Unmasking AlphaFold: integration of experiments and predictions in multimeric complexes

Claudio Mirabello[1, 2✉], **Björn Wallner**[1], **Björn Nystedt**[2, 3], **Stavros Azinas**[4], **and Marta Carroni**[4]

[1]Dept of Physics, Chemistry and Biology, Linköping University, 581 83 Linköping, Sweden
[2]National Bioinformatics Infrastructure Sweden, Science for Life Laboratory
[3]Dept of Cell and Molecular Biology, Uppsala University, Husargatan 3, SE-752 37 Uppsala, Sweden
[4]Dept of Biochemistry and Biophysics, Science for Life Laboratory, Stockholm University, Stockholm, Sweden

Since the release of AlphaFold, large efforts have been undertaken to improve its predictions and integrate it in already existing pipelines for the determination of protein structures. Though a number of functionalities have been introduced, and results at the latest CASP edition showed a marked improvement in the prediction of multimers, AlphaFold's capability of predicting large protein complexes is still limited, and integration with experimental data is not straightforward. In this study, we introduce AF_unmasked: a version of AlphaFold able to better take advantage of information from experimental structures to build high-quality protein assemblies, including very large protein complexes. We test AF_unmasked on difficult targets from the latest CASP edition as well as on large protein cryo-EM targets and show that this new approach enables generation of near-perfect structures even when little to no evolutionary information is available or when imperfect experimental structures are used as a starting point. We also show how incomplete structures can be filled in by AF_unmasked, a procedure called "structural inpainting", and how this can give insights into protein dynamics. In summary, AF_unmasked represents an easy-to-use method with much improved capacity to efficiently integrate experiments and predictions to solve large protein complexes, without any need for computationally demanding re-training.

Code: ***github.com/clami66/AF_unmasked***
Correspondence: ***claudio.mirabello@scilifelab.se***

DRAFT

Introduction

Since the release of AlphaFold (1) in 2020, part of the computational structural biology community has worked on improving AlphaFold and expanding its functionalities, sometimes in ways its creators had not envisioned initially. This is a challenging avenue of research, as it involves manipulating a complex neural network, a task that can yield unpredictable results without a good enough understanding of how it works. Interpretation of neural networks is also notoriously hard.

One interesting attempt at interpreting the AlphaFold network comes from a recent study where it is theorised that the neural network performs a sampling technique on a learned energy landscape (2). According to this theory, the first block of layers (the Evoformer module) is tasked with identifying a neighborhood within this landscape that closely approximates a reasonable minimum, while the second block of layers (the structural module), known for its ability to accurately predict the quality (1, 2), performs an energy minimisation within the identified neighborhood to generate atomic structures.

Recent efforts aimed at improving AlphaFold could be framed as strategies to improve this energy minimisation procedure. At the latest CASP15, where the main focus was on assessing the progress in the assembly category since the release of AlphaFold-Multimer (3), the best ranking groups in the assembly category that used AlphaFold either introduced stochastic noise in the neural network while generating thousands of models for each target (4), or tried to create and select better/deeper Multiple Sequence Alignments (MSAs) as input to the Evoformer module (5). If the prediction task is indeed one of energy minimisation, the first approach would make sense as a way to perform more extensive sampling of AlphaFold's energy function, while the second could be interpreted as an attempt to start the sampling procedure in a lower-energy neighborhood. Both approaches aim to maximise AlphaFold's predicted quality score in order to sift out the best models. On the other hand, there seemed to be few gains made optimising a secondary input to AlphaFold, i.e. the *templates*, containing structural information (in the form of atom coordinates) about homologous proteins structures from the Protein Data Bank (PDB). Some groups at CASP15 have gone as far as turning templates off altogether (5).

Meanwhile, in the wider user base outside of the CASP community, the consensus is that AlphaFold should be a tool to aid structure determination and density interpretation rather than a replacement for experiments (6, 7). Indeed, a lot of work is being done on developing better integrative tools. In X-ray crystallography, Phenix integrates AlphaFold within a more classical Molecular Replacement approach involving the trimming, breaking, and assembly of predicted monomers for refinement against experimental maps (8).

It has also been shown that it is possible to integrate cryo-EM and X-ray crystallography experiments with predictions by iteratively refining AlphaFold models against experimental data and inputting the refined models in AlphaFold as structural templates (9, 10). This approach shows that templates can be used as a vector to inject experimental information into a prediction pipeline. Still, AlphaFold was ini-

tially engineered for monomeric predictions and even though several multimer versions have been released, templates in AlphaFold remain monomeric in the sense that any potential information about interactions in quaternary templates is not used (see Methods). Templates will help build each monomeric subunit in a complex, but not the multimeric assembly of the subunits.

Other integrative approaches propose retraining versions of AlphaFold that are better at taking experiments into account: versions of OpenFold (11) and Uni-Fold (12) (retrainable implementations of AlphaFold) accept extra input data derived from crosslinking experiments to guide monomeric and multimeric predictions (13). Still, retraining a large neural network like AlphaFold requires a lot of resources, and results show that these new predictors are not always as good as those that participated at CASP15 (13).

In this work, we modify AlphaFold's neural network modules to leverage information from templates containing quaternary structures to its full extent without the need for retraining. The extra information derived from quaternary templates greatly reduces the complexity of the prediction problem, allowing for faster prediction times and more accurate quaternary models, in particular for large protein complexes. We call this new version of AlphaFold *AF_unmasked*.

The proposed changes involve non-parametric layers in the neural network, so no re-training is needed. Results show that this new template approach is very effective at influencing the predictions and guide the network to produce high-quality complexes, even for very large ones, and when no evolutionary information at all is available. Contrary to other approaches, where complexes are built iteratively in AlphaFold by assembling smaller subcomponents (14), here large complexes are generated in one shot. We also explore the possibility of using quaternary templates to perform better integrative modelling, for example, by using AlphaFold to resolve interfacial clashes in coarse templates, which is often what is available when manually assembling together monomers based on protein density information. When pushing this concept even further, we can even perform "structural inpainting" letting AlphaFold use evolutionary information to automatically fill in missing stretches of amino acids in experimental structures, or even using its learned energy potential to switch between different conformations when mutations are introduced.

Methods

The changes to the standard AlphaFold pipeline are depicted in Fig. 1. In summary, a new template strategy is implemented where multimeric structures, either from released structures in the PDB or produced by the user, can be input at prediction time so that the positioning of protein chains with respect to each other informs the prediction task. Novel components are are outlined more in detail below.

The current implementation of templates in AlphaFold. The data pipeline in AlphaFold can select up to four structural templates. When predicting monomers, a template input is a set of coordinates of shape $L \times 37 \times 3$, where L is the number of residues in the input sequence, 37 is the number of unique possible atom types for any type of amino acid, and 3 are the x, y, z dimensions.

When predicting multimeric targets, AlphaFold's data pipeline will perform a separate template search for each unique input sequence. The first template is selected from the search results for each chain, and coordinates are extracted from the corresponding PDB files. The coordinates are then mapped to a target sequence through a series of sequence alignments and merged across all chains so that the template input to AlphaFold contains x, y, z coordinates for all atoms covered by templates across all chains (Fig. 2a). The coordinates are merged in a single list irrespective of whether the PDB hits originate from the same or completely different PDB structures. This step is performed up to four times, which is the maximum number of templates used in AlphaFold.

For all template inputs, AlphaFold's template-embedding module computes distogram features from the coordinate inputs. Distograms contain information about distances between all pairs of $C\beta$ atoms in a template. This information might describe distances between residues that sit in the same chain (intra-chain distances) as well as distances between residues that sit in different chains (cross-chain distances). At this stage, a "multichain mask" is applied to all distograms. This is a boolean 2D array of shape $L \times L$ used to mask out sections of the distogram that contain cross-chain distances (Fig. 2a). This masking is necessary since cross-chain distances obtained by merging chains from different PDB entries with no shared frame of reference are incorrect, so they should be ignored by the neural network.

Using restraints from multi-chain templates. We introduce the possibility of either turning off multichain masking completely (Fig. 2b) or to switch to a complementary masking approach, where only cross-chain restraints are enforced in the template input while the intra-chain restraints are ignored (Fig. 2c). We accompany this modified neural network model with a set of tools to format and prepare ad-hoc multimeric templates. The templates can be either deposited PDB structures, as is done in classical homology modelling, but it could also be a draft of an experimental structure, for example when using AlphaFold as a tool to aid molecular replacement or for building structures from cryo-EM densities with different local resolution of details.

Preparing ad-hoc templates. The template search in AlphaFold's standard pipeline is not different from any other homology modelling tool, where a database of sequences in the PDB is queried with the target sequence or with a protein profile built from a target MSA. Whenever a template hit is found, the sequence alignment against the target sequence is used in the template pipeline in order to map atom coordinates from the PDB hit onto the target sequence. Gap regions in these alignments, i.e. sections of amino acids that are not covered by a given template, are masked so that they can be ignored by the neural network model at inference time.

DRAFT

The template preparation step in our pipeline replaces the template search in AlphaFold by allowing the user to align either a set of target sequences or previously built monomeric structures against the selected template structures. The possibility to perform structural, rather than sequence-based, alignments allows the user to select templates that are remote homologs and might not be detected by a simple sequence search. Structural alignments can either be performed by TM-align (15) or by lDDT-align (16), a tool we developed to perform structural alignments while maximizing the lDDT score. A superposition-free tool such as lDDT-align should be beneficial in producing reliable alignments between remote homologs, or between structures that undergo significant domain arrangements when binding to a partner.
If a template is given in PDB format rather than mmCIF, it is automatically converted to mmCIF for compatibility with AlphaFold's structure parsing step. The preparation step can be performed up to four times, either with four different templates or by repeating the same template, to fill out all of the input template slots. We have observed that using all the template slots strengthens the impact of templates.

Clipping of Multiple Sequence Alignments. The prediction of a structure is a balancing act between restraints derived from templates and those derived from MSAs. When the two are in disagreement with each other, and if the MSA-derived restraints are too strong, the Evoformer might ignore the template information, and vice-versa. Here, we introduce an option to manually reduce the number of sequences that should be included from each MSA to increase the influence of templates on the results. In this study, for example, we often clip MSAs computed from the UniProt database, as they are used in the MSA pairing step to find evolutionary couplings spanning multiple chains (3), in order to predict the pairing of subunits.
Clipping the MSA comes with the side benefit of greatly reducing the computational resources needed to generate each structure, as the time and memory cost of the Evoformer scales with the squared number of sequences in the merged MSA (1). This makes it possible to generate much larger structures than otherwise possible with the standard version of AlphaFold.

Inpainting of experimental structures. In image processing, *image inpainting* is the task of reconstructing parts of an image that are missing, for example to reconstruct damaged parts of old photos or to remove watermarks. In the past few years, a number of deep learning tools have been developed for the task of inpainting protein structures, for example to design protein scaffolds for given functional sites (17). In this study, we extend this concept to the improvement of experimental structures by leveraging AlphaFold's learned potential.
Almost all deposited protein structures have missing atoms to large stretches of amino acids, due to their flexibility or to incomplete or noisy data. We propose to use AlphaFold to fill these gaps by inputting the partial experimental structure as template along with the full target sequence and let AlphaFold model the missing parts while keeping the rest of the structure as close as possible to the template. While this is theoretically already possible with monomeric structures in AlphaFold, we are not aware of other studies attempting this, let alone with multimers.

CASP15 benchmark set. Some of the authors of this work also participated in CASP15 as predictors both in the server and human categories. The servers were built to provide the CASP15 assessors with baseline AlphaFold predictions (group IDs: NBIS-AF2-standard and NBIS-AF2-multimer), and one human group (group ID: Wallner) adopted special sampling strategies to generate more diverse sets of structures while using the same MSA inputs as they were generated by the baseline servers (4, 5). Predictions made by Wallner in the multimer category were also among the best in CASP15 overall (18).
Here, we choose to compare predictions made with AF_unmasked against predictions made by the Wallner group in CASP15. This allows to make a fair comparison against what is effectively the state of the art when predicting multimers while controlling for other variables, such as databases used for MSA generation, sampling strategy, and neural network models. It is important to note here that, even though we control for these variables to isolate the effect of the new template strategy, the goal of this set of tests is not to prove that AF_unmasked performs better than other methods on the CASP15 benchmark, but rather, to demonstrate the effectiveness of our approach in accurately predicting difficult or large targets when the right template is used. For example, we will be using information that was not accessible to the predictors at CASP15, such as partial or full native structures of targets that have been released since.

Analysis of cryo-EM macromolecular assemblies. We reason that AF_unmasked will be useful in the analysis of large macromolecular complexes obtained by cryo-EM. While AlphaFold in the current form, both in the monomer and in the multimer implementations, is able to answer structural questions about small and stable protein in a satisfactory way, it does not address the structural complexity of large and flexible macromolecules, a typical type of cryo-EM target. This method will therefore aid in the modelling of (i) poorly resolved parts of cryo-EM density maps, either because of flexibility or low occupancy, and (ii) large asymmetrical homo-oligomers which display variability in domain positioning. To test these scenarios, we consider three different large complexes: the Rubisco enzyme, the ClpB disaggregase and the Ras-regulator Neurofibromin. All these targets are large homo-oligomers that cannot be built with the standard multimer AlphaFold version. They also have flexible domains that are poorly or not resolved in the experimental cryo-EM densities.

Sampling strategy and other settings. We use the AlphaFold code as released in version 2.3, but when comparing against other predictors we use the parameters released in v2.2, which were already available for CASP15. In order

DRAFT

to model very large complexes, where no comparison is done against other predictors, we also test v2.3 parameters, which were trained on more data from large protein complexes (19). In order to stimulate additional sampling of a target's conformational space, we turn on dropout at inference time following a strategy similar to the Wallner group's in CASP15 (18). But, instead of generating thousands (6000+) of structures as in the Wallner group in CASP15, we generate up to 100 structures per neural network model (500 in total), fewer for large complexes. We also increase the number of recycles from the default for v2.2 (3) to 20, as in v2.3. Finally, we introduce an option that forces AlphaFold to use separate MSAs and template for all chains, even those with identical sequences, so that the correct templates can be used in the case of homomers. Relaxation is turned off for all experiments to isolate contributions of the neural network from those made by the relaxation protocol using the AMBER force field.

Scoring. We rank predicted models by AlphaFold's default ranking metric, which for multimers is the weighted sum of predicted TM-score (pTM) and interface pTM ($ipTM$) with weights of 0.2 and 0.8, respectively. This combination is called the *ranking confidence*. When comparing a prediction to a native structure in order to assess its quality, we use the DockQ score (20).

Results

We test AF_unmasked on a series of cases derived from a dataset made of "hard" multimeric targets from CASP15 as well as cryo-EM datasets of large protein complexes. First, we show how the multimeric template information is particularly useful in cases where the standard version of AlphaFold is unable to build the correct complex. We furthermore show that even when imperfect templates are used, e.g. multimeric templates with clashing interfaces or missing parts, AlphaFold is able to improve upon those inputs by remodelling parts or filling in the gaps with structural inpainting.

Proof of concept. The first question is whether AlphaFold is at all capable of using cross-chain information derived from multimeric templates to build protein complexes. After all, the neural network was not trained to take distances across chains in account when building assemblies. Therefore we perform a proof-of-concept test, where the same complex is modelled twice starting from the same templates, but with multichain masking turned on or off. We perform this test on target H1137 from CASP15, a large 3939 AA heteromer for which, during the 3-day CASP prediction window, the baseline server NBIS-AF2-multimer was only able to produce five models. We predict this complex by using the deposited structure of the target itself (PDB ID: $8FEF$) as template in order to keep the prediction task as simple as possible. In order to isolate the contribution of templates from that of evolutionary information, we clip all MSAs to a single sequence.
In Fig. 3, we show that the models built this way are quite similar to the template input, though some variability is present. This variability is significant, as it demonstrates that AlphaFold will not simply "copy" the template coordinates onto the prediction but rather, reconstruct the structure from distance constraints extracted from the template while maximising the neural network's learned potential.
The time elapsed to predict a model, excluding the time to generate the MSAs, is roughly 60 minutes (NVIDIA A100, 40GB RAM). If cross-chain templates are turned off, the models fail to fold altogether (results not shown). This demonstrates that, even though the templates are still informing the inference for each separate monomer in the complex, lack of cross-chain template information will make it impossible to correctly fold the complex.

Using imperfect templates. As a harder test, we use a template that is a coarse representation of a protein complex. This is a common scenario when performing molecular replacement or fitting densities in experiments, where a user might generate models for unbound monomers and manually dock them according to the data. We want to test whether AlphaFold can generate a correct structure from such an imperfect template. We perform this test with CASP15 target H1142, an antigen-antibody complex with stoichiometry A1B1. An antigen-antibody complex is a good test case, as AlphaFold is relatively weak in modelling these interactions (3), likely due to a lack of evolutionary coupling signal in the MSA (21).
We take AlphaFold predictions for each chain from incorrect models of the complex that were submitted by NBIS-AF2-multimer and position them in a roughly correct manner by superimposing each prediction to the corresponding unit in the experimental structure. Since the chains were extracted from incorrect predictions, even when they are positioned correctly with respect to each other, the interface is incorrect, with some clashes present (Fig. 5a). We input this imperfect template into AF_unmasked while clipping the MSAs to a single sequence and predict 500 structures for each test. Results show that, when using both intra-chain and cross-chain restraints, the clash is fixed but the interface is not perfect. In a second test, we turn off intra-chain restraints while keeping only the cross-chain ones active. The idea behind this test is to allow AlphaFold to rearrange each monomer wherever necessary while keeping the distances between the chains within the boundaries of the template. This results in a more diverse set of predictions (Fig. 5c), with a much-improved interface. But the best result is obtained by deleting the clashing loop from the templates and letting AlphaFold inpaint it, which results in even more diverse predictions where the prediction with highest ranking confidence is also the best overall model (Fig. 5b).

Inpainting of very large structures. A known limitation of AlphaFold is its capability to generate models for large proteins, mostly due to computational limitations in terms of GPU memory. This is, of course, a problem since many interesting protein complexes are large. For example, CASP15 target H1111 is a 27-mer with stoichiometry A9B9C9 and 8460 amino acids in total. DeepMind, who modelled the

DRAFT

complex post-CASP15 on a more efficient version (v2.3) of AlphaFold than was available to the public, could not perform the modelling of target H1111 in one go and assembled multiple structures with A3B3C3 stoichiometry instead by using a template (PDB ID: $7ALW$) as guide (19). Here, we show that it is possible to overcome this limitation with AF_unmasked while limiting the depth of the final MSA to a maximum of 512 total sequences.

We use again the deposited structure itself (PDB ID: $7QIJ$) as template. This is a harder prediction task compared to H1137, since the first 362 residues of the largest subunits (C) from the membrane-bound domain in the complex are not in the deposited structure, so AlphaFold has to inpaint this gap leveraging the evolutionary information coming from the MSA inputs. We generated 25 structure following this protocol, which takes around 10h of GPU time per predicted structure (NVIDIA A100, 80GB RAM), and select the top three models by ranking confidence. As we can see in Fig. 4, the portion of the structure that is covered by the template stays the same across the three models, while the inpainted membrane region (in green) appears in a variety of conformations, from closed to open. This is, to the best of our knowledge, the largest structure ever generated in one shot using AlphaFold.

Predicting the impact of mutations. AlphaFold is not trained to predict the effect of mutations on the folding of a protein, and it cannot predict the impact of single-point mutations on protein stability (22). This might be due to the fact that few mutations on a target sequence result in virtually identical MSAs being used as input, which might mislead the neural network into enforcing incorrect restraints. If the effect of a mutation is limited to a relatively small area of the interaction, it might be possible to delete that part of the template, and let AlphaFold inpaint the gap. In this case, it might even be beneficial to remove all evolutionary information and let AlphaFold sample the space of conformations in the area that is not covered by the template.

T1110o and T1109o are two closely related homodimeric CASP15 targets. They are, respectively, a wild-type and a mutant sequence of the isocyanide hydratase, where the single-point mutation $D183A$ causes a rearrangement of the C-terminus loops at the interface, as shown in Fig 6a/b (cyan ribbons). To test whether AlphaFold is capable of correctly switching between the two loop configurations when sampling is allowed only around the region of interest, we use a structural template from the PDB where the loops in question are missing (PDB ID: $4K2H$). The sequence identity between the target and template sequences is below 20%, so we align target and template structurally with TM-align. Again, we clip the input MSAs to a single sequence, which means that AlphaFold should follow the template wherever possible, and attempt to model the loops "ab initio" since neither structural nor evolutionary inputs are given in that area of the structure.

Results show that for both T1110o (wildtype) and T1109o (mutant), AF_unmasked correctly rearranges of the loops in the model with highest ranking confidence. In Fig 6d, we compare DockQ scores for the top 10 models by ranking confidence against the models submitted by Wallner at CASP15. The top-ranked model for the mutant is better (DockQ: 0.804) than the first-ranked Wallner model (DockQ: 0.264) as well as the best among the top 5 models (DockQ: 0.76). We also test the default intra-chain constraint setting when using the same template and find that none of the top 5 models beat the new template strategy, while two correct models are in the top 10 (best DockQ: 0.803). Overall, when using $4K2H$ as the only template, 5 out of 500 models show a correct terminal loop arrangement (DockQ > 0.5) when the default (intra-chain only) template information is used, while 36 out of 500 models obtained with the new template strategy were correct. This suggests that including cross-template information puts AlphaFold closer to the correct solution, thus allowing for better sampling of the remaining space of configurations.

In Fig 6c, we compare the models obtained for T1110o on the new template strategy against those submitted by Wallner in CASP15 and see that even in this case, the top selected model is also the one with the highest DockQ (Fig 6a), but two of the top Wallner models have higher DockQ, which makes sense when considering that the structural templates used by Wallner to build these models have the correct (wildtype-like) loop arrangement (e.g. PDB: $6NI5$, $7L9Z$).

Cryo-EM test cases. In each of the biological systems we used, Rubisco, ClpB and Neurofibromin, we aimed at inpainting missing regions and identifying areas of possible conformational variability. We interpreted the results based on previous structural and biophysical knowledge. Tens to hundreds of models were generated with AF_unmasked for each of the cases analysed, but we only display some representative examples for ease of description.

Rubisco. Rubisco plays a crucial role in CO2 fixation, making it responsible for the majority of organic carbon in the biosphere. Understanding the function and control of Rubisco remains a significant area of research, with the aim of enhancing photosynthesis efficiency in agriculture and green biotechnology. The most prevalent form of Rubisco (Form I) comprises eight large and eight small subunits, and it exists in plants, algae, and other organisms. Although the active sites of Rubisco are situated on the large subunits, the expression of the small subunit regulates the size of the Rubisco pool in plants and can impact the overall catalytic efficiency of the Rubisco complex. For this reason the small subunit is a potential target for bioengineering and biochemical studies have been performed to generate chimeras of large and small Rubisco subunits that could enhance Rubisco's performance (23).

We used AF_ummasked to predict a chimera of Rubisco composed of Arabidopsis thaliana's (24) large subunits and another organism's small subunits. We have a cryo-EM reconstruction available for this chimera and the current AlphaFold multimer version does not agree with the experimental data. Therefore, we intended to test whether AF_unmasked can predict the right arrangement of the small subunit. In order to test this, we provided the Arabidopsis

DRAFT

structure (PDB code 5IU0) as template while giving the small subunits sequence from the other organism. One loop of 20 amino acids where the initial AlphaFold predictions and the cryo-EM data where in disagreement was deleted from the small subunits in the PDB template and we let AF_unmasked inpaint this region. Results (Fig. 7) show that the inpainted models are in good agreement with the cryo-EM density. Additionally, this first inpainted model could be used further in steps of model refinement onto the cryo-EM map.

ClpB. The bacterial chaperone ClpB is able to recover proteins from large aggregates and, together with the cognate DnaKJ system (25), to refold them into their active form. ClpB plays a pivotal role in protein homeostasis of bacterial cells (reviewed in (26, 27)). ClpB works by using ATP hydrolysis power to thread aggregate into a channel made upon oligomerisation of six ClpB copies (28). A common substrate used to study ClpB activity is casein. ClpB is therefore a very dynamic complex (29, 30) that needs to recognise and then move to unfold the aggregated substrate.

A wealth of cryo-EM structural information about the ClpB hexamer is available (31–34), but large domains, such as the N-termini, are not visible in the cryo-EM densities and highly-mobile regions are poorly defined.

We gave AF_unmasked a template derived from cryo-EM data (PDB ID: $5OG1$) and performed inpainting of the missing regions. Results show that, not only AF_unmasked could inpaint the missing N-termini, but it also predicted them in multiple conformations. It has to be noted that crystal structures of the N-teminus of ClpB, both isolated in the monomer context (35, 36), are available (PDB ID: $1QVR$, $1KHY$) and were likely included in training set of AlphaFold.

Additionally, missing parts of another region were inpainted, the coiled-coil M-domain, known to be important in the regulation of ClpB and in its interaction with DnaKJ (37–39). In structural and mutational studies, the M-domain has been shown to adopt different conformations (40, 41). These are picked up in the inpainted models, together with intermediate conformational steps, showing tilting of the long coiled-coil in ways that are plausible given the existing structural information and single molecule FRET-spectroscopy data (40, 41) (Fig. 8a,b).

To investigate the ability of predicting the interaction between ClpB and the commonly used substrate casein, we used a PDB template (PDB ID: $6RN3$) where a stretch of amino acids from casein is engaged inside the ClpB pore. While most of the generated models fail to dock the casein inside ClpB (Fig. 8e), the top models by ranking confidence do show casein engaged inside the pore, and in one model it is in contact with one of the six N-termini (Fig. 8c, d). Interestingly, the interaction between casein and N-terminus is predicted via hydrophobic patches (Fig. 8d) and the interacting complexes have better pLDDT score than the ones that are not engaged with casein (Fig. 8e).

Neurofibromin. Neurofibromin is a downregulator of the oncogenic protein Ras and is ubiquitously expressed in the central neural system (42). Neurofibromin (Nf1) plays therefore a very important role in tumor growth and its mutations cause the pleomorphic disease neurofibromatosis type 1 (43) and are found up to 10% of all cancers (44).

Nf1 is an homodimer of around 600KDa and has a unique oligomeric arrangement made of a bi-lobate platform, composed of ARM and HEAT repeats (45–48). The Ras-binding domain GRD and the membrane-binding domain Sec14-PH are anchored to this helical platform by long loops that allow large movements of these domains (45, 47). Cryo-EM structural studies (45, 47, 48) show two main conformations of the GRD and Sec14-PH domain that go from a so-called closed auto-inhibited conformation to an open conformation, which can bind Ras.

The standard version of AlphaFold fails to find the right conformation of both monomeric and dimeric Nf1 arrangement. In order to test whether AF_unmasked could reproduce the movements of GRD and Sec14-PH that are visible in the experimental data, we decided to remove these domains from the available deposited PDB structure (PDB ID: $7PGU$) and perform inpainting on them. Given the considerable size of the inpainted domains, as well as the considerable scale of the movements involving these domains, we perform additional sampling and produce roughly 1000 models. Results show that AF_unmasked, using the bilobate helical part of the complex as template, was able to inpaint the missing GRD and Sec14-PH domains (Fig. 9a) as well as all the loops missing in the cryo-EM models (Fig. 9b).

However, possibly due to the complexity of this inpainting task and the fact that the closed, auto-inhibited conformation of Nf1 is stabilised by a coordinated Zn atom that cannot be taken in consideration here (Fig. 9b), none of the predicted models captures either the closed or the open conformations as seen in the deposited structures. On the other hand, the predicted conformations of GRD-Sec14-PH could be intermediate states of opening from the closed to the open Nf1 conformations found experimentally in the cryo-EM studies. Indeed, the structures predicted this way can be used to aid in generating a morphing between the closed and open positions of GRD and Sec14-PH.

Interestingly, depending on the neural network model, different types of variability are picked. For example, we were also able to model different bending states of the helical platform (Fig. 9d) that are in good agreement with the 3D variability analysis performed in cryo-EM. Some of the lower scoring dimers are similar to those obtained with the standard version of AlphaFold.

Discussion

AlphaFold has had a huge impact on the field of structural biology, but both computational and structural biologists are yet to maximise its potential, especially when it comes to modelling large protein complexes.

Approaches followed by computational biologists, providing the Evoformer module with better or more diverse sets of MSAs or performing more sampling, are promising avenues of research but only focus on the evolutionary aspect of the prediction problem while overlooking templates. But MSAs

DRAFT

and templates are complementary sources of information: the first allows to derive evolution-driven restraints between pairs of amino acids, while the second is a direct observation of 3D structures.

If the prediction task in AlphaFold is like searching through a learned energy potential in order to find the bottom of the funnel, it makes sense that restraining the search through the use of experimental models will greatly reduce the complexity of the search. Thus, a good template should help in finding the bottom of an energy funnel just as well as a deep and diverse MSA. This is especially true when it comes to protein complexes, especially larger ones or those containing a large number of protein chains, where the evolutionary signal linking different protein chains is hard or impossible to detect (e.g. in host-pathogen, antigen-antibody interactions) or where protein-protein interactions for the same polypeptide chains varies over cycles of mechanical activities. In this study, we theorise that a better integration between experimental data and predictions can be made possible by improving the template mechanism that is already present within AlphaFold. We therefore present a simple, easy to use and effective tool to integrate experiments and predictions by hijacking the mechanisms to embed structural templates within AlphaFold. We call this method AF_unmasked.

We tested AF_unmasked through a series of use cases and showed that it is effective not only when the templates are perfect (H1137), but even when they consist of rough models built from unbound monomers where interfacial amino acids are clashing (H1142), or when stretches of amino acids and even large domains, are missing in the initial structure (H1111, Rubisco, ClpB and Neurofibromin), or when a mutation causes a conformational change (T1110o, T1109o). We could also perform "structural inpainting", i.e. letting AlphaFold use evolutionary information from the MSA to fill in gaps in experimental structures (H1111, Rubisco, ClpB and Neurofibromin). In all cases, AlphaFold's own predicted quality score was able to recognise that predictions based on the right templates were better than those that were not, a further confirmation that even in cases where AlphaFold cannot produce good models on its own, it will be able to generate and recognise those models when the user can nudge it in the right direction. When inpainting large domains, we showed that AF_unmasked is able to sample the space of conformations more efficiently and predict said domains in multiple conformations. These conformations are very interesting for the interpretation of low-to-medium resolution cryo-EM and cryo-ET structures as well as for very flexible domains. We envision that intermediates states picked up by AF_unmasked could be of use for deducing correct movement paths underlying large domains' motions. From a general structural biology perspective, AF_unmasked allows to augment the available structural information in a number of ways (e.g. loops building, domain conformational variability) that can work as baseline for the design of novel biochemical and biophysical experiments.

Another important aspect of AF_unmasked is that it finally makes it possible to predict very large protein models in one go, since leveraging experimental data allows to greatly reduce the number of sequences in the MSA, thereby greatly reducing requirements in terms of GPU memory and computational time. This allowed for the first time to accurately predict the structure of very large protein complexes (up to 27 chains and 8460 amino acids) without the need for external tools to perform docking of single subunits.

We believe that AF_unmasked, which is open source and easy to install and use, will be a valuable addition to the toolbox of all structural biologists.

Data Availability

The code, installation instructions and examples are available on GitHub: github.com/clami66/AF_unmasked

All the predicted structures, along with log files and AlphaFold inputs, are publicly available: 10.17044/scilifelab.24198669

ACKNOWLEDGEMENTS

This work was done under the *BeyondFold* Technology Development Project at the Science for Life Laboratory (SciLifeLab). CM is financially supported by the Knut and Alice Wallenberg (KAW) Foundation as part of the National Bioinformatics Infrastructure Sweden at SciLifeLab. SA is financially supported by the Knut and Alice Wallenberg (KAW) Foundation grant KAW2021.0347 to MC. MC is financially supported by SciLifeLab as part of the Cryo-EM National Infrastructure Unit. BW is financially supported by KAW as part of the WASP-DDLS joint program. Computations were performed at NSC Tetralith/BerzeLiUs provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS) and the Swedish National Infrastructure for Computing (SNIC), partially funded by the Swedish Research Council through grant agreements no. 2022-06725 and no. 2018-05973.

We would also like to thank Dr. Nicholas Pearce and Dr. Yogesh Kalakoti for useful feedback throughout the study and during the development of manuscript and code.

Bibliography

1. John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.
2. James P Roney and Sergey Ovchinnikov. State-of-the-art estimation of protein model accuracy using AlphaFold. *Physical Review Letters*, 129(23):238101, 2022.
3. Richard Evans, Michael O'Neill, Alexander Pritzel, Natasha Antropova, Andrew Senior, Tim Green, Augustin Žídek, Russ Bates, Sam Blackwell, Jason Yim, et al. Protein complex prediction with AlphaFold-Multimer. *BioRxiv*, pages 2021–10, 2021.
4. Björn Wallner. AFsample: Improving multimer prediction with alphafold using massive sampling. *Bioinformatics*, page n/a, 2023.
5. CASP15: Book of abstracts. `https://predictioncenter.org/casp15/doc/CASP15_Abstracts.pdf`, 2022. Accessed: 2023-06-27.
6. Anastassis Perrakis and Titia K Sixma. AI revolutions in biology: The joys and perils of AlphaFold. *EMBO reports*, 22(11):e54046, 2021.
7. Thomas C Terwilliger, Dorothy L Leibschner, Tristan Croll, Christopher J Williams, Airlie J McCoy, Billy K Poon, Pavel Afonine, Robert D Oeffner, Jane Shelby Richardson, Randy J Read, et al. AlphaFold predictions are valuable hypotheses, and accelerate but do not replace experimental structure determination. *bioRxiv*, pages 2022–11, 2022.
8. Dorothee Liebschner. Using predicted models in Phenix. *Acta Cryst*, 75:861–877, 2019.
9. Thomas C Terwilliger, Pavel V Afonine, Dorothee Liebschner, Tristan I Croll, Airlie J McCoy, Robert D Oeffner, Christopher J Williams, Billy K Poon, Jane S Richardson, Randy J Read, et al. Accelerating crystal structure determination with iterative AlphaFold prediction. *Acta Crystallographica Section D: Structural Biology*, 79(3), 2023.
10. Thomas C Terwilliger, Billy K Poon, Pavel V Afonine, Christopher J Schlicksup, Tristan I Croll, Claudia Millán, Jane S Richardson, Randy J Read, and Paul D Adams. Improved AlphaFold modeling with implicit experimental information. *Nature Methods*, pages 1–7, 2022.
11. Gustaf Ahdritz, Nazim Bouatta, Sachin Kadyan, Qinghui Xia, William Gerecke, Timothy J O'Donnell, Daniel Berenberg, Ian Fisk, Niccolò Zanichelli, Bo Zhang, Arkadiusz Nowaczynski, Bei Wang, Marta M Stepniewska-Dziubinska, Shang Zhang, Adegoke Ojewole, Murat Efe Guney, Stella Biderman, Andrew M Watkins, Stephen Ra, Pablo Ribalta Lorenzo, Lucas Nivon, Brian Weitzner, Yih-En Andrew Ban, Peter K Sorger, Emad Mostaque, Zhao Zhang, Richard Bonneau, and Mohammed AlQuraishi. OpenFold: Retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization. *bioRxiv*, 2022. doi: 10.1101/2022.11.20.517210.
12. Ziyao Li, Xuyang Liu, Weijie Chen, Fan Shen, Hangrui Bi, Guolin Ke, and Linfeng Zhang. Uni-fold: An open-source platform for developing protein folding models beyond alphafold. *bioRxiv*, 2022. doi: 10.1101/2022.08.04.502811.

DRAFT

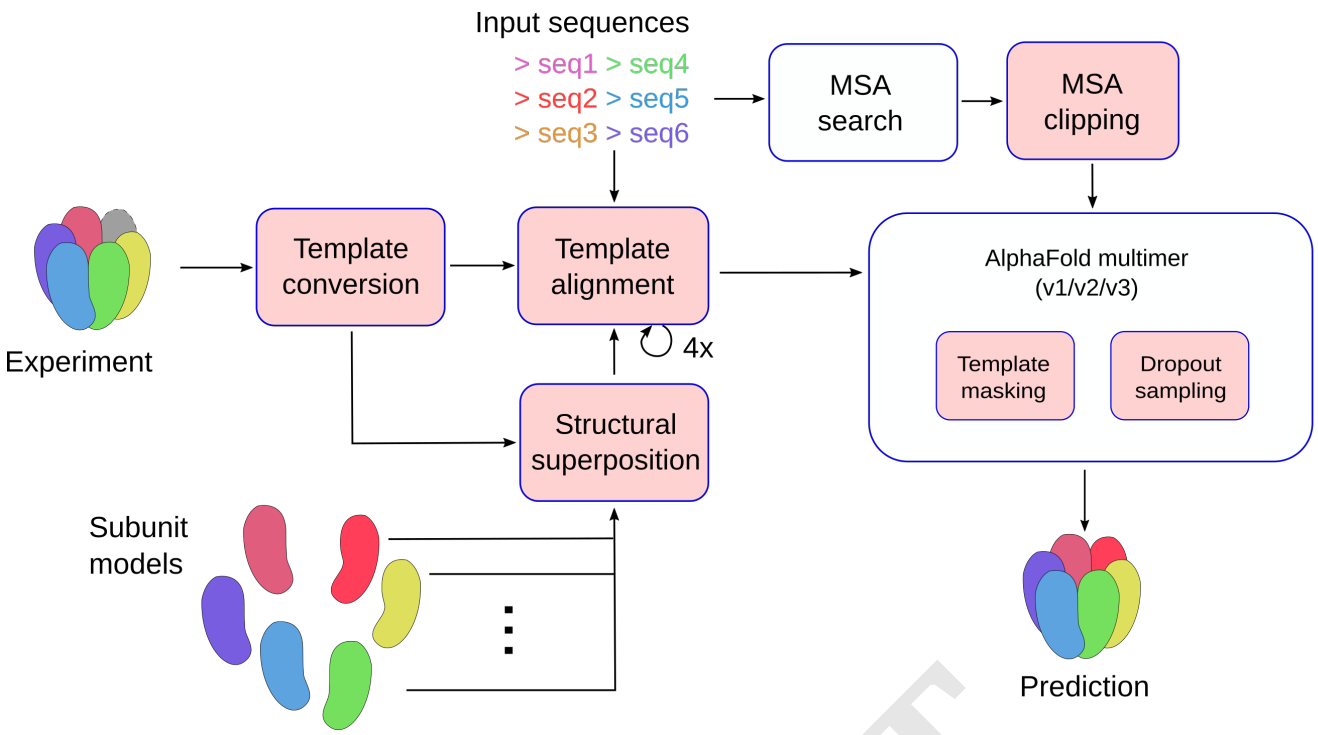

Fig. 1. Flowchart of AF_unmasked. Parts that differ from the original AlphaFold implementation are highlighted in pink. Structural data coming from an experiment is converted to the correct format (mmCIF) if necessary, then the template is aligned against the target sequences or against a set of target monomeric structures so that the template coordinates are applied to the right amino acids. The templates built this way are pushed through AlphaFold along with the evolutionary inputs (MSAs), and the cross-chain masking is turned off to enforce the desired complex conformation. In order to tune how much evolutionary or structural inputs should influence the prediction, it is possible to filter each MSA so that only a subset of alignments is used, and one to four templates can be used, either by reusing the same template or by submitting different templates. Dropout is activated at inference time to encourage sampling diversity.

DRAFT

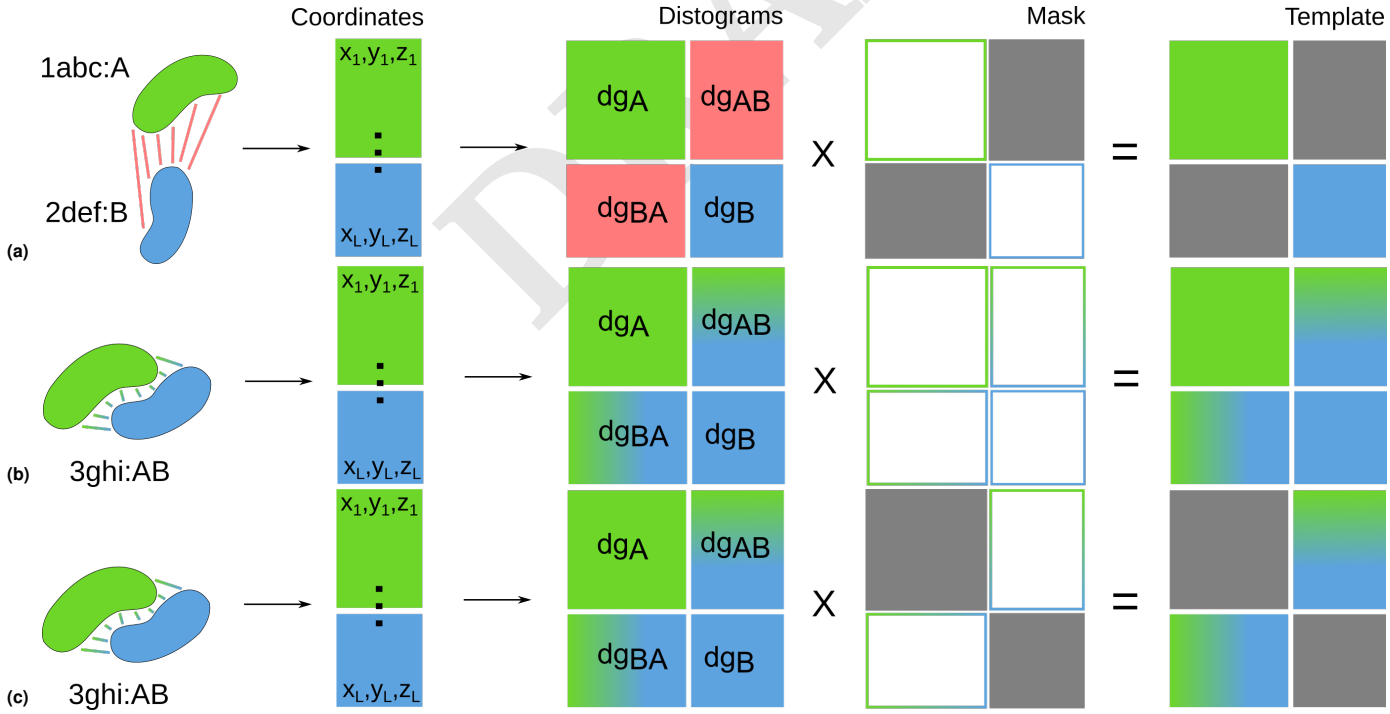

Fig. 2. (a) Each template in AlphaFold is a collection of coordinates extracted from a set of PDB structures, one for each chain. These templates are merged and distances between all atoms are calculated in the form of distograms. Information about distances across chains is masked out since they are calculated from protein chains that do not share a common reference frame. (b) We propose to disable the masking and unlock the cross-chain distances whenever a multimeric template is used in AlphaFold, where the user wants to enforce distances across chains as well as within chains. (c) Sometimes it might also be desirable to only consider cross-chains distances from templates, and discard intra-chain distances. This might be useful whenever the evolutionary information coming from MSAs is enough to correctly fold each monomer, but the MSA pairing step does not allow to extract the right cross-chain restraints.

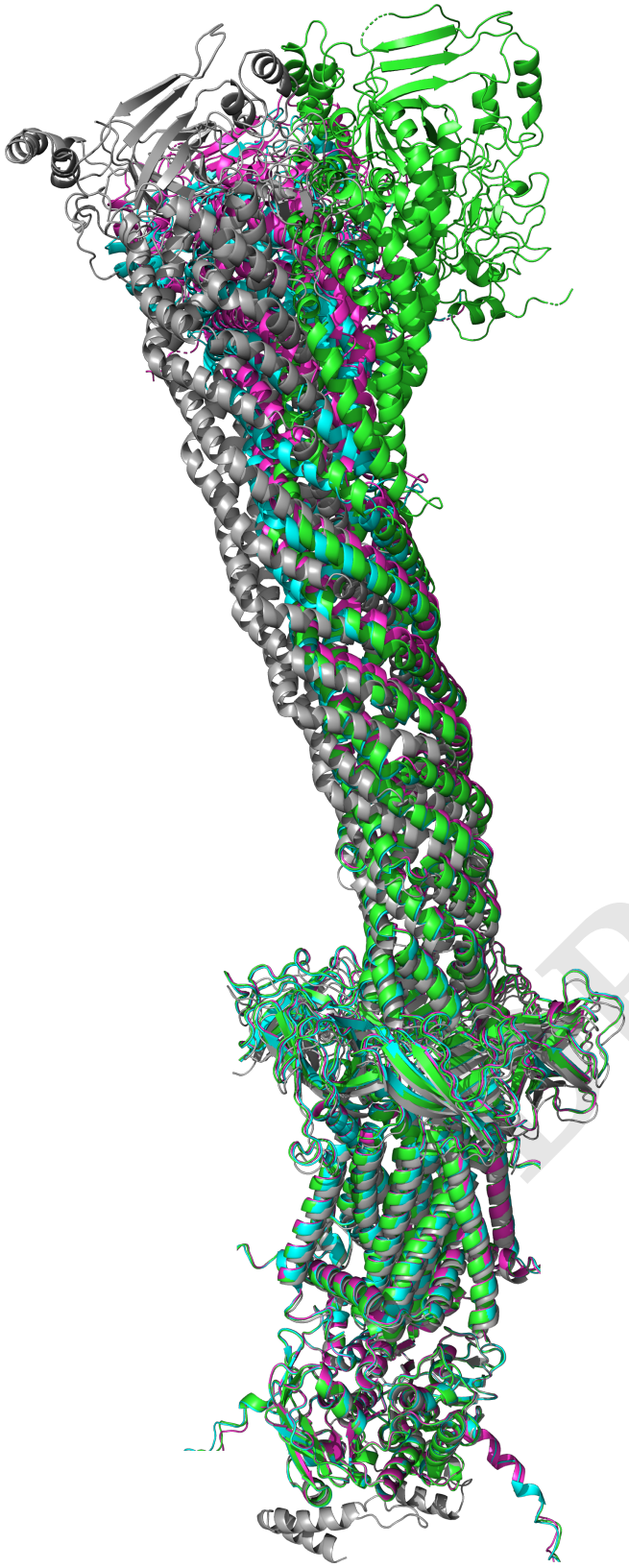

Fig. 3. CASP15 target H1137 was modelled by using the native (PDB ID: $8FEF$) as template and all MSAs were reduced to a single sequence, i.e. the target alone. Three predictions (in color) are superimposed to the native (grey), unstructured predicted regions not in the deposited model were hidden.

13. Kolja Stahl, Oliver Brock, and Juri Rappsilber. Modelling protein complexes with crosslinking mass spectrometry and deep learning. *bioRxiv*, pages 2023–06, 2023.
14. Patrick Bryant, Gabriele Pozzati, Wensi Zhu, Aditi Shenoy, Petras Kundrotas, and Arne Elofsson. Predicting the structure of large protein complexes using alphafold and monte carlo tree search. *Nature communications*, 13(1):6028, 2022.
15. Yang Zhang and Jeffrey Skolnick. TM-align: a protein structure alignment algorithm based on the tm-score. *Nucleic acids research*, 33(7):2302–2309, 2005.
16. Claudio Mirabello. lDDT-align: A tool to align protein structures while maximizing lddt. `https://github.com/clami66/lDDT_align`, 2022. Accessed: 2023-06-27.
17. Jue Wang, Sidney Lisanza, David Juergens, Doug Tischer, Joseph L Watson, Karla M Castro, Robert Ragotte, Amijai Saragovi, Lukas F Milles, Minkyung Baek, et al. Scaffolding protein functional sites using deep learning. *Science*, 377(6604):387–394, 2022.
18. Björn Wallner. Improved multimer prediction using massive sampling with alphafold in casp15. *Proteins: Structure, Function, and Bioinformatics*, n/a(n/a), 2023. doi: https://doi.org/10.1002/prot.26562.
19. AlphaFold v2.3.0 release notes and CASP15 models. `https://github.com/deepmind/alphafold/blob/main/docs/technical_note_v2.3.0.md`, 2022. Accessed: 2023-06-27.
20. Sankar Basu and Björn Wallner. DockQ: a quality measure for protein-protein docking models. *PloS one*, 11(8):e0161879, 2016.
21. Rui Yin and Brian G Pierce. Evaluation of alphafold antibody-antigen modeling with implications for improving predictive accuracy. *bioRxiv*, pages 2023–07, 2023.
22. Marina A Pak, Karina A Markhieva, Mariia S Novikova, Dmitry S Petrov, Ilya S Vorobyev, Ekaterina S Maksimova, Fyodor A Kondrashov, and Dmitry N Ivankov. Using AlphaFold to predict the impact of single mutations on protein stability and function. *Plos one*, 18(3): e0282689, 2023.
23. Yuwei Mao, Ella Catherall, Aranzazú Díaz-Ramos, George R L Greiff, Stavros Azinas, Laura Gunn, and Alistair J McCormick. The small subunit of rubisco and its potential as an engineering target. *J. Exp. Bot.*, 74(2):543–561, January 2023.
24. Karin Valegård, Dirk Hasse, Inger Andersson, and Laura H Gunn. Structure of rubisco from arabidopsis thaliana in complex with 2-carboxyarabinitol-1,5-bisphosphate. *Acta Crystallogr D Struct Biol*, 74(Pt 1):1–9, January 2018.
25. Yanting Yin, Xiang Feng, Hongjun Yu, Allison Fay, Amanda Kovach, Michael S Glickman, and Huilin Li. Structural basis for aggregate dissolution and refolding by the mycobacterium tuberculosis ClpB-DnaK bi-chaperone system. *Cell Rep.*, 35(8):109166, May 2021.
26. Sabina Kędzierska-Mieszkowska and Michal Zolkiewski. Hsp100 molecular chaperone clpb and its role in virulence of bacterial pathogens. *International Journal of Molecular Sciences*, 22(10):5319, 2021.
27. Panagiotis Katikaridis, Valentin Bohl, and Axel Mogk. Resisting the heat. 2021.
28. Mario J Avellaneda, Kamila B Franke, Vanda Sunderlikova, Bernd Bukau, Axel Mogk, and Sander J Tans. Processive extrusion of polypeptide loops by a hsp100 disaggregase. *Nature*, 578(7794):317–320, 2020.
29. Hisham Mazal, Marija Iljina, Inbal Riven, and Gilad Haran. Ultrafast pore-loop dynamics in a AAA+ machine point to a brownian-ratchet mechanism for protein translocation. *Sci Adv*, 7(36):eabg4674, September 2021.
30. Takayuki Uchihashi, Yo-Hei Watanabe, Yosuke Nakazaki, Takashi Yamasaki, Hiroki Watanabe, Takahiro Maruno, Kentaro Ishii, Susumu Uchiyama, Chihong Song, Kazuyoshi Murata, Ryota Iino, and Toshio Ando. Dynamic structural states of ClpB involved in its disaggregation function. *Nat. Commun.*, 9(1):2147, June 2018.
31. Célia Deville, Kamila Franke, Axel Mogk, Bernd Bukau, and Helen R Saibil. Two-Step activation mechanism of the ClpB disaggregase for sequential substrate threading by the main ATPase motor. *Cell Rep.*, 27(12):3433–3446.e4, June 2019.
32. Alexandrea N Rizo, Jiabei Lin, Stephanie N Gates, Eric Tse, Stephen M Bart, Laura M Castellano, Frank DiMaio, James Shorter, and Daniel R Southworth. Structural basis for substrate gripping and translocation by the ClpB AAA+ disaggregase. *Nat. Commun.*, 10 (1):1–12, June 2019.
33. Hongjun Yu, Tania J Lupoli, Amanda Kovach, Xing Meng, Gongpu Zhao, Carl F Nathan, and Huilin Li. ATP hydrolysis-coupled peptide translocation mechanism of *Mycobacterium tuberculosis* ClpB. *Proceedings of the National Academy of Sciences*, 115(41):E9560–E9569, 2018.
34. Célia Deville, Marta Carroni, Kamila B Franke, Maya Topf, Bernd Bukau, Axel Mogk, and Helen R Saibil. Structural pathway of regulated substrate transfer and threading through an hsp100 disaggregase. *Science Advances*, 3(8):e1701726, 2017.
35. Sukyeong Lee, Mathew E Sowa, Yo-Hei Watanabe, Paul B Sigler, Wah Chiu, Masasuke Yoshida, and Francis T F Tsai. The structure of ClpB: a molecular chaperone that rescues proteins from an aggregated state. *Cell*, 115(2):229–240, October 2003.
36. Cathleen Zeymer, Thomas R M Barends, Nicolas D Werbeck, Ilme Schlichting, and Jochen Reinstein. Elements in nucleotide sensing and hydrolysis of the AAA+ disaggregation machine ClpB: a structure-based mechanistic dissection of a molecular motor. *Acta Crystallogr. D Biol. Crystallogr.*, 70(Pt 2):582–595, February 2014.
37. Tobias Haslberger, Jimena Weibezahn, Regina Zahn, Sukyeong Lee, Francis T F Tsai, Bernd Bukau, and Axel Mogk. M domains couple the ClpB threading motor with the DnaK chaperone activity. *Mol. Cell*, 25(2):247–260, January 2007.
38. Yuki Oguchi, Eva Kummer, Fabian Seyffer, Mykhaylo Berynskyy, Benjamin Anstett, Regina Zahn, Rebecca C Wade, Axel Mogk, and Bernd Bukau. A tightly regulated molecular toggle controls AAA+ disaggregase. *Nat. Struct. Mol. Biol.*, 19(12):1338–1346, December 2012.
39. Rina Rosenzweig, Shoeib Moradi, Arash Zarrine-Afsar, John R Glover, and Lewis E Kay. Unraveling the mechanism of protein disaggregation through a ClpB-DnaK interaction. *Science*, 339(6123):1080–1083, March 2013.
40. Marta Carroni, Eva Kummer, Yuki Oguchi, Petra Wendler, Daniel K Clare, Irmgard Sinning, Jürgen Kopp, Axel Mogk, Bernd Bukau, and Helen R Saibil. Head-to-tail interactions of the coiled-coil domains regulate ClpB activity and cooperation with hsp70 in protein disaggregation. *Elife*, 3:e02481, April 2014.
41. Hisham Mazal, Marija Iljina, Yoav Barak, Nadav Elad, Rina Rosenzweig, Pierre Goloubinoff, Inbal Riven, and Gilad Haran. Tunable microsecond dynamics of an allosteric switch regulate the activity of a AAA+ disaggregation machine. *Nat. Commun.*, 10(1):1438, March

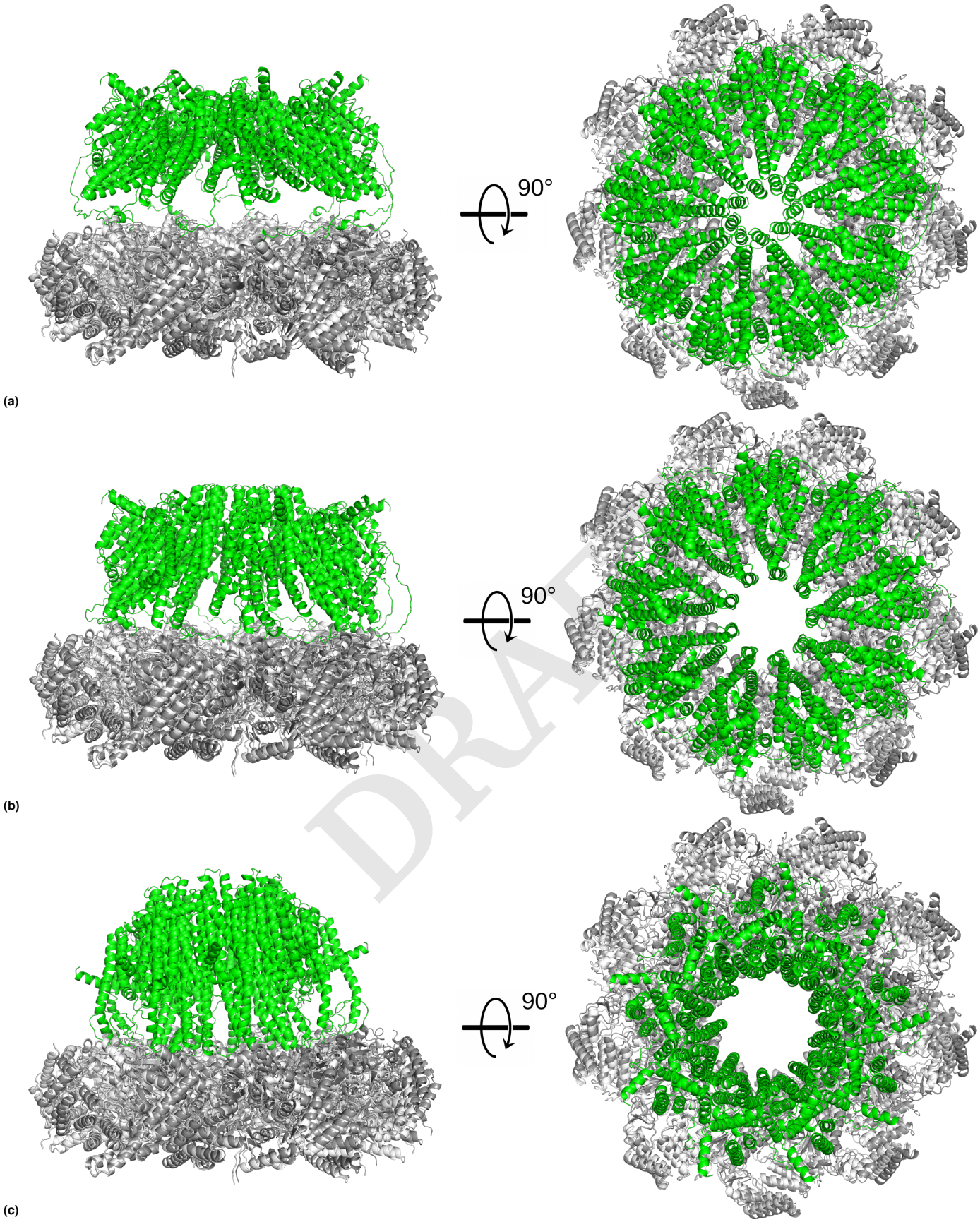

Fig. 4. CASP15 target H1111 is a very large complex (27 chains, 8460 amino acids) of a secretion export gate from *Yersinia enterocolitica*. We use the CASP15 native structure (PDB ID: $7QIJ$) as partial template (bottom ring in grey) to guide the assembly and let AlphaFold inpaint the trans-membrane region (green). The top three models by ranking confidence are all near-identical to the template in the area covered by it, while the trans-membrane region show diverse and potentially biologically relevant conformations from closed (a) to open (c).

DRAFT

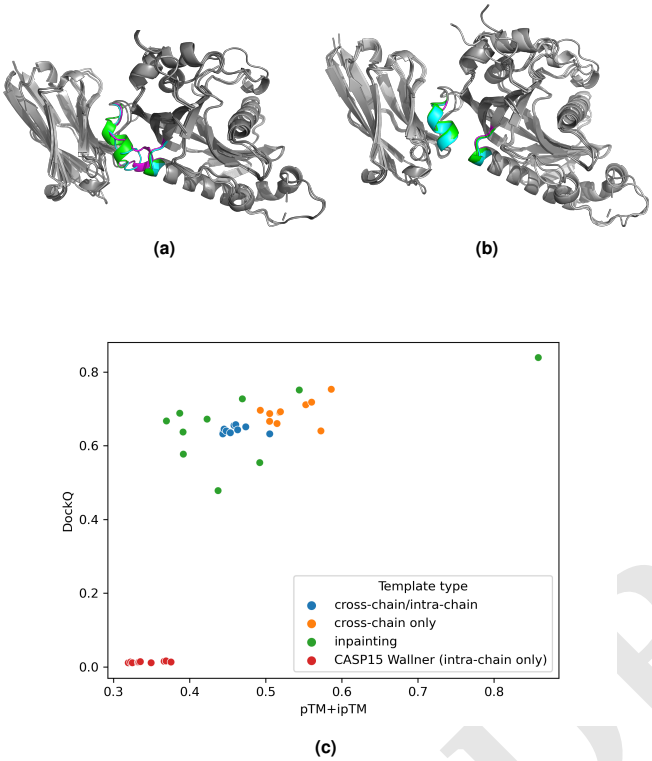

Fig. 5. CASP15 target H1142 is an antibody-antigen complex. The template was obtained by superimposing unbound structures from CASP15 predictions onto the native to simulate an imperfect template. In this case, some of the residues at the interface are clashing in the template (a, magenta). We test AF_unmasked by feeding this rough template as is (a) and using either the full set of cross- and intra-chain restraints (c, blue points) or cross-chain restraints alone (c, orange points). Then, we also cut the clashing interfacial residues (b, magenta) and performed inpainting while using again the full set of restraints (c, green points). The latter approach resulted in a seemingly more extensive sampling of the space of conformations, and the best model by ranking confidence was also the best model by DockQ.

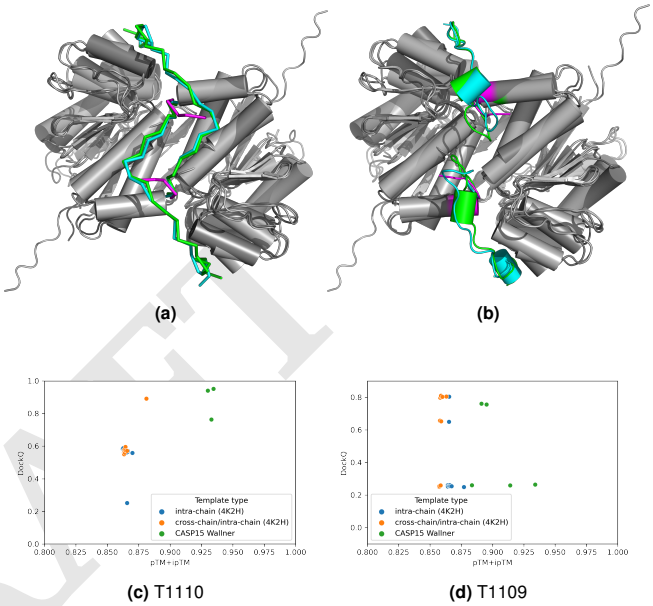

Fig. 6. CASP15 target T1110o is a homodimer of the isocyanide hydratase (a, cyan). Target T1109o is a mutant of T1110o where a single-point mutation causes a rearrangement of the C-termini (b, cyan). The template was obtained by homology against the PDB, and among a set of candidates we selected a template where the C-termini loops were missing entirely (a, b, template C-termini shown in magenta). We utilise this template as is, the mapping between target and template amino acid sequences was performed by structural superposition between unbound models and the template with TM-align. Using this template allows AF_unmasked to perform sampling of a number of different loop conformations. The top-ranking structures by confidence score show the correct loop arrangement both in T1109o (c, orange) and T1110o (d, orange) for mutant and wildtype sequences, while the default template strategy (blue dots) tend to assign to the mutant the same arrangement as in the wildtype.

DRAFT

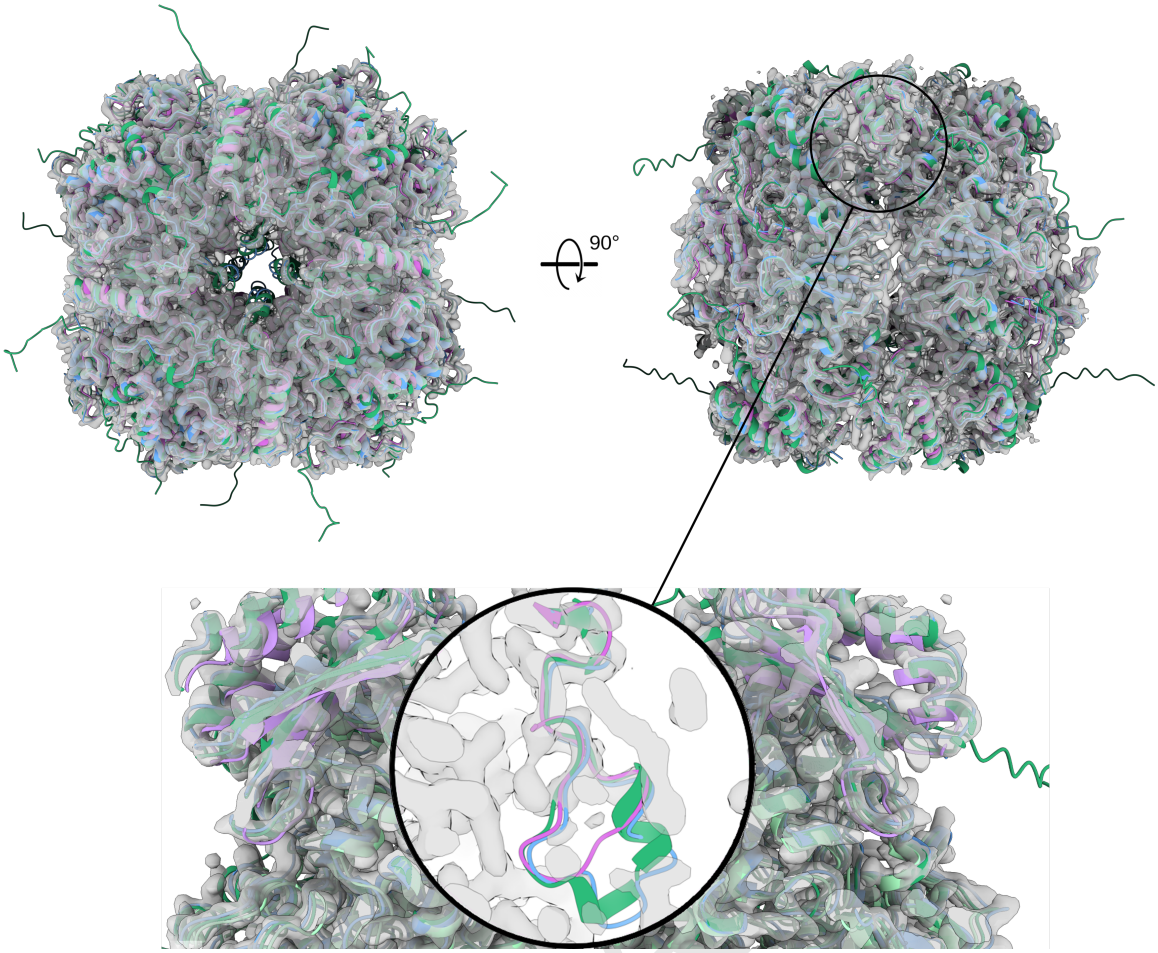

Fig. 7. CryoEM density map (grey) of an Arabidopsis thaliana Rubisco made with a hybrid small subunit with fitted atomic models. In green is the highest-scoring AF_unmasked prediction of a hexadecamer, in magenta the template used in the prediction step and in cyan the model as it was initially built and refined from the density alone. The circled area in the bottom panel shows a cross-section of the inner loop of the hybrid small subunit that was clipped from the multichain template so that AF_unmasked could inpaint it. The resulting inpainted model (green) is a closer match to the final refined model (cyan) when compared to the Arabidopsis thaliana Rubisco template (magenta).

2019.

42. Mohammed Bergoug, Michel Doudeau, Fabienne Godin, Christine Mosrin, Béatrice Vallée, and Hélène Bénédetti. Neurofibromin structure, functions and regulation. *Cells*, 9(11), October 2020.
43. D H Gutmann, D L Wood, and F S Collins. Identification of the neurofibromatosis type 1 gene product. *Proc. Natl. Acad. Sci. U. S. A.*, 88(21):9658–9662, November 1991.
44. Nancy Ratner and Shyra J Miller. A RASopathy gene commonly mutated in cancer: the neurofibromatosis type 1 tumour suppressor. *Nat. Rev. Cancer*, 15(5):290–301, May 2015.
45. Andreas Naschberger, Rozbeh Baradaran, Bernhard Rupp, and Marta Carroni. The structure of neurofibromin isoform 2 reveals different functional states. *Nature*, 599(7884):315–319, November 2021.
46. Christopher J Lupton, Charles Bayly-Jones, Laura D'Andrea, Cheng Huang, Ralf B Schittenhelm, Hari Venugopal, James C Whisstock, Michelle L Halls, and Andrew M Ellisdon. The cryo-EM structure of the human neurofibromin dimer reveals the molecular basis for neurofibromatosis type 1. *Nat. Struct. Mol. Biol.*, 28(12):982–988, December 2021.
47. Malik Chaker-Margot, Sebastiaan Werten, Theresia Dunzendorfer-Matt, Stefan Lechner, Angela Ruepp, Klaus Scheffzek, and Timm Maier. Structural basis of activation of the tumor suppressor protein neurofibromin. *Mol. Cell*, March 2022.
48. Lucy C Young, Ruby Goldstein de Salazar, Sae-Won Han, Zi Yi Stephanie Huang, Alan Merk, Matthew Drew, Joseph Darling, Vanessa Wall, Reinhard Grisshammer, Alice Cheng, Madeline R Allison, Matthew J Sale, Dwight V Nissley, Dominic Esposito, Jana Ognjenovic, and Frank McCormick. Destabilizing NF1 variants act in a dominant negative manner through neurofibromin dimerization. *Proc. Natl. Acad. Sci. U. S. A.*, 120(5):e2208960120, January 2023.

DRAFT

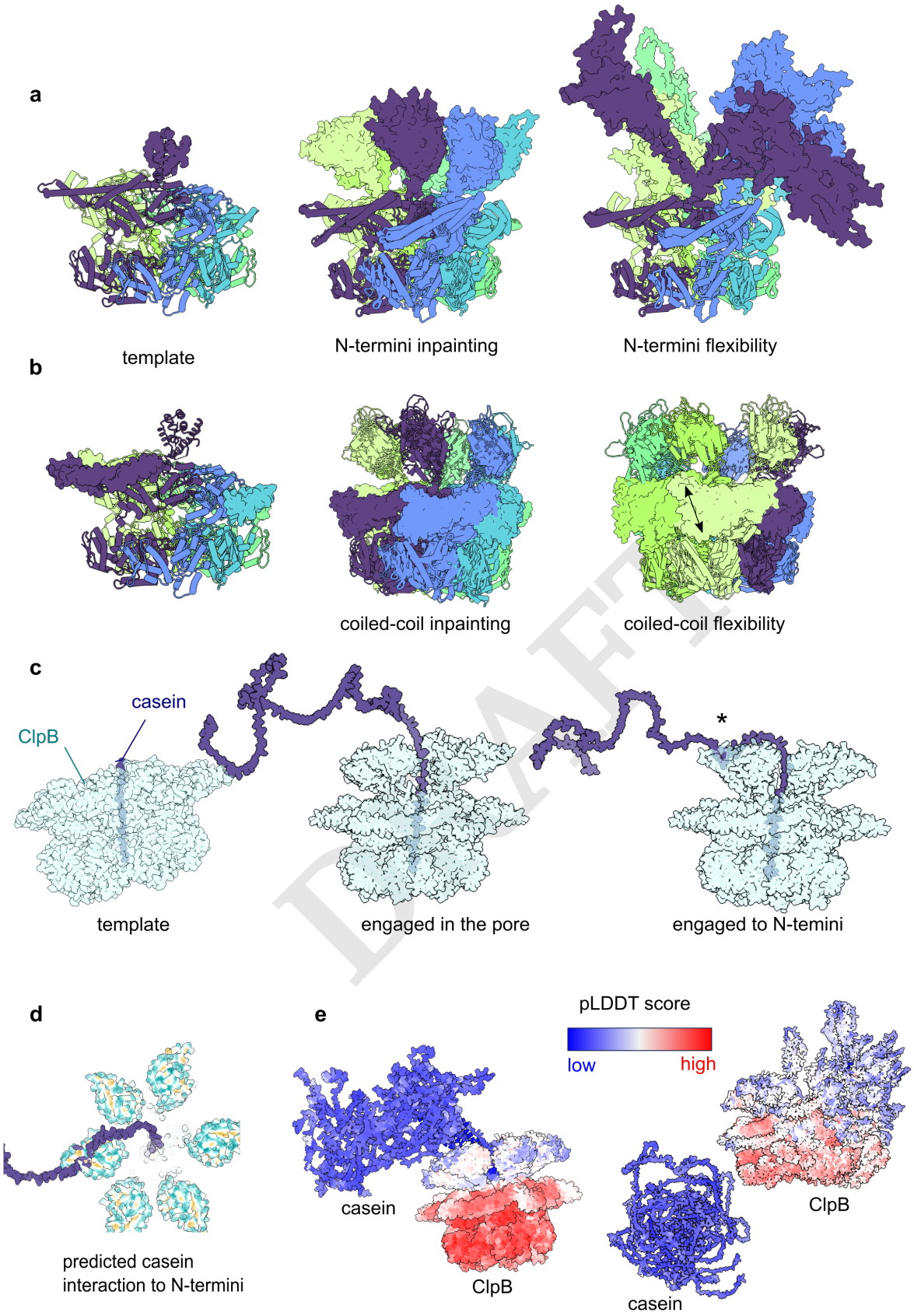

Fig. 8. Analysis of ClpB hexamer using AF_unmasked. (a) Given template and inpainted N-termini. N-termini are shown as surfaces while other domains are as cartoon. Each subunit of the hexamer is coloured differently. (b) Inpainting on the M-domains, shown as surfaces. The arrow shows the possible motion of the M-domain. (c) Inpainting of the interaction between ClpB and casein. The asterisk shows the newly predicted interaction area. (d) View of the hydrophobic regions of ClpB termini (in yellow) interacting with casein (e) Models of ClpB and casein and relative confidence scores.

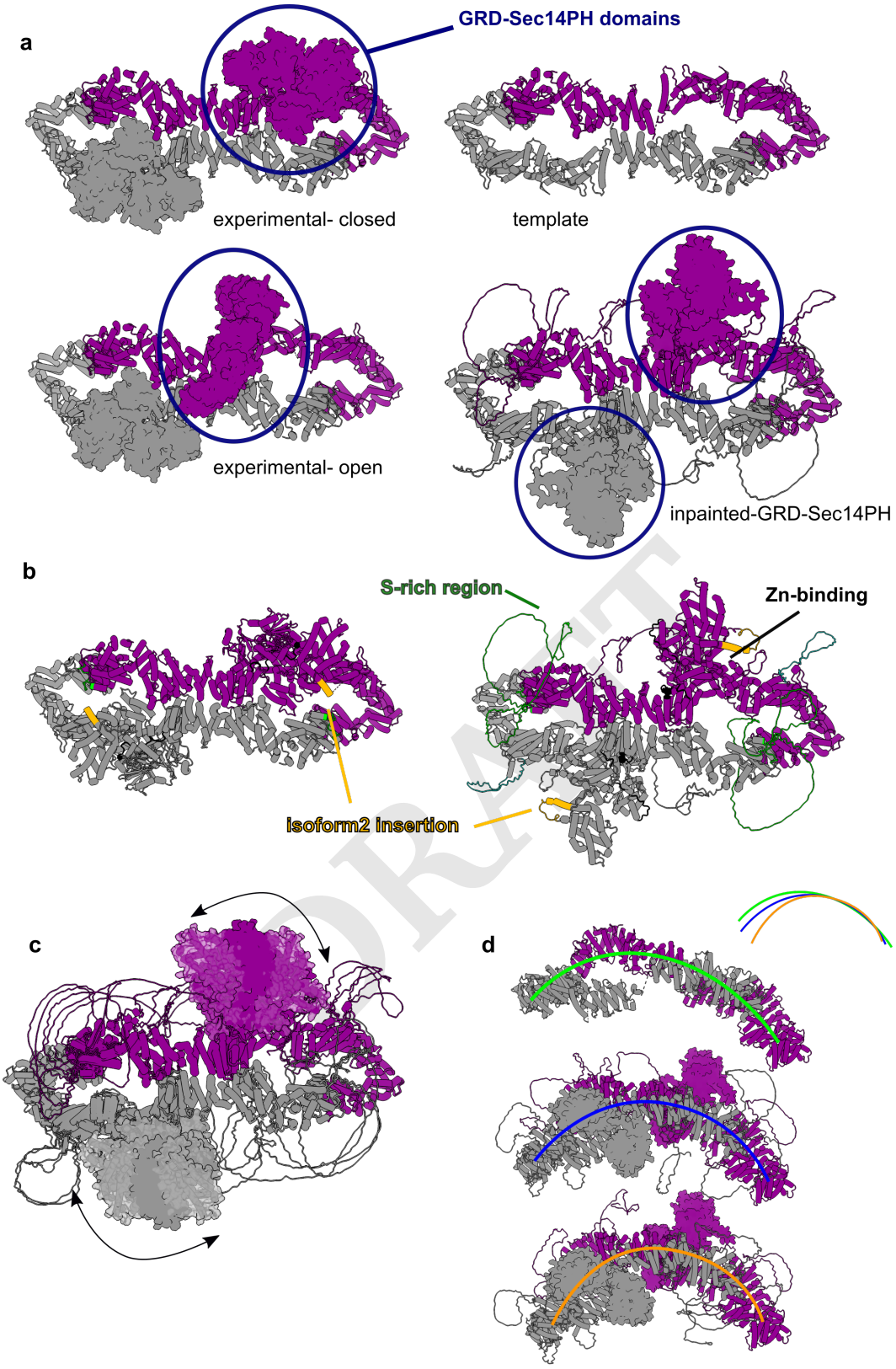

Fig. 9. Analysis of the Neurofibromin (Nf1) dimer using AF_unmasked. (a) Inpainting of the GRD and Sec14-PH domain of Nf1 in an intermediate conformation in between the experimentally observed closed and open states. (b) Comparison of the closed experimental conformation of Nf1 isoform 2 (on the left) with a AF_unmasked conformation. Important regions are highlighted. (c) Superimposition of several AF_unmasked predictions where the GRD and Sec14-PH domain were modelled in intermediate positions, suggesting a possible motion path for these domains. (d) Three different AF_unmasked predictions showing a bending of the helical Nf1 platform, also represented with differently coloured curved lines.