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

Kolja Stahl, Andrea Graziadei, Therese Dau, Oliver Brock, Juri Rappsilber, Nature Biotechnology, 2023

(from Technische Universität Berlin, Germany)

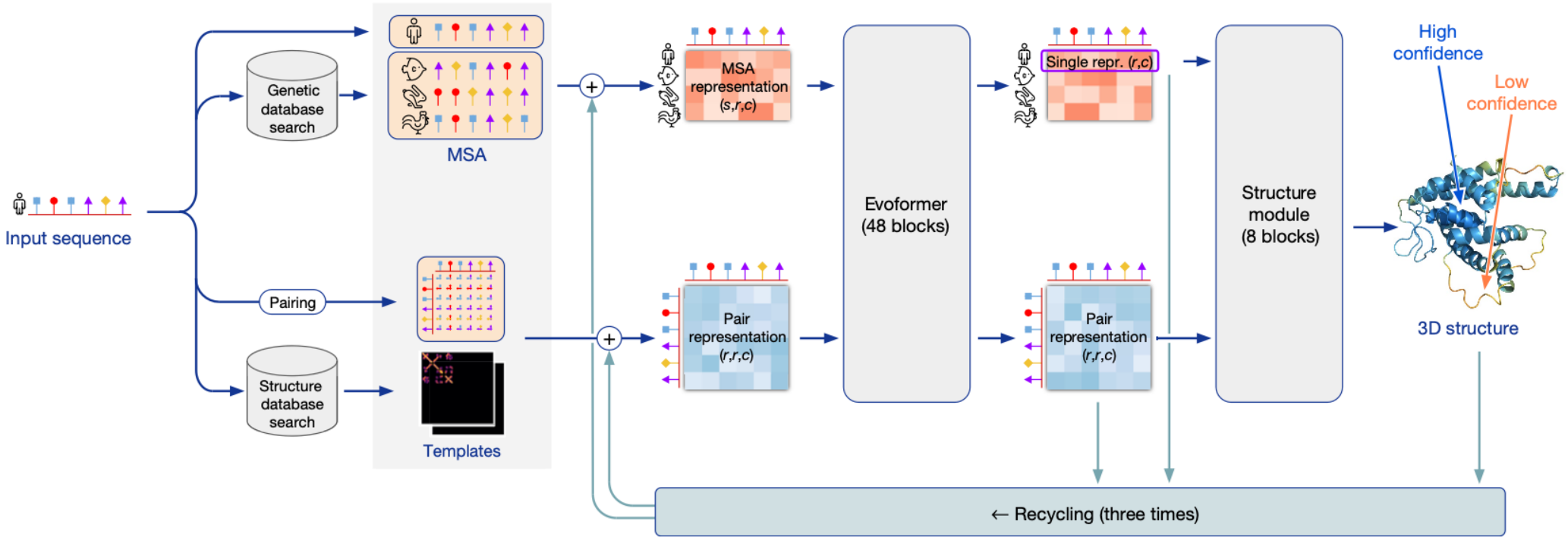
Shentong Mo

Jun 29, 2023

Presentation Line

- Background & Motivation
- AlphaLink architecture
- AlphaLink methods
- AlphaLink training data
- Main experimental results
- Discussions

Recap AlphaFold2



- Multiple Sequence Alignment (MSA) + Templates of Similar Protein Structures
- Evoformer (distance space)
- Structure Module (3D space)

Background

- AlphaFold2 predicts static models based on static input data, which is trained on two information sources, the protein structures in the Protein Data Bank (PDB) and multiple sequence alignments (MSAs).
- Challenged by targets with **insufficient evolutionary information**
  - viral proteins, proteins from understudied organisms, antibodies, and synthetic proteins have misleading information
  - structures underlying the model poorly reflect structural flexibility, multiple conformations, and dynamic interactions

Motivation

- **Structural restraints** observed on proteins in solution steer the prediction towards structural states occurring in situ under specific conditions.
- **Crosslinking mass spectrometry (MS): distance restraints**
  - photo amino acids (photo-AA) in prokaryotic and eukaryotic cells
  - photo-AA crosslinks: tight distance restraints that align well with co-evolutionary contacts.
  - Photo-leucine (photo-L): mapping conformations & binders.

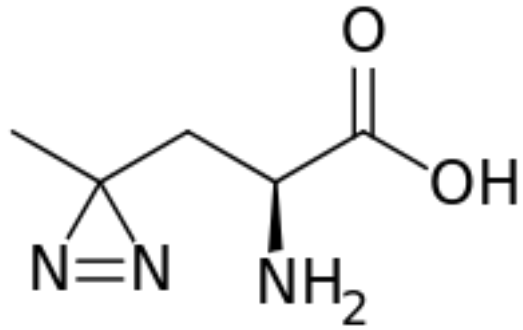
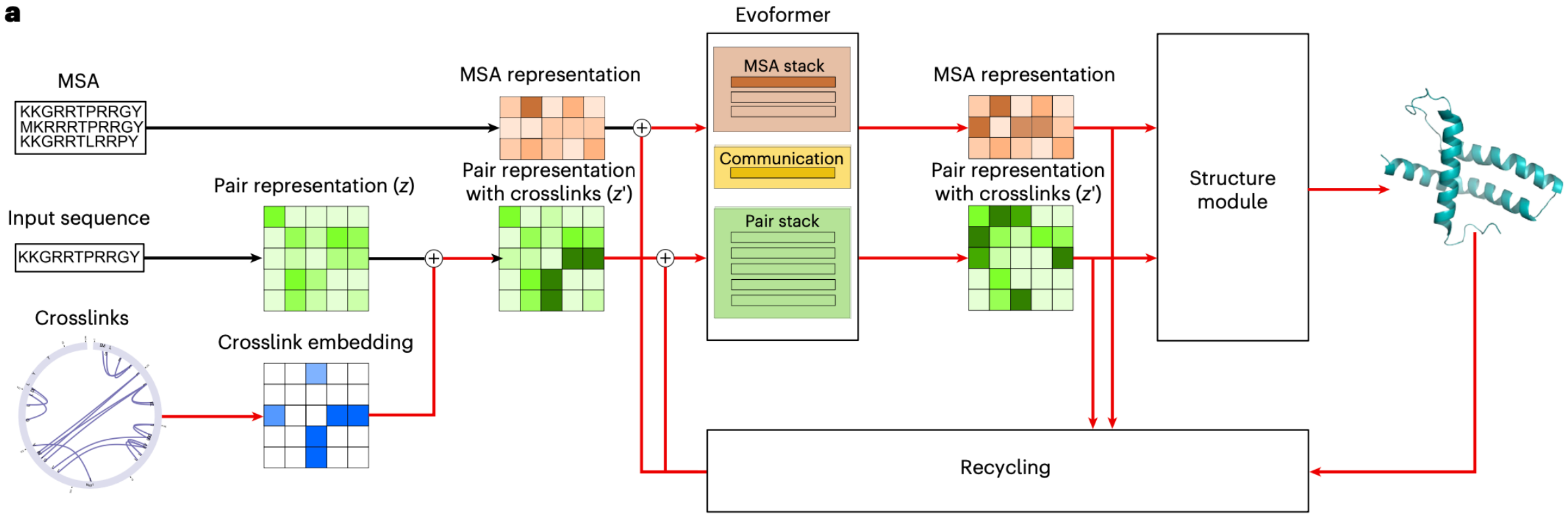


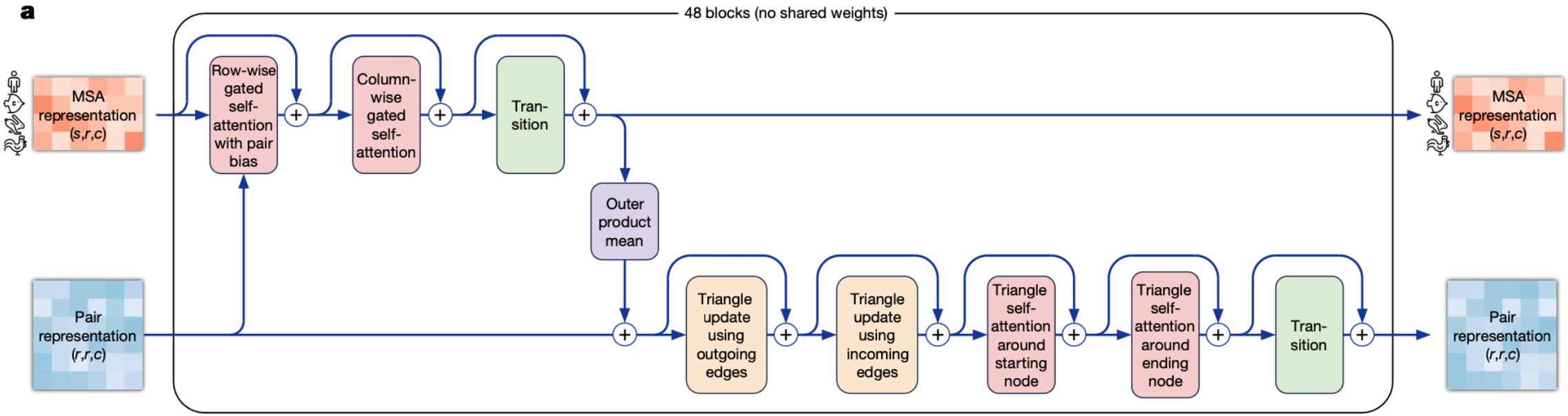
Photo-leucine, source from wikipedia

AlphaLink: integrating crosslinks into AlphaFold2 via OpenFold



- Crosslinks (blue) are embedded and added onto the pair representation (green)

Recall Evoformer

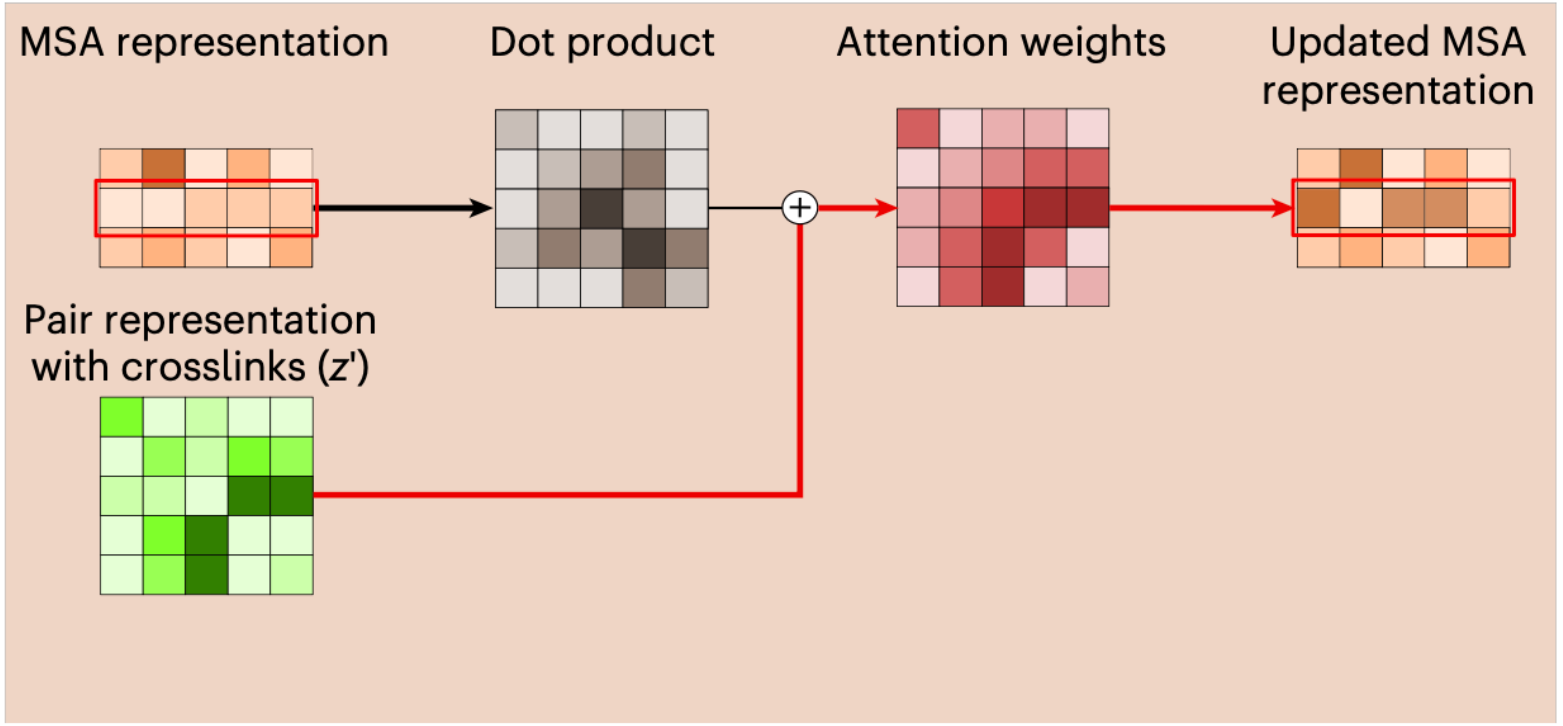


How to encode crosslink

- Two representations to encode crosslinking information
  - **soft labels**: each contact is weighted by the link-level false discovery rate (FDR) of the dataset (1-FDR)
  - **distance distributions** (distograms): uniformly distributed distograms for the given cutoff
- Use the same binning for the first 64 bins in Evoformer and extend the distogram further to 128 bins, spanning from 2.3125 Å to 42 Å.

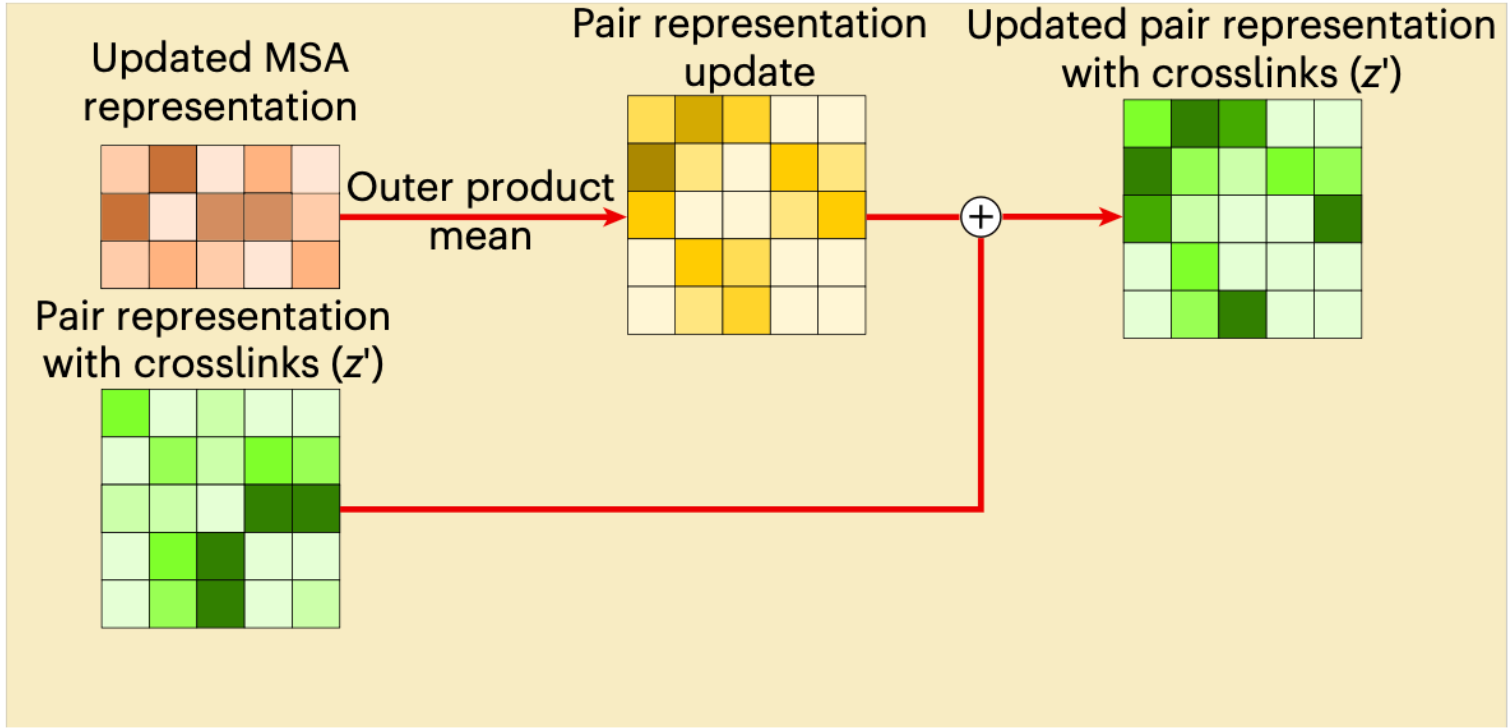
Crosslink as a bias in MSA transformer

- Crosslinks influence the retrieval of co-evolutionary information.

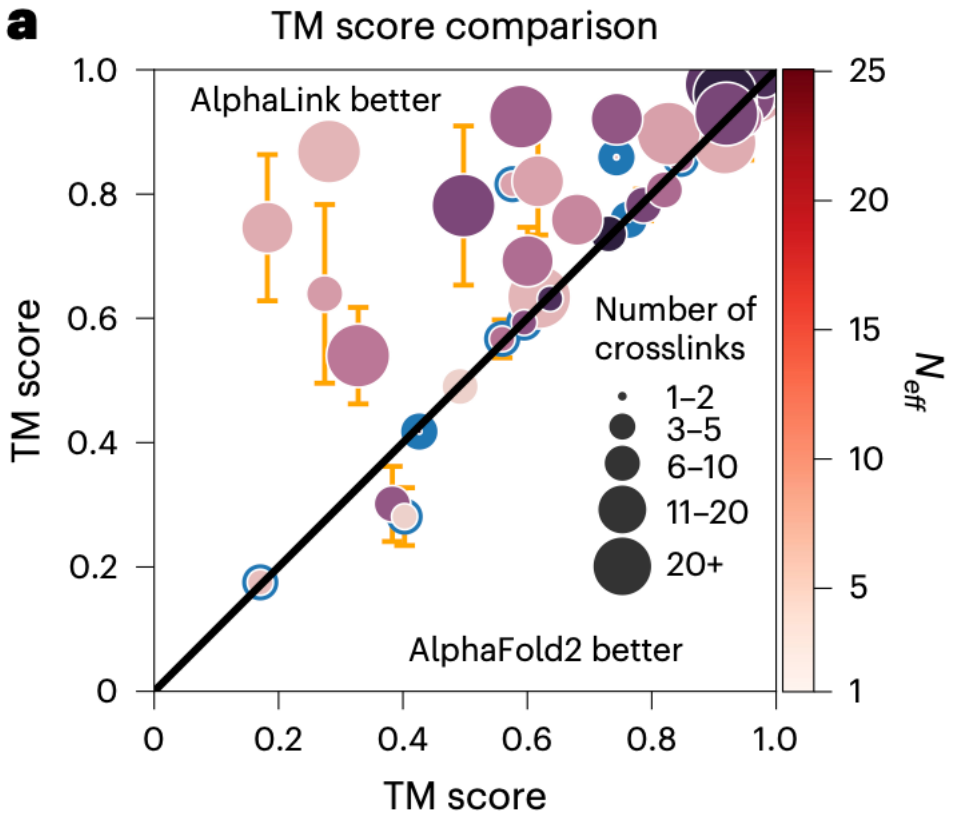


Crosslink as a bias in pair representations

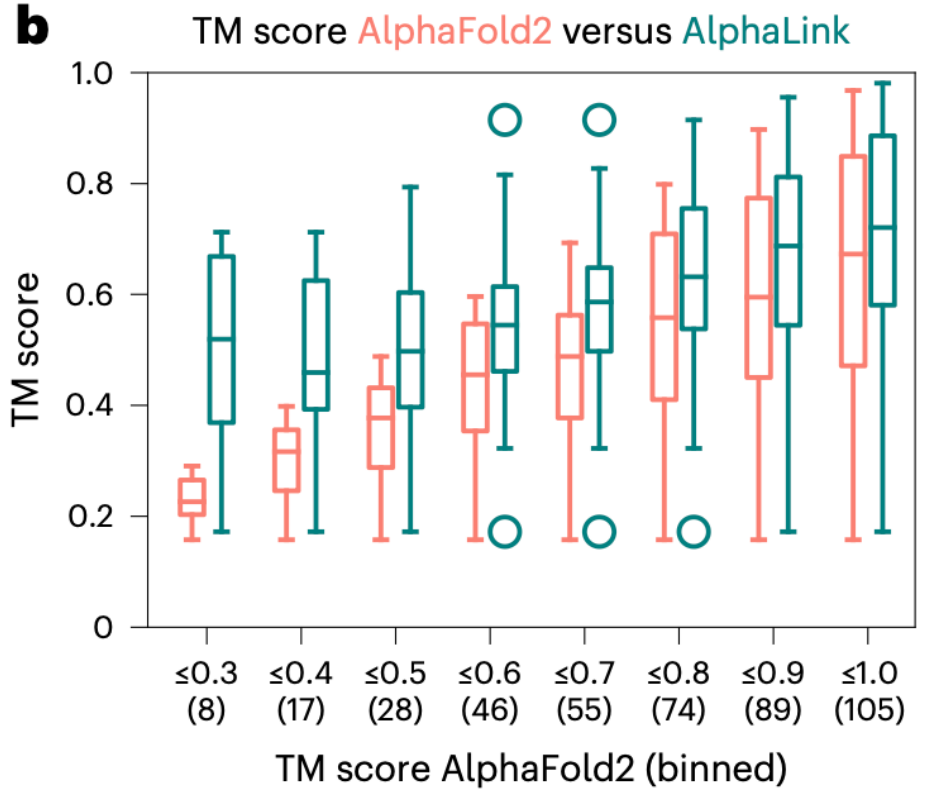
- The pair representation is updated with information from the MSAs that have been biased with the crosslinks.



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



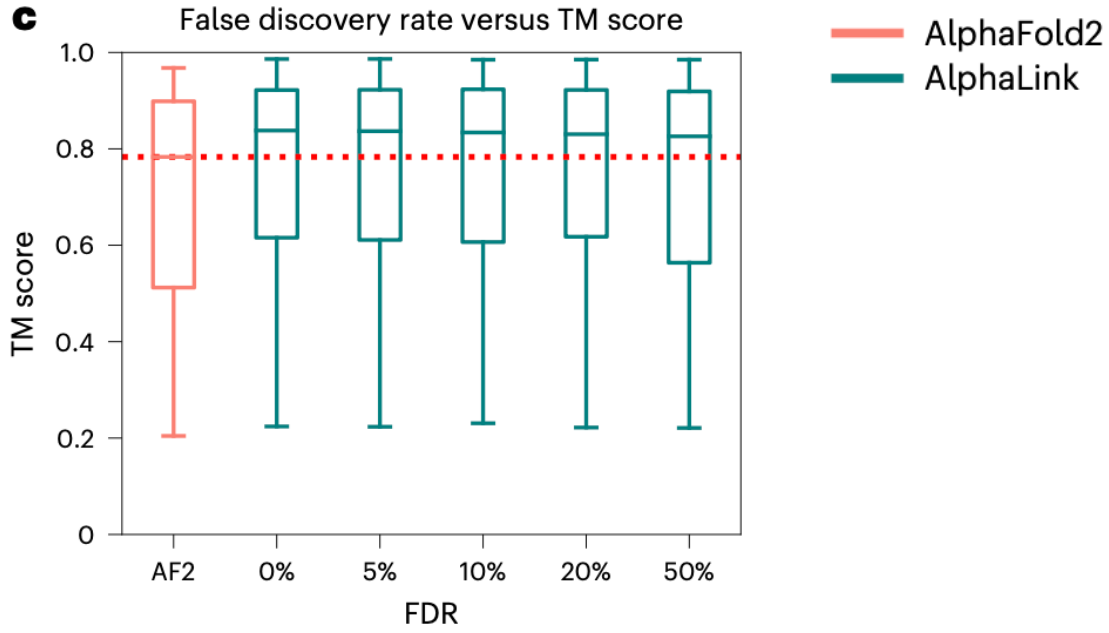
- TM score comparison on 49 CAMEO targets. TM score improves on average by 19.2%.



- TM score performance on 60 CASP14 and 45 CAMEO targets (15.2% improvement)

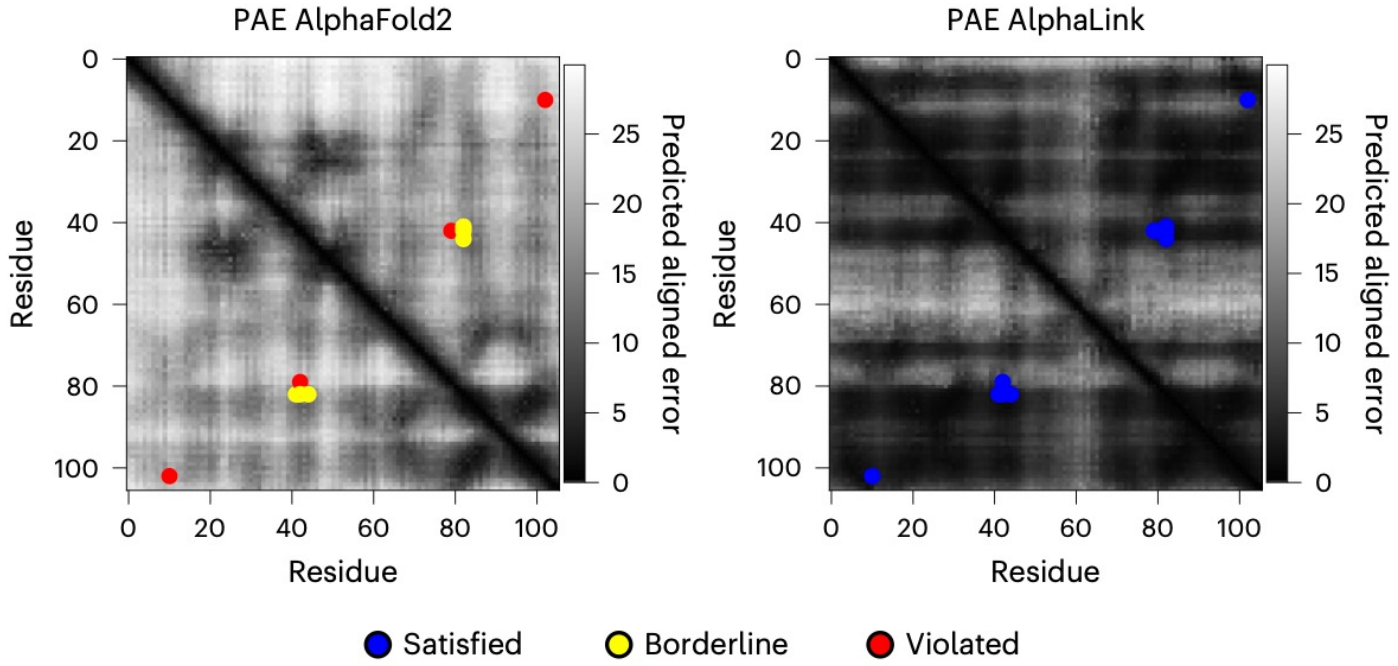
Performance on different noise levels

- Performance on 60 CASP14 targets 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.

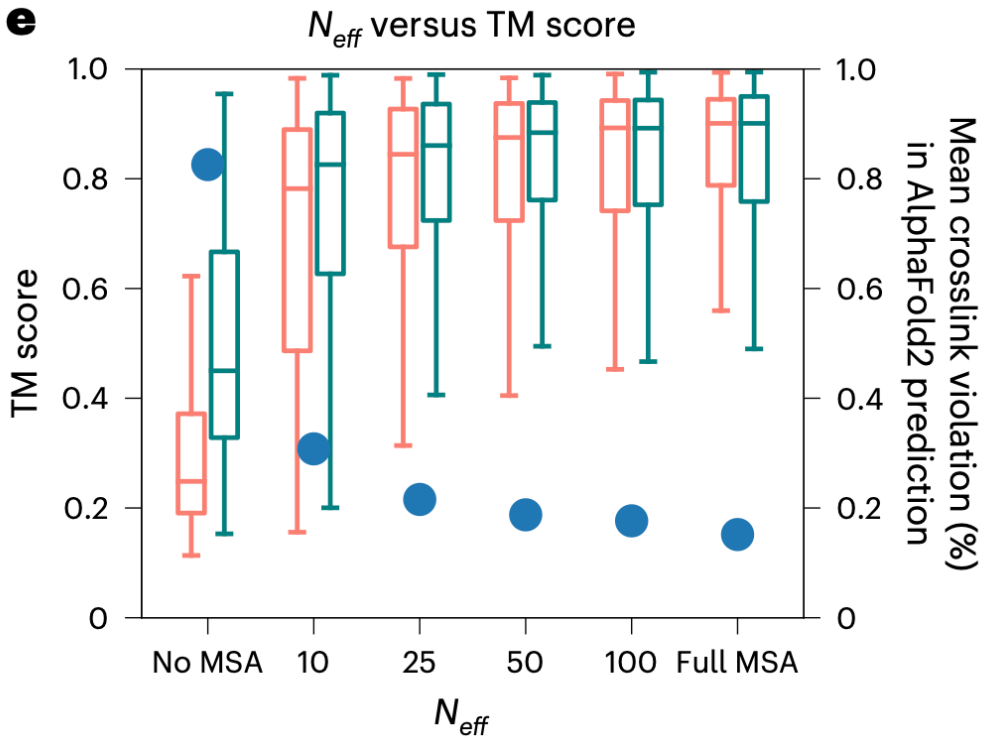


Sparse restraints decrease uncertainty across the whole protein

- Predicted aligned error of AlphaFold2 (left) and AlphaLink (right) on T1064.
- Satisfied crosslinks <10 Å Cα–Cα, borderline crosslinks (10–15 Å Cα–Cα), violated crosslinks >15 Å Cα–Cα



Crosslink diminishes with increasing MSA size



- Performance on 60 CASP14 targets as a function of MSA size (N = 100, 10 MSAs and 10 crosslink sets)
- Blue dots represent the mean percentage of nonsatisfied crosslinks (>10 Å Cα–Cα) in the AlphaFold2 prediction.

Performance without MSAs

- 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).

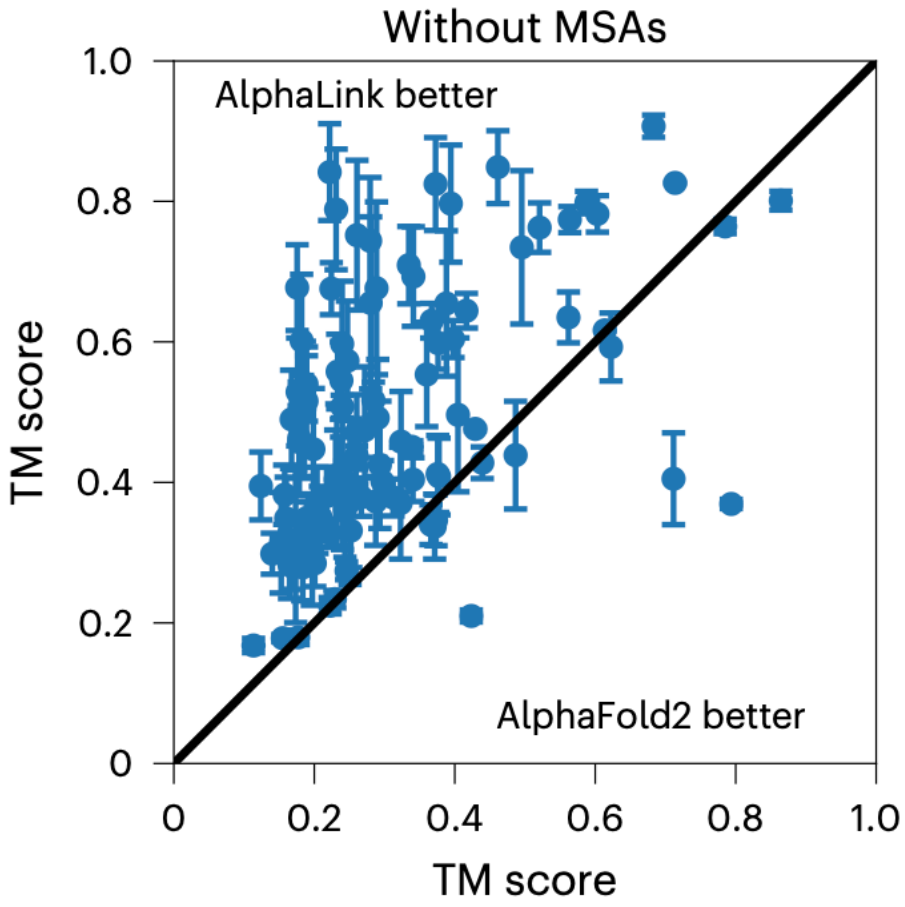
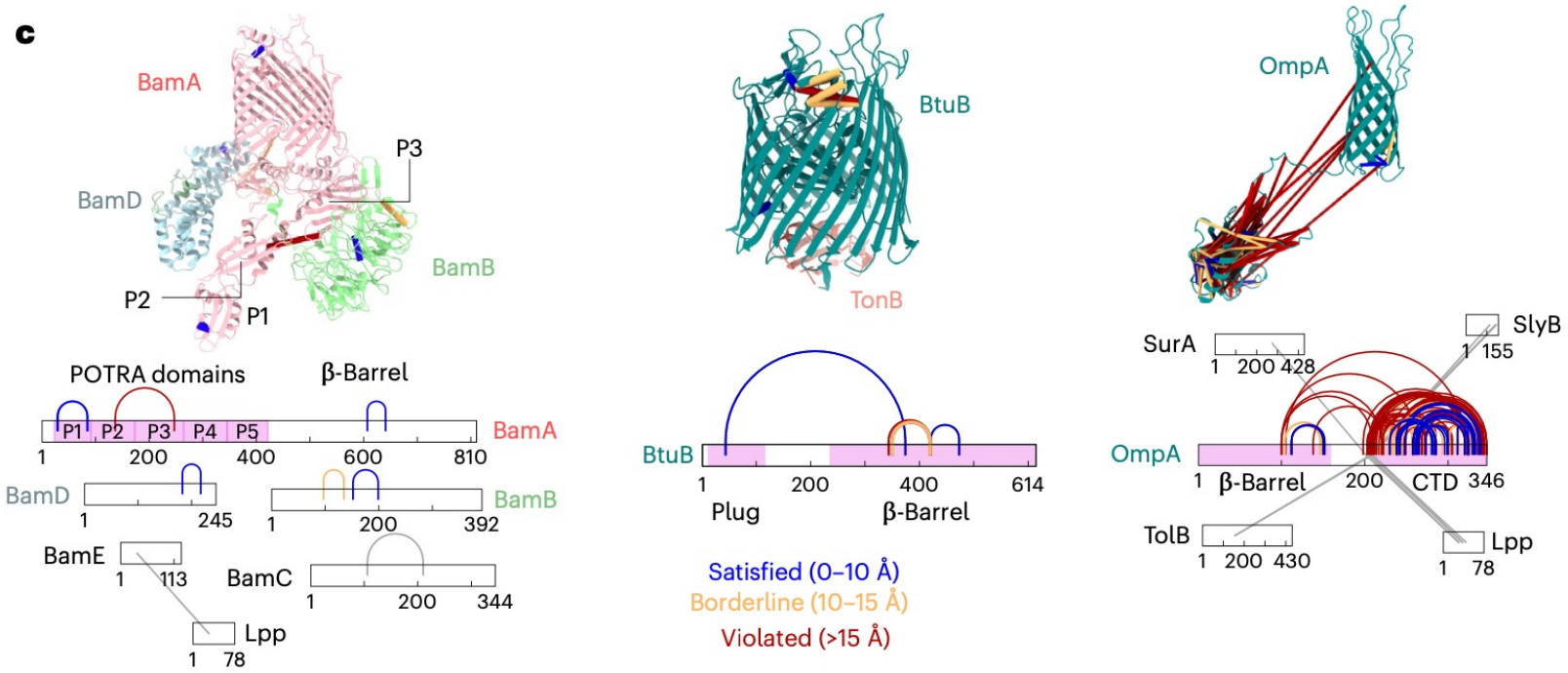


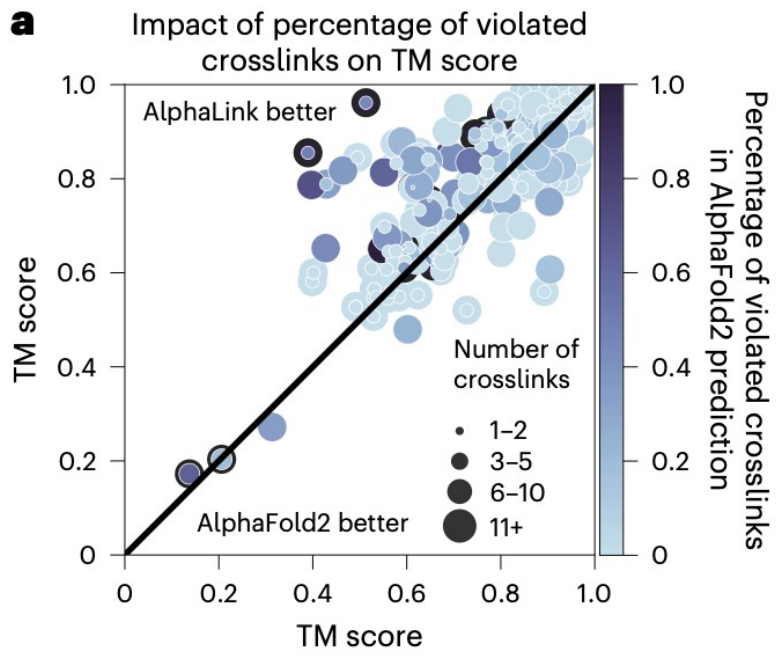
Photo-Leucine as an in situ structural probe

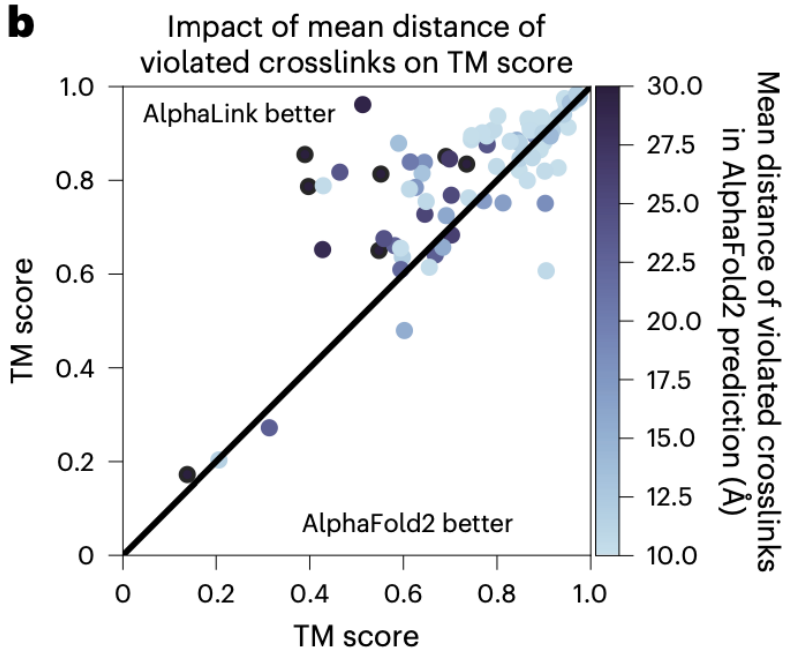
- large-scale experimental photo-AA dataset with in situ structural restraints on the *Escherichia coli* membrane fraction
- 615 residue pairs involving 112 proteins at 5% link-level FDR
- Photo-Leucine provides in situ conformation of multiprotein complexes



Structure prediction with Photo-Leucine data

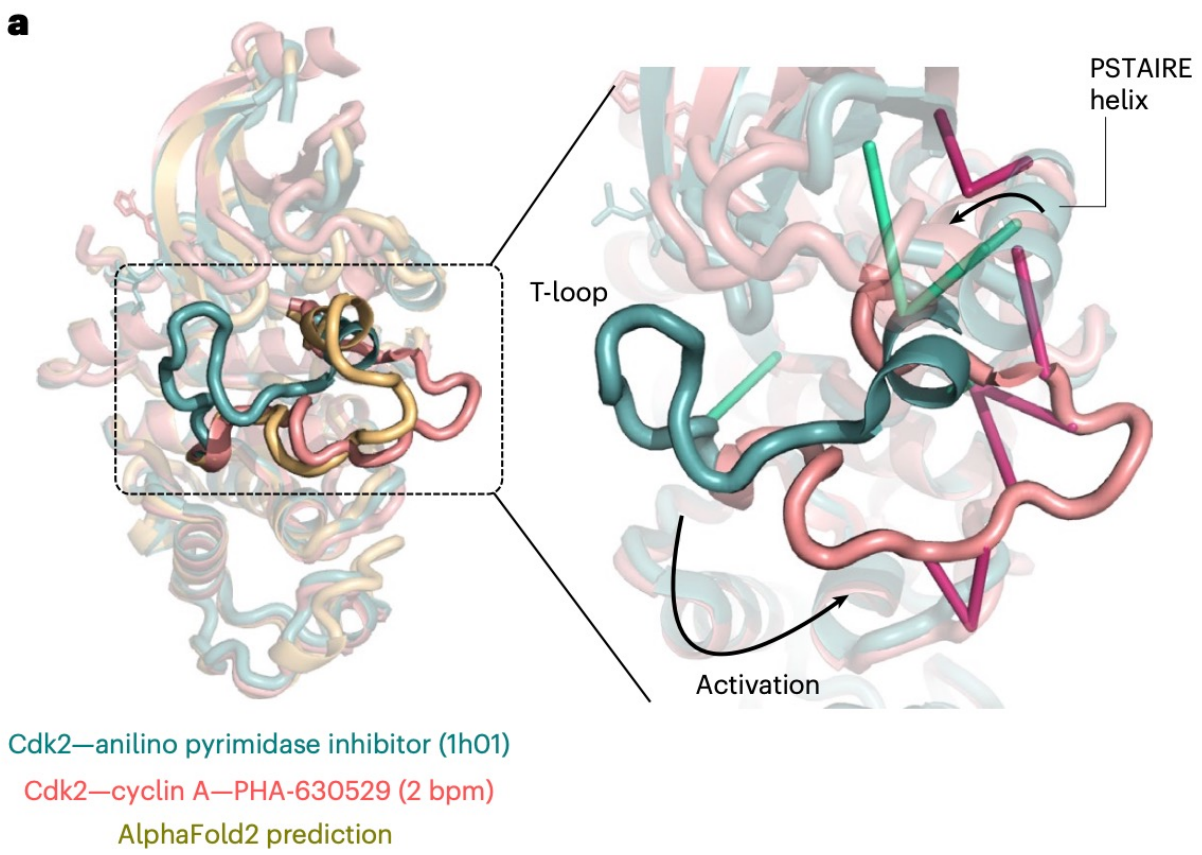
- 31 targets with high-resolution structures
- **a**: performance improvement is bigger for targets with a higher percentage of nonsatisfied crosslinks in the base prediction
- **b**: predictions that improved the most have unsatisfied crosslinks with large distances





Probing conformational dynamics in situ

- a proof-of-concept experiment on the human cyclin-dependent protein kinase Cdk2, a drug target in cancer therapy.
- 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.

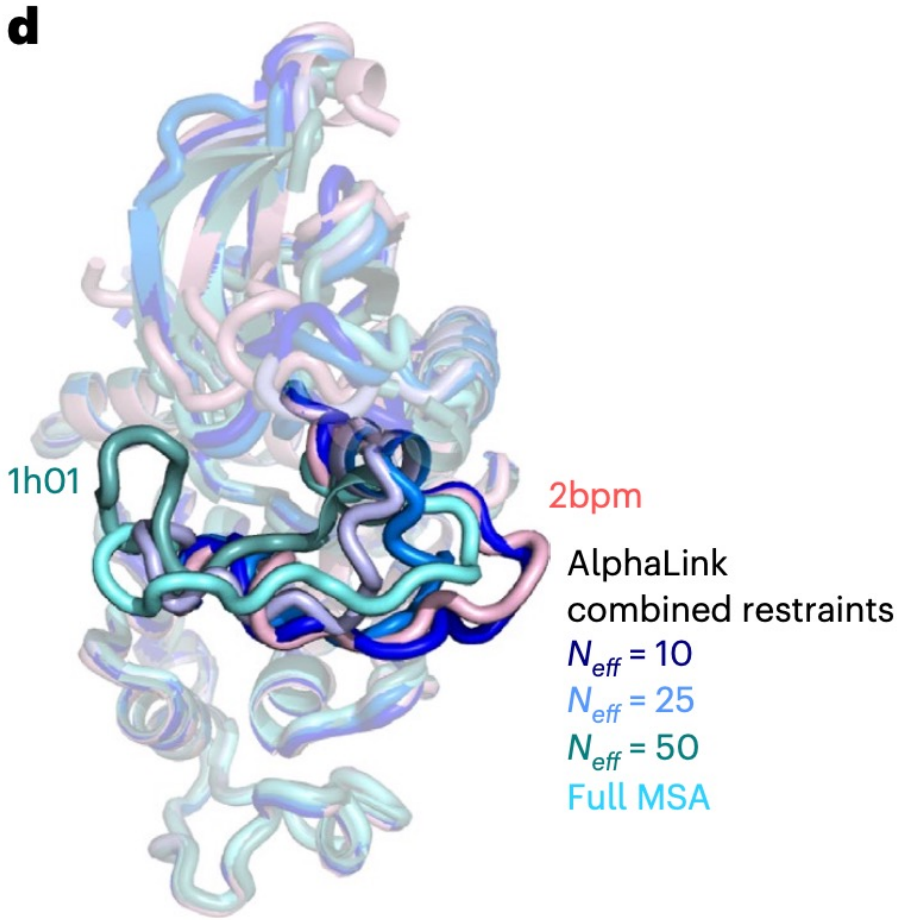


Comparison on active/inactive conformation

- photo-crosslinking: protein was acquired in either its inhibited or in its cyclin A-bound states.
- Cdk2 structure using AlphaLink with these restraints, showing that the loop structure is driven towards the appropriate conformation.
- all AlphaFold2 predictions converge to the cyclin A-bound state, failing to predict the inactive conformation.

Outcome with a combined set of restraints

- At low N_{eff} (*number of effective sequences*) 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.
- Crosslinking is weighted against the MSA depending on the information content and size of both strands of information.



Conclusion

- AlphaLink integrates experimental data from photo-AA crosslinking directly into the AlphaFold2 architecture to merge co-evolutionary relationships and crosslinking data in distance space, exploiting the complementary nature of the data.
- AlphaLink can leverage noisy experimental contacts to improve predictions.
- AlphaLink performs a large-scale crosslinking MS study with photo-Leucine, identifying 615 in situ residue-residue contacts in *Escherichia coli* membrane fractions.
- Even sparse crosslinking MS data can anchor predictions to particular conformational states, opening up the possibility of hybrid experimental/deep learning approaches.

Questions