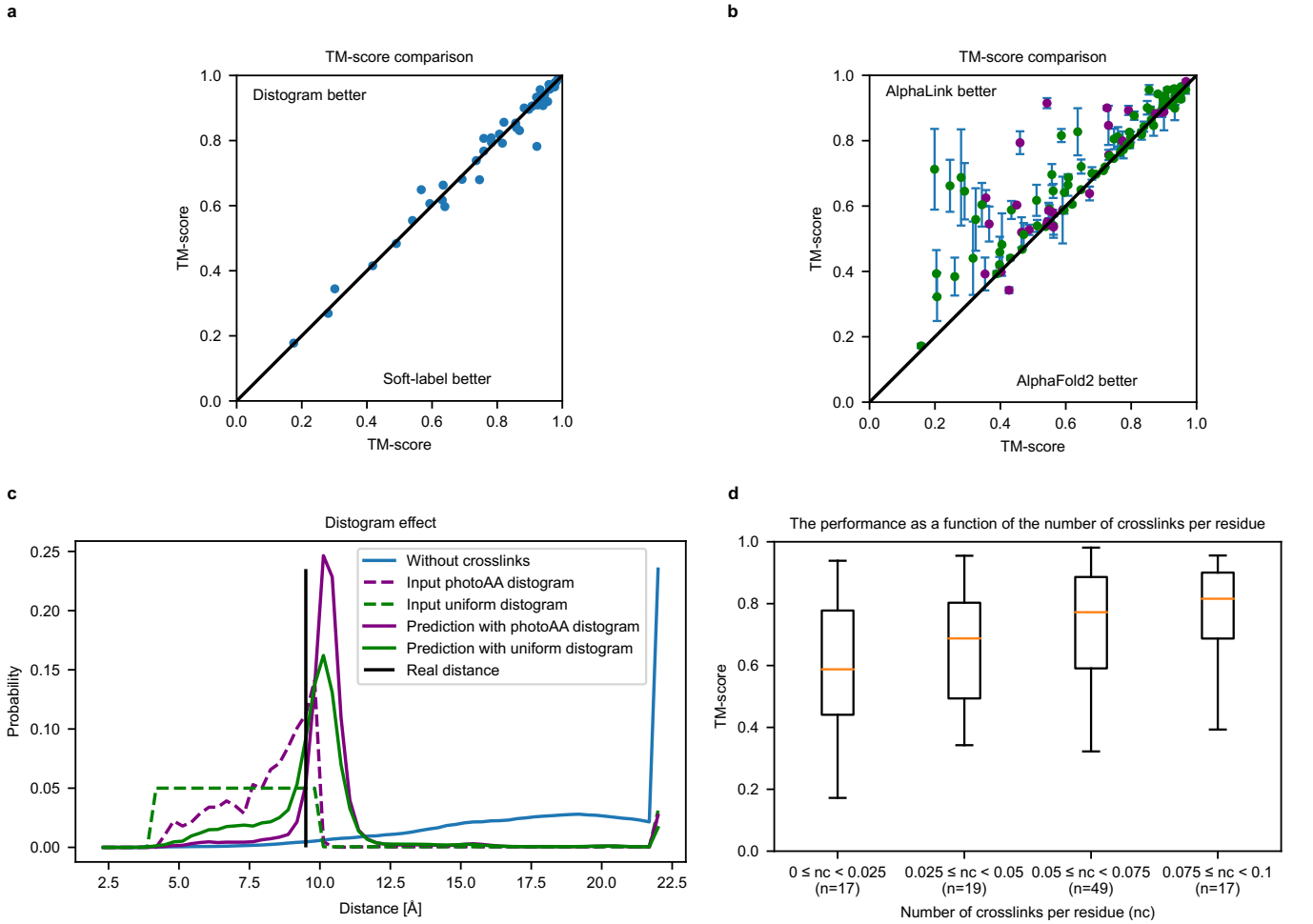
Extended data is available for this paper at https://doi.org/10.1038/s41587-023-01704-z.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-023-01704-z.

Correspondence and requests for materials should be addressed to Oliver Brock or Juri Rappsilber.

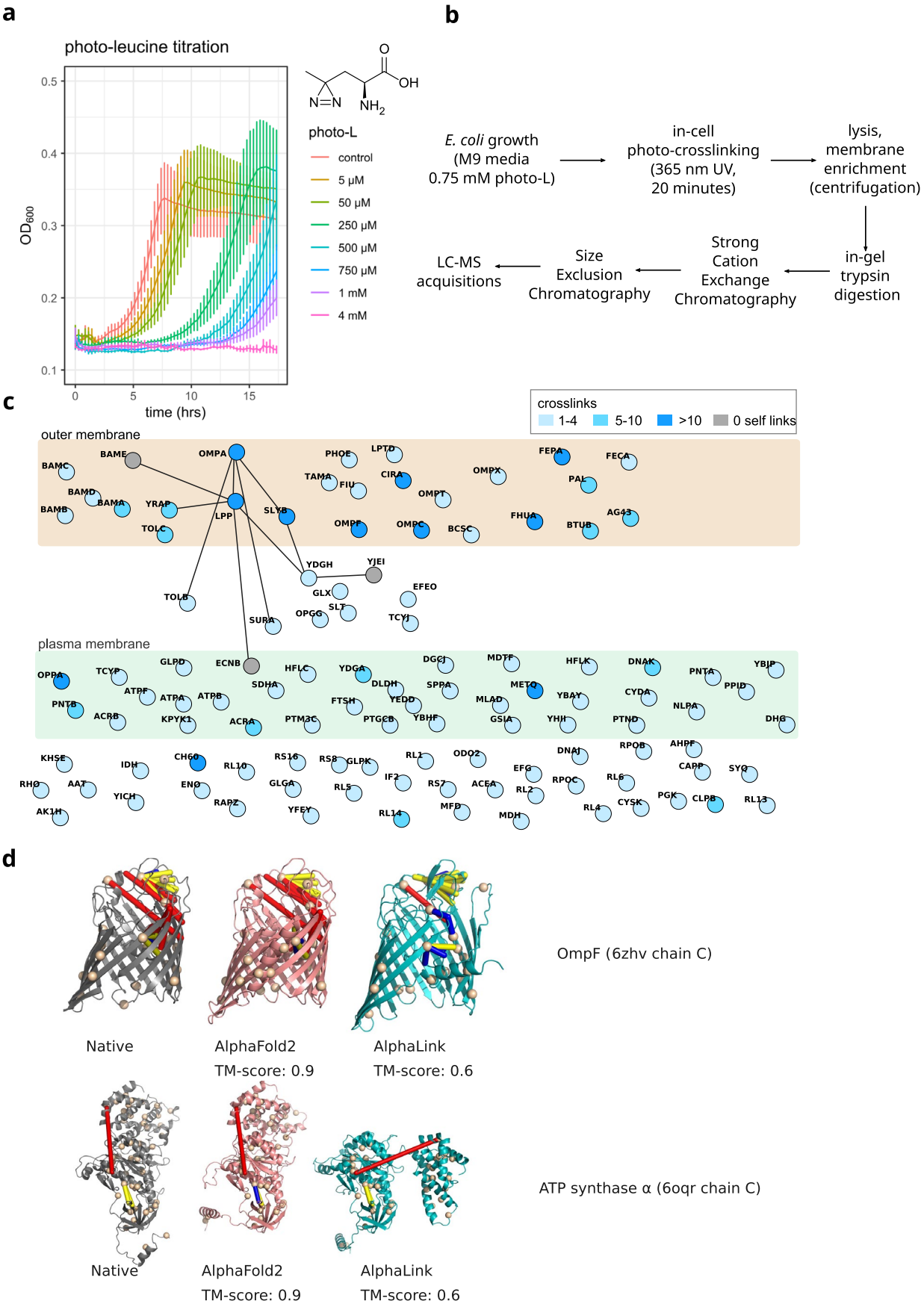
Peer review information *Nature Biotechnology* thanks Alexander Leitner and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

https://doi.org/10.1038/s41587-023-01704-z



Extended Data Fig. 1 | Performance on benchmark data set and distogram effect. a- TM-score comparison ($N = 10$) 49 challenging CAMEO targets for the network trained with the soft-label representation and the network trained with the distogram representation. Each target was predicted with 10 randomly subsampled crosslink sets. Scatter points show the mean. The photo-L crosslinks were represented as a uniformly distributed distogram. The performance is on par. b- Performance on 60 CASP14 and 45 CAMEO ($N_{eff} = 10$). AlphaLink improves the TM-score on average by 15.2%. The error bars represent the 95% confidence interval ($N = 10$). Points show the mean. Purple highlights multi-domain targets, green highlights single-domain targets. c- The effect two different distogram inputs (dashed) have on the distogram AlphaLink predicts (solid) between residues 11 and 103 for T0164. The solid black line designates the real distance. Blue shows the distogram without using crosslinks. The green distogram mimics an upper bound contact restraint, while the purple distogram is the expected distance distribution for simulated photo-L and photo-K crosslinks on the training set. Predicted distogram based on the expected photoAA distance distribution (purple, solid) is more narrow compared to the prediction with the uniform distance distribution (green, solid). Using a distance distribution improves the prediction from TM-score of 0.68 to 0.7. We show the first 64 bins of the input distograms and sum up the probabilities for the rest. In the absence of this restraint, the prediction has a TM-score of 0.28 ($N_{eff} = 10$). d- Performance of AlphaLink as a function of the number of crosslinks per residue ($N_{eff} = 10$). 2 bins (nc > 0.1) were omitted because they only contained 1 and 2 samples. Performance generally increases with more crosslinks per residue. The line shows the median and the whiskers represent the 1.5x interquartile range.

https://doi.org/10.1038/s41587-023-01704-z

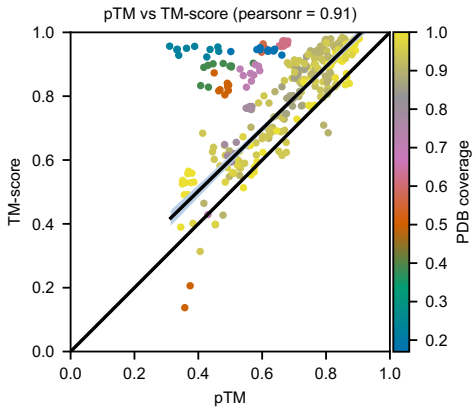


Extended Data Fig. 2 | See next page for caption.

https://doi.org/10.1038/s41587-023-01704-z

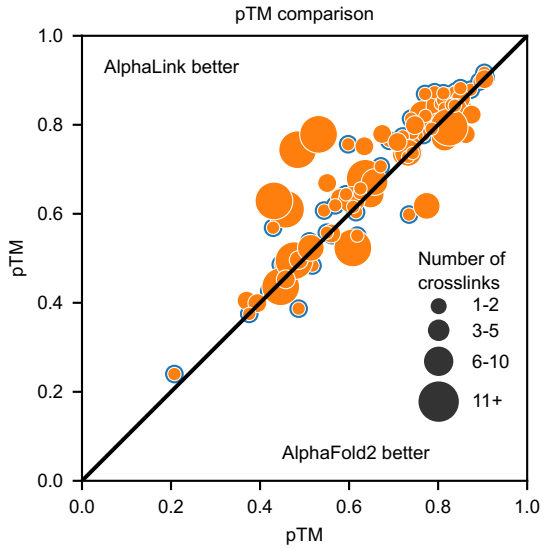
Extended Data Fig. 2 | Photo amino acid crosslinking MS workflow and membrane outliers. a- E. coli K12 grown with increasing concentration of photo-L in M9 minimal media and structure of photo-L. Average cell density with standard error across 3 individual colonies is plotted. b- Workflow for two-dimensional fractionation of peptides from E. coli crosslinked with photo-L in cell. Error bars representing standard error. c- Nodes represent proteins detected with at least one crosslink, and edges indicate the presence of at least a crosslinked residue pair between proteins. Nodes are coloured according to the number of crosslinks detected in that protein. d- Two examples where crosslinks negatively impact the performance in AlphaLink. The crosslink sets contain overlength links (shown in red in the native structure) which cause movements in AlphaLink (for example, domain movement in ATP synthase - likely a false positive link). OmpF is homo-multimeric, the overlength links might stem from links between different subunits.

https://doi.org/10.1038/s41587-023-01704-z



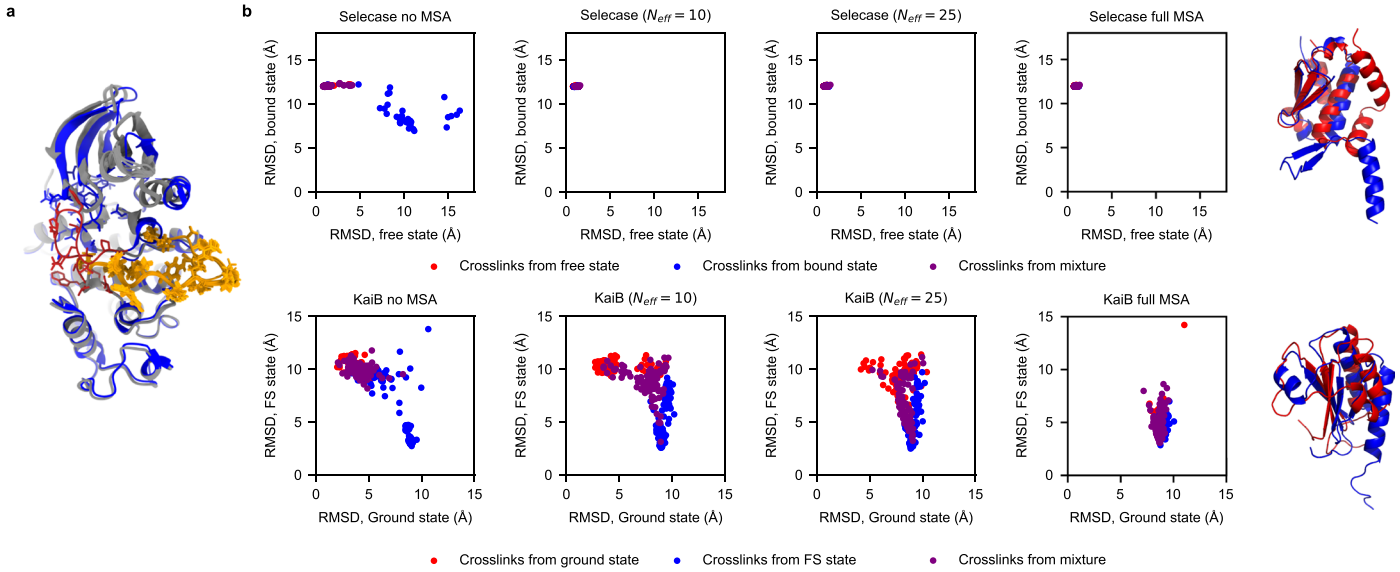
Extended Data Fig. 3 | AlphaFold2 pTM correlation. Calibration of the predicted TM-score (pTM) on $N = 320$ predictions of the E. coli membrane fraction dataset. On predictions that are at least 80% covered by the crystal structure, the correlation is 0.91. The true TM-score is generally underestimated, meaning that the pTM-score of AlphaFold2 is a conservative estimate. The shaded area corresponds to the 95% confidence interval. Line shows the linear fit.

https://doi.org/10.1038/s41587-023-01704-z



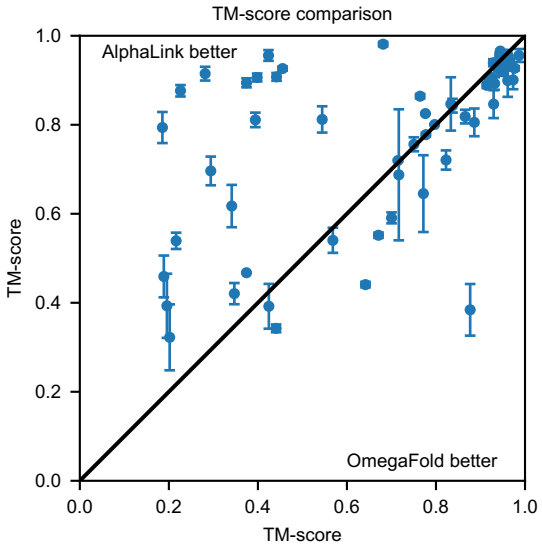
Extended Data Fig. 4 | pTM comparison on 96 proteins from the E. coli membrane fraction dataset. pTM comparison on $N = 96$ proteins from the E. coli membrane fraction dataset. Each points is one protein with one MSA subsample ($N_{eff} = 10$). The higher pTM of AlphaLink indicates improved structures.

https://doi.org/10.1038/s41587-023-01704-z



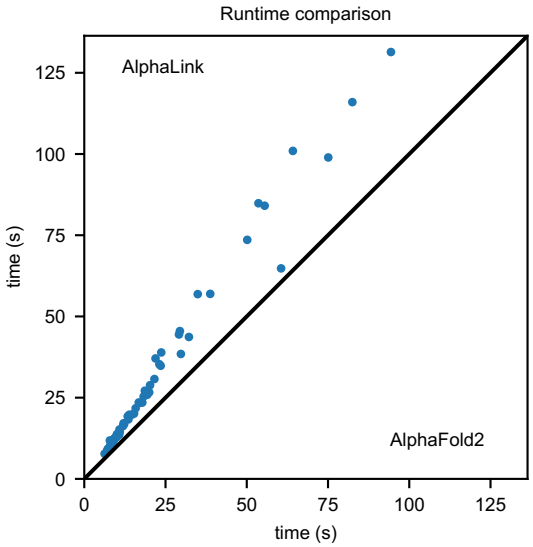
Extended Data Fig. 5 | Predicting multiple conformations. a- Cdk2 predictions by AlphaFold2.2 with default settings (full MSA, 5 models predicted, 5 random seeds per model), in gray, overlaid with the structure of the inhibited conformation of Cdk2 (blue, pdb 1h01). The T-loop is highlighted in orange in the AlphaFold2 predictions and in red in the inhibited conformation structure. All 25 AlphaFold2 predictions converge on the cyclin A-bound conformation of the T-loop and PSTAIRE helix of Cdk2. b- Each points is one prediction with a randomly sampled crosslink (FDR = 5%) set which includes additional links from the highlighted conformation. 100 samples per conformation. For Selecase, AlphaLink always predicts the free state with high accuracy if we use MSAs. Without MSAs, some predictions are driven towards the bound state. For KaiB, AlphaLink is able to predict the ground state with good precision without MSAs. Introducing co-evolutionary information leads to better clustering, although some of the ground state predictions now end up in the average state. This movement proceeds with more co-evolutionary information. With full MSAs, almost all predictions have moved to the fold-switched state.

https://doi.org/10.1038/s41587-023-01704-z



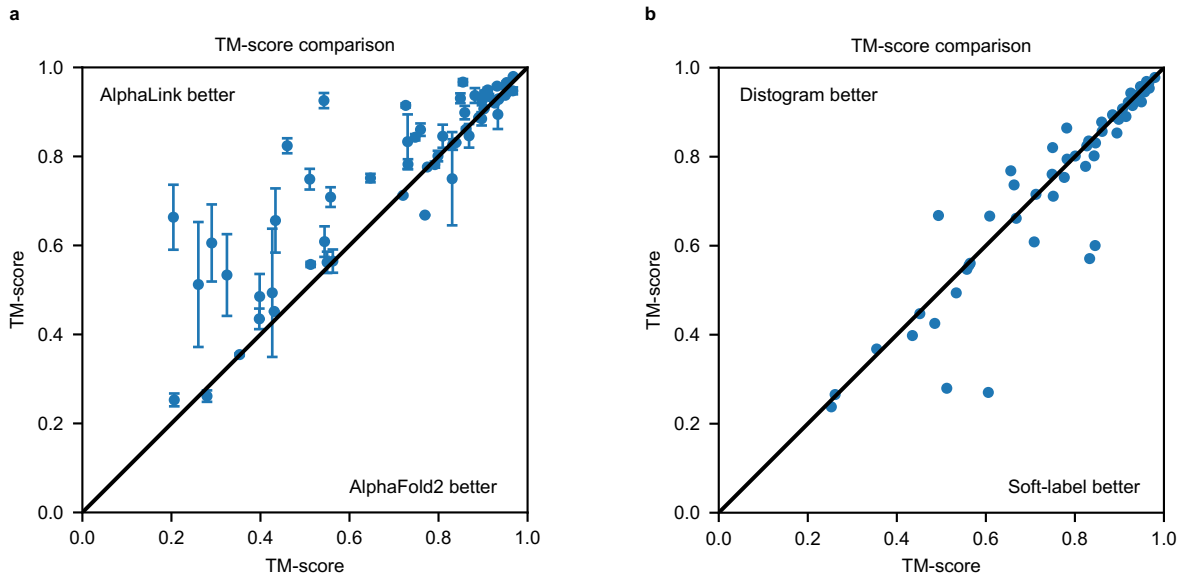
Extended Data Fig. 6 | AlphaLink vs OmegaFold. AlphaLink vs OmegaFold performance on 58 CASP14 targets with $N_{eff} = 10$ (test set) with simulated photo-L crosslinks. AlphaLink improves the TM-score on average by 36.7% ± 76.6. AlphaLink improves 10 additional targets past a TM-score > 0.5. Shown here is the mean and 95% confidence interval ($N = 10$) for AlphaLink and the corresponding performance of OmegaFold.

https://doi.org/10.1038/s41587-023-01704-z



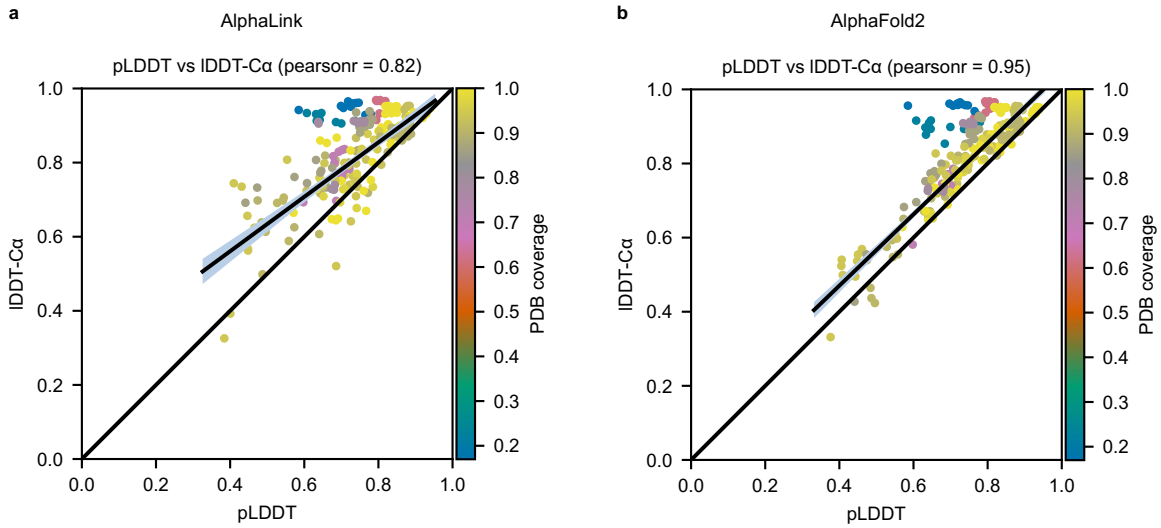
Extended Data Fig. 7 | AlphaLink vs AlphaFold2 timings. Tested on $N = 60$ CASP14 targets with $N_{eff} = 10$. We ran inference, including relaxation on pre-computed features with model_5_ptm. Timed on a node with a single Nvidia A100 80GB GPU and 2 Intel XEON Gold 5118 CPUs (2 × 24 cores) with 2.3 GHz. Mean running time (s) for AlphaLink is 31.7 ± 7.4 s and 22.14 ± 5.16 s for AlphaFold2.

https://doi.org/10.1038/s41587-023-01704-z



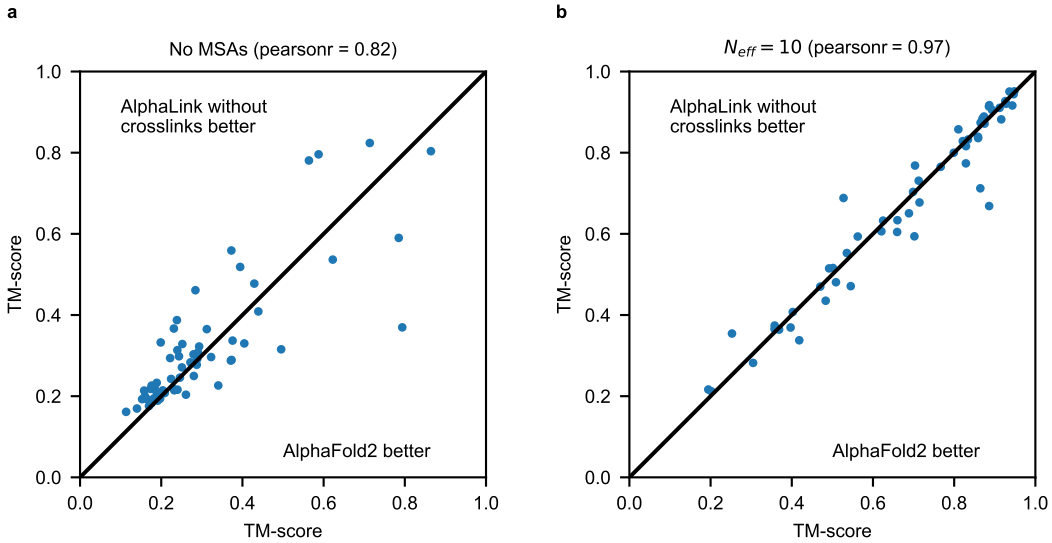
Extended Data Fig. 8 | AlphaLink with soluble crosslinker on CASP14. a- AlphaLink vs AlphaFold2 performance on 59 CASP14 targets with $N_{eff} = 10$ (test set) with simulated SDA crosslinks. AlphaLink improves the TM-score on average by 16.4% ± 9.7. AlphaLink predicts 5 additional targets with a TM-score > 0.5. Shown here is the mean and 95% confidence interval ($N = 10$) for AlphaLink and the corresponding performance of AlphaFold2. b- We compare the mean TM-score ($N = 10$) per target on 59 CASP14 targets for the network trained with the soft-label representation vs the network trained with the distogram representation. Each target was predicted with a single MSA subsample ($N_{eff} = 10$) and 10 randomly subsampled crosslink sets. The sulfo-SDA crosslinks were represented as a uniformly distributed distogram. The soft-label representation outperforms the distogram representation on average by 5%.

https://doi.org/10.1038/s41587-023-01704-z



Extended Data Fig. 9 | AlphaLink and AlphaFold2 pLDDT correlation. a- We show the calibration of the predicted lDDT-Cα score (pLDDT) for AlphaLink on *N* = 220 predictions of the E. coli membrane fraction dataset. On predictions that are at least 80% covered by the crystal structure, the correlation is 0.82. The true lDDT-Cα score is generally underestimated, meaning that the pLDDT-score is a conservative estimate. b- We show the calibration of the predicted lDDT-Cα score (pLDDT) for AlphaFold2 on N = 220 predictions of the E. coli membrane fraction dataset. On predictions that are at least 80% covered by the crystal structure, the correlation is 0.95.The shaded area corresponds to the 95% confidence interval. Line shows the linear fit.

https://doi.org/10.1038/s41587-023-01704-z



Extended Data Fig. 10 | Refining on low N_{eff} targets doesn't change results substantially. a- Performance improvements we observed are due to adding crosslinking information, not additional fine-tuning of the AlphaFold2 weights on low N_{eff} targets. There are few outliers on both sides. The performance is virtually identical for targets without MSAs (TM-score average: AlphaLink = 0.322, AlphaFold2 = 0.308, Z-statistic = −0.072). b- Performance improvements we observed are due to adding crosslinking information, not additional fine-tuning of the AlphaFold2 weights on low Neff targets. There are few outliers on both sides. The performance is virtually identical for N_{eff} = 10 (TM-score average: AlphaLink = 0.701, AlphaFold2 = 0.702, Z-statistic = 0.033). Points show the mean (N = 10) over 10 MSA subsamples.

nature portfolio

Corresponding author(s): Juri Rappsilber

Last updated by author(s): Jan 27, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Thermo XCalibur 4.4 and Tune 3.4 on Orbitrap Fusion Lumos instrument
Data analysis	MaxQuant 1.6.17, MSconvert MSfragger 3.4, ProteoWizard MSConvert 3.0.22, xiSEARCH 1.7.6.4, xiFDR 2.1.5.5, AlphaFold v2.1 AlphaFold v2.2, OpenFold v0.1.0, AlphaLink (https://github.com/lhatsk/AlphaLink, described in this manuscript and included in deposition)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Crosslinking Mass spectrometry data is deposited in ProteomeXChange JPOST42 with accession code JPST001851 (reviewer link
Crosslinking Mass spectrometry data is deposited in ProteomeXChange JPOST42 with accession

code JPST001851 (reviewer link
https://repository.jpostdb.org/preview/14305031816331a59080810 Access key 2354).
Models are deposited in ModelArchive (accession ma-rap-alink)
Integrative models from AlphaLink with experimental crosslinks are deposited in PDB-Dev (awaiting accession number)

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	A single data set of crosslinks from E. Coli membranes was used due to obtain crosslinks to evaluate AlphaFold and AlphaLink's performance against experimental data. This results in a sample size of 474 crosslinks at a 5% false discovery rate that could be mapped against 96 experimental structures. In total, 615 crosslinks were detected at a 5% false discovery rate. Due to the extensive acquisition time required for these experiments and their non-quantitative nature (crosslinks are only identified and not quantified), in-cell, large-scale crosslinking MS experiments are not easily repeatable. Benchmark sets for figures 2 and 4 were selected from CASP14 and CAMEO from 2020 to include all protein targets that were part of modeling evaluations but not part of AlphaFold's training. This resulted in a sample size of 154 protein targets.
Data exclusions	no data was excluded
Replication	Neuronal network replication is described in detail in the methods section. Simulated crosslinks were subsampled 10-times to obtain error bars for predicting structures based on simulated data (Figure 2). Experimental data in Figure 3 is analyzed without replication as the data comes from a single E. Coli membrane fraction. Multiple sequence alignments in figure 4 were subsampled 10 times and all replicas are shown.
Randomization	Crosslinks and multiple sequence alignments were subsampled randomly and the sampling repeated as described whenever comparing the performance of AlphaLink and AlphaFold2 (Fig.2,4 and 5). No randomization is applied when analysing exprimental mass spectrometry data (Fig. 3).
Blinding	The analysis of crosslinking mass spectrometry involved "blinding" in determining false discovery rates by supplying the program with decoy sequences to determine the false positive rate of database-spectral matching.

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	*Briefly describe the study type including whether data are quantitative, qualitative, or mixed-methods (e.g. qualitative cross-sectional, quantitative experimental, mixed-methods case study).*
Research sample	*State the research sample (e.g. Harvard university undergraduates, villagers in rural India) and provide relevant demographic information (e.g. age, sex) and indicate whether the sample is representative. Provide a rationale for the study sample chosen. For studies involving existing datasets, please describe the dataset and source.*

Sampling strategy	*Describe the sampling procedure (e.g. random, snowball, stratified, convenience). Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient. For qualitative data, please indicate whether data saturation was considered, and what criteria were used to decide that no further sampling was needed.*
Data collection	*Provide details about the data collection procedure, including the instruments or devices used to record the data (e.g. pen and paper, computer, eye tracker, video or audio equipment) whether anyone was present besides the participant(s) and the researcher, and whether the researcher was blind to experimental condition and/or the study hypothesis during data collection.*
Timing	*Indicate the start and stop dates of data collection. If there is a gap between collection periods, state the dates for each sample cohort.*
Data exclusions	*If no data were excluded from the analyses, state so OR if data were excluded, provide the exact number of exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.*
Non-participation	*State how many participants dropped out/declined participation and the reason(s) given OR provide response rate OR state that no participants dropped out/declined participation.*
Randomization	*If participants were not allocated into experimental groups, state so OR describe how participants were allocated to groups, and if allocation was not random, describe how covariates were controlled.*

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	*Briefly describe the study. For quantitative data include treatment factors and interactions, design structure (e.g. factorial, nested, hierarchical), nature and number of experimental units and replicates.*
Research sample	*Describe the research sample (e.g. a group of tagged Passer domesticus, all Stenocereus thurberi within Organ Pipe Cactus National Monument), and provide a rationale for the sample choice. When relevant, describe the organism taxa, source, sex, age range and any manipulations. State what population the sample is meant to represent when applicable. For studies involving existing datasets, describe the data and its source.*
Sampling strategy	*Note the sampling procedure. Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient.*
Data collection	*Describe the data collection procedure, including who recorded the data and how.*
Timing and spatial scale	*Indicate the start and stop dates of data collection, noting the frequency and periodicity of sampling and providing a rationale for these choices. If there is a gap between collection periods, state the dates for each sample cohort. Specify the spatial scale from which the data are taken*
Data exclusions	*If no data were excluded from the analyses, state so OR if data were excluded, describe the exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.*
Reproducibility	*Describe the measures taken to verify the reproducibility of experimental findings. For each experiment, note whether any attempts to repeat the experiment failed OR state that all attempts to repeat the experiment were successful.*
Randomization	*Describe how samples/organisms/participants were allocated into groups. If allocation was not random, describe how covariates were controlled. If this is not relevant to your study, explain why.*
Blinding	*Describe the extent of blinding used during data acquisition and analysis. If blinding was not possible, describe why OR explain why blinding was not relevant to your study.*

Did the study involve field work? ☐ Yes ☐ No

Field work, collection and transport

Field conditions	*Describe the study conditions for field work, providing relevant parameters (e.g. temperature, rainfall).*
Location	*State the location of the sampling or experiment, providing relevant parameters (e.g. latitude and longitude, elevation, water depth).*
Access & import/export	*Describe the efforts you have made to access habitats and to collect and import/export your samples in a responsible manner and in compliance with local, national and international laws, noting any permits that were obtained (give the name of the issuing authority, the date of issue, and any identifying information).*

Disturbance | *Describe any disturbance caused by the study and how it was minimized.*

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used | *Describe all antibodies used in the study; as applicable, provide supplier name, catalog number, clone name, and lot number.*

Validation | *Describe the validation of each primary antibody for the species and application, noting any validation statements on the manufacturer's website, relevant citations, antibody profiles in online databases, or data provided in the manuscript.*

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s) | *State the source of each cell line used and the sex of all primary cell lines and cells derived from human participants or vertebrate models.*

Authentication | *Describe the authentication procedures for each cell line used OR declare that none of the cell lines used were authenticated.*

Mycoplasma contamination | *Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination.*

Commonly misidentified lines (See ICLAC register) | *Name any commonly misidentified cell lines used in the study and provide a rationale for their use.*

Palaeontology and Archaeology

Specimen provenance | *Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information). Permits should encompass collection and, where applicable, export.*

Specimen deposition | *Indicate where the specimens have been deposited to permit free access by other researchers.*

Dating methods | *If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.*

☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Ethics oversight | *Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Animals and other research organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research, and Sex and Gender in Research

Laboratory animals	*For laboratory animals, report species, strain and age OR state that the study did not involve laboratory animals.*
Wild animals	*Provide details on animals observed in or captured in the field; report species and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.*
Reporting on sex	*Indicate if findings apply to only one sex; describe whether sex was considered in study design, methods used for assigning sex. Provide data disaggregated for sex where this information has been collected in the source data as appropriate; provide overall numbers in this Reporting Summary. Please state if this information has not been collected. Report sex-based analyses where performed, justify reasons for lack of sex-based analysis.*
Field-collected samples	*For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.*
Ethics oversight	*Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	*Provide the trial registration number from ClinicalTrials.gov or an equivalent agency.*
Study protocol	*Note where the full trial protocol can be accessed OR if not available, explain why.*
Data collection	*Describe the settings and locales of data collection, noting the time periods of recruitment and data collection.*
Outcomes	*Describe how you pre-defined primary and secondary outcome measures and how you assessed these measures.*

Dual use research of concern

Policy information about dual use research of concern

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

No	Yes	
☒	☐	Public health
☒	☐	National security
☒	☐	Crops and/or livestock
☒	☐	Ecosystems
☒	☐	Any other significant area

Experiments of concern

Does the work involve any of these experiments of concern:

No	Yes	
☒	☐	Demonstrate how to render a vaccine ineffective
☒	☐	Confer resistance to therapeutically useful antibiotics or antiviral agents
☒	☐	Enhance the virulence of a pathogen or render a nonpathogen virulent
☒	☐	Increase transmissibility of a pathogen
☒	☐	Alter the host range of a pathogen
☒	☐	Enable evasion of diagnostic/detection modalities
☒	☐	Enable the weaponization of a biological agent or toxin
☒	☐	Any other potentially harmful combination of experiments and agents

ChIP-seq

Data deposition

☐ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☐ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links *May remain private before publication.*	*For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.*
Files in database submission	*Provide a list of all files available in the database submission.*
Genome browser session (e.g. UCSC)	*Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.*

Methodology

Replicates	*Describe the experimental replicates, specifying number, type and replicate agreement.*
Sequencing depth	*Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.*
Antibodies	*Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.*
Peak calling parameters	*Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.*
Data quality	*Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.*
Software	*Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.*

Flow Cytometry

Plots

Confirm that:

☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☐ All plots are contour plots with outliers or pseudocolor plots.

☐ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	*Describe the sample preparation, detailing the biological source of the cells and any tissue processing steps used.*
Instrument	*Identify the instrument used for data collection, specifying make and model number.*

Software

Describe the software used to collect and analyze the flow cytometry data. For custom code that has been deposited into a community repository, provide accession details.

Cell population abundance

Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.

Gating strategy

Describe the gating strategy used for all relevant experiments, specifying the preliminary FSC/SSC gates of the starting cell population, indicating where boundaries between "positive" and "negative" staining cell populations are defined.

☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type

Indicate task or resting state; event-related or block design.

Design specifications

Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.

Behavioral performance measures

State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects).

Acquisition

Imaging type(s)

Specify: functional, structural, diffusion, perfusion.

Field strength

Specify in Tesla

Sequence & imaging parameters

Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle.

Area of acquisition

State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined.

Diffusion MRI ☐ Used ☐ Not used

Preprocessing

Preprocessing software

Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).

Normalization

If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.

Normalization template

Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.

Noise and artifact removal

Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).

Volume censoring

Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.

Statistical modeling & inference

Model type and settings

Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).

Effect(s) tested

Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.

Specify type of analysis: ☐ Whole brain ☐ ROI-based ☐ Both

Statistic type for inference
(See Eklund et al. 2016)

Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.

Correction

Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).

Models & analysis

n/a	Involved in the study
☐	☐ Functional and/or effective connectivity
☐	☐ Graph analysis
☐	☐ Multivariate modeling or predictive analysis

Functional and/or effective connectivity

Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information).

Graph analysis

Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).

Multivariate modeling and predictive analysis

Specify independent variables, features extraction and dimension reduction, model, training and evaluation metrics.