Soft disorder modulates the assembly path of protein complexes

Beatriz Seoane
Departamento de Física Teórica,
Universidad Complutense, 28040 Madrid, Spain

Alessandra Carbone
Sorbonne Université, CNRS, IBPS, Laboratoire de Biologie Computationnelle
et Quantitative - UMR 7238, 4 place Jussieu, 75005 Paris, France.

We ordered the entire set of X-ray crystallographic structures in the Protein Data Bank into hierarchies of progressive interactions involving identical or very similar protein chains, obtaining 40 205 hierarchies of protein complexes with an increasing number of partners. We then studied these hierarchies as a proxy for assembly pathways in protein complex formation. The database was used to show statistically that new interfaces, upon protein complex oligomerization, tend to be observed on residues that have been characterized as softly disordered (flexible, amorphous, or intrinsically disordered) in simpler complexes, i.e. structures with fewer partners and preceding the newly oligomerized complex in the hierarchy. Furthermore, we observe that the location of soft disorder changes place in the sequence as the number of partners increases. Our results unveil the existence of a general mechanism of protein assembly involving soft disorder that modulates and orders the way protein complexes are assembled. Since soft disorder regions are highly correlated with interactions in protein complexes, this work highlights the difficulty of predicting the structure of large protein complexes from sequence, and underscores the importance of coupling soft disorder predictors with the next generation of complex structure predictors. Finally, we investigate the relationship between Alphafold2's structure prediction confidence metric pLDDT on unbound structures and soft disorder, showing a strong correlation between Alphafold2's low-confidence residues and the union of all soft disorder regions observed throughout the interaction hierarchy, paving the way for the use of the pLDDT metric as a proxy for interface prediction and assembly paths.

INTRODUCTION

Structural biology is witnessing a complete revolution thanks to modern deep learning algorithms. Among all them, Alphafold2 (AF2) is, by the first time in History, able to predict with atomic accuracy the three-dimensional structure of protein amino-acid (AA) sequences [2, 3]. Furthermore, the power of these tools goes beyond single proteins: also protein complexes are now accessible from sequence [4–6]. These results will surely dramatically enlarge the amount of protein and protein complex structures at hand. Not in vain, thanks to the fast and cheap modern genome sequencing techniques, the number of candidates for viable protein sequences is several orders larger than the number of experimentally validated structures. Yet, all this concerns mostly the well-structured proteins. What happens then with all those proteins that are known to be totally or partially disordered under physiological conditions [7, 8]? At present, AF2 predictions leave these intrinsically disordered proteins (IDPs) and protein regions (IDPRs) unstructured, or predict structures for regions that undergo a disorder-to-order transition when interacting with some partners [5, 9].

After decades of studies, it seems clear now that IDPs/IDPRs play a major role in encouraging and tuning protein interactions with other partners, anticipating that knowledge on disorder will be crucial to automatically predict the structure of new or large protein complexes. Indeed, IDPs and IDPRs possess a large capacity to bind to multiple partners as compared to well-structured proteins [10–13]. For instance, IDPs/IDPRs are known to be rich in molecular recognition features or motifs that are used for protein-protein interactions [14–16]. Moreover, many of them are observed to undergo disorder-to-order transitions to build an interface with other partners [17–19], or even folding in alternative structures depending on the partner involved in the interaction, often resulting in unrelated or even opposite protein function [20]. In summary, IDPs and IDPRs are believed to encourage disorder-based-mechanisms that could determine the assembly of protein complexes [21].

Yet, despite the ubiquity of IDPs/IDRPs, the role of disorder in protein interaction networks is normally discussed only in terms of the phenomenology observed in a narrow number of IDPs/IDPRs. Even though several predictors of binding disordered regions have been developed [17–19], very little statistical evidence unveils the general role of the structural disorder in complex assembly. Recently, we provided a large-scale analysis of disorder in known structures which establishes the statistical basis above mentioned: namely, statistical correlation between the location of disorder and interfaces. Proofs of correlation between the location of binding and disorder regions were investigated in [22] using the entire set of X-ray experimental structures stored in the Protein Data Bank (PDB). The results clearly highlighted that once all the alternative structures containing a particular

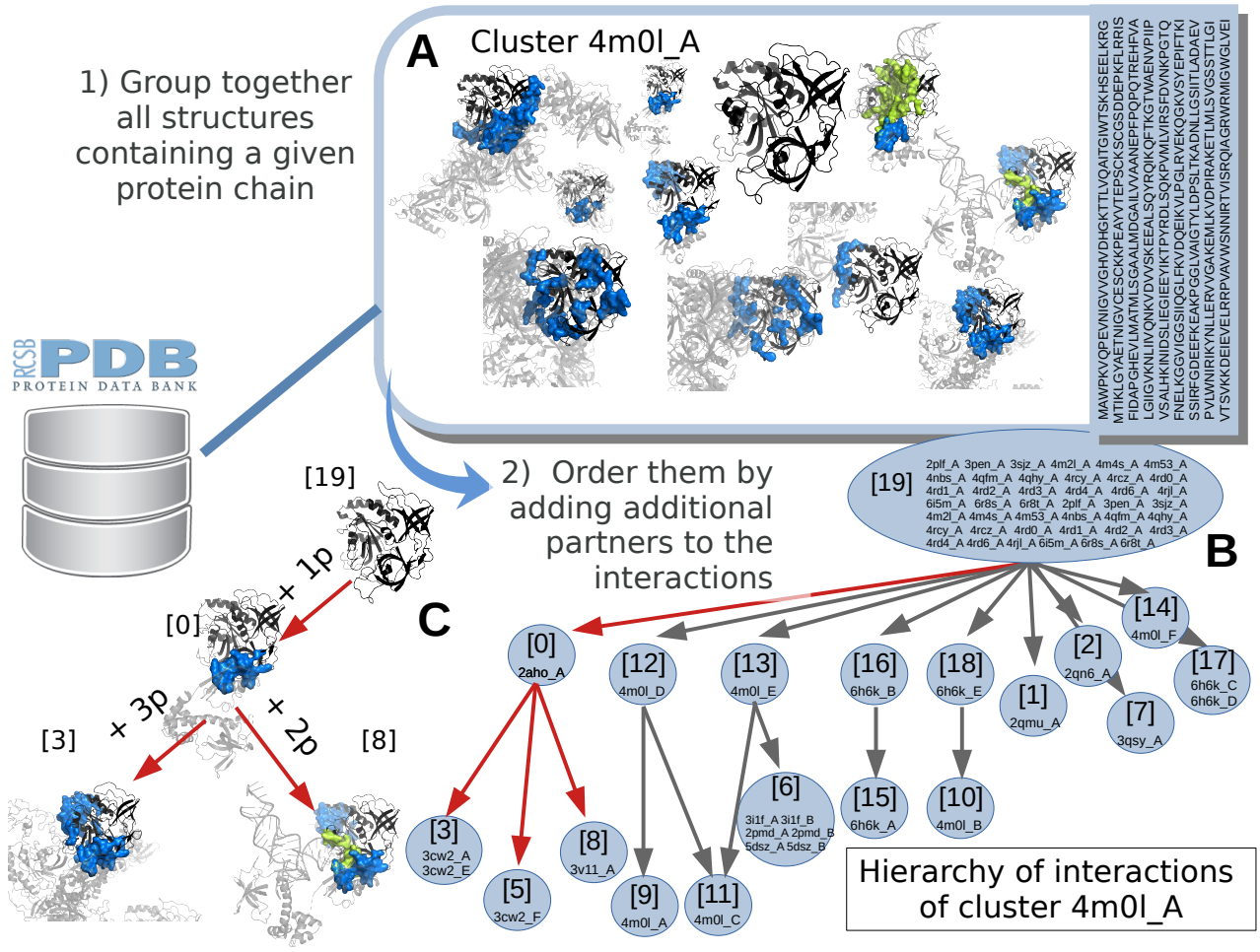

FIG. 1. **Sketch describing the construction of the interaction hierarchy of the protein cluster 4m0l_A.** At the first stage, we group together all the PDB structures containing a given protein chain or very similar ones (with respect to sequence; see the reference sequence chain displayed with a blue background in **A**). In this example we use the sequence 4m0l_A[1]. We list the PDBID of each of these structures, together with the label of the chain that contains our selected protein chain. We label this joint information in the form of PDBID_chain. We identify all the different interactions of this protein with distinct partners observed in our cluster of structures and compute the interfaces. We use this information to build a hierarchy of progressive assembly as the one shown in the pale blue graph in **B**. In this graph, each node gathers together all the PDBID_chains where our protein interacts with the same partner through the same (or very similar) interface region. The list of PDBID_chain combined for each node are shown. Then, oriented edges connect to nodes where new partners and interfaces are observed that sum up to those of the parent node. We show the structures and interface regions along one of these branches of the graph in **C**.

(or a very similar) protein sequence are cross analyzed, interfaces occur with a statistically significant preference in those AAs that are characterized as disordered, typically in different PDB structures. Yet, it was also shown that in order to make this correlation sharper, it was necessary to enlarge the definition of structural disorder from the standard *missing residues* in the PDB structure, to all those residues that are poorly resolved in the experiments, that is, hot loops, flexible or even spatially amorphous regions of the protein [23, 24]. This softened version of disorder has been called *soft disorder*. These results suggest that new interfaces tend to accommodate themselves in the floppy parts of a protein, and points to the idea that soft disorder could actually mediate the order in which protein complexes are assembled. With this intuition in mind, several examples of progressive assembly were discussed in [22], describing a disorder-mediated interaction mechanism very similar to some hypothetical ones proposed previously in the IDP literature [21]. In this work, we go beyond previous studies, and provide statistical proofs of the central role of soft disorder in the progressive assembly of

protein complexes. Upon oligormerization, the location of soft disorder regions may change place in the protein structure, and we observe that this new location is correlated with the regions where we observe new binding at higher oligomerization degrees. A similar outcome is observed in unbound structures, who seem to carry information about all the alternative new binding regions, though differentiate the signals of different binding interfaces from the sequence might be a difficult problem. This work highlights, on the one hand, the importance of correctly predicting the flexible/disordered regions of a particular protein complex in order to predict where new partners will be accommodated. But also, on the other hand, that soft disorder depends on the structure of the intermediate complexes. In other words, predicting selectively, based on sequences, the position of the interface region (IR) in a specific complex (among all the possible interactions of a given protein) is a difficult problem.

We organize the paper as follows. We first introduce the notion of soft disorder and explain the construction of directed graphs describing hierarchies of progressive complex assembly. We then test the notion of soft disorder as an interface predictor for new interactions in the hierarchy. We conclude with a comparative analysis of soft disorder and AF2's low confidence on some protein complexes.

RESULTS

Soft disorder

In this work, we restrict our analysis of disorder to the residues that are poorly resolved in the PDB X-ray structure, in other words, to the floppy, highly flexible, fluctuating or amorphous protein regions. These residues can be detected through their anomalously high B (or temperature) factor [24], which allows us defining a softer notion of disorder than the missing residues, what we refer as *soft disorder* (SD) in [22]. The B-factor gives a measure of the error committed in the estimation of the atom's coordinates, hence its scale is mostly fixed by the resolution of the experiment. This means that, in order to compute the SD regions (SDRs), we are not interested in the absolute value reached by the B-factor in a particular experiment, but only in the atoms presenting an anomalously high B-factor when compared to the rest of the protein chain. In the following, we will consider the B-factor of a residue i, B_i, as the B-factor of its C_α atom. Then, in order to detect the floppy regions of a protein, we will rely on a normalized version of the PDB's B-factor, namely *b-factor* (with b in lower case)

$$b_i = \frac{B_i - \langle B \rangle}{\sigma_B}, \tag{1}$$

with $\langle B \rangle$ and σ_B being the mean and the standard deviation of all the B_i in the entire protein chain (that is, B is normalized in the chain, not the protein complex). Then, in order to define the SDR, we need to fix a static threshold for b to label the difference between ordered and disordered AA. The effects of this threshold were discussed in detail in [22]. In particular, high thresholds, for instance $b > 3$ (only the AAs with a B-factor larger than $3\sigma_B$ are considered), are more likely to form an interface in alternative structures of a given protein than regions with $b > 0.5$. Yet, being the $b > 3$ regions much shorter than the $b > 0.5$ regions, it is far more likely that new interfaces are covered by the SDR defined with $b > 0.5$ than with $b > 3$. We found that the best approximate compromise between positive predictive value and sensitivity is obtained using a threshold $b > 1$. For this reason, for now on, we will say that an AA is soft disordered if its $b_i > 1$ (which concerns in average to the 16.7% of the AAs of a chain) or if its structure is missing in the PDB (note that in [22], both kinds of residues were studied separately).

Hierarchy of the complexity of interactions in the PDB

In order to test the effect of SD in the progressive assembly of a protein complex, we need to order all the protein structures available in the PDB by oligomerization degree. In particular, we have used all information available in the bank up to the 7th of January 2022 and selected those structures obtained via X-ray diffraction experiments, totaling 155749 structures. In practice, the first step to build our hierarchies of interactions is to gather together all the PDB structures (i.e. their PDBID) containing a given protein sequence, together with its identification within the complex (i.e., their chain name). In practice, we have considered two slightly different sequences as equal (and thus included in the same protein cluster), if they were equal up to a 90% of sequence identity and a 90% of length. For details concerning the clustering procedure, see Materials and Methods. In total, we have analyzed the *interaction complexity* of 51332 different clusters (40205 have more than 1 structure) of very similar protein sequences. With interaction complexity of a given cluster, we refer to all the multiple interactions where our cluster protein is involved, that is, all the interactions with different partners (even if they share the same binding region), or with identical partners but at different locations on its structure, whether partially overlapping or totally separated.

As in [22], we label each cluster by the PDBID and chain label of one of its constituent chain structures. In fact, the protein chain that gives the name to the group is taken as the *representative sequence* of the cluster, and it is used to map the information observed in alternative structures to the same sequence for the cluster. We show an example of a cluster construction in the box of Figure 1A. Once the cluster is built, for each of its chain structures and each of its partners (the same protein, different proteins, or DNA/RNA), we compute the in-

terface region (IR), i.e. the binding residues belonging to the interface of the cluster protein and its partners, and keep track of the identity of the partners. In order to identify if our protein interacts with the same partners (having identical or very similar sequences too) in alternative complexes of the clusters, we label partner chains by the name of the cluster to which they belong (that is, the same set we use to build our hierarchy). At this point, we use this information (location of IRs in the sequence and the identity of the partners) to build a *directed acyclic graph (DAG)* of increasing complexity in the interactions of a given protein chain, i.e., an hierarchy of the progressive assembly of that protein; see the graph in Figure 1B. We begin by grouping together all the chains containing very similar IRs and the same exact partners (for technical details, see Materials and Methods). Each of these different groups constitute a node (or a vertex) of the graph. We will say that a node has K partners, if our protein has binding sites with K chains in the complex. We stress that these partners' chains can be either identical or different, or protein or DNA/RNA too. Once having all the nodes created, we introduce directed edges (or arrows) in the graph, pointing to the increase of partners and IRs. All the nodes with $K' > K$ connected to a node with K partners are then called *descendants* of that common *ancestor* node. In our construction, two nodes can only be connected (related) if the descendant has more partners than the ancestor, and if all the partners and more than a 75% of the IR observed in the ancestor are also present in the descendant. In addition, two nodes satisfying these conditions are related only if no third complex structure in the cluster, with an intermediate number of partners K'', can be placed in between them in the genealogical reconstruction. See an example of this construction in Figure 1C. Notice that a node will be a root of the DAG if it is a common ancestor to all the structures in the cluster, which is always the case for nodes with $K = 0$ (that is, nodes composed of *unbound structures*). Yet, this is not necessarily the case for clusters with no unbound structure, where the whole interaction DAG can be composed by multiple disconnected graphs sharing no (known) common ancestor.

By construction, our DAG edges connect nodes with variable K, the only constraint is that the younger node's number of partners $K' > K$. This means that edges do not necessarily add up just +1 new partner. For the subsequent analysis, we find useful, as we will discuss later, to track the *degree* of each relation in the DAG, that is, the number added partners at the two sides of the edge. We label this degree by k. In Figure 2, we show the total number of edges relating a node with K partners and a node with $K + k$ partners.

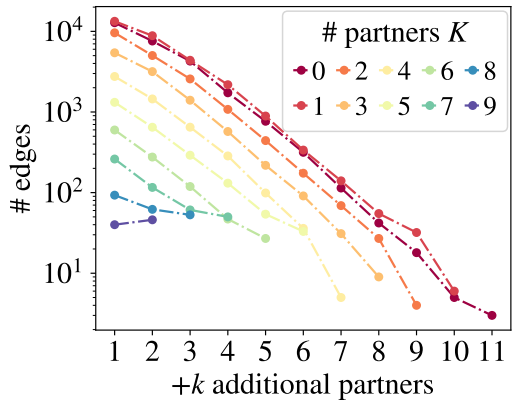

FIG. 2. **Distribution of edges from the PDB.** Distribution of the total number of edges in our database connecting nodes of structures with K partners (in different color lines), and $K + k$ partners (k is shown in the horizontal axis).

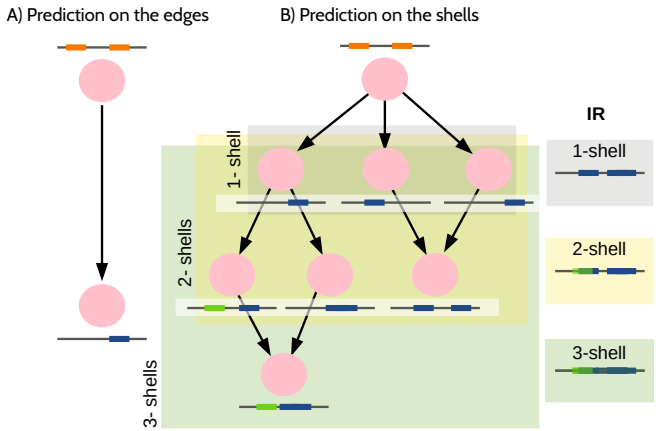

FIG. 3. **Schema of the two different procedures to test the predictions.** A) We compare the soft disorder (sketched as orange residues on a black sequence) of the current node with the interface observed in the next node of the graph (sketched as blue or green interface regions on the sequence). B) The soft disorder of the current node is compared with the union of all the interfaces where the protein interacts either with $k = 1$ extra partner (the first shell of interactions), or with $k \leq 2$ extra partners (the second shell), or $k \leq 3$ (the third shell), and so on. The union of the interface regions in each shell is sketched on the right. Thus, each new shell contains all interface residues of the lower shells. This means that the total interface in the shell (IR) tend to be larger than in single nodes with the same number of partners K.

Comparison between soft disorder in ancestors and interfaces in descendants

Once the hierarchy of interactions for each cluster of protein chains is built, we need to assign an IR and an SDR to each node of the graph (as the union of all the

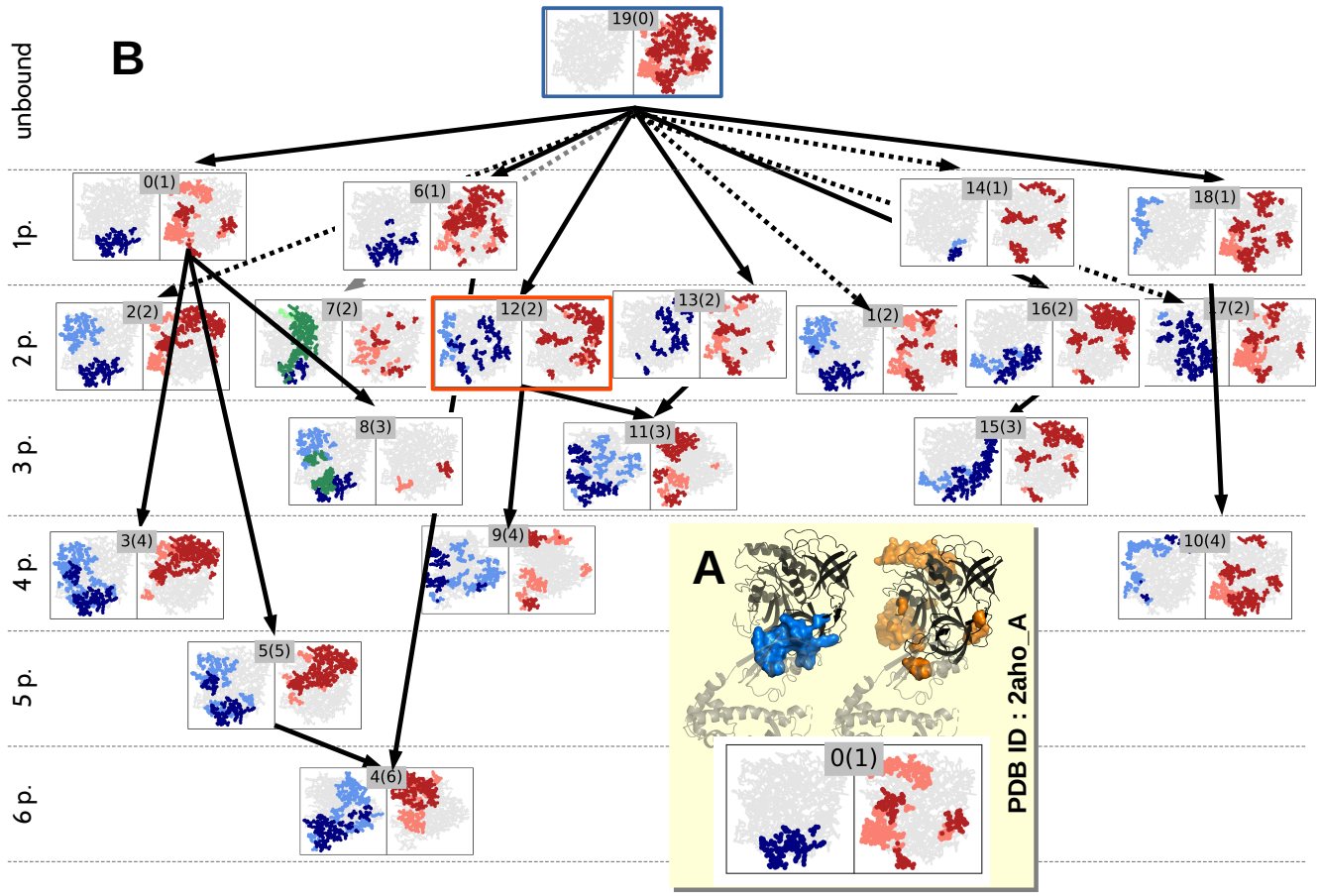

FIG. 4. **Complexity of the interactions (and its associated soft disorder) for the protein sequence associated to cluster 4m0l_A discussed in Fig. 1.** With the goal of enhancing the visualization, we have projected the 3-dimensional structures in a 2-dimensional sketch. Interface regions are displayed in blue or green depending on the interaction being with another protein or with DNA, respectively, soft disorder regions are colored in red, and residues located forward are shown in a darker shade of color than those placed backwards. In the inset **A**, we show an equivalence between the 3-dimensional representation of the protein and the sketch in 2-dimensions for the structure of node #0. In the 3D. structures, interface residues are in blue (top left) and soft residues in orange (top right). In **B**, we show the entire hierarchy of interactions collected from the PDB. Horizontal lines (Kp.) collect all protein complexes with a fixed number of partners K. Arrows point out the increase of the number of partners, from top to bottom. With respect to the unbound structure in node #19 (in a blue square), all interactions included in the row 1p., define the 1-shell of interactions in Fig. 3. In the same way, all the interfaces inside the 1p. and 2p. rows, define the 2-shell of interactions, and so on. The union of all the interfaces in this graph forms its *all*-shell. In case another ancestor node is considered as origin, as for instance node #12 (in an orange square), its interaction shells would only be composed by those nodes with a higher K that are linked to it. That is, node #11 for the 1-shell, and nodes #9 and #11 for the 2-shell.

IRs/SDRs in the structures included of that node). To do so, we align the sequences of each structure to the cluster's representative one, and tag a residue as being part of the IR or the SDR if it was labeled as IDR/SDR in at least one of the structures of the node. Interactions with proteins and DNA/RNA are treated separately, as two different kinds of IRs. As an example, in Fig. 4, we graphically show the IRs (in blue for protein interactions, and in green for DNA interactions) and the SDRs (in red) at each node for the hierarchy discussed in Figure 1B. For the sake of clarity, we have projected the 3D structures in two dimensions as sketched in Figure 4A. Furthermore, in order to distinguish forwards and backwards in the 3D structure, we have highlighted in darker shades the residues having a larger depth than the protein's geometrical center.

Now that we have all the information about IRs and SDRs ordered along the hierarchies of interface complexity, we can test the hypothesis of SD modulating the location of new interfaces upon complex oligomerization. In practice, we compare the location (residue by residue in the sequence) of the ancestor's SDR, with the new IRs

observed in its descendant (with new IR we mean the IR residues that were not already labelled as IR in the ancestor). If the same residue is characterized as SDR in an ancestor, and new IR in a descendant, we say the prediction is a true positive (TP). If the residue is not new IR in the descendant, it counts as a false positive (FP). On the other way around, residues labeled as new IRs in the descendant, but not SDR in the ancestor are false negatives (FN), and true negatives (TN) if the residue is neither new IR nor SDR in the two related nodes. We can combine these numbers to build different estimators of the goodness of the prediction:

$$\text{Sensitivity (Sen)} = \frac{\text{TP}}{\text{TP+FN}}, \quad (2)$$

$$\text{Specificity (Spe)} = \frac{\text{TN}}{\text{FP+TN}}, \quad (3)$$

$$\text{Accuracy (Acc)} = \frac{\text{TP+TN}}{\text{TP+FP+TN+FN}}, \quad (4)$$

$$\text{Precision or Positive Predictive Value (PPV)} = \frac{\text{TP}}{\text{TP+FP}}, \quad (5)$$

$$\text{Negative Predictive Value (NPV)} = \frac{\text{TN}}{\text{TN+FN}}. \quad (6)$$

The Sen quantifies the portion of the new IR that is correctly predicted by the ancestor's SDR and the Spe, the same but for the non-IR. The Acc tells us the ratio of the total residues whose role is correctly predicted in the descendant. Finally, the PPV tells us the fraction of the IR predictions are actually new IR in the descendant and the NPV, the fraction of the non-IR predictions that are indeed not IR in the descendant.

In a chain containing L residues, a totally random prediction of $N_{\rm D}$ SDR residues, would predict correctly (in average) a new IR residue with probability $r_{\rm I} = N_{\rm I}/L$, where $N_{\rm I}$ is the number of new IR residues in the descendant. Similarly, in a random guess scenario, one expects the following values for the above estimators: $\langle \text{Sen}^{\rm r} \rangle = r_{\rm D}$, $\langle \text{Spe}^{\rm r} \rangle = 1 - r_{\rm D}$, $\langle \text{Acc}^{\rm r} \rangle = r_{\rm D}(2r_{\rm I} - 1) + (1 - r_{\rm I})$ $\langle \text{PPV}^{\rm r} \rangle = r_{\rm I}$, $\langle \text{NPV}^{\rm r} \rangle = 1 - r_{\rm I}$, with $\langle r_{\rm D} \rangle = N_{\rm D}/L$). In other words, $\langle \text{Sen}^{\rm r} \rangle = 1 - \langle \text{Spe}^{\rm r} \rangle$ and $\langle \text{PPV}^{\rm r} \rangle = 1 - \langle \text{NPV}^{\rm r} \rangle$. We quantify the interface prediction power of the SDR of a node in two distinct manners, either we compare it with the new IRs found in a given direct descendant (namely *edge prediction*, see Figure 3A) or with the union of the new IRs observed in all its descendants *shell prediction*. For this latter, we can compute the union only up to a fixed number of new additional partners $+k$, in which case we call it a k-shell prediction, see Figure 3B, or extend it to union of the entire set of descendants of a given ancestor, to which we refer as the *all*-shell prediction.

In Fig. 5–A, we show the rate of TP (Sen) versus the rate of FP (1-Spe) for all our predictions of new IRs up to 2−shells. We have colored the points according to the local density in the graph to highlight populated regions. We observe that the 74% of our predictions are better than random, even if the knowledge of all the possible interactions of a given protein in the PDB is still extremely incomplete. It is important to stress that the total size of the new IRs grow significantly if two related nodes have very different K, and that the size of the non-IR regions decreases as the protein interacts with more and more partners. For this reason, it is important to analyze separately the prediction of the ancestor as a function of K, or as a function of the number of additional partners $+k$ of the nodes used to compute the new IRs. In Fig. 5-B, we show the percentage of the predictions that are better than random, averaged over different $k-$shells (for different but fixed values of $+k$) of ancestors nodes with K partners. In Figs. 5-C,D, we show the quotients of the Sen/(1-Spe) and PPV/(1-NPV), respectively, averaged using the same groups of predictions. In all the figures, the random expectation is highlighted as a horizontal black dotted line. In all cases, predictions are better than random. They improve for larger values $+k$, thus agreeing with the hypothesis that SDR is a quantitative indicator of the propensity in interface formation. We also observe that prediction power improves as K increases, which is mostly related to the size of the available regions for new interfaces, who decreases as the number of new partners arrive. We will discuss and try to remove this effect, later in the text. In Figs. 5-E,F,G, we compare the *all*-shell curves with the equivalent ones computed using each edge prediction in the database. Again, we see that both tests agree with a statistically meaningful correlation between SDRs and new IRs, even if shell predictions seem to be slightly better.

We study in more detail the shell predictions in Fig. 6. In Fig. 6-A, we show the average of the different goodness predictors obtained in each of the *all*-shell predictions of our dataset conditioned to ancestors with K partners. The real data tests are shown in solid lines and the randomization tests predictions in dotted ones. In Fig. 6-B we compare the quotient Sen/Senr (which is equal to PPV/PPV^r by construction) for the same tests. The difference between the real and the random expectations become clearer for all the estimators when we separate predictions by shell-degree level k. In Fig. 6-C, we compare the averaged value of the predictors in shells of K layers (K shown in the horizontal axis). It is clear that the PPV (both for data and random tests) grows sharply as K increases, which is nothing but a direct consequence of the fact that the total new IR region grows as we add partners. In order to get rid of this effect, we compare the real data and the random expectation through the averaged quotient of Sen/Senr for each prediction in Fig. 6-D, showing that the ratio is rather stable, and even slightly decreasing with K. This is somehow expected because as we add up shells, we also add up errors. In both kinds of analysis, we find that the predictions from unbound nodes ($K = 0$) are significantly worse than the rest of the predictions, and this effect seems to be related to the existence of quite different SDRs in distinct unbound structures, which could be precursors of specific individual branches of the graph. Such an analysis would require considering different nodes associated to unbound structures and selecting preferred branches for the statistical analysis. This analysis goes beyond the scope of this work.

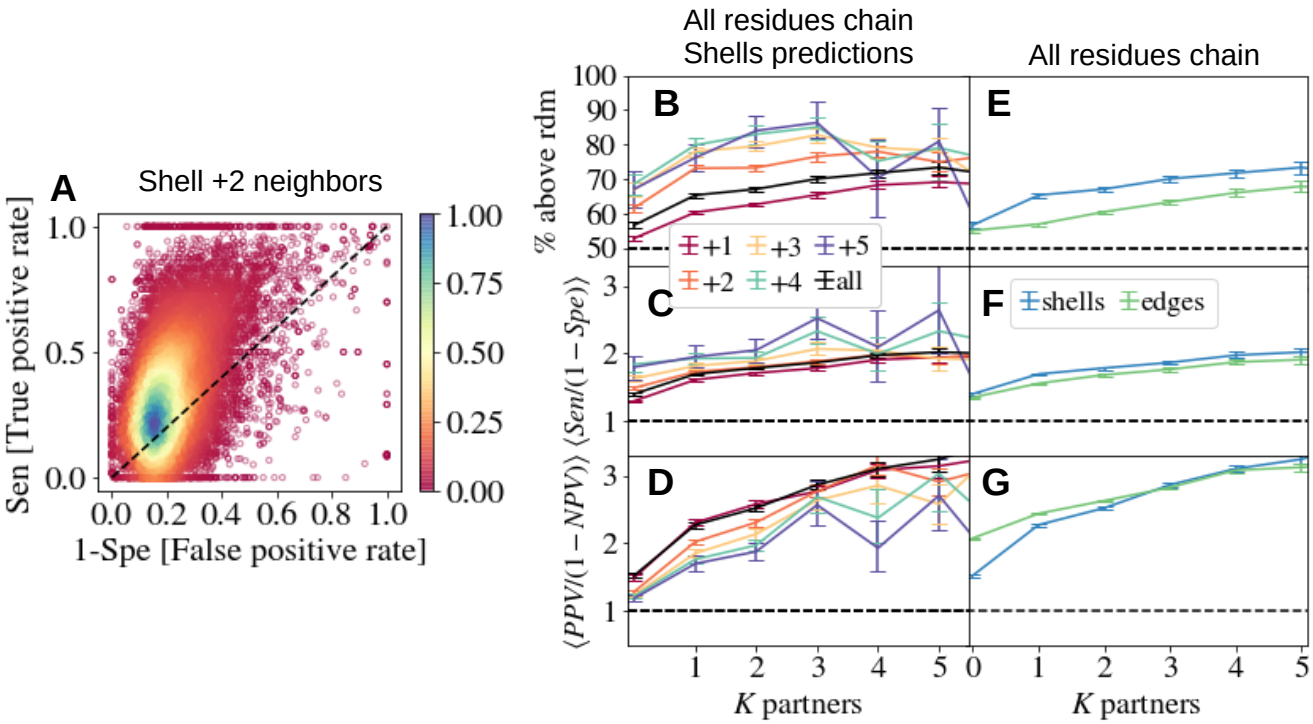

FIG. 5. In **A**, we show the true positive rate (Sen) versus the false positive rate (1-Spe) of the predictions of new IRs in 2-shells (the union of all the IRs of nodes having 2 or 1 additional partners than the ancestor), based on the knowledge of the SDR in the ancestors nodes, for all the hierarchies in our dataset (reaching at least a 2-degree of interactions). Each dot corresponds to a prediction, and color codes the local density of points in that region. The random guess behavior is highlighted by the diagonal dotted line. In this case, a 74% of the predictions are above the diagonal. In **B** we show this ratio conditioned to shells of $+k$ new partners (in different colors) and ancestors nodes of K neighbors. In black we show the statistics obtained when comparing the SDR of an ancestor node with the union of the IRs of all its descendants, no matter their degree (the *all*-shell). In **C** and **D**, we show the same analysis but for the quotients Sen/(1-Spe) and PPV/(1-NPV), respectively. In **E**, **F** and **G**, we compare the *all*-shell prediction curves of figures **B**, **C** and **D**, with the analogous curves we obtain using all the edges predictions from nodes with K partners. In all cases, the horizontal black dotted line marks the random guess expectation.

We repeat the same analysis of Fig. 6, but this time for the edges predictions in Fig. 7. The results are qualitatively similar to those of the shell predictions, with fewer effects associated to the sharp increase of the IR regions as the shell degree increases. In the case of the edges, $+k$ only marks the difference between the number of partners of the nodes joined by an edge. It is remarkable that the typical size of the IR predicted is extremely similar in both tests (around a 30-40% of the total new IR), while PPV (and the quotient PPV/PPV^r) is higher for the shell tests. This combination supports the idea that an important part of the new IRs accommodate in regions that were SDR in the ancestor, but also that a given ancestor carries information about possibly multiple new interfaces.

Finally, as mentioned several times before, by construction, one expects that the size of the protein surface available for new interfaces decreases as we go down in the DAG, so does the size of the total IR. In parallel, IR residues tend to have lower b-factors and high b-factors tend to locate at the protein surface. This means that we need to ensure that the correlation between SDR and new interfaces is not just a consequence of a decrease of the surface available for new interactions and a rigidification of the "older" interaction regions. With this goal in mind, we consider three more tests to compare the data and random expectations for the predictors for each node. Namely, instead of using the entire protein chain, as done up to now, (i) we exclude from the analysis the residues belonging to the IR of a node (namely, C-IR), (ii) we consider only the residues in the surface of the complex (namely, S-all), or (iii) both, we include in the comparison analysis the residues that belong to the surface of the complex and are not part of the ancestor's interface (namely, S-IR). Clearly, (ii) and (iii) are very similar because interfaces tend to be grounded. We show, in Fig. 8, the statistics of Fig. 5 for these new tests. In Fig. 8-A, we compare the percentage of all-shell prediction points above the diagonal as a function of the number of partners of the ancestor node for the 4 tests (that is, the old

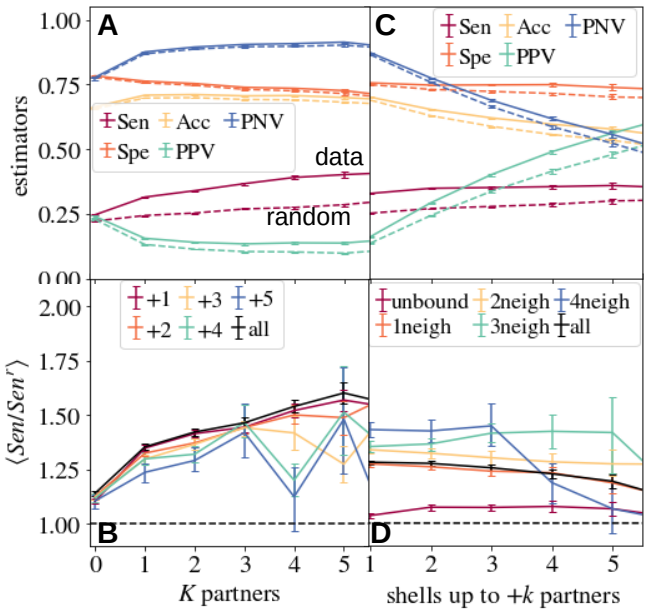

FIG. 6. **Global SDR analysis of K-shells.** In **A**, we show the average predictor estimators when comparing the SDR of an ancestor node with the union of the SDRs of all its descendants (the *all*-shell) as function of the number of partners of the ancestor node. Solid lines represent the data points obtained in our dataset, and the dotted lines, the random guess expectation. In **B**, we show the quotient between Sen and Senr again as function of K averaged over shells of the same degree K (in colors) or the total shell of descendants, that is, between the solid and the dotted crimson line in **A**. In **C**, we show the same estimators than in **A**, but this time averaged over shells of identical degree K no matter the value of K of the ancestor node. In **D**, we show the quotient Sen/Senr, again as function of the shell degree K, but this time also conditioned to ancestor nodes with a fixed number of neighbors (in different colors).

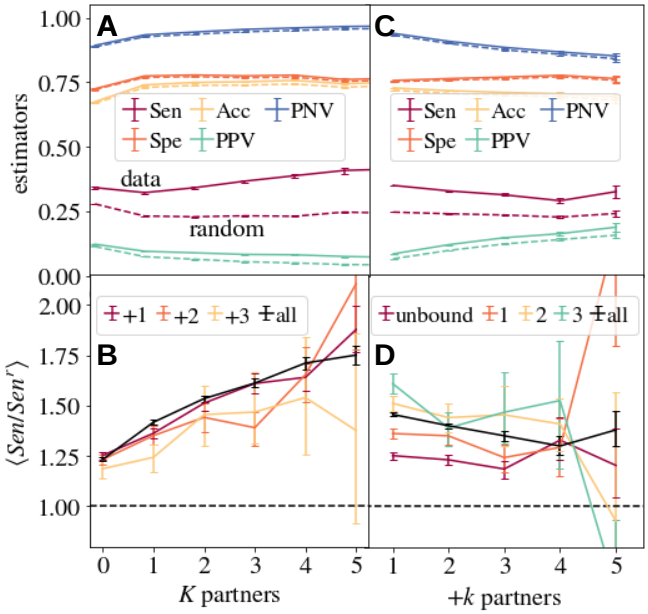

FIG. 7. **Local SDR analysis of K-shells (prediction on the edges).** We repeat Fig. 6 but restricting the test condition to edges for a localized analysis. In **A** we show the average predictor estimators when comparing the ancestor's SDR with the descendant's new IRs as a function of the number of partners of the ancestor node K. Solid lines represent the data points obtained in our dataset, and the dotted lines, the random guess expectation. In **B**, we show the quotient between Sen and Senr again as function of K averaged over all the edges that linked two nodes with $+k$ partners (in colors) or with any degree in black, that is, between the solid and the dotted crimson line in **A**. In **C**, we show the same estimators as in **A**, but this time averaged over edges of relation degree k, no matter the number of partners K of the ancestor node. In **D**, we show the quotient Sen/Senr, again as function of the edge relation degree k, but this time also conditioned to ancestors with a fixed number of neighbors (in different colors).

test using the entire protein chain plus, the 3 new tests). In Figs. 8-B and C, we show the averaged value of the quotients Sen/(1-Spe) and PPV/(1-NPV), respectively, , as a function of K too. We observe that, even if the predictions are in general worse than those computed using the entire chain, there is still an important correlation between SDR and new IRs, and that all tests follow very similar trends. In Figs. B, D and F, we show the dependence with K in the most restricted test (the one using only the S-IR residues), where the main difference with Figs. 5 D-F is the disappearance of the strong dependence in K, as previously anticipated.

SOFT DISORDER AND AF2 LOW CONFIDENCE REGIONS

Very low values of the AF2 confidence metric pLDDT has been reported to correlate well with intrinsic disorder [3]. We have observed that this connection is much stronger and extends to SDR and IRs, with some nuances. As done for our notion of SDR, we label as "low-confidence" regions (LCR) the residues with pLDDT below the protein's backbone average. In Fig. 9B, we consider two examples of crystal structures and their associated hierarchy of complexes in the PDB. For these two examples, we predicted the AF2 model structures of the unbound form and a complexified conformation. LCR in AF2 structural models of the unbound form correspond well to SDR measured on crystal structures over the entire hierarchy of interactions. When the complexified AF2 model is considered, one observes that its LCR remains essentially unchanged compared to the LCR computed in the unbound form, in contrast to SDR measured on the corresponding crystal structures. Then, we compare AF2 LCR measured on the unbound structure with the union of all the SDRs measured either in the protein hierarchy or in the unbound structure only in

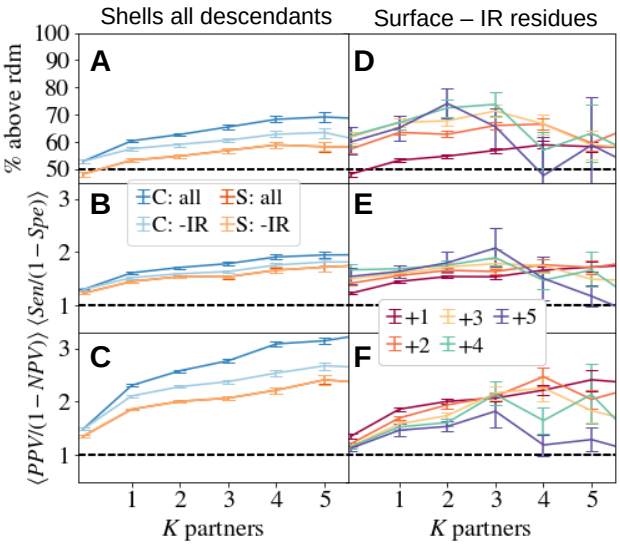

FIG. 8. We compare the statistics of Figs. 5 after removing trivial correlations between SDR and new IRs, associated to both kinds of regions preferring the protein surface and avoiding old interface regions. We consider 4 different tests for the comparisons between SDR and new IRs: (C:all) we use all the chain residues, (C:-IR) we remove the IR residues in the ancestor from the test, (S:all) we use only the residues in the node's complex surface and (S:-IR), we use only those surface residues that did not belong to the IR in the ancestor. For all these tests, we show the percentage of *all*-shell prediction points having Sen > 1-Spe in **A** and the average of the quotients Sen/1-Spe in **B** and PPV/(1-NPV) in **C**, all them computed after grouping together predictions from nodes with the same number of partners K. In **D, E, F**, we study the dependence of the shells K degree in these statistics for the most restricted test, the "S:-IR".

Fig. 9A. Clearly, the relative pLDDT seems to correlate well with the SDR computed over the hierarchy, in contrast to the SDR computed on the unbound structure. Most importantly, when we compare the spatial location of AF2 LCR with the union of all the IRs present in the hierarchy, we observe an excellent correspondence in all the clusters tested. This means that the relative uncertainty in the structure predictions of AF2 carries information about the interaction network of a given chain, even if it remains unclear how to use it to predict a particular assembly path yet. To conclude, we want to stress that LCRs may exist in model structures that are predicted with high confidence by AF2, that is with a global pLDDT > 80 or 90 over all residues. Finally, we stress that all this last discussion about the AF2 LCRs rely on the investigation of some particular examples and it should be tested statistically in future works.

DISCUSSION

We have shown that there is a significant correlation between the position occupied by soft disorder residues (flexible, missing or amorphous) in a protein complex, and the location of new interfaces as more and more partners are added to that complex. These results are supported by a large scale analysis of all the structures in the PDB that occur in hierarchies where similar protein chains interact with an increasing number of partners. These hierarchies can be considered as an incomplete proxy of the pathway of complex assembly. The picture emerging from their analysis is that a large part of the new interface regions are located on the more floppy or amorphous parts of the simpler complex' surface. The existence of a similar disordered based directional and sequential mechanism for complex assembly, had been previously proposed in [21] under the name of the binding change model, our work provides statistical proofs of the generality of this phenomenon.

The determination of the next assembly step during complex formation guided by soft regions has a number of direct implications in the design of computational strategies aimed at reconstructing the full path of assembly: when the partner is available, one can identify where it will bound; when no partner is available, information on soft disordered regions helps to highly reduce the number of potential interactors by restricting the search on specific areas of the surface. This is particularly useful when the set of potential partners have been identified experimentally but their interactions are yet unknown.

The knowledge of soft disorder for pairs of protein structures, could be systematically employed in protein docking experiments to highly reduce the conformational search space, as it was done by considering predicted interfaces before [25, 26].

Exciting hypotheses can be advanced on how to cross the knowledge of soft disordered regions with predictions from AF2. In [3, 9], it has been observed that the regions where intrinsic disorder is present correspond to those with high uncertainty for AF2. In this work, we have shown several examples that suggest that regions with relatively low pLDDT coincide with the union of all soft disordered regions in our protein interaction hierarchies. AF2 can be used to reduce the space of search in 3D complex prediction to specific regions of the protein surface and specific partners. If AF2 signal on unbound forms could be disentangled to predict the new/next? soft disordered regions upon binding, then AF2 low confidence regions identified on unbound forms could be useful to define suitable strategies to sort the next interacting region during the assembly.

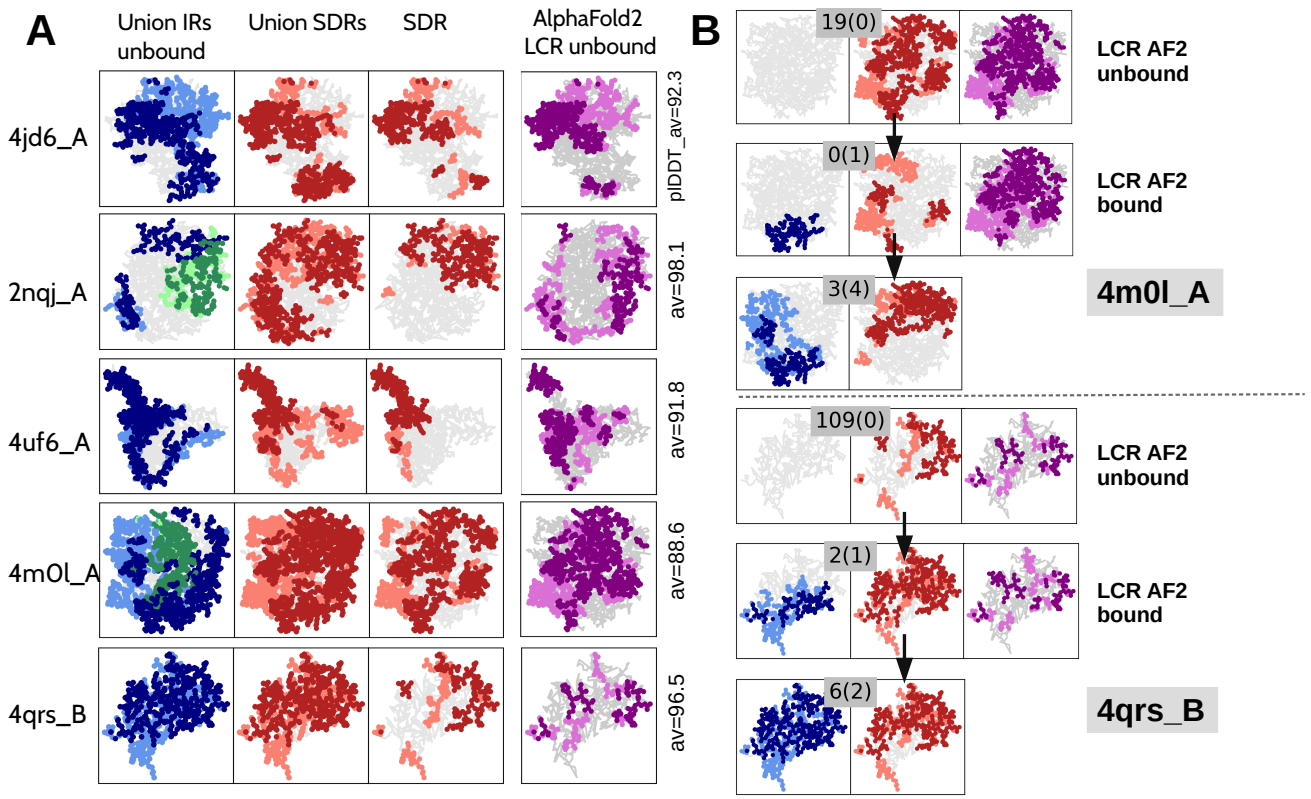

FIG. 9. **A**. We compare the position of the union of all interface regions (IR, in blue for proteins and green for DNA and RNA) and all soft disorder regions (SDR, in red) identified by graph analysis of five protein clusters (4jd6_A,2nqj_A, 4uf6_A, 4m0l_A and 4qrs_B) with those regions where the AF2 model structure is predicted with high uncertainty (that is, the pLDDT is below the protein's average; in purple). For each chain, the average pLDDT value for all residues is reported on the right. We also show SDR computed on the unbound chain. Structures and regions were projected in two dimensions following the same procedure as in Fig. 4. **B**. Two hierarchies, for 4m0l_A and 4qrs_B, are described. The AF2 LCR is computed for the unbound and bound forms, and in both cases, one sees the same result, while the corresponding SDR changes from the unbound to the bound form.

I. MATERIALS AND METHODS

Error bars

Error bars of Figs. 5-8 represent the 90% confidence interval, and are computed using the Bootstrap method with 500 re-samplings.

Clustering procedure

Clusters of similar protein chains in the PDB are created using the MMseqs2 method[27, 28]. The representative chain of the hierarchy gives its name and is chosen as the PDB_ID of the experiment with higher resolution or R-value.

Interface computation

Protein-protein binding residues for each structure node are computed using the INTerface Builder method[29] (two AAs are considered in contact as long as their two C_α are at a distance $\leq 5\AA$). To obtain the protein-DNA/RNA binding sites, we search the residues whose relative surface area (RASA) decreases after binding. The change in RASA is computed with naccess[30] (with a probe size of 1.4Å).

Hierarchy construction

We have considered that two or more structures have the same IR and can be thus grouped together in a single node if two conditions are fulfilled: (i) the number and identity of their partners is identical (that is, they have the same K, and the name of all the hierarchies

to which their partners belong are identical) and (ii) the intersection between all the IR of the structures in the node is larger than the 75% of each of these IRs. For this analysis, protein and DNA interfaces are treated as completely different kinds of interfaces, which means that (ii) condition must be satisfied separately for each kind of interface. The identity of all RNA and DNA partners has been considered as equal.

Two nodes are joined by a directed edge if: (i) the "younger" node has all the partners of the "older node" (same identity) in addition to the new ones, and (ii) if the descendant's IR contains at least the 75% of the total ancestor's IR. Again, DNA and protein interfaces are considered separately.

Alignment to the representative chain

The IR and SDR residues computed at each of the protein structures belonging to a hierarchy are mapped to the protein chain representative via sequence alignment using the Biopython's [31] PAIRWISE2.ALIGN.GLOBALXX routine.

AF2 structure predictions and pLDDT

AF2 and AF2-multimer have proven their ability to accurately predict the 3D structure of single proteins and protein complexes from the amino acid sequence, respectively. Each residue in a model structure is accompanied by a measure of its confidence on a scale of 0 to 100, based on the pLDDT metric. Regions with pLDDT $>$ 90 are thus modeled with very high accuracy, whereas regions with pLDDT $<$ 50 are advised as uninterpreted, and are known to be strongly correlated with the presence of intrinsic disorder [3].

In order to compare the pLDDT values with our definition of SD (which is normalized over chains), we did not focus on absolute values but on values that are relative to the mean over the chain. Namely, the purple regions shown in Fig. 9 are defined by residues with confidence values lower than the mean. We call these regions *low confidence regions* or LCR for short. In practice, we used the ColabFold [32] web server to obtain the AF2 predictions and calculated the average pLDDT in the protein chain considering the 5 different AF2 models provided. The list of regions with the lowest confidence is obtained as the union of all residues with a pLDDT lower than the average in the five models.

II. ACKNOWLEDGMENTS

The Comunidad de Madrid and the Complutense University of Madrid through the Atracción de Talento program (2019-T1/TIC-12776) (BS); partial support by MINECO (PGC2018-094684-B-C21 partially funded by FEDER, Spain) (BS); and ANRS–AAP2021CSS12-D21342 (AC).

[1] Pdb_id:4m0l, https://www.rcsb.org/structure/4m0l (2013).

[2] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, *et al.*, Highly accurate protein structure prediction with alphafold, Nature **596**, 583 (2021).

[3] K. Tunyasuvunakool, J. Adler, Z. Wu, T. Green, M. Zielinski, A. Žídek, A. Bridgland, A. Cowie, C. Meyer, A. Laydon, *et al.*, Highly accurate protein structure prediction for the human proteome, Nature **596**, 590 (2021).

[4] J. Ko and J. Lee, Can alphafold2 predict protein-peptide complex structures accurately?, bioRxiv (2021).

[5] M. Akdel, D. E. Pires, E. P. Pardo, J. Jänes, A. O. Zalevsky, B. Mészáros, P. Bryant, L. L. Good, R. A. Laskowski, G. Pozzati, *et al.*, A structural biology community assessment of alphafold 2 applications, bioRxiv (2021).

[6] R. Evans, M. O'Neill, A. Pritzel, N. Antropova, A. Senior, T. Green, A. Žídek, R. Bates, S. Blackwell, J. Yim, O. Ronneberger, S. Bodenstein, M. Zielinski, A. Bridgland, A. Potapenko, A. Cowie, K. Tunyasuvunakool, R. Jain, E. Clancy, P. Kohli, J. Jumper, and D. Hassabis, Protein complex prediction with alphafold-multimer, bioRxiv 10.1101/2021.10.04.463034 (2021).

[7] H. J. Dyson and P. E. Wright, Intrinsically unstructured proteins and their functions, Nature reviews Molecular cell biology **6**, 197 (2005).

[8] P. Tompa, *Structure and function of intrinsically disordered proteins* (CRC press, 2009).

[9] K. M. Ruff and R. V. Pappu, Alphafold and implications for intrinsically disordered proteins, Journal of Molecular Biology **433**, 167208 (2021).

[10] P. Tompa, C. Szász, and L. Buday, Structural disorder throws new light on moonlighting, Trends in biochemical sciences **30**, 484 (2005).

[11] A. K. Dunker, M. S. Cortese, P. Romero, L. M. Iakoucheva, and V. N. Uversky, Flexible nets: the roles of intrinsic disorder in protein interaction networks, The FEBS journal **272**, 5129 (2005).

[12] T. Mittag, L. E. Kay, and J. D. Forman-Kay, Protein dynamics and conformational disorder in molecular recognition, Journal of Molecular Recognition: An Interdisciplinary Journal **23**, 105 (2010).

[13] R. Van Der Lee, M. Buljan, B. Lang, R. J. Weatheritt, G. W. Daughdrill, A. K. Dunker, M. Fuxreiter, J. Gough, J. Gsponer, D. T. Jones, *et al.*, Classification of intrinsically disordered regions and proteins, Chemical reviews **114**, 6589 (2014).

[14] V. Neduva and R. B. Russell, Linear motifs: evolutionary interaction switches, FEBS letters **579**, 3342 (2005).

[15] M. Fuxreiter, P. Tompa, and I. Simon, Local structural disorder imparts plasticity on linear motifs, Bioinformatics **23**, 950 (2007).

[16] N. E. Davey, K. Van Roey, R. J. Weatheritt, G. Toedt, B. Uyar, B. Altenberg, A. Budd, F. Diella, H. Dinkel, and T. J. Gibson, Attributes of short linear motifs, Molecular BioSystems **8**, 268 (2012).

[17] B. Mészáros, I. Simon, and Z. Dosztányi, Prediction of protein binding regions in disordered proteins, PLoS computational biology **5**, e1000376 (2009).

[18] F. M. Disfani, W.-L. Hsu, M. J. Mizianty, C. J. Oldfield, B. Xue, A. K. Dunker, V. N. Uversky, and L. Kurgan, Morfpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins, Bioinformatics **28**, i75 (2012).

[19] D. T. Jones and D. Cozzetto, Disopred3: precise disordered region predictions with annotated protein-binding activity, Bioinformatics **31**, 857 (2015).

[20] V. N. Uversky, C. J. Oldfield, and A. K. Dunker, Showing your id: intrinsic disorder as an id for recognition, regulation and cell signaling, Journal of Molecular Recognition: An Interdisciplinary Journal **18**, 343 (2005).

[21] M. Fuxreiter, A. Tóth-Petróczy, D. A. Kraut, A. T. Matouschek, R. Y. Lim, B. Xue, L. Kurgan, and V. N. Uversky, Disordered proteinaceous machines, Chemical reviews **114**, 6806 (2014).

[22] B. Seoane and A. Carbone, The complexity of protein interactions unravelled from structural disorder, PLOS Computational Biology **17**, 1 (2021).

[23] G. Rhodes, *Crystallography made crystal clear: a guide for users of macromolecular models* (Elsevier, 2010).

[24] Z. Sun, Q. Liu, G. Qu, Y. Feng, and M. T. Reetz, Utility of b-factors in protein science: interpreting rigidity, flexibility, and internal motion and engineering thermostability, Chemical reviews **119**, 1626 (2019).

[25] A. Lopes, S. Sacquin-Mora, V. Dimitrova, E. Laine, Y. Ponty, and A. Carbone, Protein-protein interactions in a crowded environment: an analysis via cross-docking simulations and evolutionary information, PLoS computational biology **9**, e1003369 (2013).

[26] C. Dequeker, Y. Mohseni Behbahani, L. David, E. Laine, and A. Carbone, From complete cross-docking to partners identification and binding sites predictions, PLoS computational biology **18**, e1009825 (2022).

[27] M. Steinegger and J. Söding, Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets, Nature biotechnology **35**, 1026 (2017).

[28] M. Steinegger and J. Söding, Clustering huge protein sequence sets in linear time, Nature Communications **9**, 2542 (2018).

[29] C. Dequeker, E. Laine, and A. Carbone, Interface builder: A fast protein–protein interface reconstruction tool, Journal of Chemical Information and Modeling **57**, 2613 (2017).

[30] S. J. Hubbard and J. M. Thornton, Naccess, Computer Program, Department of Biochemistry and Molecular Biology, University College London **2** (1993).

[31] P. J. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, and M. J. De Hoon, Biopython: Freely available Python tools for computational molecular biology and bioinformatics, Bioinformatics **25**, 1422 (2009).

[32] M. Mirdita, K. Schütze, Y. Moriwaki, L. Heo, S. Ovchinnikov, and M. Steinegger, Colabfold - making protein folding accessible to all, bioRxiv 10.1101/2021.08.15.456425 (2021).