High-throughput prediction of peptide structural conformations with AlphaFold2

Alexander M. Ille[1,2], Christopher Markosian[1,2], Stephen K. Burley[3,4,5,6,7], Paul C. Whitford[8,9], José N. Onuchic[10,11,12,13], Renata Pasqualini[1,2,15,*], and Wadih Arap[1,14,15,*]

[1]Rutgers Cancer Institute, Newark, NJ, USA
[2]Division of Cancer Biology, Department of Radiation Oncology, Rutgers New Jersey Medical School, Newark, NJ, USA
[3]Research Collaboratory for Structural Bioinformatics Protein Data Bank, Institute for Quantitative Biomedicine, Rutgers, The State University of New Jersey, Piscataway, NJ, USA
[4]Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, Piscataway, NJ, USA
[5]Rutgers Data Science and Artificial Intelligence (RAD) Collaboratory, Rutgers, The State University of New Jersey, Piscataway, NJ, USA
[6]Rutgers Cancer Institute, New Brunswick, NJ, USA
[7]Research Collaboratory for Structural Bioinformatics Protein Data Bank, San Diego Supercomputer Center, University of California-San Diego, La Jolla, San Diego, CA, USA
[8]Center for Theoretical Biological Physics, Northeastern University, Boston, MA, USA
[9]Department of Physics, Northeastern University, Boston, MA, USA
[10]Center for Theoretical Biological Physics, Rice University, Houston, TX, USA
[11]Department of Biosciences, Rice University, Houston, TX, USA
[12]Department of Chemistry, Rice University, Houston, TX, USA
[13]Department of Physics & Astronomy, Rice University, Houston, TX, USA
[14]Division of Hematology/Oncology, Department of Medicine, Rutgers New Jersey Medical School, Newark, NJ, USA
[15]These authors jointly supervised the work.
*Correspondence to Renata Pasqualini (renata.pasqualini@rutgers.edu) or Wadih Arap (wadih.arap@rutgers.edu)

ABSTRACT

Protein structure prediction *via* artificial intelligence/machine learning (AI/ML) approaches has sparked substantial research interest in structural biology and adjacent disciplines. More recently, AlphaFold2 (AF2) has been adapted for the prediction of multiple structural conformations—beyond the original scope of predicting single-state structures. This is accomplished by using multiple random seeds and subsampling the multiple sequence alignment (MSA). Research using this novel approach has focused on proteins (typically 50 residues in length or greater), while multi-conformation prediction of shorter peptides has not yet been explored in this context. Here, we report AF2-based structural conformation prediction of a total of 557 peptides (ranging in length from 10 to 40 residues) for a benchmark dataset with corresponding nuclear magnetic resonance (NMR)-determined conformational ensembles. *De novo* structure predictions were accompanied by structural comparison analyses to assess prediction accuracy. We found that the prediction of conformational ensembles of peptides with AF2 varied in accuracy *versus* NMR data, with average root-mean-square deviation (RMSD) among structured regions under 2.5 Å and average root-mean-square fluctuation (RMSF) differences under 1.5 Å for the entire set of 557 peptides. Our results reveal notable capabilities of AF2-based structural conformation prediction for peptides but also highlight considerable limitations, underscoring the necessity for interpretation discretion and the need for improved conformational ensemble prediction approaches.

INTRODUCTION

AlphaFold2 (AF2) is an artificial intelligence/machine learning (AI/ML) model capable of predicting the three-dimensional (3D) structures of proteins from amino acid sequence alone with accuracy comparable to lower-resolution experimentally determined protein structures [1, 2]. AF2 was trained on multiple sequence alignment (MSA) of metagenomic sequencing data in combination with protein structures determined by X-ray crystallography and cryo-electron microscopy (cryoEM) from the Protein Data Bank (PDB) [3]. Since its release, AF2 has garnered widespread use for various applications across the biological sciences and was recognized with a share of the 2024 Nobel Prize in Chemistry [4, 5]. More recently, AF2 has been adapted for predicting the structures of multiple protein conformations, going beyond the original scope of single static structure prediction [6-9]. This is accomplished by using multiple random seeds, *i.e.,* stochastic prediction initializations, and subsampling of the input MSA, resulting in an ensemble of predicted structures [8]. MSA subsampling, in particular, introduces sequence-based variability by dedicating unique subsets of the MSA for the different structures in a predicted conformational ensemble. The prediction of multiple conformations with AF2 has been demonstrated to align with experimentally determined conformational ensemble data for certain proteins as determined by nuclear magnetic resonance (NMR) spectroscopy [6, 7], which is notable considering that the training data for AF2 did not contain NMR structures [1]. However, AF2-based multi-conformation prediction is a novel area of research, and—as with the prediction of static structures—accuracy is expected to vary and should be considered with discretion [8]. Furthermore, the focus of this research has been on regular-sized proteins, while prediction of structural conformations for peptides in this context has not yet been explored.

Accurate prediction of peptide structures is of great relevance for elucidating biological processes as well as for the development of peptide-based therapeutics, and AF2 is a subject of ongoing research and evaluation for application towards these purposes [10, 11]. McDonald et al. [12] previously benchmarked the conventional prediction capabilities of AF2 for a rich dataset of 588 peptides with NMR-determined 3D structures in the PDB. These peptides range from 10 to 40 amino acid residues in length, and were categorized into various structural groupings, including α-helical membrane-associated peptides (AH MP), α-helical soluble peptides (AH SL), β-hairpin peptides (BHPIN), disulfide-rich peptides (DSRP), mixed secondary structure membrane-associated peptides (MIX MP), and mixed secondary structure soluble peptides (MIX SL). However, the benchmark study by McDonald et al. did not explore the more recent approach for predicting multiple structural conformations, but rather utilized the standard approach which produces five structural models per individual prediction. The inherent conformational variability of peptides merits particular consideration if AI/ML approaches are to be utilized for the discovery and characterization of peptide functions and peptide-protein interactions—both of which are structurally dynamic in nature. Advances in this domain are anticipated to facilitate experimental research involving peptides in contexts ranging from biomedical to industrial, including in vivo phage display for the discovery of therapeutically-relevant peptides and their target receptors [13-17], the design of peptides for agricultural disease control and sustainability [18-21], and peptide-based biomaterial development [22, 23], among others.

Herein, we report AF2-based multi-conformation prediction of peptide structures from the aforementioned dataset using multiple random seeds and MSA subsampling. High-throughput prediction of conformational ensembles was performed for n = 557 peptides from the McDonald et al. dataset with n = 80 individual structures computed per peptide, providing a combined total of 44,560 individual structures. Computational analyses involving various metrics, including root-mean-square deviation (RMSD), root-mean-square fluctuation (RMSF), and MSA sequence depth comparisons, were used to evaluate the predicted structures. Our analyses revealed that AF2-based conformational ensemble prediction of peptides varied in accuracy compared to NMR data, with overall RMSD under 2.5 Å and overall RMSF differences under 1.5 Å for the entire set of 557 peptides. The current work demonstrates the feasibility of peptide conformational ensemble prediction, extending beyond conventional AF2 static structure prediction using a high-throughput approach. While these results highlight notable capabilities

of this AF2-based multi-conformation prediction approach, substantial limitations are also present, underscoring the need for discretion in prediction interpretation and improved prediction methodologies.

RESULTS

Prediction of peptide conformational ensembles

AF2 predictions rely on the model's pre-trained weights as well as the MSA generated for a given input sequence [1, 8]. In order to predict conformational ensembles using the novel AF2 multi-conformation approach [6-9] **(Fig. 1a)**, we utilized the ColabFold implementation of AF2 which provides a specialized workflow for introducing structural variability via random seed initializations and MSA subsampling [8]. The number of seeds was set to 16, with five structures per seed (n = 80 structures per prediction), and MSA subsampling parameters were set to 16 for number of sequence cluster points and 32 for extra sequences (see Methods for additional information). An example of a conformational ensemble prediction and its assessment is shown in **Figs. 1b and 1c** for an antimicrobial peptide derived from *Amaranthus caudatus* (PDB ID 1ZUV) [24], which highlights an area of conformational variability **(Fig. 1b)** and depicts per-residue RMSF differences (ΔRMSF) *versus* the NMR-determined conformational ensemble **(Fig. 1c)**. To provide conformational variability while avoiding substantial reduction in sample size across peptide categories, the original McDonald et al. dataset was filtered to include only peptides for which the NMR-determined ensemble contained five or more structural conformers for a total of n = 557 peptides **(Fig. 1d and Supplementary Table 1)**.

The six peptide categorizations of (1) AH MP, (2) AH SL, (3) BHPIN, (4) DSRP, (5) MIX MP, and (6) MIX SL from the McDonald et al. dataset were also adopted in the current study in order to assess how conformational prediction accuracy might vary among these groupings. The average depth of MSAs generated by ColabFold (*via* MMseqs2) [8, 25] was 1,102 sequences across all peptides **(Fig. 1e)**. RMSD analysis of alpha carbons (Cα) of structured regions, as previously annotated [12], between the AF2-predicted and NMR-determined conformational ensembles of the peptides was performed for each individual peptide, and subsequently averaged across all peptides (See Methods). The structured region RMSD was found to be 2.294 Å averaged across all peptides **(Fig. 1f)**. To gain insight into conformational similarity between AF2-predicted and NMR-determined conformational ensembles in terms of structural flexibility, RMSF differences were assessed using a global ΔRMSF metric **(eqn. 1)**, where i is the residue position, n_r is the number of residues, and ΔRMSF **(Fig. 1c)** is the per-residue difference between the individual RMSF values of two conformational ensembles, with each ΔRMSF taken as an absolute value (see Methods). The global ΔRMSF was found to be 1.247 Å overall across all peptides **(Fig. 1g)**.

$$global\ \Delta RMSF = \frac{\sum_{i=1}^{n_r} |\Delta RMSF_i|}{n_r}$$

eqn. 1

While certain trends appear to be present across the peptide groupings among the RMSD, RMSF, and MSA depth metrics **(Fig. 1e-g)**, it is important to consider that the number of peptides across these groups vary substantially—for example, the DSRP group consists of 261 peptides while the MIX MP group consists of only 12 peptides. With this in consideration, the BHPIN and MIX SL groups had the greatest MSA depth. Additionally, predictions for the DSRP peptide group appear to have been the most accurate in terms of having the lowest of both structured region RMSD and global ΔRMSF, *i.e.* when considering both metrics, while predictions for the MIX MP and AH SL peptide groups were the least accurate with respect to structured region RMSD and global ΔRMSF, respectively **(Figs. 1f and 1g)**. The distribution plotted for the structured region RMSD metric paired with the global ΔRMSF metric for each peptide did not indicate any observable differences among the six groupings, but overall revealed that 54.2% and 73.8% of peptides had values less than 2 Å and 3 Å for both metrics, respectively **(Fig. 1h)**. Comparison of MSA depth with the structured region RMSD and global ΔRMSF metrics in combination (treated here

as the product of the two metrics) indicates an essentially negligible negative correlation, both for peptides across the entire MSA depth range (R^2 = 0.0076) and for peptides with an MSA depth of < 250 sequences (R^2 = 0.021) **(Fig. 1i)**, suggesting that greater MSA depth does not necessarily contribute to improved prediction accuracy for peptide conformational ensemble prediction.

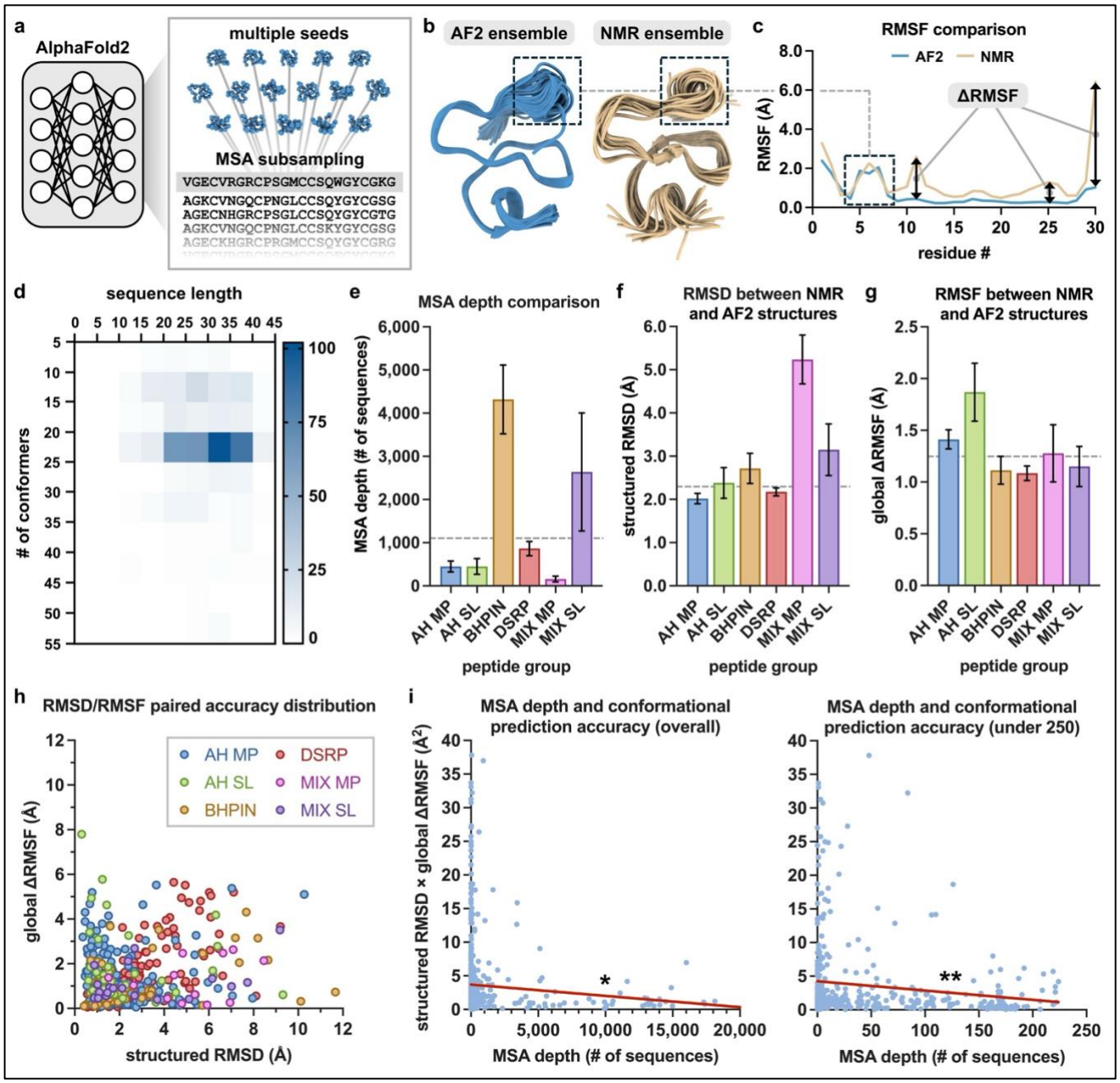

Fig. 1: Peptide structural conformation prediction and analysis. (a) Overview of the AF2-based multiple conformation prediction approach with multiple seeds and MSA subsampling. **(b)** AF2-predicted and NMR-determined conformational ensembles of an antimicrobial peptide (PDB ID 1ZUV). **(c)** Comparisons of the Cα RMSF values of the conformational ensembles from (b). A flexible region present in both ensembles is outlined with a dashed box. Examples of individual ΔRMSF values are also indicated with arrows. **(d)** Heatmap of the *n* = 557 peptide NMR dataset displaying sequence length *versus* number of conformational structures per ensemble, with the scale bar representing the number of PDB entries. **(e-g)** Among the six different peptide groupings: (e) MSA depth (number of sequences within the MSA) for each peptide, (f) structured region Cα RMSD between NMR-determined and AF2-predicted conformational ensembles, and (g) Cα global ΔRMSF **(eqn. 1)** between NMR-

determined and AF2-predicted conformational ensembles. Data represented as means ± SEM, with dashed lines indicating overall mean. AH MP *n* = 172 peptides; AH SL *n* = 37 peptides; BHPIN = 56 peptides; DSRP *n* = 261 peptides; MIX MP *n* = 12 peptides; MIX SL *n* = 19 peptides. **(h)** Distribution of structured region RMSD paired with global ΔRMSF between AF2-predicted and NMR-determined conformational ensembles among the six different peptide groupings. **(i)** MSA depth *versus* the product of structured region RMSD and global ΔRMSF between NMR-determined and AF2-predicted conformational ensembles across all peptides considering the entire MSA depth range (left graph) or peptides with an MSA depth less than 250 sequences (right graph). Pearson correlation R^2 = 0.0076 (entire MSA depth, *n* = 557 peptides) and R^2 = 0.021 (MSA depth < 250, *n* = 415 peptides), * p = 0.0394, ** p = 0.003. Lines determined by linear regression are depicted in red. Two data-points were excluded from the graphs in panel i for eligibility and space preservation but were not excluded from statistical analyses, as described in the Methods section.

Example predictions and insights based on NMR conformer number differences

Examples of AF2-predicted conformational ensembles consistent with NMR-based conformational ensembles are shown in Fig. 2, with structural depictions for each of the six peptide groupings **(Fig. 2a)** and corresponding RMSF comparisons **(Fig. 2b)**. These examples highlight well-predicted structural variation along with maintenance of overall structure similarity. Agreement between the RMSF pattern of AF2-predicted and NMR-determined conformational ensembles varied, however, as exemplified with global ΔRMSF statistical outliers from each of the six peptide groupings. Outliers were found to have either drastically lower or greater per-residue RMSF values across the entire peptide length **(Fig. 3a-c)**, revealing pronounced predictive limitations. The examples shown here, along with the considerable variability in terms of RMSD and RMSF-based accuracy metrics, emphasize the need for discretion when interpreting AF2-based conformational predictions for peptides.

Another aspect we considered was the number of individual conformers from NMR-determined ensembles, which varied between peptides **(Fig. 1d)**. We reasoned that NMR ensembles with greater numbers of individual conformers, *i.e.* greater structural heterogeneity, may provide a more comprehensive assessment of prediction accuracy given that AF2 ensembles consist of *n* = 80 conformers per ensemble. When considering all peptides collectively, neither structured region RMSD **(Fig. 4a)** nor global ΔRMSF **(Fig. 4b)** metrics varied significantly based on the number of NMR conformers. However, certain trends were noted upon independent examination of the six peptide groupings **(Fig. 4c and d)**. Both structured region RMSD and global ΔRMSF metrics of BHPIN, DSRP, and MIX SL peptide groups trended towards improved accuracy with greater NMR conformer count, while AH MP and AH SL peptide groups either trended towards decreased accuracy or exhibited mixed results. A discernable trend was not apparent for the MIX MP group, which is comprised mostly of NMR ensembles with low conformer count. These trends may be interpreted to indicate that some peptide groups fare poorly in terms of conformational ensemble prediction accuracy while others fare well, particularly in the case of the DSRP group—which also exhibited favorable accuracy in the generalized assessment **(Fig. 1f and g)**.

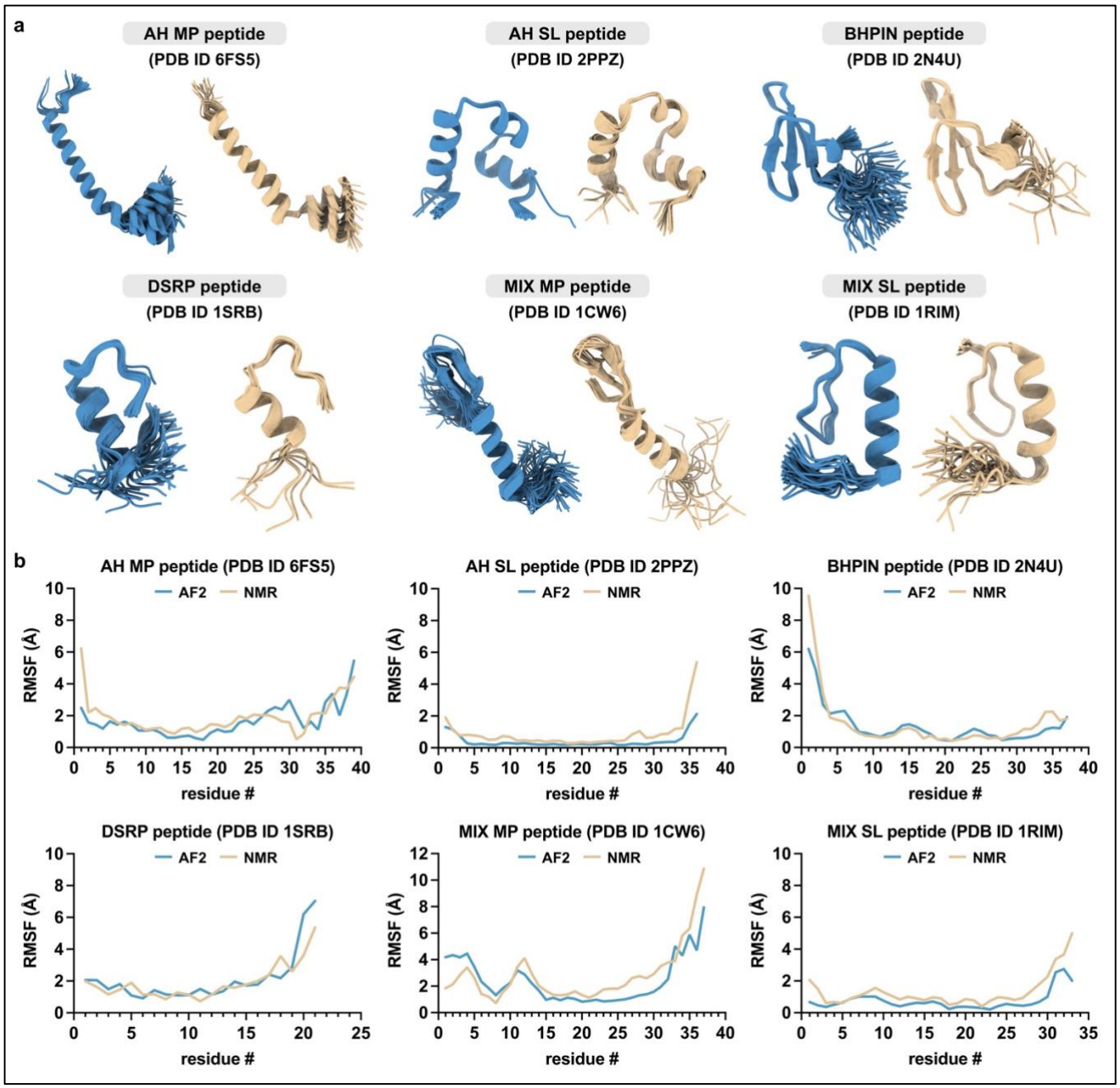

Fig. 2: Exemplary peptide structural conformation ensemble predictions. (a) AF2-predicted (blue) and NMR-determined (beige) conformational ensembles of exemplary peptides from each of the six peptide groupings: AH MP (PDB ID 6FS5) [26], AH SL (PDB ID 2PPZ) [27], BHPIN (PDB ID 2N4U) [28], DSRP (PDB ID 1SRB) [29], MIX MP (PDB ID 1CW6) [30], and MIX SL (PDB ID 1RIM) [31]. **(b)** Cα RMSF comparisons between AF2-predicted (blue lines) and NMR-determined (beige lines) conformational ensembles corresponding to the peptides depicted in (a). Global ΔRMSF values for each peptide are as follows: AH MP, 0.607 Å; AH SL, 0.455 Å; BHPIN, 0.433 Å; DSRP, 0.506 Å; MIX MP, 1.038 Å; MIX SL, 0.561 Å. Structured region Cα RMSD values for each peptide are as follows: AH MP, 1.759 Å; AH SL, 2.152 Å; BHPIN, 1.212 Å; DSRP, 2.621 Å; MIX MP, 2.633 Å; MIX SL, 1.410 Å.

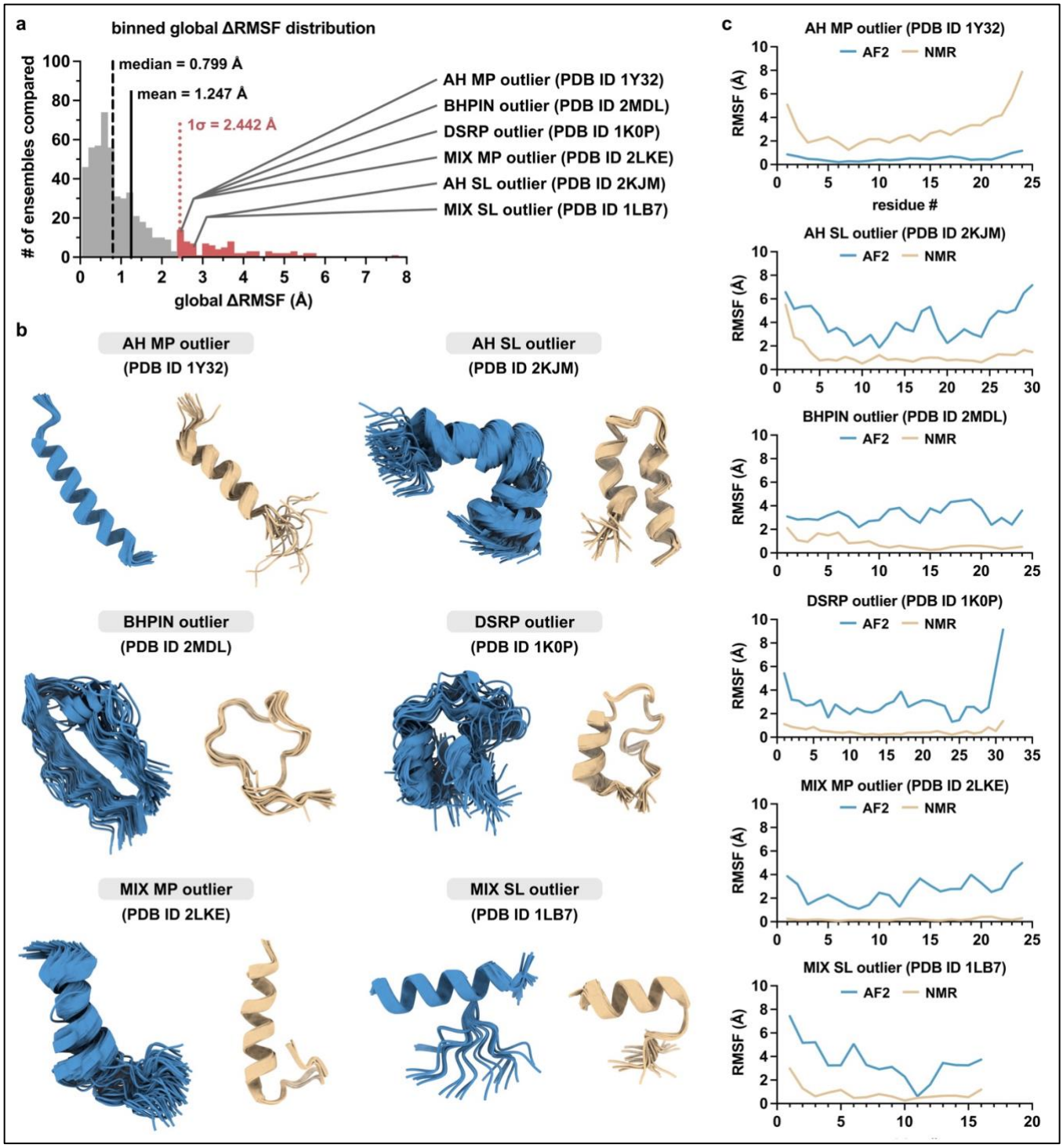

Fig. 3: Outlier analysis depicting AF2-predicted ensembles which do not resemble corresponding NMR-determined ensembles. (a) Global ΔRMSF distribution of the predicted peptides plotted across 50 bins, along with the mean (black solid line), median (black dashed line), and one standard deviation above the mean, 1σ (red dotted line). For each of the six peptide groups, the peptide with the lowest global ΔRMSF value above 1σ was selected as an outlier for representation. **(b)** AF2-predicted (blue) and NMR-determined (beige) conformational ensembles of the outlier peptides from each of the six peptide groupings: AH MP (PDB ID 1Y32) [32], AH SL (PDB ID 2KJM) [33], BHPIN (PDB ID 2MDL), DSRP (PDB ID 1K0P) [34], MIX MP (PDB ID 2LKE) [35], and MIX SL (PDB ID 1LB7) [36]. **(c)** Cα RMSF comparisons between AF2-predicted (blue lines) and NMR-determined (beige lines) conformational ensembles corresponding to the peptides depicted in (b). Global ΔRMSF values for each peptide are as follows: AH MP, 2.468 Å; AH SL, 2.783 Å; BHPIN, 2.459 Å; DSRP, 2.488 Å; MIX MP, 2.472 Å; MIX SL, 2.692

Å. Cα RMSD values for each peptide are as follows: AH MP, 1.268 Å; AH SL, 6.389 Å; BHPIN, 5.732 Å; DSRP, 5.691 Å; MIX MP, 4.828 Å; MIX SL, 2.677 Å.

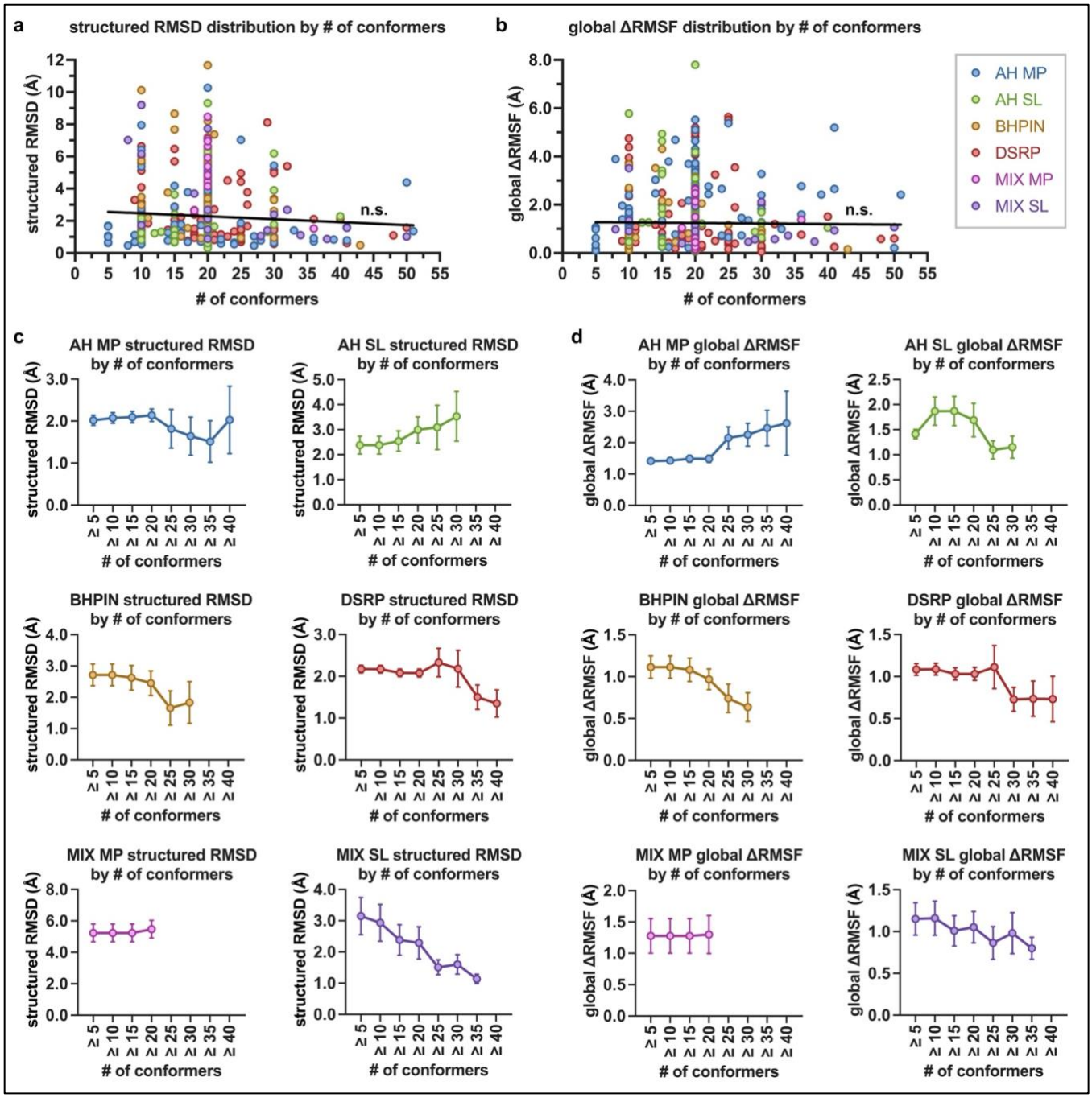

Fig. 4: AF2 Prediction accuracy with respect to number of NMR-determined conformers. (a-b) Distribution of structured region Cα RMSD **(a)** and Cα global ΔRMSF **(b)** relative to the number of NMR-determined structures per conformational ensemble across all peptides. Pearson correlation $R^2 = 0.0042$, $p = 0.1251$ (structured region RMSD distribution), and $R^2 = 0.0002$, $p = 0.7713$ (global ΔRMSF distribution), n.s., not significant. Lines determined by linear regression are depicted in black. **(c-d)** Number of NMR-determined structures per conformational ensemble *versus* structured region RMSD **(c)** and global ΔRMSF **(d)** for each of the six peptide groups. Data represented as means ± SEM, with each conformer count cutoff containing at least $n = 3$ NMR-determined conformational ensembles, *i.e.* any conformer count cutoff containing less than three NMR-determined conformational ensembles for a given peptide group is not depicted in panels c and d.

DISCUSSION

In the current work, we utilized AF2 for high-throughput prediction of peptide conformational ensembles by multiple seed initialization and MSA subsampling. This approach was applied to predict structural conformations of a dataset of 557 peptides categorized into six distinct structural groups, originating from a curated peptide dataset with NMR-determined conformational ensembles previously used for conventional AF2-based prediction [12]. The study from which the peptide dataset originates used the standard AF2 prediction approach (single seed without MSA subsampling) which results in five structural models per peptide [12], whereas the adapted multi-conformation approach [6-9] which we employed generates ensembles consisting of 80 structures per peptide. We performed structural analyses to compare the AF2-predicted and NMR-determined ensembles using metrics of RMSD of structured regions and global ΔRMSF, with the former serving to measure rigid structure prediction accuracy and the latter for assessing structural flexibility prediction accuracy. The predicted conformations were often structurally similar to NMR-determined conformations, with over 50% of the predictions differing by no more than 2 Å for both RMSD and RMSF metrics, and nearly 75% differing by less than 3 Å. Additionally, we found that the contribution of increased MSA depth to prediction accuracy was essentially negligible, which potentially negates the need of an extensive amount of homologous sequences for the current approach, at least for peptides.

There are, however, various limitations that merit consideration. Firstly, while NMR is arguably the most conformationally sensitive experimental structure determination approach, it does not necessarily capture the breadth of conformational space, especially when an NMR-determined ensemble consists of a limited number of structural conformers. While time-resolved X-ray crystallography [37] and time-resolved cryoEM [38] are also conformationally sensitive, these experimental approaches have yielded even fewer diverse structural conformers that are deposited in the PDB. However, we did gain some insight in this regard, with trends toward greater accuracy for certain peptide groups (BHPIN and DSRP), but also trends toward reduced accuracy for other peptide groups (AH MP and AH SL). Secondly, RMSD and RMSF are useful metrics both independently and in combination but only partially characterize structural differences between conformational ensembles. As conformational ensemble prediction approaches continue to evolve, more comprehensive evaluation metrics should be formulated. It is also important to consider that while NMR structures were not included for AF2 model training [1], corresponding X-ray crystallography or cryoEM structures with identical or highly similar sequences from the PDB that may have been included in the training data of AF2 would be expected to influence prediction accuracy—though it is unclear how such structures might inform conformational variability.

In conclusion, AI/ML-based structural conformation prediction holds promise as a novel avenue for the prediction of peptide conformational ensembles. AF2 exhibits notable capabilities in this regard, as documented in the current work with peptides, and by others with larger proteins [6-9]. This innovation not only reinforces the long-standing principle that protein structure is determined by amino acid sequence [39], as is the basis upon which AF2 was trained [1], but also suggests that sequence encodes conformational dynamics—especially given the extensive amount of sequence data used to train AF2. However, the novelty of such approaches and the current variability in terms of prediction accuracy underscore the necessity for discretion and careful interpretation of results. Further research is needed to better characterize conformational ensemble predictions and to advance the current state of multi-conformation prediction methodology, including the development of novel AI/ML models trained specifically for this purpose.

METHODS

AF2 Predictions

Predictions were performed using the ColabFold [8] implementation of AF2 in a Google Colab Jupyter Notebook environment with Hardware Accelerator set to T4 GPU. The ColabFold conformation prediction protocol was used, as previously described [8]. This involved adjustment of the num_seeds parameter to 16 and adjustment of the max_msa parameter to 16:32. The original protocol calls for max_msa to be set to 32:64 for a larger protein [8], which was reduced to 16:32 in the current work in consideration of the shorter sequence length of peptides compared to proteins. All other parameters were left unmodified from their default settings, *i.e.* num_relax: 0; template_mode: none; pair_mode: unpaired_paired; msa_mode: mmseqs2_uniref_env; num_recycles: 3; recycle_early_stop_tolerance: auto; use_dropout: unselected. By default, AlphaFold provides five structural models per individual prediction . As such, five structures were computed per seed, and with num_seeds set to 16 a total of $n = 80$ structures per peptide. In total, 44,560 individual structural conformers were predicted across the entire set of 557 peptides.

Structural Analyses

RMSD and RMSF values were computed using custom implementations of Python code available online from the CHARMM-GUI lecture series [40, 41]. Whole-structure alignments were performed for both RMSD and RMSF calculations. RMSD calculations included only Cα atoms of structured regions, as annotated in the original peptide dataset by McDonald et al. for regions with secondary structures and stretches with multiple disulfide bonds [12]. Each structure in a given AF2-predicted conformational ensemble was compared to each structure in the corresponding NMR-determined conformational ensemble **(Supplementary Fig. 1a)**. The average of all individual structure comparisons between one AF2 ensemble and the corresponding NMR ensemble provided the structured region RMSD value for each individual peptide. Subsequently, the structured region RMSD averaged across all peptides was calculated using the averaged RMSD values for each individual peptide. It should be noted that, because of the intrinsic structural heterogeneity within any given conformational ensemble, the RMSD within the ensemble holds a non-zero value. In other words, when comparing identical ensembles or comparing each structure within an ensemble to every other structure within the same ensemble, there is an inherent non-zero RMSD. Therefore, when comparing two non-identical ensembles, *i.e.* an AF2-predicted ensemble *versus* an NMR-determined ensemble, the calculated RMSD contains both 'within-ensemble' variability as well as 'between-ensemble' variability.

For RMSF determination, the individual structures in a given AF2-predicted conformational ensemble were aligned to the top-ranked (highest pLDDT score) structure within the ensemble, while the individual structures of the NMR-determined conformational ensembles were aligned to the initial structure in the PDB entry. Per-residue RMSF values were calculated based on Cα atoms for each residue across all structures within a given conformational ensemble. Per-residue RMSF values were used to calculate the global ΔRMSF metric **(eqn. 1)** between the two corresponding conformational ensembles, AF2-predicted and NMR-determined, for each peptide **(Supplementary Fig. 1b)**. The global ΔRMSF metric first considers the absolute value of the difference between the RMSF of each individual residue pair (AF2-predicted *versus* NMR-determined) and then takes the average of the summed per-residue differences to provide a 'global' (overall) RMSF comparison between the two ensembles. Taking the absolute value of each per-residue RMSF difference is necessary in order to account for cancellation of positive and negative differences, which would otherwise skew the average of the summed differences towards 0 Å and thereby artificially overestimate accuracy.

Statistical Analyses and Graphical Representation

Statistical analyses of standard error of the mean, Pearson correlation, and linear regression, as well as graphical representations shown in Figs. 1-4, were performed using GraphPad Prism 10. In Fig. 1i, for

the graph depicting entire MSA depth range (left graph), two data-points (PDB ID 2RR0 with MSA depth of 23,764 and RMSD × ΔRMSF product of 1.635 $Å^2$; and PDB ID 6A8Y with MSA depth of 436 and RMSD × ΔRMSF product of 52.410 $Å^2$) are not shown on the graph but were not excluded from statistical analyses. The complete set of data used for statistical analyses is presented in Supplementary Table 1.

ACKNOWLEDGEMENTS

This work was supported by the Levy-Longenbaugh Donor-Advised Fund (to R.P. and W.A.). RCSB Protein Data Bank is jointly funded by the National Science Foundation (DBI-2321666, PI: S.K.B.), the US Department of Energy (DE-SC0019749, PI: S.K.B.), and the National Cancer Institute, the National Institute of Allergy and Infectious Diseases, and the National Institute of General Medical Sciences of the National Institutes of Health (R01GM157729, PI: S.K.B.). This work was also supported by the National Science Foundation (PHY-2210291, PI: J.N.O.), National Institutes of Health (R35GM153502-01, PI: P.C.W.) and the Welch Foundation (C-1792, to P.C.W.). Work at the Center for Theoretical Biological Physics was supported by the National Science Foundation (PHY-2019745). J.N.O. is a Cancer Prevention Research in Texas Scholar in Cancer Research.

The annotated peptide dataset curated by McDonald et al. [12] was a foundational aspect of this work. AF2 predictions were performed by using ColabFold, an open-source platform developed by the research groups of Dr. Sergey Ovchinnikov at Massachusetts Institute of Technology and of Dr. Martin Steinegger at Seoul National University. Molecular analyses were performed with custom implementations of Python code available online from the CHARMM-GUI lecture series developed by the research group of Dr. Wonpil Im at Lehigh University. Molecular graphics were visualized with UCSF ChimeraX, developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco, with support from the National Institutes of Health R01-GM129325 and the Office of Cyber Infrastructure and Computational Biology, National Institute of Allergy and Infectious Diseases [42].

AUTHOR CONTRIBUTIONS

A.M.I., C.M., S.K.B., P.C.W., J.N.O., R.P., and W.A. conceptualization; A.M.I. and C.M. methodology; A.M.I. investigation; A.M.I., C.M., S.K.B., P.C.W., J.N.O., R.P., and W.A. formal analysis; A.M.I. and C.M. writing—original draft; S.K.B., P.C.W., J.N.O., R.P., and W.A. writing—review & editing; S.K.B., P.C.W., J.N.O., R.P., and W.A. funding acquisition; R.P. and W.A. overall project supervision.

COMPETING INTERESTS

A.M.I. is a founder and partner of North Horizon, which is engaged in the development of artificial intelligence-based software. R.P. and W.A. are founders and equity shareholders of PhageNova Bio. R.P. is Chief Scientific Officer and a paid consultant of PhageNova Bio. R.P. and W.A are founders and equity shareholders of MBrace Therapeutics. R.P. and W.A. serve as paid consultants for MBrace Therapeutics. R.P. and W.A. have Sponsored Research Agreements (SRAs) in place with both PhageNova Bio and MBrace Therapeutics; this study falls outside of the scope of these SRAs. These arrangements are managed in accordance with the established institutional conflict-of-interest policies of Rutgers, The State University of New Jersey. C.M., S.K.B., P.C.W., and J.N.O declare no competing interests.

REFERENCES

1. Jumper, J., et al., Highly accurate protein structure prediction with AlphaFold. *Nature*, 2021. 596(7873): p. 583-589. doi:10.1038/s41586-021-03819-2
2. Kryshtafovych, A., et al., Critical assessment of methods of protein structure prediction (CASP)-Round XIV. *Proteins*, 2021. 89(12): p. 1607-1617. doi:10.1002/prot.26237
3. Burley, S.K. and H.M. Berman, Open-access data: A cornerstone for artificial intelligence approaches to protein structure prediction. *Structure*, 2021. 29(6): p. 515-520. doi:10.1016/j.str.2021.04.010
4. Kovalevskiy, O., J. Mateos-Garcia, and K. Tunyasuvunakool, AlphaFold two years on: Validation and impact. *Proc Natl Acad Sci U S A*, 2024. 121(34): p. e2315002121. doi:10.1073/pnas.2315002121
5. The Nobel Foundation, Press release: Nobel Prize in Chemistry 2024. 2024. https://www.nobelprize.org/prizes/chemistry/2024/press-release/
6. Wayment-Steele, H.K., et al., Predicting multiple conformations via sequence clustering and AlphaFold2. *Nature*, 2024. 625(7996): p. 832-839. doi:10.1038/s41586-023-06832-9
7. Monteiro da Silva, G., et al., High-throughput prediction of protein conformational distributions with subsampled AlphaFold2. *Nat Commun*, 2024. 15(1): p. 2464. doi:10.1038/s41467-024-46715-9
8. Kim, G., et al., Easy and accurate protein structure prediction using ColabFold. *Nat Protoc*, 2024. doi:10.1038/s41596-024-01060-5
9. Del Alamo, D., et al., Sampling alternative conformational states of transporters and receptors with AlphaFold2. *Elife*, 2022. 11. doi:10.7554/eLife.75751
10. Burley, S.K., W. Arap, and R. Pasqualini, Predicting Proteome-Scale Protein Structure with Artificial Intelligence. *N Engl J Med*, 2021. 385(23): p. 2191-2194. doi:10.1056/NEJMcibr2113027
11. Agarwal, V. and A.C. McShan, The power and pitfalls of AlphaFold2 for structure prediction beyond rigid globular proteins. *Nat Chem Biol*, 2024. 20(8): p. 950-959. doi:10.1038/s41589-024-01638-w
12. McDonald, E.F., et al., Benchmarking AlphaFold2 on peptide structure prediction. *Structure*, 2023. 31(1): p. 111-119 e2. doi:10.1016/j.str.2022.11.012
13. Pasqualini, R. and E. Ruoslahti, Organ targeting in vivo using phage display peptide libraries. *Nature*, 1996. 380(6572): p. 364-6. doi:10.1038/380364a0
14. Arap, W., et al., Steps toward mapping the human vasculature by phage display. *Nat Med*, 2002. 8(2): p. 121-7. doi:10.1038/nm0202-121
15. Pasqualini, R., W. Arap, and D.M. McDonald, Probing the structural and molecular diversity of tumor vasculature. *Trends Mol Med*, 2002. 8(12): p. 563-71. doi:10.1016/s1471-4914(02)02429-2
16. Staquicini, F.I., et al., Vascular ligand-receptor mapping by direct combinatorial selection in cancer patients. *Proc Natl Acad Sci U S A*, 2011. 108(46): p. 18637-42. doi:10.1073/pnas.1114503108
17. Mandelin, J., et al., Selection and identification of ligand peptides targeting a model of castrate-resistant osteogenic prostate cancer and their receptors. *Proc Natl Acad Sci U S A*, 2015. 112(12): p. 3776-81. doi:10.1073/pnas.1500128112

18. Krouk, G., W. Szponarski, and S. Ruffel, Unleashing the potential of peptides in agriculture and beyond. *Trends Plant Sci*, 2023. 28(7): p. 734-736. doi:10.1016/j.tplants.2023.03.025

19. Ormancey, M., et al., Complementary peptides represent a credible alternative to agrochemicals by activating translation of targeted proteins. *Nat Commun*, 2023. 14(1): p. 254. doi:10.1038/s41467-023-35951-0

20. Patel, P., et al., Cell-penetrating peptides for sustainable agriculture. *Trends Plant Sci*, 2024. 29(10): p. 1131-1144. doi:10.1016/j.tplants.2024.05.011

21. Montesinos, E., Functional Peptides for Plant Disease Control. *Annu Rev Phytopathol*, 2023. 61: p. 301-324. doi:10.1146/annurev-phyto-021722-034312

22. Grove, T.Z. and L. Regan, New materials from proteins and peptides. *Curr Opin Struct Biol*, 2012. 22(4): p. 451-6. doi:10.1016/j.sbi.2012.06.004

23. Sinha, N.J., et al., Peptide Design and Self-assembly into Targeted Nanostructure and Functional Materials. *Chem Rev*, 2021. 121(22): p. 13915-13935. doi:10.1021/acs.chemrev.1c00712

24. Chavez, M.I., et al., On the importance of carbohydrate-aromatic interactions for the molecular recognition of oligosaccharides by proteins: NMR studies of the structure and binding affinity of AcAMP2-like peptides with non-natural naphthyl and fluoroaromatic residues. *Chemistry*, 2005. 11(23): p. 7060-74. doi:10.1002/chem.200500367

25. Steinegger, M. and J. Soding, MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol*, 2017. 35(11): p. 1026-1028. doi:10.1038/nbt.3988

26. Mercurio, F.A., et al., The antimicrobial peptides casocidins I and II: Solution structural studies in water and different membrane-mimetic environments. *Peptides*, 2019. 114: p. 50-58. doi:10.1016/j.peptides.2018.09.004

27. Gronwald, W., T. Hohm, and D. Hoffmann, Evolutionary Pareto-optimization of stably folding peptides. *BMC Bioinformatics*, 2008. 9: p. 109. doi:10.1186/1471-2105-9-109

28. Maisuradze, G.G., et al., Preventing fibril formation of a protein by selective mutation. *Proc Natl Acad Sci U S A*, 2015. 112(44): p. 13549-54. doi:10.1073/pnas.1518298112

29. Atkins, A.R., R.C. Martin, and R. Smith, 1H NMR studies of sarafotoxin SRTb, a nonselective endothelin receptor agonist, and IRL 1620, an ETB receptor-specific agonist. *Biochemistry*, 1995. 34(6): p. 2026-33. doi:10.1021/bi00006a024

30. Wang, Y., et al., Solution structure of carnobacteriocin B2 and implications for structure-activity relationships among type IIa bacteriocins from lactic acid bacteria. *Biochemistry*, 1999. 38(47): p. 15438-47. doi:10.1021/bi991351x

31. Liu, Y., et al., Design and characterization of helical peptides that inhibit the E6 protein of papillomavirus. *Biochemistry*, 2004. 43(23): p. 7421-31. doi:10.1021/bi049552a

32. Benaki, D., et al., Solution structure of humanin, a peptide against Alzheimer's disease-related neurotoxicity. *Biochem Biophys Res Commun*, 2005. 329(1): p. 152-60. doi:10.1016/j.bbrc.2005.01.100

33. Zhang, M., et al., Interaction of the histone mRNA hairpin with stem-loop binding protein (SLBP) and regulation of the SLBP-RNA complex by phosphorylation and proline isomerization. *Biochemistry*, 2012. 51(15): p. 3215-31. doi:10.1021/bi2018255

34. Evanics, F., et al., Nuclear magnetic resonance structures of the zinc finger domain of human DNA polymerase-alpha. *Biochim Biophys Acta*, 2003. 1651(1-2): p. 163-71. doi:10.1016/s1570-9639(03)00266-8

35. Chua, G.L., et al., Structures and interaction analyses of integrin alphaMbeta2 cytoplasmic tails. *J Biol Chem*, 2011. 286(51): p. 43842-43854. doi:10.1074/jbc.M111.280164

36. Deshayes, K., et al., Rapid identification of small binding motifs with high-throughput phage display: discovery of peptidic antagonists of IGF-1 function. *Chem Biol*, 2002. 9(4): p. 495-505. doi:10.1016/s1074-5521(02)00129-1

37. Branden, G. and R. Neutze, Advances and challenges in time-resolved macromolecular crystallography. *Science*, 2021. 373(6558). doi:10.1126/science.aba0954

38. Amann, S.J., et al., Frozen in time: analyzing molecular dynamics with time-resolved cryo-EM. *Structure*, 2023. 31(1): p. 4-19. doi:10.1016/j.str.2022.11.014

39. Anfinsen, C.B., Principles that govern the folding of protein chains. *Science*, 1973. 181(4096): p. 223-30. doi:10.1126/science.181.4096.223

40. Im Lab, Lehigh University. CHARMM-GUI lecture series: RMSD Between Two Structures. https://www.charmm-gui.org/?doc=lecture&module=scientific&lesson=11.

41. Im Lab, Lehigh University. CHARMM-GUI lecture series: Average and RMSF from Multiple Structures. https://www.charmm-gui.org/?doc=lecture&module=scientific&lesson=12.

42. Meng, E.C., et al., UCSF ChimeraX: Tools for structure building and analysis. *Protein Sci*, 2023. 32(11): p. e4792. doi:10.1002/pro.4792