Impact of protein conformational diversity on AlphaFold predictions

Tadeo Saldaño[1,2,*], Nahuel Escobedo[1,2,*], Julia Marchetti[1,2], Diego Javier Zea[3], Juan Mac Donagh[1,2], Ana Julia Velez Rueda[1,2], Eduardo Gonik[4,2], Agustina García Melani[5], Julieta Novomisky Nechcoff[1], Martín N. Salas[1], Tomás Peters[6], Nicolás Demitroff[6,2], Sebastian Fernandez Alberti[1,2], Nicolas Palopoli[1,2], Maria Silvina Fornasari[1,2] and Gustavo Parisi[1,2,#].

[1] Departamento de Ciencia y Tecnología, Universidad Nacional de Quilmes, Bernal, Argentina. [2] Consejo Nacional de Investigaciones Científicas y Técnicas, Buenos Aires, Argentina. [3] Independent researcher, 14 rue Léo Lagrange, 31400, Toulouse, France. [4] INIFTA (CONICET-UNLP) - Fotoquímica y Nanomateriales para el Ambiente y la Biología (nanoFOT), La Plata, Argentina. [5] IMBICE (CONICET - UNLP). Laboratorio de Electrofisiología, La Plata, Argentina. [6] Fundación Instituto Leloir-Instituto de Investigaciones Bioquímicas de Buenos Aires, Buenos Aires, Argentina.
* Authors contributed equally, # To whom correspondence should be addressed.

Abstract

Motivation: After the outstanding breakthrough of AlphaFold in predicting protein 3D models, new questions appeared and remain unanswered. The ensemble nature of proteins, for example, challenges the structural prediction methods because the models should represent a set of conformers instead of single structures. The evolutionary and structural features captured by effective deep learning techniques may unveil the information to generate several diverse conformations from a single sequence. Here we address the performance of AlphaFold2 predictions obtained through ColabFold under this ensemble paradigm.
Results: Using a curated collection of apo-holo pairs of conformers, we found that AlphaFold2 predicts the holo form of a protein in ~70% of the cases, being unable to reproduce the observed conformational diversity with the same error for both conformers. More importantly, we found that AlphaFold2's performance worsens with the increasing conformational diversity of the studied protein. This impairment is related to the heterogeneity in the degree of conformational diversity found between different members of the homologous family of the protein under study. Finally, we found that main-chain flexibility associated with apo-holo pairs of conformers negatively correlates with the predicted local model quality score plDDT, indicating that plDDT values in a single 3D model could be used to infer local conformational changes linked to ligand binding transitions.
Availability: Data and code used in this manuscript are publicly available at https://gitlab.com/sbgunq/publications/af2confdiv-oct2021
Contact: Gustavo Parisi. Email: gusparisi@gmail.com
Supplementary Information: Supplementary data is available at the journal's web site.

1 Introduction

The first ideas to predict protein structures from their sequences came in the early sixties, after Anfinsen's experiment showed that the structure of a protein is encoded in its amino-acid sequence (Anfinsen *et al.*, 1961). After decades of extensive experimentation and efforts, the practical demonstration of Anfinsen's motto came from deep learning techniques taking advantage of evolutionary information. In the last year, the computational tool AlphaFold2 (Jumper *et al.*, 2021) developed by DeepMind, reached an impressive performance in predicting protein structures with an accuracy similar to experimental techniques (Kinch *et al.*, 2021; Pearce and Zhang, 2021). AlphaFold2 uses a novel neural network architecture with some attention-based components to take advantage of the evolutionary information codified in a multiple sequence alignment. These neural networks create novel representations of the protein sequence and the inter-residue relative distances that are iteratively improved. The output of AlphaFold2 is a set of highly accurate structural models with accompanying residue-specific estimates of model reliability.

This outstanding achievement is not only conceptual, in the sense of the advancement of novel deep learning techniques and protein science, but also practical. It provides the scientific community with a method for fast, reliable, and cheap determination of structural models that can be applied at a large scale. Recently, DeepMind and EMBL-EBI have jointly released the database of AlphaFold2 predictions for the whole human proteome (Tunyasuvunakool *et al.*, 2021) and other key organisms (https://alphafold.ebi.ac.uk/). Furthermore, an easy-to-use and fast version of the AlphaFold2 pipeline was introduced by modifying the time-consuming step of multiple sequence alignments generation with almost identical results (Mirdita *et al.*, 2021). These exceptional endeavors will soon contribute to filling the gap between proteomes and structuromes, triggering the blooming of almost every related biology field involving both wet-lab practices and computational-based approaches.

The AlphaFold2 neural network is trained using structures derived from crystallization and X-ray diffraction experiments. It is thus expected that the 3D models obtained will reproduce "regular" PDB structures (Jumper *et al.*, 2021). How much do regular PDB structures resemble the native state of proteins? It is widely accepted that protein function relies on a conformational ensemble describing the native state of proteins (Wei *et al.*, 2016; Boehr *et al.*, 2009; Tsai *et al.*, 1999; Motlagh *et al.*, 2014) that is not entirely captured in the PDB (Marino-Buslje *et al.*, 2019). Structural differences between conformers promote ligand binding (Gunasekaran and Nussinov, 2007), transport (Gora *et al.*, 2013), or catalysis (Gutteridge and Thornton, 2004; Callender and Dyer, 2015). These differences are also relevant for signal transduction (Tompa, 2016) and define metabolic regulation by mechanisms like cooperativity and allosterism (Motlagh *et al.*, 2014, 2012; Donovan *et al.*, 2016; del Sol *et al.*, 2009). Conformers in the native ensemble could be identical in their backbones but differ just in the conformations of some residues, defining open and close transitions of tunnels (Monzon, Zea, Fornasari, *et al.*, 2017; Kingsley and Lill, 2015) and/or volume variations in their cavities (Barletta *et al.*, 2018; Hasenahuer *et al.*, 2017). Increasing differences involve backbone movements comprising loops, secondary structural elements rearrangements, and relative domains movements (Gerstein *et al.*, 1994; Gerstein and Krebs, 1998; Gu *et al.*, 2015). Extreme cases of conformational diversity are represented by intrinsically disordered proteins which lack tertiary structure and form complex ensembles with high ratios of interchange between conformers (Tompa, 2011).

Given the ensemble nature of proteins we explored the impact of conformational diversity in the AlphaFold2 performance prediction. Firstly, we relied on a hand-curated set of proteins with different extents of experimentally estimated conformational diversity, defined by an apo conformer and the corresponding holo form bound to a biologically relevant ligand. We obtained structural models of each protein in our dataset through the ColabFold implementation of AlphaFold2 and then studied if AlphaFold2 can reproduce both known conformers among their resulting top-scoring models. We also explored how AlphaFold2's performance is affected by the degree of conformational diversity of the protein under study. Additionally, as AlphaFold2's predictions heavily rely on evolutionary information, we used families of homologous proteins with different extents of conformational diversity among its members to test whether this heterogeneity affects prediction.

2 Results

2.1 Description of the dataset

We selected 91 proteins (Supplementary Table 1) with different degrees of conformational diversity expressed as the range of pairwise global Cα-RMSD between their conformers in the PDB (Figure 1). All the pairs of conformers for each protein are apo-holo pairs selected from the CoDNaS database (Monzon *et al.*, 2016) and bibliography. Manual curation for each protein confirmed that structural deformations were associated with a given biological process based on experimental evidence. This step is essential to ensure that conformational diversity is not associated with artifacts, misalignments, missing regions, or the presence of flexible ends. When more than two conformers were known, we selected the apo-holo pair showing the maximum Cα-RMSD (maxRMSD). Other considerations were absence of disorder, PDB resolution, absence of mutations, and sequence differences. We previously observed that when conformational diversity is derived from experimentally-based conformers, different ranges of RMSD are obtained between them depending on the structure determination method (Monzon, Zea, Fornasari, *et al.*, 2017). Here we considered a continuum of protein flexibility measured as the RMSD between apo and holo forms as shown in Figure 1.

2.2 AlphaFold2 does not reproduce conformational diversity

We have predicted the structure of each protein in the dataset using ColabFold (AlphaFold2.ipynb - Colaboratory), running AlphaFold2 with MMSeq2 (Mirdita *et al.*, 2021) without the use of templates and with the option to obtain relaxed models with Amber force fields (Eastman *et al.*, 2017), gathering the first five top models according to the average of the plDDT (predicted local Distance Difference Test) scores (Mariani *et al.*, 2013). Supplementary Figure 1 shows the distribution of the plDDT scores for all the models. We found that 90% of the models scored higher than 85, reaching 89 if only the best model for each protein is considered, evidencing the good quality of the models obtained.

All AlphaFold2 models of each protein in the dataset were structurally aligned to the experimentally resolved apo and holo conformations and the RMSD value was calculated for each alignment. Figure 2A shows the relationships of RMSD values against the apo and holo forms for all the obtained AlphaFold2 models, while Figure 2B is limited to the best model for each protein, defined as the model with the highest plDDT global quality score.

We found that for 67% of the proteins the models are obtained with the lowest RMSD to the holo form, while only 33% of the proteins are modeled with lowest RMSD to the apo form instead. In the larger subset, with proteins modeled closer to the holo form, the RMSD against the apo form is significantly higher than to the holo form (right-tailed Wilcoxon signed-rank test p-value = 0.0033). Here we used the simple rule to decide if a model is better than another by choosing the model with the lowest RMSD regardless of the RMSDs being compared. Using different RMSD cutoffs to filter out models of lower quality (for example, limiting the analysis to models with RMSD <3 or <4Å) did not introduce significant variations in our results (see Supplementary Figure 2).

In Figure 3 we plot the distributions of the average RMSD of the five models of a protein against their apo and holo forms, discriminating between proteins according to which of both forms their models resemble the most. For proteins modeled closer to apo, the average RMSD against their apo form is significantly lower (mean = 2.58Å) than against their holo form (mean = 4.17Å) (Figure 3, left panel). On the contrary, for proteins modeled closer to holo, the average RMSD to the holo form is 1.87Å and climbs to 3.24Å against the apo form (Figure 3, right panel). We conclude that most of the proteins are modeled with a bias towards a given conformer. It is then impossible to estimate the degree of conformational diversity captured in apo and holo pairs with the same precision that can be estimated for a single representative conformation of a given protein. As expected, the error in the estimation of the conformational diversity is highly correlated with the structural differences between the apo and the holo forms, with a Pearson correlation coefficient of 0.97 (p-value < 0.001) between the RMSD of apo-holo pairs and the average RMSD of the models to the unfavored form (see Supplementary Figure 3).

The preference for a single conformer, whether apo or holo, is not associated with the AlphaFold2 predictive performance: the median pIDDT scores for all models that resemble the holo or the apo forms are 95.47 and 94.08, respectively, or 96.33 and 95.62 using the best models only (Wilcoxon Rank Sum test, p-value = 0.27). Figure 4 shows three examples that illustrate the model preference for the apo or the holo form, or its lack of, taken from the results described above.

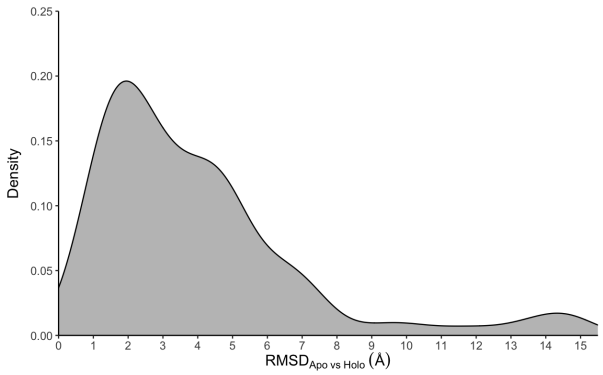

Figure 1: Distribution of RMSD between apo-holo pairs. The average of the distribution is 4.00Å.

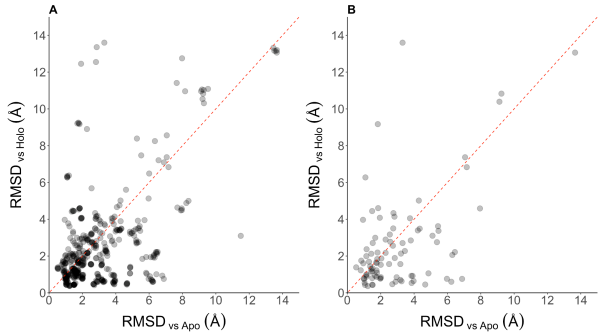

Figure 2: Comparison of RMSD values derived from the alignment of each AlphaFold2 model to the apo (*x*-axis) and the holo (*y*-axis) conformations of the corresponding protein. **Panel A** shows the distribution of RMSD for all models, while **panel B** is limited to the best model per protein according to pIDDT scores.

2.3 AlphaFold2 predictions worsen with increasing conformational diversity of the protein

Given that AlphaFold models better resemble the holo conformation of proteins, in this section we study how the conformational diversity of the protein, measured as the structural difference between apo and holo forms, affects performance predictions. We found that proteins are less predictable as the RMSD between their apo and holo form increases, with a predictive performance (measured as the lowest RMSD of a model to the apo or holo forms) highly dependent on the conformational diversity of the protein (Pearson correlation coefficient = 0.76, p-value <0.001) (Figure 5A). This tendency was also observed when we studied the correlation of the global plDDT for the best model with the level of conformational diversity (Pearson correlation coefficient = -0.65, p-value <0.001) (Figure 5B).

The model error shows low dependency on the protein length (Pearson correlation coefficient = -0.23, p-value <0.05), and non-significant correlation with the total number of sequences in the input alignment (Pearson correlation coefficient = 0.07, p-value = 0.51) or the number of effective sequences per alignment (Pearson correlation coefficient = -0.03, p-value = 0.74). To study how the error in the model depends on the type of protein movements between the apo and holo forms, we classified the pairs in our dataset in two broad categories (Kempner, 1993): according to the presence of domains and hinges movements using the DynDom software (Taylor *et al.*, 2013) and to the presence of flexible loops in just one domain. We found that the prediction error, measured as the lowest RMSD to apo or holo, does not depend on the type of movement (Wilcoxon Rank Sum test, p-value = 0.22), with median values of 0.98Å and 1.6Å for the domain movements group and the subset of proteins with loop movements, respectively.

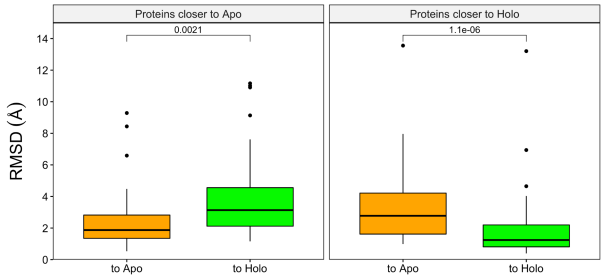

Figure 3: Distribution of the average RMSD between AlphaFold2 models and the apo or holo forms. The left panel corresponds to the subset of proteins that were modeled closer to the apo form, while the right panel is limited to proteins modeled towards the holo form. In each panel, 'to Apo' and 'to Holo' present the distribution of average RMSD values between all models towards apo and holo forms, respectively. Significance values are obtained with a Wilcoxon Rank Sum test.

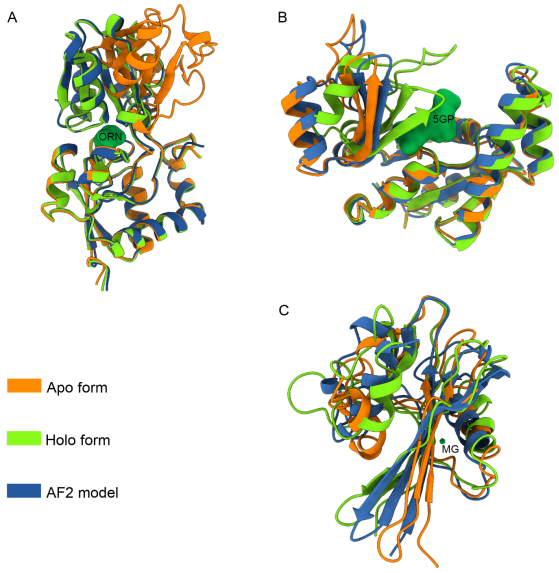

Figure 4: Three different examples of AlphaFold2 models (blue) and experimental structures (apo, orange; holo, green) among proteins in our dataset. **A.** Periplasmic lysine-, arginine-, ornithine-binding protein (LAO) from *Salmonella typhimurium (Oh et al., 1994)*. The AlphaFold2 model closest to an experimental structure is closer to the holo (Cα-RMSD = 0.45Å, PDB ID = 1LAH_E) than to the apo form (4.67Å, 2LAO_A). **B.** Guanylate Kinase from yeast. The best AlphaFold2 model showed a better match with the apo form (Cα-RMSD = 0.94Å, PDB ID = 1EX6_B) than with the holo form (3.97Å, 1EX7_A) (Blaszczyk *et al.*, 2001). **C.** Nucleoside triphosphate pyrophosphohydrolase from *E. coli* (Abeygunawardana *et al.*, 1995). A case where the AlphaFold2 model is different to both the apo (Cα-RMSD = 3.76Å, PDB ID = 1MUT-11_A) and holo (4.05Å, 1PUN-7_A) forms. Proteins are shown as cartoons while biologically significant ligands are labeled and shown in surface representation.

2.4 Fuzzy evolutionary information could affect AlphaFold2 prediction

To explain the impairment observed in AlphaFold2 prediction capacity with increasing protein conformational diversity, we hypothesized that the evolutionary information in the input multiple sequence alignment could be fuzzy due to the conformational diversity heterogeneity in the protein family. Previously we found that families with highly flexible proteins (Monzon, Zea, Fornasari, *et al.*, 2017) heavily affect homology modeling due to a noisy relationship

between sequence and structure divergence (Monzon, Zea, Marino-Buslje, *et al.*, 2017). Moreover, we have also characterized that the inter-residue contacts predicted using coevolutionary methods are the consensus ones, independently of the structural variations among family members (Zea *et al.*, 2018). Taking into account our finding that protein dynamical behavior is mostly not conserved in protein families (Monzon, Zea, Marino-Buslje, *et al.*, 2017), families that include highly flexible and rigid proteins could have confounding mixtures of sequence signatures.

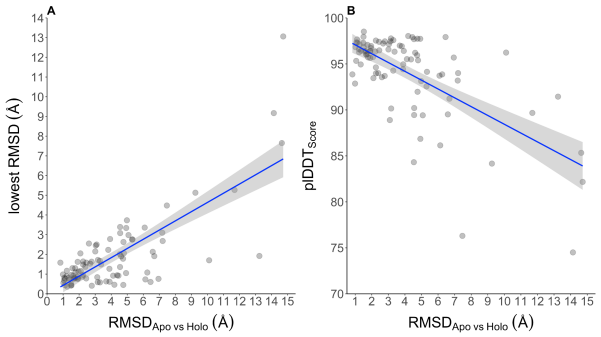

Figure 5: Quality prediction of AlphaFold2. **Panel A**: As conformational diversity increases between apo and holo forms, the lowest RMSD to any of the forms increases as well (Pearson correlation 0.76, p-value <0.001). **Panel B**: Likewise, the global plDDT scores decrease with larger protein conformational diversity (Pearson correlation -0.65, p-value <0.001).

To test this hypothesis, we explore families of homologous proteins with experimentally based conformational diversity. These families were obtained from the CoDNaS database using a sequence-based clustering with 40% sequence identity and 70% coverage. Each of the ~29000 protein entries in CoDNaS has an associated maximum Cα-RMSD derived from comparing all the conformers belonging to a given protein. This maxRMSD is taken as the maximum conformational diversity the protein could have. Clusters were further classified into homogeneous or heterogeneous according to the range between the minRMSD and the maxRMSD for each protein in each family (range < 4Å for homogeneous families, or heterogeneous otherwise).

Mapping our 91 proteins to these clusters only retrieved 20 proteins distributed in 10 homogeneous and 10 heterogeneous clusters. For each of these 20 proteins we studied the correlation between their corresponding error in AlphaFold2 predictions (estimated again as the lowest RMSD to any experimental conformer) and the dispersion of the conformational diversity in all proteins from the same family. We found that the heterogeneous clusters performed worse than the homogeneous ones (average lowest RMSD values for hetero and homogeneous families were 2.03Å and 1.31Å, respectively; Wilcoxon p-value <0.005).

To further test this hypothesis, we repeated the estimation with 175 chosen proteins, one for each of the most populated clusters described above (23.53 homologous proteins on average per cluster). For each of these 175 proteins, we ran AlphaFold2 using ColabFold and estimated the lowest RMSD, comparing the top obtained models with the corresponding structure of the protein. We observed the same trend as shown in Figure 6A (average lowest RMSD with mean values of 1.96Å and 1.54Å for hetero and homogeneous families, respectively; Wilcoxon p-value < 0.005).

To further explore if the model estimation is affected by the flexibility of each family, we further classified the clusters in 'flexible' and 'rigid', with a flexible cluster defined with an average maxRMSD >1.0Å, and a rigid cluster otherwise (Monzon, Zea, Fornasari, *et al.*, 2017). The same conclusions observed for heterogeneous and homogeneous families can be derived using this classification of rigidity (Figure 6B) (average lowest RMSD with mean values 1.88Å and 1.44Å for flexible and rigid families, respectively; Wilcoxon p-value < 0.001).

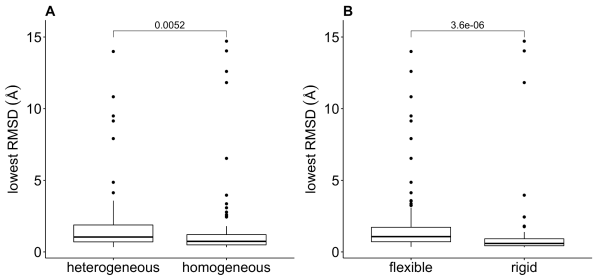

Figure 6: Distribution of model error, estimated as the lowest RMSD to the apo and holo forms, as a function of the evaluation of the distribution of the conformational diversity in 175 homologous families. Panel A contains families classified as heterogeneous or homogeneous using the range of the proteins in each family (range < 4Å for homogeneous families, or heterogeneous otherwise). Panel B contains families that were classified as flexible or rigid (average maxRMSD >1.0Å for flexible clusters, and rigid< otherwise).

2.5 High flexibility regions are anti correlated with plDDT score

We mentioned that structural differences between conformers could be so tiny as the rotation of the side chains to large movements of loops and domains. This section studies how more flexible regions between apo and holo conformations are related to the plDDT score.
Using RMSF to measure protein flexibility between apo and holo conformers, we studied how this parameter correlates with plDDT. Taking only the models with the lowest RMSD to apo and holo forms, we found that the correlation between RMSF and plDDT is -0.44 (Pearson p-value < 0.001). However, the absolute value of this correlation increased when we used windows of different widths, reaching a strongest correlation of -0.48 (Pearson p-value < 0.001) with a window of 15 residues. In Figure 7 we represent these trends as box plot representations, grouping observations in four intervals of quality according to their plDDT scores. RMSF captures the flexibility of the protein per position, as derived from the comparison between apo and holo conformers. To study how the intrinsic flexibility of each conformer relates with the plDDT, we used the profile of the normalized α carbons B-factors obtained by performing normal mode analysis for the apo form of the protein as described in Methods. In this way a similar correlation of -0.42 (Pearson p-value <0.001) was obtained.

Low values of plDDT have been related to the occurrence of disordered regions (Jumper *et al.*, 2021). However, according to our results, low-scoring regions could also represent flexible regions connecting ordered conformers, as observed for most of the proteins in our dataset.

3 Discussion

AlphaFold's breakthrough in predicting protein 3D models has certainly changed the way we study the protein structure-function relationship. Full structuromes of key organisms have been made available recently, along with easy-to-use utilities to run predictions. It is also outstanding to note that most of the predictions made are of the highest quality, mostly comparable with crystallographic resolution. In this work we have studied how the conformational diversity of the native state could be available through predictions and how in turn this key feature of protein biology affects the performance of predictions.

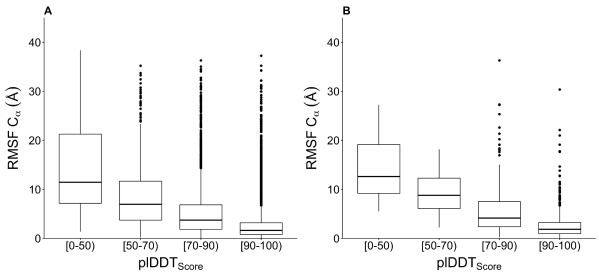

Figure 7: Box plots showing the RMSF between apo and holo forms as a function of the plDDT score, calculated by site (Panel A, left) or averaged in 15-residue windows (Panel B, right). plDDT scores are grouped in four intervals of quality (0-50, 50-70, 70-90 and 90-100) as used by the original DeepMind paper. Models were selected according to their lowest RMSD.

The first purpose is a practical one: Can we consider the top predicted models as snapshots of the conformational ensemble that describes the native state of proteins? Unfortunately, we can't. Only 2 out of 91 (~2%) proteins in our dataset showed models resembling the holo and the apo forms with similar error, measured as the best RMSD to a given form (Figure 2). For the rest of the proteins, it is not possible to model both the apo and holo forms simultaneously with the same low error as when considering a single conformer (Figure 3). Far from disappointing, this observation was expected since a large set of redundant protein structures (conformers) would have been required during the neural network training processes (Tunyasuvunakool *et al.*, 2021) in order to predict conformational diversity.

We also found that most of the predictions made resemble and are mostly indistinguishable from the holo form of the studied protein (67% of the dataset). This is an exciting result because holo forms of proteins describe the binding capacity to a substrate or any other biologically relevant ligand (see Figure 4). Jumper et al. mentioned that AlphaFold could infer structures when the presence of a ligand is predictable from the sequence (Jumper *et al.*, 2021). We thought that this finding could be explained due to a bias in the training process of AlphaFold2. However, exploring the BioLip database (Yang *et al.*, 2013) to estimate the relative presence of apo and holo forms in PDB shows about 64% of apo forms (a similar proportion observed in CoDNaS, from which our dataset was obtained). We hypothesized that holo forms could have a higher number of inter-residue contacts and that these could have influenced the modeling process in AlphaFold2. However, we did not detect differences in the number of contacts when comparing holo and apo forms (the median number of contacts are 3.40 and 3.44 for the apo and holo forms, respectively; Wilcoxon p-value >0.5) (Supplementary Figure 4). Additionally we did not detect differences between the radius of gyration in apo and holo forms (Wilcoxon p-value >0.5, Supplementary Figure 5). Apparently, differences in the number of directional polar interactions in contrast to interactions between nonpolar residues can explain differential flexibility patterns between holo and apo forms (Clark *et al.*, 2019; Gunasekaran and Nussinov, 2007). At this point, further work is required to understand this bias fully.

Does conformational diversity affect AlphaFold predictions? This second purpose of our work was a conceptual one, related to the capability to recover evolutionary information associated with protein flexibility and encoded in multiple alignments. We have found that the AlphaFold prediction capacity worsens with the increasing conformational diversity of the protein being studied (Figure 5). We showed that this impairment is related with the heterogeneous dynamic behaviors in families of homologous proteins. Additionally, proteins from flexible homologous families are also difficult to predict (Figure 6). Several works showed that conformational diversity modulates the evolutionary process imprinting sequence information with dynamic behavior (Zea *et al.*, 2013; Parisi *et al.*, 2015; Jeon *et al.*, 2011; Liu and Bahar, 2012; Morcos *et al.*, 2013; Saldaño *et al.*, 2016). Due to functional divergence, protein families could show different degrees of conformational diversity, making it difficult to extract specific sequence information from multiple sequence alignments for a given conformational motion. After the early and well-established observation that structures are very well conserved during evolution, it became evident that this conservation imposes structural constraints on sequence divergence (Lesk and Chothia, 1980; Chothia and Lesk, 1986; Panchenko *et al.*, 2005; Illergård *et al.*, 2009; Williams and Lovell, 2009). However, and more recently, we showed that this sequence-structure relationship becomes fuzzy within families with significant degrees of conformational diversity (Monzon, Zea, Marino-Buslje, *et al.*, 2017). Moreover, in protein families with complex dynamical behavior (i.e., different degrees of conformational diversity), coevolutionary analysis allowed to infer inter residue contacts representing the most populated contacts among the family's different structures, challenging the extraction of sequence features characterizing specific conformational patterns (Zea *et al.*, 2018). It is then expected that proteins belonging to families with heterogeneous flexibility behavior would be difficult to predict from the evolutionary information used by AlphaFold2 (Figure 6 A and B). During the revision of this manuscript two reprints suggested that filtering and/or changing the alignment information could be useful to predict different conformers using AlphaFold (del Alamo *et al.*, 2021; Heo and Feig, 2021). Although their approaches were applied to very few examples, they support that sampling of alignment information could be a promising resource to predict different conformations.

Finally, our results suggest that the plDDT score can be used to scan flexible regions between ordered conformers. It was pointed out that plDDT could be helpful to predict disordered regions, but we can speculate that, as there is a continuum in ordered-disordered proteins (Davey, 2019), there could exist a range of plDDT thresholds to detect different sorts of protein flexibility. All the proteins in the main dataset are mostly ordered, with regions of different flexibility and less than 15% of disordered segments. The trend shown in Figure 7 indicates that plDDT could capture the presence of flexible regions, defining the conformational plasticity between apo and holo forms. We think that our results provide useful information to further improve 3D model prediction using AlphaFold2.

4 Materials and Methods

4.1 Description of the dataset

The set of apo and holo structures was obtained from the database of Conformational Diversity in the Native State of proteins (CoDNaS) (Monzon *et al.*, 2016). CoDNaS is a redundant collection of PDB structures for the same sequence that can be taken as snapshots of protein dynamism. The conformational diversity for each protein was estimated as the Cα-RMSD between apo and holo forms. In order to obtain a well-curated dataset containing protein motions related to a given biological activity we followed several specific quality criteria: (i) Only crystal structures with resolution < 3.9Å were considered; (ii) structures must not have missing residues; (iii) there must be 100% sequence identity between the conformers; (iv) structural deformations between pairs of conformers were associated with a given biological process based on experimental evidence; (v) no reported mutations; (vi) no disordered regions; and (vii) visual inspection was used to confirm an existing conformational diversity (e.g., movements should not be limited to flexible ends or arise from

errors in the structural alignment). This allowed us to finally obtain apo-holo pairs of conformers for a total of 91 protein structures.

4.2 Predictions and comparison of structures

Predicted models for each protein in the dataset were obtained using ColabFold v1.0 (Mirdita *et al.*, 2021) due to its easy access through Google Colab Notebooks without a significant decrease in prediction performance. Runs were performed using no templates, automatic alignments, Amber energy minimization and num_recycles = 3 (default value). For each run we used the 5 top models derived from the energy minimization. Each model was structurally compared between each other and against the correspondent apo and holo structures. As sequences between conformers and models are identical, the alignments are straightforward. We then quantified the structural similarity using the Cα Root Mean Square Deviation (Cα-RMSD).

4.3 Evolutionary information

We sequentially clustered the CoDNaS database into homologous families containing sequences with more than 40% sequence identity and 70% coverage using CD-HIT. Each protein in CoDNaS has an associated maximum RMSD (maxRMSD) derived from the pairwise comparison of all its conformers. The maxRMSD is taken as the extent of the protein conformational diversity. A total of 175 well-populated clusters were taken (>8 proteins per cluster). A random protein from each cluster was modeled using ColabFold following the procedure mentioned above. The error of this model estimation was calculated as the lowest RMSD obtained from the comparison of any of the top 5 models with the crystallographic structures of the protein.

4.4 B-factors analysis

Temperature factors or B-factors (B_i) have been obtained performing normal mode analysis (NMA) using the coarse-grained Elastic Network Model (Tirion, 1996; Atilgan *et al.*, 2001) that considers the protein as an elastic network with nodes linked by springs within a cutoff distance r_c. Herein the Cα are taken as nodes, and the value of r_c is varied from 7Å to 15Å for X-ray structures in order to optimize the correlation between theoretical and experimental B-factors, while r_c= 11Å is used for NMR structures. We perform the NMA for the apo form of the protein on the basis that normal modes obtained with the apo form of a given protein give a better description of the conformational change than those obtained with the holo form (Tama and Sanejouand, 2001) . The normalized B-factor B_i' of atom i is obtained as $B_i' = (B_i - \langle B \rangle)/\sigma(B)$, being $\langle B \rangle$ and $\sigma(B)$ the average and standard deviation of the B-factor distribution for the corresponding protein structure, respectively. Each B_i' was averaged over the neighbors of the ith residue within a radius of 7Å.

4.5 Inter residue contacts and Rg analysis

Inter residue contacts have been obtained using the RING 2.0 web server (Piovesan *et al.*, 2016). Interacting pairs were identified following the closest contact strategy, i.e., all atoms are included to measure distances between residue pairs. While every pair of residues forms multiple interactions, the most energetic interaction per pair was considered. Interactions were defined distinguishing disulfides, salt bridges, hydrogen bonds, and aromatic interactions from generic van-der-Waals contacts. Radii of gyration were calculated using Pymol (http://pymol.org).

4.6 Motions classification

We have used the DynDom software v1.5 (Taylor *et al.*, 2013) to classify our dataset into proteins with “domain movements” (two or more domains presenting hinge movements) and proteins with “loop movements” (one domain, movements due to loops).

4.7 Apo-holo characterizations

Classification of the 91 proteins in the dataset was done by manual curation following the bibliography. In parallel, the database of biological ligands BioLip (Yang *et al.*, 2013) in its most recent version (October 01, 2021) was used to crosslink all the chains of the PDB (October 2021, total chains 661494) and CoDNaS v3 (March 2021, total chains 430151). If at least one biological ligand is found for a chain, it is assigned in the holo category; otherwise, it is considered as an apo conformer.

Data and code availability

The data and code used in this manuscript are publicly available at https://gitlab.com/sbgunq/publications/af2confdiv-oct2021.

Funding

This work has been supported by grants from Universidad Nacional de Quilmes (PUNQ 1309/19), Agencia Nacional de Promoción de la Investigación, el Desarrollo Tecnológico y la Innovación (PICT-2018 3457) and Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) (PIP-2015-2017 11220150100853CO) from Argentina, and the European Union's Horizon 2020 Research and Innovation Staff Exchange program (grant agreements 778247 and 823886). NP, MSF, SFA, and GP are researchers, TS, JM, AJVR are postdoctoral fellows, and NE, JMD, ND and EG are PhD fellows, all from CONICET. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Abeygunawardana,C. *et al.* (1995) Solution structure of the MutT enzyme, a nucleoside triphosphate pyrophosphohydrolase. *Biochemistry*, **34**, 14997–15005.

del Alamo,D. *et al.* (2021) Sampling the conformational landscapes of transporters and receptors with AlphaFold2. *BioRxiv*.

Anfinsen,C.B. *et al.* (1961) The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proc Natl Acad Sci USA*, **47**, 1309–1314.

Atilgan,A.R. *et al.* (2001) Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys. J.*, **80**, 505–515.

Barletta,G.P. *et al.* (2018) Dynamics fingerprints of active conformers of epidermal growth factor receptor kinase. *J. Comput. Chem.*, **39**, 2472–2480.

Blaszczyk,J. *et al.* (2001) Crystal structure of unligated guanylate kinase from yeast reveals GMP-induced conformational changes. *J. Mol. Biol.*, **307**, 247–257.

Boehr,D.D. *et al.* (2009) The role of dynamic conformational ensembles in biomolecular recognition. *Nat. Chem. Biol.*, **5**, 789–796.

Callender,R. and Dyer,R.B. (2015) The dynamical nature of enzymatic catalysis. *Acc. Chem. Res.*, **48**, 407–413.

Chothia,C. and Lesk,A.M. (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J.*, **5**, 823–826.

Clark,J.J. *et al.* (2019) Inherent versus induced protein flexibility: Comparisons within and between apo and holo structures. *PLoS Comput. Biol.*, **15**, e1006705.

Davey,N.E. (2019) The functional importance of structure in

unstructured protein regions. *Curr. Opin. Struct. Biol.*, **56**, 155–163.

Donovan,K.A. *et al.* (2016) Conformational dynamics and allostery in pyruvate kinase. *J. Biol. Chem.*, **291**, 9244–9256.

Eastman,P. *et al.* (2017) OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput. Biol.*, **13**, e1005659.

Gerstein,M. and Krebs,W. (1998) A database of macromolecular motions. *Nucleic Acids Res.*, **26**, 4280–4290.

Gerstein,M. *et al.* (1994) Structural mechanisms for domain movements in proteins. *Biochemistry*, **33**, 6739–6749.

Gora,A. *et al.* (2013) Gates of enzymes. *Chem. Rev.*, **113**, 5871–5923.

Gunasekaran,K. and Nussinov,R. (2007) How different are structurally flexible and rigid binding sites? Sequence and structural features discriminating proteins that do and do not undergo conformational change upon ligand binding. *J. Mol. Biol.*, **365**, 257–273.

Gutteridge,A. and Thornton,J. (2004) Conformational change in substrate binding, catalysis and product release: an open and shut case? *FEBS Lett.*, **567**, 67–73.

Gu,Y. *et al.* (2015) Decoding the mobility and time scales of protein loops. *J. Chem. Theory Comput.*, **11**, 1308–1314.

Hasenahuer,M.A. *et al.* (2017) Pockets as structural descriptors of EGFR kinase conformations. *PLoS ONE*, **12**, e0189147.

Heo,L. and Feig,M. (2021) Multi-state Modeling of G-protein Coupled Receptors at Experimental Accuracy. *BioRxiv*.

Illergård,K. *et al.* (2009) Structure is three to ten times more conserved than sequence--a study of structural response in protein cores. *Proteins*, **77**, 499–508.

Jeon,J. *et al.* (2011) Molecular evolution of protein conformational changes revealed by a network of evolutionarily coupled residues. *Mol. Biol. Evol.*, **28**, 2675–2685.

Jumper,J. *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**, 583–589.

Kempner,E.S. (1993) Movable lobes and flexible loops in proteins. Structural deformations that control biochemical activity. *FEBS Lett.*, **326**, 4–10.

Kinch,L.N. *et al.* (2021) Topology evaluation of models for difficult targets in the 14th round of the critical assessment of protein structure prediction. *Proteins*.

Kingsley,L.J. and Lill,M.A. (2015) Substrate tunnels in enzymes: structure-function relationships and computational methodology. *Proteins*, **83**, 599–611.

Lesk,A.M. and Chothia,C. (1980) How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *J. Mol. Biol.*, **136**, 225–270.

Liu,Y. and Bahar,I. (2012) Sequence evolution correlates with structural dynamics. *Mol. Biol. Evol.*, **29**, 2253–2263.

Mariani,V. *et al.* (2013) lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, **29**, 2722–2728.

Marino-Buslje,C. *et al.* (2019) On the dynamical incompleteness of the Protein Data Bank. *Brief. Bioinformatics*, **20**, 356–359.

Mirdita,M. *et al.* (2021) ColabFold - Making protein folding accessible to all. *BioRxiv*.

Monzon,A.M. *et al.* (2016) CoDNaS 2.0: a comprehensive database of protein conformational diversity in the native state. *Database (Oxford)*, **2016**.

Monzon,A.M., Zea,D.J., Fornasari,M.S., *et al.* (2017) Conformational diversity analysis reveals three functional mechanisms in proteins. *PLoS Comput. Biol.*, **13**, e1005398.

Monzon,A.M., Zea,D.J., Marino-Buslje,C., *et al.* (2017) Homology modeling in a dynamical world. *Protein Sci.*, **26**, 2195–2206.

Morcos,F. *et al.* (2013) Coevolutionary signals across protein lineages help capture multiple protein conformations. *Proc Natl Acad Sci USA*, **110**, 20533–20538.

Motlagh,H.N. *et al.* (2012) Interplay between allostery and intrinsic disorder in an ensemble. *Biochem. Soc. Trans.*, **40**, 975–980.

Motlagh,H.N. *et al.* (2014) The ensemble nature of allostery. *Nature*, **508**, 331–339.

Oh,B.H. *et al.* (1994) Structural basis for multiple ligand specificity of the periplasmic lysine-, arginine-, ornithine-binding protein. *J. Biol. Chem.*, **269**, 26323–26330.

Panchenko,A.R. *et al.* (2005) Evolutionary plasticity of protein families: coupling between sequence and structure variation. *Proteins*, **61**, 535–544.

Parisi,G. *et al.* (2015) Conformational diversity and the emergence of sequence signatures during evolution. *Curr. Opin. Struct. Biol.*, **32**, 58–65.

Pearce,R. and Zhang,Y. (2021) Deep learning techniques have significantly impacted protein structure prediction and protein design. *Curr. Opin. Struct. Biol.*, **68**, 194–207.

Piovesan,D. *et al.* (2016) The RING 2.0 web server for high quality residue interaction networks. *Nucleic Acids Res.*, **44**, W367-74.

Saldaño,T.E. *et al.* (2016) Evolutionary conserved positions define protein conformational diversity. *PLoS Comput. Biol.*, **12**, e1004775.

del Sol,A. *et al.* (2009) The origin of allosteric functional modulation: multiple pre-existing pathways. *Structure*, **17**, 1042–1050.

Tama,F. and Sanejouand,Y.H. (2001) Conformational change of proteins arising from normal mode calculations. *Protein Eng.*, **14**, 1–6.

Taylor,D. *et al.* (2013) Classification of domain movements in proteins using dynamic contact graphs. *PLoS ONE*, **8**, e81224.

Tirion,M.M. (1996) Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis. *Phys. Rev. Lett.*, **77**, 1905–1908.

Tompa,P. (2016) The principle of conformational signaling. *Chem. Soc. Rev.*, **45**, 4252–4284.

Tompa,P. (2011) Unstructural biology coming of age. *Curr. Opin. Struct. Biol.*, **21**, 419–425.

Tsai,C.J. *et al.* (1999) Folding funnels, binding funnels, and protein function. *Protein Sci.*, **8**, 1181–1190.

Tunyasuvunakool,K. *et al.* (2021) Highly accurate protein structure prediction for the human proteome. *Nature*, **596**, 590–596.

Wei,G. *et al.* (2016) Protein ensembles: how does nature harness thermodynamic fluctuations for life? the diverse functional roles of conformational ensembles in the cell. *Chem. Rev.*, **116**, 6516–6551.

Williams,S.G. and Lovell,S.C. (2009) The effect of sequence evolution on protein structural divergence. *Mol. Biol. Evol.*, **26**, 1055–1065.

Yang,J. *et al.* (2013) BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic Acids Res.*, **41**, D1096-103.

Zea,D.J. *et al.* (2018) How is structural divergence related to evolutionary information? *Mol. Phylogenet. Evol.*, **127**, 859–866.

Zea,D.J. *et al.* (2013) Protein conformational diversity correlates with evolutionary rate. *Mol. Biol. Evol.*, **30**, 1500–1503.