Assessment of AlphaFold2 for Human Proteins via Residue Solvent Exposure

Bæk, Kristoffer T.; Kepp, Kasper P.

Published in:
Journal of Chemical Information and Modeling

Link to article, DOI:
10.1021/acs.jcim.2c00243

Publication date:
2022

Document Version
Peer reviewed version

Assessment of AlphaFold2 for human proteins via residue solvent exposure

Kristoffer T. Bæk and Kasper P. Kepp*

DTU Chemistry, Technical University of Denmark, Building 206, 2800 Kgs. Lyngby, Denmark

* Correspondence: E-mail: kpj@kemi.dtu.dk Tel: +45 45252409

ORCID
Kristoffer T. Bæk: 0000-0003-3243-3962
Kasper P. Kepp: 0000-0002-6754-7348

Abstract

As only 35% of human proteins feature (often partial) PDB structures, the protein structure prediction tool AlphaFold2 (AF2) could have massive impact on human biology and medicine fields, making independent benchmarks of interest. We studied AF2's ability to describe the backbone solvent exposure as a functionally important and easily interpretable "natural coordinate" of protein conformation, using human proteins as test case. After screening for appropriate comparative sets, we matched 1818 human proteins predicted by AF2 against 7585 unique experimental PDBs, and after curation for sequence overlap, we assessed 1264 comparative pairs comprising 115 unique AF2-structures and 652 unique experimental structures. AF2 performed markedly worse for multimers, whereas ligands, cofactors and experimental resolution were interestingly not very important for performance. AF2 performed excellently for monomer proteins. Challenges relating to specific groups of residues and multimers were analyzed. We identify larger deviations for lower-confidence scores (pLDDT) and exposed residues, and polar residues (Asp, Glu, Asn e.g.) being less accurately described than hydrophobic residues. Proline conformations were the hardest to predict, probably due to common location in dynamic solvent-accessible parts. In summary, using solvent exposure as a metric, we quantify the performance of AF2 for human proteins and provide estimates of the expected agreement as a function of ligand presence, multimer/monomer status, local residue solvent exposure, pLDDT, and amino acid type. Overall performance was found to be excellent.

Introduction

The deep-learning protein-structure prediction tool AlphaFold2 (AF2)[1] has generated enormous excitement in broad areas of biology and medicine[2] due to its accurate prediction of protein structure from sequence, an outstanding challenge in biology [3]. With very many protein structures still experimentally undetermined, such high-quality predicted structures are relevant to drug design, understanding protein evolution and pathogenic mutations, protein-or-peptide-based therapeutics, and structure-guided rational protein design[4–10].

AF2 combines multi-sequence analysis with advanced structural pattern recognition that accounts for intra-protein epistasis (amino acid correlations in space not captured by the sequences directly that modify each amino acid's structure and substitution likelihood[11,12]). With additional vast amounts of structural and sequence data available for training, such deep learning of the 3D residue correlations has created unprecedented descriptive performance[5,6,13]. Critical assessment of AF2 requires curated test sets that e.g., separate multimer and monomer proteins, compare the same structures (i.e., without gaps or substitutions e.g.), account for data-bias towards proteins more represented by experimental structures, and quantify performance on a per residue-level for focused application. As independent assessments emerge, a high accuracy of AF2 seems to be confirmed.[14] Performance is naturally best for structured parts, but this is not critical since dynamic coils and loops are likely less-well described by any static structure. The method is more accurate for protein structures best covered empirically, and work remains to determine confidence for structurally rare proteins with less coverage, membrane proteins, multimeric proteins[15], proteins with ligands, cofactors or modifications, and intrinsically disordered proteins[5].

Given its importance to human health applications, the AF2-predicted structures of the human proteome were recently fully documented.[16] The present work constitutes a complimentary external validation of these data focusing on the local solvent exposure of a residue, which provides important information on protein compactness, function, site evolution rates, and protein stability effects, which directly depend on residue solvent exposure[17–19]. Accurate residue solvent exposure via AF2 could lead to improved estimates of structure-informed substitution frequencies and protein stability effects[19] making this parameter a natural and interpretable coordinate in comparison to e.g. XYZ coordinates or Ramachandran torsion angles, and perhaps more suitable for external validation as not directly considered in the training of the method.[1] Below we report this analysis for human proteins.

Methods

Structures used in the test set

There are three reasons for our focus on human proteins: 1) They are of particular interest from a health perspective (the motivation in the AF2 paper on the human proteome, Tunyasuvunakool et al.[16]); 2) the human proteome has a high coverage of experimental structures making validation meaningful, and mixing organism orthologs would potentially add noise and representation bias; 3) we wanted a focused external validation of the fully published AF2 data.[16]

The AF2 predictions for 20,504 proteins of the *Homo sapiens* reference proteome (UniProt proteome ID UP000005640) were downloaded from https://AlphaFold.ebi.ac.uk/download on September 16, 2021 (version v1, Data availability statement)[20]. AF2 predictions are improved continuously based on feedback from the community, producing new versions regularly. The version tested is published and documented,[16] in contrast to newer versions. Our study should thus be seen as a test/validation in direct comparison to Tunyasuvunakool et al.[16] In addition, PDB files for all human protein structures solved by X-ray crystallography with a resolution ≤ 2.0 Å (19,710 PDB structures) were retrieved from the protein data bank[21,22] (https://www.rcsb.org) in .pdb format on October 1, 2021. These structures were then further curated into relevant non-biased test sets and subsets to account for performance confounders, as described below.

Matching AF2 structures to experimental structures

To compare protein structures predicted by AF2 to protein structures solved experimentally, we first needed to match each experimental structure to the AF2 prediction of the protein. This was done in a two-step process in which each AF2 structure was first paired to one or more experimental PDBs by matching UniProt numbers, if possible, and then the correct experimental chain structure was identified by sequence alignment as outlined in **Figure 1**. We disregarded all structures with R > 2 Å, as lower-resolution structures could impact the benchmarking.

To match by UniProt number, the unique UniProt number of each AlphaFold structure was used to search through the retrieved PDBs described above. This resulted in 3930 AF2-structures being matched to 18,389 unique experimental PDBs. Some AF2-structures were matched to several PDBs, and some PDBs were matched to more than one AF2-structure, if the PDB represented the structure of a heteromultimer (**Figure 1A**). All PDBs with multiple chains were classified as multimers, since even shorter chains (peptide binders etc.) would make the chemical composition of the compared states different and potentially confound comparison.

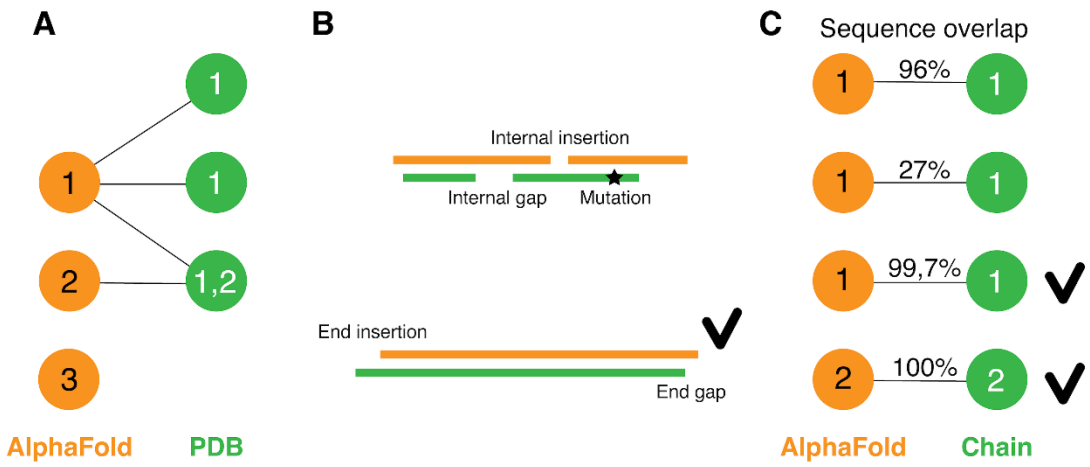

Figure 1. Protocol for matching AlphaFold structures to experimental structures. (A) Each AlphaFold structure was matched to experimental PDBs by UniProt number. AlphaFold structures can match none, one or more PDBs. PDBs can match one or more (i.e., heteromultimer) AlphaFold structures. **(B)** Examples of alignments of AlphaFold and experimentally solved sequences (from the ATOM field of the PDB file). Pair alignments with internal gaps, insertions or mutations (upper panel) were discarded, while gaps or insertions at ends were tolerated. **(C)** Four different pairs with different degrees of sequence overlap (exemplified by indicated percentages). Numbers inside circles indicate UniProt ID. Checkmarks indicate pairs that are part of the final dataset.

To 1) identify the correct experimental chain structures for each AF2-structure and 2) assess how much of the protein chain was actually solved in the crystal structure, the resolved amino acid sequences were extracted from the PDB files using the residue information in the ATOM field, not the SEQRES field, of the PDB. The resolved amino acid sequence of each chain was then aligned by pairwise alignment to its UniProt-paired amino acid sequence. Alignments were performed with the pairwise2 BioPython module with the settings: global, no penalization of end gaps, +2 for identity, -1 for mismatch, -2 for gap opening, and -1 for gap extension. Experimental sequences that had either internal gaps, internal insertions, or mutations compared to the UniProt (and thus AF2) sequence were excluded from further analyses (**Figure 1B**). This resulted in 16,036 pairs, each pair consisting of one AF2 structure and one experimental chain structure (in the following referred to as data pairs).

The set of comparable pairs resulting from this protocol comprised 2249 unique AF2 structures, 12,520 unique experimental chains, and 7585 unique experimental PDBs. In the AF2 dataset, proteins longer than 2700 amino acid residues are split into 1400 amino acid long,

overlapping structure fragments that may all match the same experimental chain given the described matching protocol[20]. Therefore, in this dataset, the number of data pairs exceeds the number of experimental chains, and the number of unique AF2 structures exceeds the number of unique human proteins (1818). This dataset thus emphasizes the experimental structures with either 100% sequence identity to the AF2 sequence or with gaps or insertions only in the beginning or end of the sequence compared to the AF2 sequence (**Figure 1C**). We consider this curation important, since experimental PDB structures with missing or modified parts could erroneously affect performance.

Next, we reduced this dataset further to include only comparable pairs of structures for which the overlap between sequences comprised >99% of the experimental and the AF2/canonical sequence length. The proportion of sequence overlap was calculated as the length of overlap divided by the length of sequence, and the data pairs were restricted to those whose length of overlap constituted >99% of both the UniProt sequence and the experimental sequence. This last exclusion step resulted in a final dataset consisting of 1264 data pairs comprising 115 unique AF2 structures and 652 unique PDBs (**Table S1**, provided as a separate file). The length of overlapping sequence in the data pairs ranged from 68 to 888 residues with an average length of 197 residues. The above procedure was performed with Python using BioPython v.1.78 modules[23].

Calculation of solvent accessible surface

The relative solvent accessible area (RSA) is a descriptor of local conformation in the context of the overall protein structure that is functionally relevant for e.g., substitution frequencies and protein stability.[17–19] An additional reason for using this metric is that we wanted a strong external validation of the AF2 human proteome data[16], and local backbone XYZ coordinates and Ramachandran torsion angles were trained by the AF2 procedure or refined in the force field step.[1] RSA was computed using the open-source software FreeSASA available as a Python module using the classifier configuration option "naccess"[24]. We only used the solvent accessibility of the backbone (i.e., including the C_α atom) without the side-chain, as the prediction of side-chain orientation will have less certainty[1] and confound the analysis unfairly, as side chain conformations are inherently dynamic and not necessarily important to be accurately described vs. a crystal structure (the solution dynamics of surface residue side chains can be studied by other non-static methods such as molecular dynamics and NMR).

The RSA is defined as the ratio of the absolute solvent accessible area of a residue divided by that within an Ala-X-Ala tripeptide, where X is the residue studied. For each data pair, the RSA

values for each residue in the AF2-structure (RSA_{AF}) and the experimental structure (RSA_{Exp}), respectively, were computed with FreeSASA. Then, each residue in the overlapping portion of the two sequences were matched using residue number and residue type, and the RSA_{AF} and RSA_{Exp} values were assigned to each common residue. Because the numbering of residues in the experimental PDBs may not be consistent with the numbering in the AF2-structures, a temporary common number, derived from the pairwise sequence alignments, was assigned to each residue to allow unambiguous matching of residues. The first and last residue in each structure, both AF2 and experimental, were excluded from the analyses due to the risk of overestimated RSA values.

The typical RSA is between 0 and 1, i.e., a residue with RSA = 0 is fully buried with no access to water molecules, and RSA = 1 is highly exposed. However, since the RSA is calculated as the SASA relative to a reference SASA for X in Ala-X-Ala tripeptides in a stretched conformation[24], a residue X may sometimes be more exposed in a specific protein structure, and thus RSA values will sometimes be larger than 1[25]. This effect would not affect the benchmark except by a simple scaling of the RSA of X in both the AF2 and experimental PDB, and thus also scale the deviation correspondingly, typically up to 20%[25], although for a very few cases, the effect can be larger.

AlphaFold per-residue confidence values

AlphaFold predicts residue coordinates with different confidence using a per-residue estimate called predicted lDDT-C_α (pLDDT), which has numbers on a scale from 0 to 100[1,26]. In the AlphaFold DB, pLDDT values for each residue are stored in the B-factor fields of PDB files[20], and to assess the effect of pLDDT, the B-factor fields were extracted from the PDB files using BioPython v.1.78 modules.

Ligands

Experimental structures can have non-covalently bound ligands such as metabolites, inhibitors, drugs, cofactors, and ions[27]. To assess the effect of such co-solutes on performance, the experimental structures in the dataset were classified according to ligands. Ligand information was extracted from the HET field in the PDB files using Python, and experimental structures with no ligands or only water or the small ions Na^+, Cl^-, K^+, F^-, Li^+, Br^-, I^-, SO_4^{2-}, or PO_4^{3-} (HET codes HOH, NA, CL, K, F, LI, BR, I, SO4, and PO4) were classified as having no ligands.

Dataset subsets

The dataset (**Table S1**) was divided into six disjointed subsets based on 1) proportion of sequence overlap (100 %, or >99% and <100%), 2) resolution of the experimental structure (<= 1.5 Å, or >1.5 Å and <=2.0 Å), and 3) whether the experimental structure is a monomer. Since the number of pairs with 100% sequence identity was small, this group was not split based on resolution. The six data subsets are shown in **Table 1**.

Identification of chain-interface residues in multimers

In multimeric proteins, residues from one chain may be in close contact to residues from another chain, potentially affecting the solvent accessibility of these residues. To analyze the effect of these chain-interface residues, all experimental multimeric structures in the dataset were analyzed with BioPython to identify chain-interface residues in the chains that were paired to an AF2-structure. For each atom of each residue in the chain, the distance to atoms belonging to other chains in the structure was calculated, and residues that contained an atom that were less than 3.5 Å from atoms in other chains were classified as chain-interface residues.

Results and discussion

Performance of AF2 as measured by relative solvent accessibility

Since only 35% of human proteins feature (often partial) PDB structures[16], AF2 could have major impact for understanding human diseases in protein-structural context. There are multiple ways of assessing a structural prediction by AF2, including various metrics of the accuracy and precision, and various target descriptors[14]. In this work we emphasized RSA as a directly interpretable endpoint of evaluation, or a "natural" coordinate" of each residue in the context of the protein fold with direct implication for function and evolution, with the further advantage from an external validation perspective that it was not directly optimized by the AF2 procedure.

We compared AF2 predictions of proteins in the human proteome as fully documented previously[16] to their experimentally solved structures. We limited the comparison to structures solved by X-ray crystallography with a resolution of 2.0 Å or better. Of the 23,391 human protein chains predicted by AF2[16,20], 115 were successfully matched to 1264 experimental chain structures from 652 PDBs. In order to avoid erroneous effects, we excluded experimental structures with internal gaps, insertions or mismatches (mutations) compared to the UniProt sequence (i.e. only 100% sequence identity or experimental structures with gaps or insertions only in the beginning or end of the sequence) (**Figure 1B**). For pairs of AF2- and experimental structures with less than 100% sequence identity, we only included overlaps between sequences of more than 99% of the sequence length (**Figure S1**). The final dataset included 1264 pairs comprising 115 unique AF2-structures and 652 unique PDBs (**Table S1**). The overlapping sequences ranged from 68 to 888 residues with an average length of 197 residues.

We then calculated the RSA for each residue backbone in both the AF2-generated (RSA_{AF}) and experimental structures (RSA_{Exp}), and compared the RSA values for each matched residue in each structure pair. AF2 generally predicts side-chain coordinates with less accuracy than backbone coordinates[1], but the dynamics expected in a solution state makes the exact coordinates of exposed side-chains in a static PDB structure less significant; accordingly, we compared only backbone RSA values. Scatter plots showing the correlation between RSA_{AF} and RSA_{Exp} for each of the 1264 analyzed data pairs are shown in **Figure S2.** In order to understand AF2's performance in detail, we divided the data pairs into six non-overlapping groups based on 1) length of sequence overlap, 2) resolution of experimental structure, and 3) monomer vs. multimer structures. We compared the RSA for each matched residue in the structure pairs belonging to the six groups (**Figure 2A**). For each group, we calculated the mean absolute error (MAE), the mean signed deviation (MSD), and the standard deviation of the absolute errors (SA_{Abs}, **Table 1**).

Table 1. Characteristics of the six groups of data pairs

	Sequence overlap (%)	Resolution (Å)	Mer	$N_{Residues}$	N_{Pairs}	N_{AF2}	MAE	MSD	SD_{Abs}
Not weighted	100%	<=2.0	Mono	10597	55	14	0.044	0.003	0.077
	100%	<=2.0	Multi	15071	100	23	0.068	0.033	0.126
	>99% <100%	<= 1.5	Mono	13370	59	19	0.047	0.001	0.077
	>99% <100%	<= 1.5	Multi	35876	214	30	0.067	0.032	0.124
	>99% <100%	>1.5 <=2.0	Mono	38888	129	31	0.043	0.000	0.071
	>99% <100%	>1.5 <=2.0	Multi	133113	707	66	0.070	0.037	0.132
Weighted	100%	<=2.0	Mono	3167	-	14	0.047	0.003	0.082
	100%	<=2.0	Multi	3556	-	23	0.072	0.032	0.127
	>99% <100%	<= 1.5	Mono	4213	-	19	0.044	0.000	0.072
	>99% <100%	<= 1.5	Multi	6249	-	30	0.059	0.021	0.106
	>99% <100%	>1.5 <=2.0	Mono	7936	-	31	0.045	0.000	0.074
	>99% <100%	>1.5 <=2.0	Multi	15428	-	66	0.062	0.026	0.112

The MAE, the MSD, and the spread of absolute deviations are very dependent on the monomer-multimer status of the experimental structure, but not on the length of the sequence overlap or the resolution of the experimental structure (**Table 1**). The higher MAE, MSD, and spread of differences for multimeric structures is likely caused mainly by residues that have low RSA values in the experimental structure but high RSA values in the AF2 structure (**Figure 2A,** lower panels). This illustrates why caution should be exercised when deducing surface structure for multimeric proteins based on AF2 predictions for a single monomer, especially at the interfaces. After establishing this, we restricted further analysis mainly to monomer comparisons.

There was very large variation in the number of experimental structures matching individual AF2-structures (**Table S1**). Some AF2-structures matched only one experimental structure whereas a few (such as hemoglobin subunit alpha, UniProt ID P69905) matched hundreds of experimental structures (**Figure S3**). This variation may potentially skew the error metrics (MAE, MSD and SD_{Abs}) towards those of highly represented structures. To adjust for this bias, we calculated the average RSA_{Exp} for each residue that were paired to the same AF2 structure if the pair belonged to the same overlap-, resolution- and monomer/multimer group described above **(Figure 2B)**. Similar to the first analysis (**Table 1**, upper half), the absolute and signed deviations calculated from the averaged RSA_{Exp} values resulted in MAE, MSD, and SD_{Abs} values that depended highly on the monomer-multimer status, but not on the length of sequence overlap or resolution of the experimental structure (**Table 1**, lower half).

Table 2. The effect of ligands in the experimental structure. Ligands are heteromolecules that are not H_2O, Na^+, Cl^-, K^+, F^-, Li^+, Br^-, I^-, SO_4^{2-}, or PO_4^{3-}.

Sequence overlap (%)	Resolution (Å)	Mer	Ligands	$N_{Residues}$	N_{Pairs}	N_{AF2}	MAE	MSD	SD_{Abs}
100	<=2.0	Mono	No	1306	7	6	0.044	0.000	0.079
100	<=2.0	Mono	Yes	9291	48	10	0.044	0.003	0.077
100	<=2.0	Multi	No	1624	11	4	0.080	0.043	0.153
100	<=2.0	Multi	Yes	13447	89	21	0.066	0.032	0.122
>99 <100	<= 1.5	Mono	No	558	4	3	0.040	0.005	0.065
>99 <100	<= 1.5	Mono	Yes	12812	55	19	0.047	0.001	0.077
>99 <100	<= 1.5	Multi	No	603	4	4	0.075	0.022	0.139
>99 <100	<= 1.5	Multi	Yes	35273	210	26	0.067	0.032	0.124
>99 <100	>1.5 <=2.0	Mono	No	3980	25	7	0.044	0.002	0.075
>99 <100	>1.5 <=2.0	Mono	Yes	34908	104	26	0.043	-0.001	0.071
>99 <100	>1.5 <=2.0	Multi	No	8165	39	12	0.070	0.039	0.137
>99 <100	>1.5 <=2.0	Multi	Yes	124948	668	60	0.070	0.037	0.132

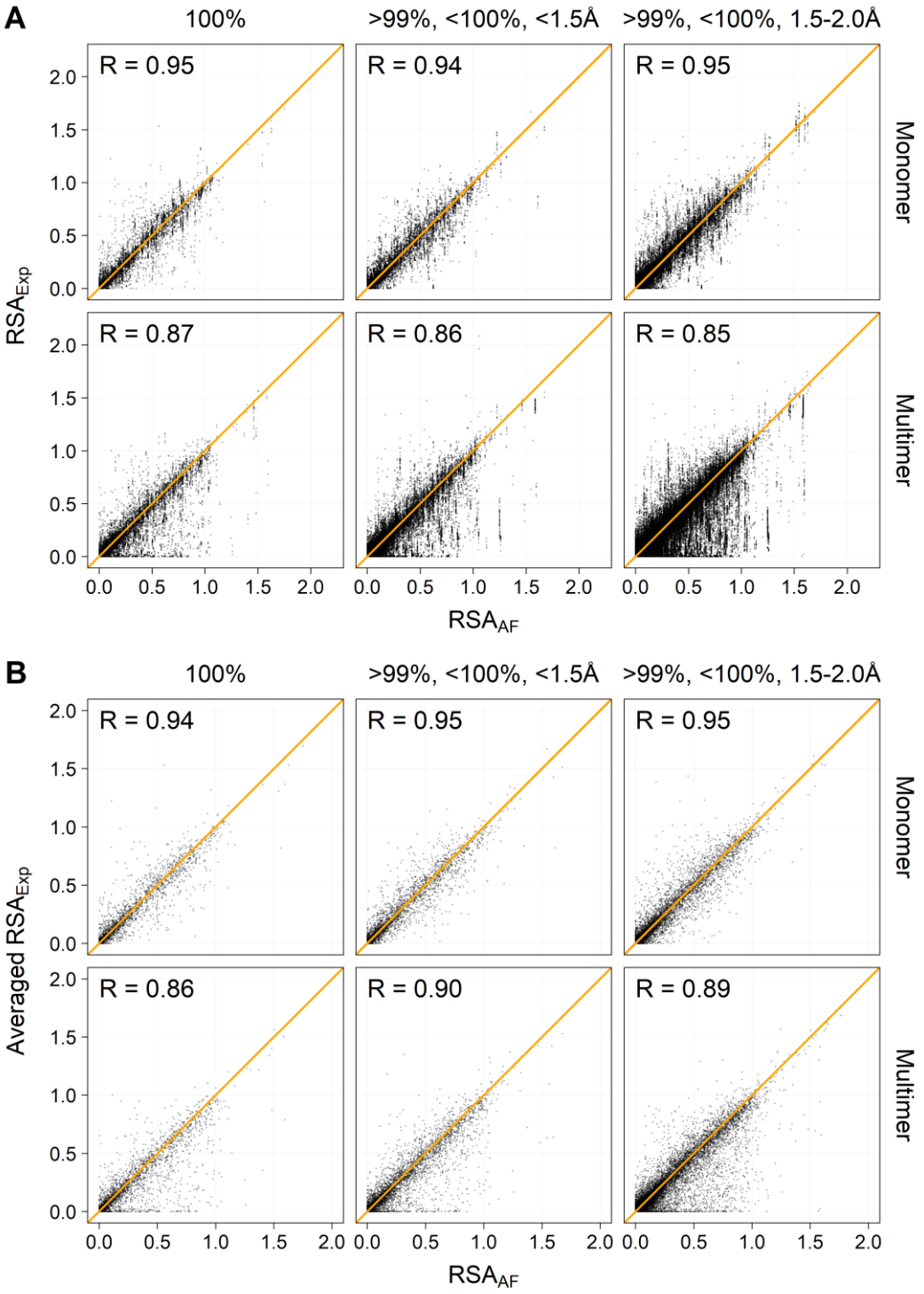

Figure 2. Experimental vs. AF2 RSA values. Each panel shows results grouped based on sequence overlap, resolution of the experimental structure, and monomer-multimer status of the experimental structure (indicated above and to the right). The six groups are disjointed from each other. The orange line represents the ideal where the RSA_{AF} are equal to RSA_{Exp}. The correlation coefficient R is indicated for each group. **(A)** Each dot represents a residue belonging to a data pair. **(B)** Each dot represents the average of experimental residues belonging to the same AF2-structure and the same overlap-, resolution- and monomer/multimer group.

The effect of non-covalently bound ligands on prediction performance

Many experimental protein structures in the PDB have non-covalently bound ligands such as metabolites, inhibitors, drugs, cofactors, ions, and solvent[27]. We therefore tested whether the observed differences between RSA_{AF} and RSA_{Exp} correlated with the presence of ligands, by first classifying the experimental structures in the dataset as being with or without ligands, and then measuring the RSA prediction accuracy separately on these two subsets. For this analysis, water or the small ions, Na^+, Cl^-, K^+, F^-, Li^+, Br^-, I^-, SO_4^{2-}, or PO_4^{3-} were not counted as ligands. We found no apparent correlation between the MAE, the MSD, or SD_{Abs} and the presence of ligands (excluding water and small ions, **Table 2**); the performance for the monomer structures was very similar regardless of ligand presence. Since ligands vary substantially in their chemical structure, size and binding pockets, we did not want to speculate on the reason, especially as such analysis would be biased towards the ligands favored in the human proteins and PDB, but suggest that this is explored in future work. We therefore continue our analysis below without separating out ligand-containing structures.

Larger disagreement for lower-confidence scores and exposed residues

In order to determine the factors that affect the accuracy of RSA predicted by AF2, we first hypothesized that residues with high prediction confidence or residues buried in the core of the protein were predicted more accurately. To test this hypothesis, we only analyzed data pairs in which the experimental structure is a monomer. AF2 produces an estimate of confidence for each residue, predicted lDDT-C_α (pLDDT) based on the Local Distance Difference test, on a scale from 0–100, where 100 indicates the highest possible confidence[1,16,26]. Regions with pLDDT > 90 are expected to be modelled to high accuracy, whereas regions with pLDDT < 50 are most likely either unstructured or only structured as part of a complex[20]. We hypothesized that the confidence of AF2-predicted residue coordinates correlates with the ability to predict the RSA of a residue. If so, the MAE or the SD_{Abs} between RSA_{AF} and RSA_{Exp} would be lower for residues with high pLDDT values and *vice versa*. To test this hypothesis, we divided the residues of all AF2-structures that were matched to a monomeric experimental structure into pLDDT bins as shown in **Figure 3A**, and calculated the MAE, MSD, SD_{Abs}, and SD_{Signed} for each group (**Table S2**). As can be seen in **Figure 3A** (and **Figure S4** and **S5)**, the MAE and the SD_{Abs} are indeed dependent on pLDDT with both values increasing substantially for residues with <90 pLDDT scores. The MSD is less dependent on pLDDT, i.e., there is no tendency for RSA_{AF} to be either under- or overestimated for residues with low pLDDT values.

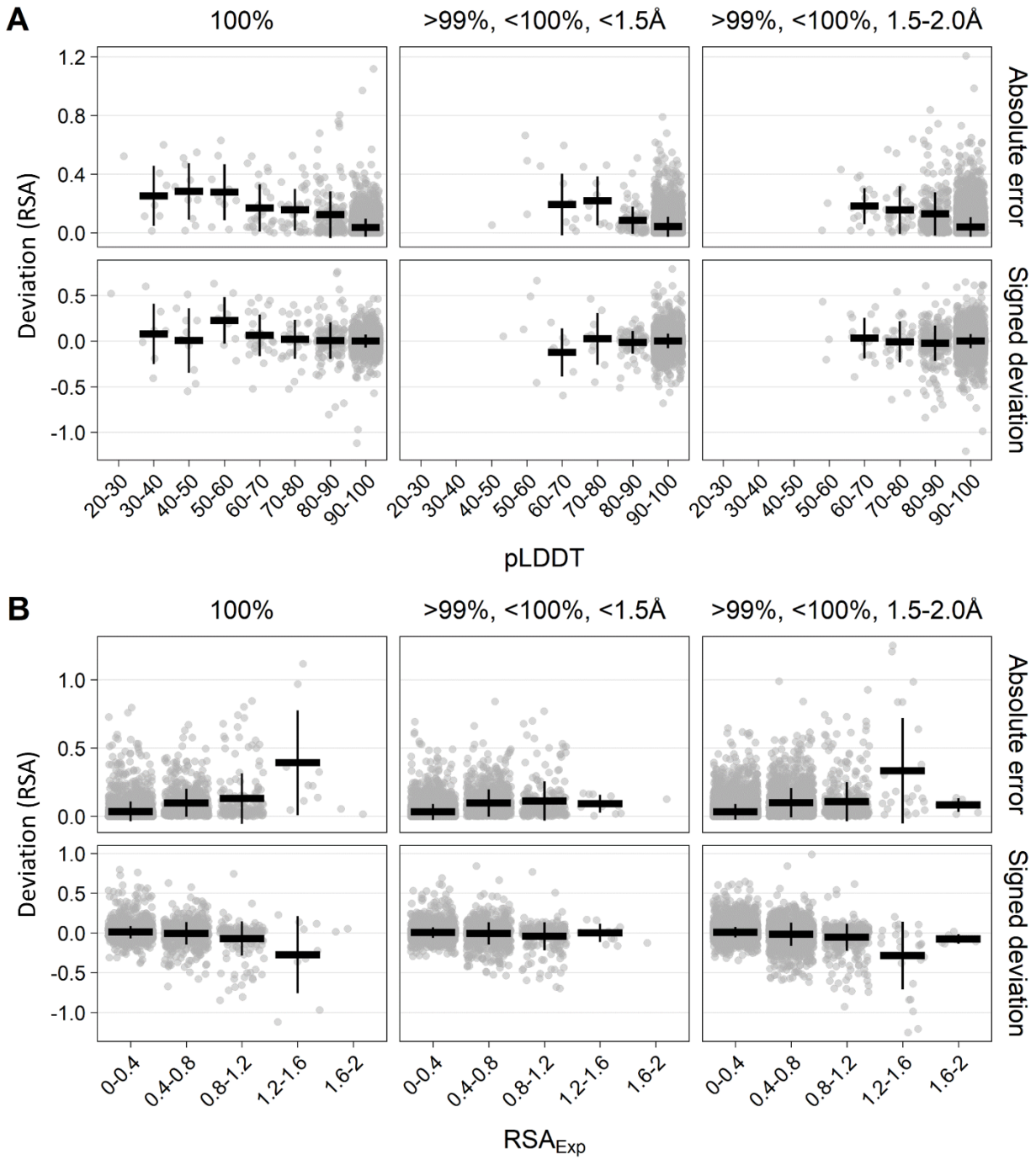

Figure 3. Deviations in relation to pLDDT and RSA. Data pairs are grouped by sequence overlap and resolution of the experimental structure (only monomers). Gray circles indicate the absolute or signed deviation for the average RSA_{Exp} for each unique AF2-structure belonging to the same overlap-, resolution- and monomer/multimer group. The mean deviations for each pLDDT or RSA_{Exp} group are shown as horizontal black bars and one standard deviation (SD) of the mean errors as vertical black bars. **(A)** Deviation as a function of pLDDT. The pLDDT values are grouped in 10 percent-point bins. **(B)** Deviation as a function of RSA_{Exp}. The experimental RSA values are grouped in 0.4 RSA bins. Data for this plot are in **Table S2** and **Table S3**.

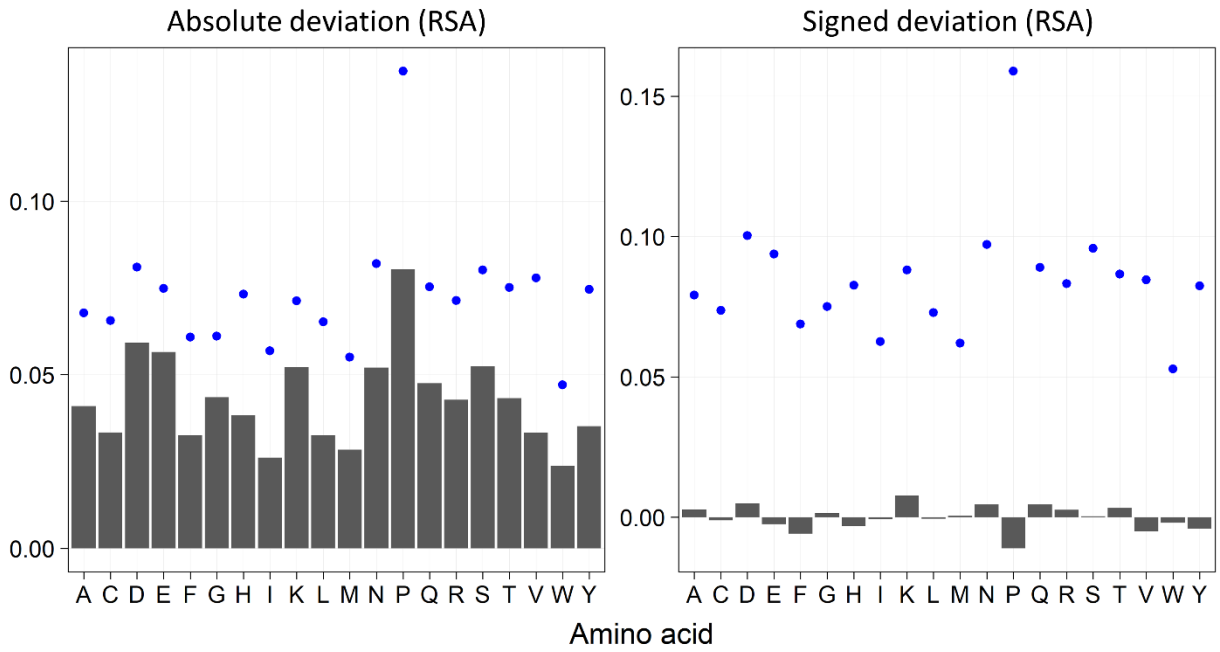

Figure 4. Agreement between experimental and AF2 structures depending on residue type. Left: MAE. Right: MSD ($RSA_{AF} - RSA_{Exp}$)**.** Only pairs with monomeric experimental structures included. The standard deviations are shown as blue dots.

Next, to test if the ability of AF2 to predict RSA depended on experimentally determined solvent accessibility of the residues (RSA_{Exp}), we divided the residues of all AF2-structures that were matched to a monomeric experimental structure into RSA bins, as shown in **Figure 3B**, and calculated the MAE, MSD, SD_{Abs}, and SD_{Signed} for each group (**Table S3**). As can be seen in **Figure 3B** (and **Figure S6** and **S7)**, the MAE and the SD_{Abs} are dependent on the experimental RSA values: Residues with low RSA, i.e. buried residues were predicted more accurately by AF2 than surface exposed residues. There is a tendency for the MSD to be more negative for high RSA residues, meaning that AF2 tends to underestimate the RSA for more surface exposed residues. The observation that disagreement is larger for exposed residues, and for residues with lower confidence, is of course correlated and expected, but the numbers in **Figure 3** gives an estimate of the magnitude of the deviations that can be expected for each class of residue. The outliers are relatively few in both cases, and the tendency of many outliers at low exposure or high confidence relates to the fact that there are many data points in total of these types.

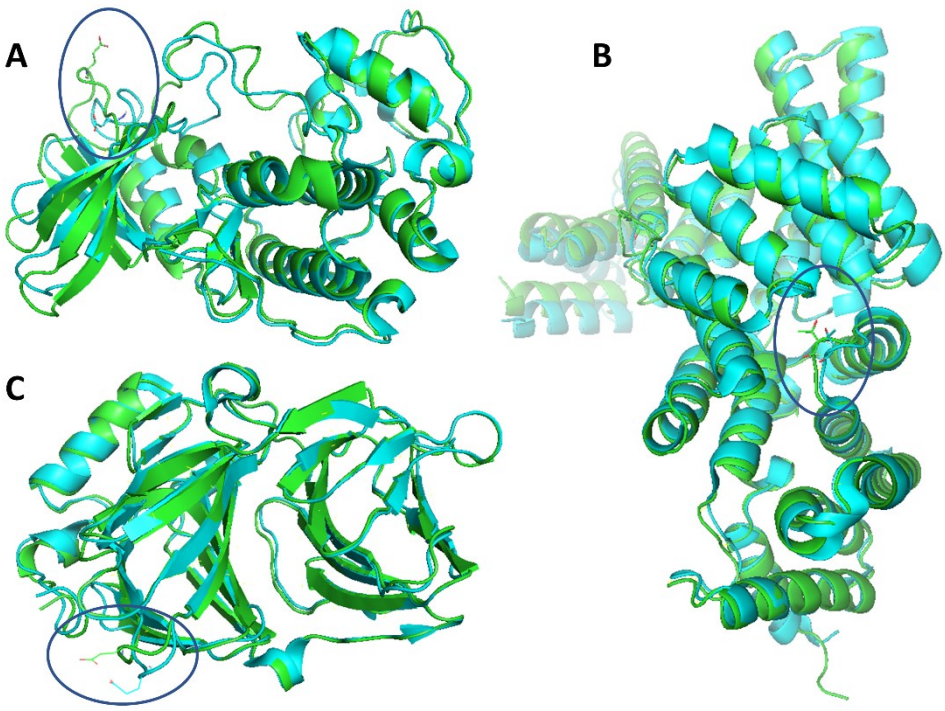

Figure 5. Examples of proteins with larger disagreements of residue solvent exposure. Green: AF2. Cyan: experimental. **(A)** Glu-40 of Human cyclin dependent kinase 2 (UNIPROT: P24941; PDB: 2B54). **(B)** Thr-47 of human interferon-induced protein with tetratricopeptide repeats (IFIT5) (UNIPROT: Q13325; PDB: 4HOR). **(C)** Glu-108 of human pirin (UNIPROT: O00625; PDB: 4EWE).

Larger disagreement for polar residues and proline

The accuracy of AF2-predicted RSA could also depend on amino acid type. In order to understand this, the same analysis as above (**Figure 3**) was performed but with all groups combined, and errors divided into amino acid type (**Figure 4**). The effect of amino acid type was remarkably large, more than 100% even after averaging errors across all pair comparisons. The best-described amino acids tend to be isoleucine, leucine, methionine, phenylalanine, tryptophan, cysteine, and valine, which are hydrophobic. The worst-described amino acids are polar, such as aspartate and glutamate, lysine, asparagine, and serine. However, proline is clearly hardest to predict. These observations can be explained by polar residues and proline being more often located in less-well-described surface areas of the proteins, i.e., with a correlation to the pLDDT/RSA in **Figure 3**.

To put these tendencies into context, **Figure 5** shows examples from some PDB structures of residues in proteins that displayed largest disagreement with AF2, marked in blue circles. These tended to locate in loops either on the surface or in cavities of the proteins. **Figure 5A** shows Glu-40 of Human cyclin dependent kinase 2 (UNIPROT: P24941; PDB: 2B54) in a loop-dominated

part of the protein that is generally less well-described (absolute RSA deviation: 0.90). **Figure 5B** shows Thr-47 in a partly solvated cavity of the RNA-binding protein human interferon-induced protein with tetratricopeptide repeats (IFIT5) (UNIPROT: Q13325; PDB: 4HOR); giving an absolute RSA deviation of 0.52, and **Figure 5C** shows Glu-108 in human pirin (UNIPROT: O00625; PDB: 4EWE), which produced an RSA deviation of 0.38.

While expected, **Figure 3** and **Figure 4** quantify the expected differences vs. experimental PDB structures in a prediction for each type of residue, and **Figure 5** gives examples of large errors, of interest to applications. Despite these deviations, it is encouraging to see that the overall magnitude of the MAEs is in the range of 0.02-0.08. With RSA-values averaging to ~0.2, i.e., the expected deviation in RSA from an AF2 prediction is typically of the order of 20%, which we consider a relevant "natural" measure of the conformational uncertainty. A corresponding analysis including multimers (**Figure S8**), although confirming the challenge of proline conformations, produced more variable results due to some hydrophobic residues poorly described due to location on multimer interfaces (i.e., exposed hydrophobic residues).

AF2 prediction of residues in multimeric structures

The AF2 predictions are all single chains whereas most of the experimental structures are multimeric, either homomultimers or heteromultimers. We showed above that the correlation between RSA_{AF} and RSA_{Exp} values is substantially stronger for experimental monomers compared to multimers, and that the difference seems to be caused by residues that have low RSA_{Exp} but high RSA_{AF} (**Figure 2**). This makes intuitive sense as residues located in the interface between two chains may have lower solvent accessibility.

To test if the lower correlation between RSA_{AF} and RSA_{Exp} values observed in multimeric data pairs are caused by residues in chain interfaces, we identified such residues in the experimental structures and removed them from analysis. Chain interface residues were defined as residues whose atoms had a distance <3.5 Å to atoms in other chains. As can be seen in the left panels of **Figure 6,** the correlation between RSA_{AF} and RSA_{Exp} for interface residues is weak. The average RSA_{AF} for all identified interface residues in the multimer dataset (n = 2751) is 0.323 whereas the average RSA_{Exp} (using the per-AF2-averaged RSA_{Exp} values) is 0.135.

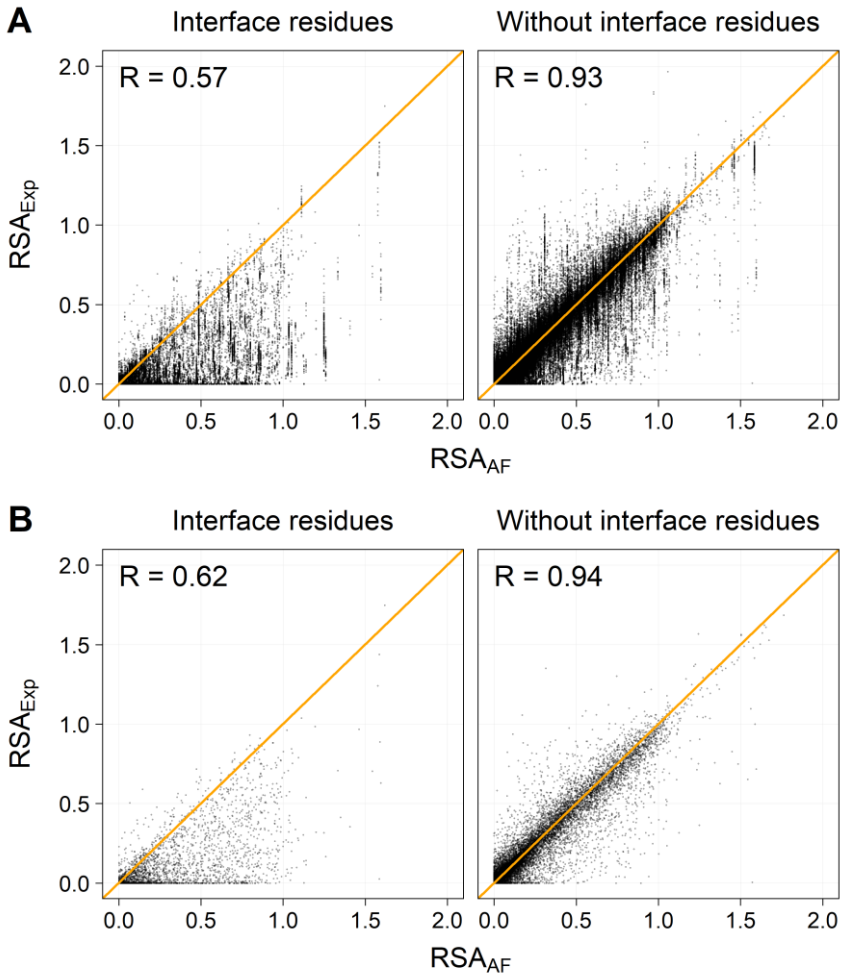

Figure 6. Experimental vs. AF2 RSA values in multimeric experimental structures. Correlation of RSA_{AF} and RSA_{Exp} for interface residues (left) and for non-interface residues (right). Orange lines represent the ideal where the RSA_{AF} are equal to RSA_{Exp}. **(A)** Each dot represents a residue belonging to a data pair. **(B)** Each dot represents the average of experimental residues belonging to the same AlphaFold structure.

Thus, AF2 predictions on single chains substantially overestimate the solvent accessibility of these residues. If interface residues are removed from the multimeric data pairs, the average RSA_{AF} of the remaining residues (n = 17,895) becomes 0.199 compared to an average RSA_{Exp} (using the per-AF2 averaged RSA_{Exp} values) of 0.190. This also results in a stronger correlation between RSA_{AF} and RSA_{Exp} that mimics the correlation of monomeric data pairs (compare right panels of **Figure 6** to upper panels of **Figure 2**). In summary, because the AF2 predictions are based on single protein chains, it overestimates the solvent accessibility of the ~10% of residues in the dataset that are in close contact (by the above definition) to other protein chains, and this should always be accounted for in predictions.

Variation in RSA values among experimental structures of the same protein chain

Some AF2-structures are matched to many experimental structures (**Figure S3**). The different experimental structures representing one particular protein chain (and therefore one AF2-structure) can be compared to each other to determine the amount of variation in RSA values that exists for each residue among experimental structures. To evaluate the variation within the experimental dataset used to assess the prediction accuracy of AF2, we calculated the per-residue spread in RSA_{Exp} and compared it to the per-residue MAE for AF2-structures matched to five or more monomeric experimental structures (n = 10).

An example is shown in **Figure S9** for peptidyl prolyl *cis/trans* isomerase A (UniProt ID P62937) for which there are 52 monomeric experimental structures in the dataset. **Figure S9A** shows the SD of RSA_{Exp} (SD_{Exp}) for each residue in the chain. Although the experimental structures are all monomers, and variation caused by chain interactions are therefore not relevant, some residues have very different RSA values in the 52 structures while others only show little RSA variation. For comparison, **Figure S9B** shows the absolute deviations between RSA_{AF} and RSA_{Exp} for each residue along the protein chain.

To better understand how experimental variation affects the performance benchmark, **Figure 7A** shows correlation between the per-residue variation in RSA_{Exp} (measured as SD_{Abs}) and the per-residue MAE for the 2636 compared residues. The outliers in **Figure 7A** can be interpreted in terms of confidence in the RSA_{AF} prediction. Residues for which the RSA is similar among the experimental structures (low SD_{Abs}) but different from the AF2-predicted RSA (high MAE) can be interpreted as residues that are unlikely to be predicted accurately. In contrast, residues for which there is a high variation in RSA among the experimental structures (high SD_{Abs}) but a low MAE when compared to the AF2-prediction, indicate residues for which some experimental structures may be problematic.

Furthermore, **Figure 7B** shows that the variation in RSA among experimental structures depends to some extent on RSA_{Exp} with more surface exposed residues exhibiting more variation. This structural heterogeneity in experimental structures not only puts a limit on the accuracy expected from a prediction method but also emphasizes the need to consider the experimental structure quality and heterogeneity in an assessment, e.g., by sensitivity analysis or precision estimates using multiple experimental structures in the study.

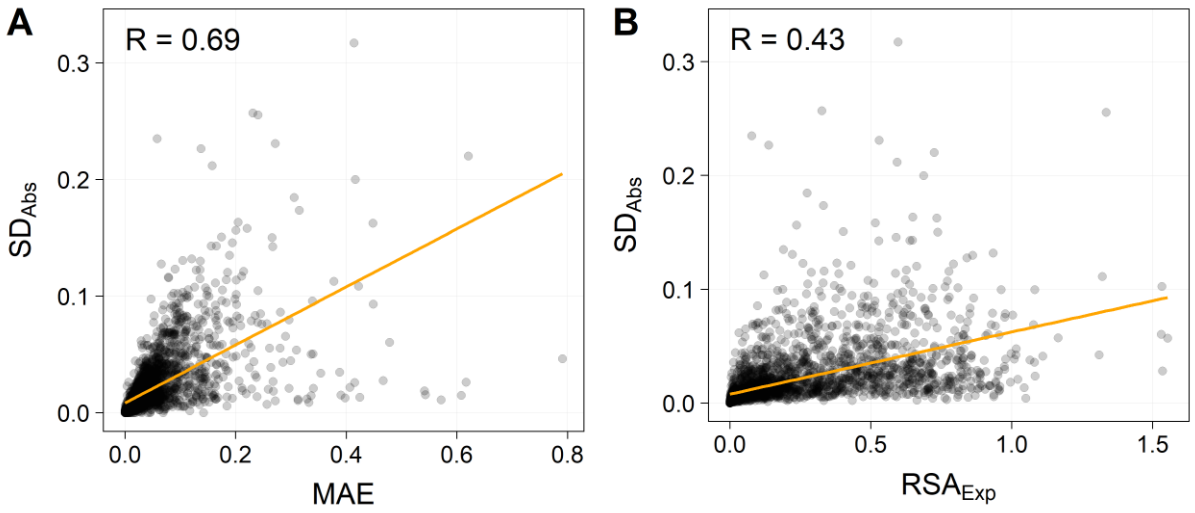

Figure 7. Internal variation among experimental structures. Each point indicates one residue in the monomeric data pairs for which there are 5 or more experimental structures per AlphaFold structure **A)** Correlation of SD_{Abs} and MAE, **B)** correlation between SD_{Abs} and RSA_{Exp}. Linear regressions and correlation coefficients (R) are shown.

Concluding remarks

We analyzed the performance of AF2 applied to human proteins, using the local residue's relative solvent exposure as a "natural" residue coordinate of functional and evolutionary importance. We carefully curated data sets by sequence overlap to avoid incomplete or erroneous comparions and explored the dependence of AF2 performance on monomer/multimer status (important), presence of cofactors and ligands, experimental resolution, exposure, and amino acid type.

AF2 performed excellently once comparing specifically to monomer proteins. However, notable challenges persist relating especially to proline and exposed residues. We identify larger disagreement for lower-confidence scores (pLDDT) and exposed residues on average, which also correlates with polar residues (Asp, Glu, Asn e.g.) being substantially less well described than hydrophobic residues. Polarity correlates with solvent exposure which correlates with being more dynamic and less well-described, but effects of the electrostatic treatment in the force field relaxation step of AF2 are also possible, since small variations in electrostatic point charge models of the polar residues affect hydration free energies[28] and thus interaction energies and conformations. In the paper, we provide estimates of such expected deviations divided into amino acid type, exposure, and pLDDT value, and also emphasize the structural heterogeneity and representation bias of experimental data themselves in such benchmarks.

An important point of caution is that in the application to real proteomes, the predictions of AF2 are likely worse due to training on usually unmodified or specifically modified proteins from the PDB, whereas many eukaryotic protein chains are heavily modified (truncated, glycosylated, etc.) sometimes in very diverse ways in their physiologically relevant forms[29]. Also, the protein structures used for training and assessment represent crystal state *in vitro* conditions not necessarily applicable to the pH, temperature, and macromolecular cellular environment where the protein is located. Thus, we are still far from *in vivo* predictability of protein conformational ensembles - but we are possibly close to understanding in vitro protein conformational ensembles, as this study has hopefully helped to document.

Acknowledgement. The Danish Council for Independent Research (Grant case: 8022-00041B) is gratefully acknowledged for supporting this work.

Supporting Information. Table S1 (separate file) contains data and performance metrics. The supporting information pdf file contains additional tables and figures with information relevant to analysis. This information is available free of charge at http://pubs.acs.org

Data availability. The structures benchmarked correspond to the first AF2 prediction release for the human proteome which is fully documented (in contrast to later version) by Tunyasuvunakool et al.[16] and thus serves as a complementary external validation of these data, in direct comparison.

https://ftp.ebi.ac.uk/pub/databases/alphafold/v1/UP000005640_9606_HUMAN_v1.tar

All data analyses were performed with Python v.3.8 or R v.4.0.5 and the code is available at https://github.com/ktbaek/AlphaFold.

References

(1) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction

with AlphaFold. *Nature* **2021**, *596*, 583–589 DOI: 10.1038/s41586-021-03819-2.

(2) Thornton, J. M.; Laskowski, R. A.; Borkakoti, N. AlphaFold Heralds a Data-Driven Revolution in Biology and Medicine. *Nat. Med.* **2021**, *27*, 1666–1669.

(3) Marks, D. S.; Hopf, T. A.; Sander, C. Protein Structure Prediction from Sequence Variation. *Nat. Biotechnol.* **2012**, *30*, 1072–1080.

(4) Jones, D. T.; Thornton, J. M. The Impact of AlphaFold2 One Year On. *Nat. Methods* **2022**, *19*, 15–20 DOI: 10.1038/s41592-021-01365-3.

(5) Perrakis, A.; Sixma, T. K. AI Revolutions in Biology. *EMBO Rep.* **2021**, *22*, e54046 DOI: https://doi.org/10.15252/embr.202154046.

(6) Skolnick, J.; Gao, M.; Zhou, H.; Singh, S. AlphaFold 2: Why It Works and Its Implications for Understanding the Relationships of Protein Sequence, Structure, and Function. *J. Chem. Inf. Model.* **2021**, *61*, 4827–4831 DOI: 10.1021/acs.jcim.1c01114.

(7) Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G. R.; Wang, J.; Cong, Q.; Kinch, L. N.; Schaeffer, R. D.; Millán, C.; Park, H.; Adams, C.; Glassman, C. R.; DeGiovanni, A.; Pereira, J. H.; Rodrigues, A. V; van Dijk, A. A.; Ebrecht, A. C.; Opperman, D. J.; Sagmeister, T.; Buhlheller, C.; Pavkov-Keller, T.; Rathinaswamy, M. K.; Dalwadi, U.; Yip, C. K.; Burke, J. E.; Garcia, K. C.; Grishin, N. V; Adams, P. D.; Read, R. J.; Baker, D. Accurate Prediction of Protein Structures and Interactions Using a Three-Track Neural Network. *Science* **2021**, *373*, 871–876 DOI: 10.1126/science.abj8754.

(8) Tong, A. B.; Burch, J. D.; McKay, D.; Bustamante, C.; Crackower, M. A.; Wu, H. Could AlphaFold Revolutionize Chemical Therapeutics? *Nat. Struct. Mol. Biol.* **2021**, *28*, 771–772.

(9) Monzon, V.; Haft, D. H.; Bateman, A. Folding the Unfoldable: Using AlphaFold to Explore Spurious Proteins. *Bioinforma. Adv.* **2022**, *2*, vbab043.

(10) Tsaban, T.; Varga, J. K.; Avraham, O.; Ben-Aharon, Z.; Khramushin, A.; Schueler-Furman, O. Harnessing Protein Folding Neural Networks for Peptide–Protein Docking. *Nat. Commun.* **2022**, *13*, 176 DOI: 10.1038/s41467-021-27838-9.

(11) Mumenthaler, C.; Braun, W. Predicting the Helix Packing of Globular Proteins by Self-correcting Distance Geometry. *Protein Sci.* **1995**, *4*, 863–871.

(12) Ortiz, A. R.; Kolinski, A.; Skolnick, J. Nativelike Topology Assembly of Small Proteins Using Predicted Restraints in Monte Carlo Folding Simulations. *Proc. Natl. Acad. Sci.*

1998, *95*, 1020–1025.

(13) Xu, J. Distance-Based Protein Folding Powered by Deep Learning. *Proc. Natl. Acad. Sci.* **2019**, *116*, 16856–16865.

(14) Akdel, M.; Pires, D. E. V; Pardo, E. P.; Jänes, J.; Zalevsky, A. O.; Mészáros, B.; Bryant, P.; Good, L. L.; Laskowski, R. A.; Pozzati, G.; Shenoy, A.; Zhu, W.; Kundrotas, P.; Ruiz Serra, V.; Rodrigues, C. H. M.; Dunham, A. S.; Burke, D.; Borkakoti, N.; Velankar, S.; Frost, A.; Lindorff-Larsen, K.; Valencia, A.; Ovchinnikov, S.; Durairaj, J.; Ascher, D. B.; Thornton, J. M.; Davey, N. E.; Stein, A.; Elofsson, A.; Croll, T. I.; Beltrao, P. A Structural Biology Community Assessment of AlphaFold 2 Applications. *BioRxiv* **2021,** 2021.09.26.461876. DOI: 10.1101/2021.09.26.461876

(15) Evans, R.; ONeill, M.; Pritzel, A.; Antropova, N.; Senior, A.; Green, T.; Žídek, A.; Bates, R.; Blackwell, S.; Yim, J.; Ronneberger, O.; Bodenstein, S.; Zielinski, M.; Bridgland, A.; Potapenko, A.; Cowie, A.; Tunyasuvunakool, K.; Jain, R.; Clancy, E.; Kohli, P.; Jumper, J.; Hassabis, D. Protein Complex Prediction with AlphaFold-Multimer. *bioRxiv* **2021,** 2021.10.04.463034. DOI: 10.1101/2021.10.04.463034.

(16) Tunyasuvunakool, K.; Adler, J.; Wu, Z.; Green, T.; Zielinski, M.; Žídek, A.; Bridgland, A.; Cowie, A.; Meyer, C.; Laydon, A.; Velankar, S.; Kleywegt, G. J.; Bateman, A.; Evans, R.; Pritzel, A.; Figurnov, M.; Ronneberger, O.; Bates, R.; Kohl, S. A. A.; Potapenko, A.; Ballard, A. J.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Clancy, E.; Reiman, D.; Petersen, S.; Senior, A. W.; Kavukcuoglu, K.; Birney, E.; Kohli, P.; Jumper, J.; Hassabis, D. Highly Accurate Protein Structure Prediction for the Human Proteome. *Nature* **2021**, *596*, 590–596 DOI: 10.1038/s41586-021-03828-1.

(17) Bloom, J. D.; Drummond, D. A.; Arnold, F. H.; Wilke, C. O. Structural Determinants of the Rate of Protein Evolution in Yeast. *Mol. Biol. Evol.* **2006**, *23*, 1751–1761 DOI: 10.1093/molbev/msl040.

(18) Zhou, T.; Drummond, D. A.; Wilke, C. O. Contact Density Affects Protein Evolutionary Rate from Bacteria to Animals. *J. Mol. Evol.* **2008**, *66*, 395–404 DOI: 10.1007/s00239-008-9094-4.

(19) Caldararu, O.; Blundell, T. L.; Kepp, K. P. Three Simple Properties Explain Protein Stability Change upon Mutation. *J. Chem. Inf. Model.* **2021**, *61*, 1981–1988.

(20) Varadi, M.; Anyango, S.; Deshpande, M.; Nair, S.; Natassia, C.; Yordanova, G.; Yuan, D.; Stroe, O.; Wood, G.; Laydon, A.; Žídek, A.; Green, T.; Tunyasuvunakool, K.; Petersen,

S.; Jumper, J.; Clancy, E.; Green, R.; Vora, A.; Lutfi, M.; Figurnov, M.; Cowie, A.; Hobbs, N.; Kohli, P.; Kleywegt, G.; Birney, E.; Hassabis, D.; Velankar, S. AlphaFold Protein Structure Database: Massively Expanding the Structural Coverage of Protein-Sequence Space with High-Accuracy Models. *Nucleic Acids Res.* **2022**, *50*, D439–D444 DOI: 10.1093/nar/gkab1061.

(21) wwPDB consortium. Protein Data Bank: The Single Global Archive for 3D Macromolecular Structure Data. *Nucleic Acids Res.* **2019**, *47* (D1), D520–D528.

(22) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242 DOI: 10.1093/nar/28.1.235.

(23) Cock, P. J. A.; Antao, T.; Chang, J. T.; Chapman, B. A.; Cox, C. J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; De Hoon, M. J. L. Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics. *Bioinformatics* **2009**, *25*, 1422–1423 DOI: 10.1093/bioinformatics/btp163.

(24) Mitternacht, S. FreeSASA: An Open Source C Library for Solvent Accessible Surface Area Calculations. *F1000Research* **2016**, *5*, 189.

(25) Tien, M. Z.; Meyer, A. G.; Sydykova, D. K.; Spielman, S. J.; Wilke, C. O. Maximum Allowed Solvent Accessibilites of Residues in Proteins. *PLoS One* **2013**, *8*, e80635.

(26) Mariani, V.; Biasini, M.; Barbato, A.; Schwede, T. IDDT: A Local Superposition-Free Score for Comparing Protein Structures and Models Using Distance Difference Tests. *Bioinformatics* **2013**, *29*, 2722–2728 DOI: 10.1093/bioinformatics/btt473.

(27) Sen, S.; Young, J.; Berrisford, J. M.; Chen, M.; Conroy, M. J.; Dutta, S.; Di Costanzo, L.; Gao, G.; Ghosh, S.; Hudson, B. P.; Igarashi, R.; Kengaku, Y.; Liang, Y.; Peisach, E.; Persikova, I.; Mukhopadhyay, A.; Narayanan, B. C.; Sahni, G.; Sato, J.; Sekharan, M.; Shao, C.; Tan, L.; Zhuravleva, M. A. Small Molecule Annotation for the Protein Data Bank. *Database* **2014**, *2014*, bau116. DOI: 10.1093/database/bau116.

(28) Jensen, K. P. Improved Interaction Potentials for Charged Residues in Proteins. *J. Phys. Chem. B* **2008**, *112*, 1820–1827 DOI: 10.1021/jp077700b.

(29) Bagdonas, H.; Fogarty, C. A.; Fadda, E.; Agirre, J. The Case for Post-Predictional Modifications in the AlphaFold Protein Structure Database. *Nat. Struct. Mol. Biol.* **2021**, *28*, 869–870.

For table of contents use only