Assessing PDB Macromolecular Crystal Structure Confidence at the Individual Amino Acid Residue Level

Chenghua Shao[1,2,*], Sebastian Bittrich[3], Sijian Wang[2,4], and Stephen K. Burley[1,2,3,5,6,*]

[1]Research Collaboratory for Structural Bioinformatics Protein Data Bank, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA.

[2]Institute for Quantitative Biomedicine, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA.

[3]Research Collaboratory for Structural Bioinformatics Protein Data Bank, San Diego Supercomputer Center, University of California San Diego, La Jolla, CA 92093, USA.

[4]Department of Statistics, Rutgers, The State University of New Jersey, New Brunswick, NJ, 08903, USA.

[5]Rutgers Cancer Institute of New Jersey, Robert Wood Johnson Medical School, New Brunswick, NJ, 08903, USA.

[6]Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA.

*Correspondence: Dr. Chenghua Shao (chenghua.shao@rcsb.org) and Dr. Stephen K. Burley (stephen.burley@rcsb.org)

Lead Contact: Dr. Chenghua Shao (chenghua.shao@rcsb.org)
Mailing Address: Institute for Quantitative Biomedicine,
Rutgers, The State University of New Jersey
174 Frelinghuysen Road
Piscataway, NJ 08854 USA.

Keywords

Macromolecular structure quality, Structure confidence, Real-space-correlation-coefficient, RSCC, Protein Data Bank, PDB, RCSB PDB, AlphaFold2, AlphaFoldDB, Predicted local distance difference test, pLDDT

Summary

Approximately 87% of the more than 190,000 atomic-level, (three-dimensional) 3D biostructures in the Protein Data Bank (PDB) were determined using macromolecular crystallography (MX). Agreement between 3D atomic coordinates and experimental data for >100 million individual amino acid residues occurring within ~150,000 PDB MX structures was analyzed in detail. The Real-Space-Correlation-Coefficient (RSCC) calculated using the 3D atomic coordinates for each residue and experimental-data-derived electron density enables outlier detection of unreliable atomic coordinates (particularly important for poorly-resolved sidechain atoms) and ready evaluation of local structure quality by PDB users. For human protein MX structures in PDB, comparisons of the per-residue RSCC metric with AlphaFold2 computed structure model confidence (pLDDT-predicted local distance difference test) document (i) that RSCC values and pLDDT scores are correlated (median correlation coefficient~0.41), and (ii) that experimentally-determined MX structures (3.5 Å resolution or better) are more reliable than AlphaFold2 computed structure models and should be used preferentially whenever possible.

Introduction

The PDB was established in 1971 as the first digital data resource in biology (Protein Data Bank, 1971). It has grown more than 27,000-fold to become the only freely accessible global archive of 3D structures of proteins, nucleic acids, and their complexes with one another and small-molecule ligands experimentally determined using macromolecular crystallography (MX), nuclear magnetic resonance (NMR) spectroscopy, and electron microscopy (3DEM). Open access to well-validated, expertly-biocurated PDB structures enables scientific advances across fundamental biology, biomedicine, energy sciences, and biotechnology/bioengineering (Burley et al., 2018; Westbrook and Burley, 2019; Westbrook et al., 2020; Goodsell et al., 2020; Goodsell and Burley, 2022). PDB structures also played critical roles in efforts aimed at predicting (or computing) atomic-level 3D structure models from protein sequence alone (Burley and Berman, 2021; Burley et al., 2021). Today, AlphaFold2 (Jumper et al., 2021; Tunyasuvunakool et al., 2021) and RoseTTAFold (Baek et al., 2021) support computation of structure models of globular proteins with accuracies comparable to those of lower-resolution experimental methods.

The Worldwide Protein Data Bank (wwPDB, wwpdb.org; (Berman et al., 2003; wwPDB consortium, 2019)) manages the PDB archive according to the FACT (Fairness-Accuracy-Confidentiality-Transparency (van der Aalst et al., 2017)) and FAIR (Findable-Accessible-Interoperable-Reusable (Wilkinson et al., 2016)) Principles underpinning responsible data stewardship in the modern era. Current wwPDB members include the US-funded RCSB Protein Data Bank or RCSB PDB (Berman et al., 2000; Burley et al., 2022; Rose et al., 2021); Protein Data Bank in Europe or PDBe (Mir et al., 2018); and Protein Data Bank Japan or PDBj (Kinjo et al., 2018); plus two specialist data resources (3DEM data resource: Electron Microscopy Data Bank or EMDB (Abbott et al., 2018); NMR data resource: Biological Magnetic Resonance Bank or BMRB (Ulrich et al., 2008)). The wwPDB OneDep system for global deposition, validation, and biocuration of PDB structures (Young et al., 2017; Gore et al., 2017; Young et al., 2018; Feng et al., 2021) serves tens of thousands of structural biologists working on all permanently

inhabited continents. wwPDB validation reports generated within OneDep for every PDB structure provide comprehensive quality assessments, calculated using community-standard software tools. For MX structures, wwPDB validation reports summarize individual residue quality using the local electron density goodness-of-fit metric RSCC $= corr(\rho_{obs}, \rho_{calc})$ as Pearson's correlation coefficient between observed and calculated electron densities of ρ_{obs} and ρ_{calc}, respectively (Tickle, 2012; Brändén and Jones, 1990). Within the wwPDB validation pipeline, RSCC is calculated using EDS (Kleywegt et al., 2004).

Relentless growth in the number of MX structures in PDB since 1971 has yielded an enormous body of open access data for biomedical research. It has also created considerable challenges for some PDB data consumers, who may encounter difficulties when discerning which part (or parts) of a given PDB structure are not to be trusted, and, consequently, may not be useful for interpreting experimental results or generating hypotheses. Herein, we describe use of per-residue RSCC values to identify well-resolved *versus* less well-resolved regions of MX structures in PDB. The robustness of RSCC was then systematically cross-examined with another recently developed per-residue confidence measure pLDDT used by AlphaFold2 (Jumper et al., 2021; Varadi et al., 2022). Unlike RSCC that reflects quality of fitting to experimental data, pLDDT is calculated with an entirely different algorithm utilizing superposition-free pairwise distance test (Mariani et al., 2013) calibrated against the PDB archive (Jumper et al., 2021). Comparing RSCC and pLDDT furthers understanding of reliability of both PDB MX structures and AlphaFold2 computed structure models, and sheds light on what information to use from hundreds of thousands of computed structure models of proteins freely available from AlphaFoldDB (Varadi et al., 2022) and the ModelArchive (Schwede et al., 2009).

Results

RSCC Distribution

RSCC Distribution and Identification of Outliers in PDB MX Structures at the Individual Residue Level

The distribution of RSCC values for >100 million standard amino acid residues and nucleotides from ~150,000 PDB MX structures is illustrated in Figure 1A (mean and median values 0.935 and 0.955, respectively). Markedly different from a normal distribution, the skewed RSCC distribution is heavily tailed on its left-side (lower RSCC values) and bounded between the values -1.0 and 1.0. Therefore, neither standard deviation (σ) nor interquartile range (IQR) can be used to accurately characterize statistical dispersion of the RSCC distribution or identify outliers. For example, >4% of residues in PDB have RSCC values below μ(mean)-2σ, *versus* only 2.5% for a normal distribution. Assuming normal distribution on non-normally distributed data may yield improper classification of outliers. For example, in the current wwPDB validation report, residues with poor fit of atomic coordinates to local electron density are identified as outliers with RSRZ>2, where RSRZ is the normalized Z score of the per-residue real-space R value (RSR) (Kleywegt et al., 2004). Because RSR values do not follow a normal distribution, use of this criterion (RSRZ>2) can lead to overestimation on the number of outliers (see STAR Methods). An alternative criterion for outlier classification frequently used for non-normal symmetrical distributions is 1.5 IQR below the first quartile. Because the RSCC distribution is so heavily skewed, opting for this metric would classify >7% of residues as RSCC outliers, which would overestimate the number of outliers.

Transformation of the RSCC distribution into a regular parametric distribution was attempted unsuccessfully. By way of explanation, the Pearson-correlation coefficient bounded between -1.0 and +1.0 can be transformed into normal distribution through Fisher's Z-transformation, but only when the data used to generate the correlation coefficient have a bivariate normal distribution. This condition is not

met for either experimentally-observed electron density (experimental-data-derived) or calculated electron density based on 3D structure atomic coordinates (see STAR Methods).

An outlier is defined as an observation that deviates so much from the other observations as to arouse suspicion that it originated from a different mechanism (Hawkins, 1980). Using a probability density-based approach, outliers are the least probable observations associated with the lowest estimated probability density from the data distribution (Shao et al., 2018). RSCC values follow a unimodal distribution with monotonically increasing probability density from the lowest value to the mode or peak value (Figure 1A). Consequently, for data values less than the mode probability density ranking is consistent with ranking of RSCC values from low to high, permitting identification of one-sided outliers with a percentile cut off for poorly resolved amino acids in MX structures of proteins. The lowest 1% of RSCC values (Figure 1A, to the left of the vertical 1% line) or the next 4% of RSCC values (Figure 1A, between the 1% and 5% lines) have lowest probability rankings. Both 1% and 5% probability cutoffs are commonly used thresholds for classifying data values as outliers. The remainder of the first quartile (between 5% and 25%) can be considered as having intermediate probability. The observations between 25% and 100% have high probability.

RSCC Distribution versus Structure Residue Type and MX Resolution

RSCC values are influenced by the chemical structure of the biopolymer component. Comparison of RSCC distributions for proteins and nucleic acids (Figure 1B) documented that RSCC values for nucleotides are typically lower than for amino acid residues, demonstrating that the fit of atomic coordinates to experimental-data-derived electron density for individual nucleotides is generally inferior to that observed for individual amino acid residues in PDB MX structures. Figure 1C illustrates separate RSCC distributions for each amino acid residue type (hereafter residue). In general, residues with non-polar sidechains occurring more frequently within the hydrophobic cores of globular proteins (*e.g.*, valine, leucine, isoleucine, phenylalanine, and tryptophan) have higher median RSCC values than residues with

polar sidechains, which are more frequently found on the surfaces of globular proteins (*e.g.*, serine, asparagine, aspartate, glutamine, glutamate, lysine, histidine, and arginine). The relative paucity of steric constraints on surface residue atomic positions permits adoption of multiple sidechain and even backbone conformations that may not be well resolved in MX experiments, particularly at resolution worse than ~3.5 Å (*i.e.*, higher value). Analyses presented below were limited to PDB MX structures of proteins.

Resolution of the experimental diffraction data represents an important determinant of MX structure quality, both globally and locally. Figure 1D illustrates RSCC distributions *versus* resolution. As expected, the median value of RSCC decreases with resolution, because experimental-data-derived electron density is not as well resolved at lower resolution. This trend is most evident when comparing higher- and lower-resolution ranges (1.0 to 1.5 Å *versus* 3.0 to 3.5 Å). RSCC distributions for each residue type were analyzed as a function of resolution (see Supplementary Data S1 for tabulated summary statistics). These results enable identification of lowest and low probability RSCC value cutoffs for each residue type as a function of resolution. For each residue type and resolution range, we examined both high probability residues (*i.e.*, those with RSCC values near the peak of the distributions illustrated in Figure 1) and lowest probability or outlier residues (*i.e.*, those with RSCC values falling within the far-left tails of distributions illustrated in Figure 1). PDB structures are generally well resolved, with very few structures that are grossly wrong. In constructing the RSCC distribution, we excluded the small number of MX structures that do not agree with the underlying experimental data (~2.6%). Short segments of poorly resolved residues within an otherwise well-resolved structure are characterized by consecutive RSCC outliers.

For the avoidance of doubt, MX is an extremely powerful experimental tool for determining 3D structures of well-ordered, globular proteins. Atomic coordinates of residues with RSCC values greater than 0.85 can be trusted for all residue types. The fraction of individual residues in PDB MX structures with RSCC>0.85 is ~95% for those with resolutions better than 2 Å (~50% of PDB MX structures) and ~93% for those with resolution better than 3.5 Å (~98% of PDB MX structures), documenting the power of the method. In contrast, conventional MX method is not well suited to the challenge of studying

conformational heterogeneity within proteins, whatever its origins. The atomic coordinates of poorly-resolved individual residues present in a given PDB MX structure (*i.e.*, outlier residues with lowest or low probability RSCC values) should not be trusted.

Comparing RSCC Distributions with AlphaFoldDB pLDDT Distributions

Comparing RSCC and pLDDT for PDB MX Structures of Human Proteins

To assess 3D structure prediction confidence quantitatively, AlphaFold2 provides per-residue pLDDT (Jumper et al., 2021) scores (scaled between 0 and 100): Very high confidence if pLDDT≥90; Confident if 90>pLDDT≥70; Low confidence if 70>pLDDT≥50; Very low confidence if pLDDT<50. Artificial intelligence/deep learning approaches outperform physicochemical based methods for predicting intrinsically-disordered regions (IDRs) of proteins (Necci et al., 2021). Lower pLDDT scores are relatively good predictors of protein disorder (Ruff and Pappu, 2021).

More than 23,000 AlphaFold2 predicted 3D structures of human proteins (computed structure models or CSMs) were downloaded from AlphaFoldDB (Varadi et al., 2022) for analysis. The pLDDT distribution of ~15 million individual residues contained within the downloaded CSMs is illustrated in Figure 2 (dashed line). It is bimodal (major peak~95, minor peak~35). Approximately 28% of these CSM residues have pLDDT<50, indicating Very low confidence in their predicted atomic coordinates, which is consistent with earlier observations (Thornton et al., 2021) and independent estimates of IDRs in the human proteome (Ruff and Pappu, 2021; Tunyasuvunakool et al., 2021).

Amino acid sequences of each human protein AlphaFoldDB CSM were used to query and align with all PDB protein structure sequences using the RCSB PDB 1D coordinate server Application Programming Interface or API (https://1d-coordinates.rcsb.org/) (Segura et al., 2020). Approximately 7,500 unique human protein sequences were detected in more than 53,000 PDB structures. Approximately, 64% of human protein structures in PDB encompass only part of the full-length polypeptide chain (<95%

sequence coverage), whereas ~36% of PDB structures of human proteins have ≥95% sequence coverage.

To compare RSCC and pLDDT at the individual residue level, RSCC values for complete residues (*i.e.*, excluding all residues with missing atoms, and/or partial atomic occupancy, and/or multiple conformations) in more than 41,000 PDB MX structures of ~5,300 unique human proteins were selected and compared to per-residue pLDDT scores of the corresponding AlphaFoldDB CSMs in pairwise fashion. Figure 2 dotted line illustrates the pLDDT distribution for residues occurring within CSMs of human protein sequences or sequence regions present within experimental samples (*i.e.*, protein studied by MX) in PDB. This subset has a markedly different pLDDT distribution *versus* that of all human protein CSMs. Only 9% of residues have pLDDT<50, and there is no minor peak in the distribution at ~35 as seen for AlphaFoldDB CSMs of all human proteins. Figure 2 solid line illustrates the pLDDT distribution of CSM residues corresponding to completely resolved residues in all PDB human proteins structures determined by MX (median pLDDT score~96; ~2.4% residues have pLDDT<70 and ~0.6% residues have pLDDT<50). The narrow-peaked Figure 2 solid line distribution reflects the fact that MX performs best with relatively compact globular proteins (lacking poorly ordered N- or C-termini and/or long surface loops). Considerable efforts in expression construct design are frequently required before MX can be employed for high-resolution structural studies of human protein domain structures (Gao et al., 2005).

Comparing RSCC and pLDDT for Human Protein MX Structures in PDB

The distribution of overall correlation coefficients between RSCC values and pLDDT scores (RSCC/pLDDT-CC) for every residue represented in a human protein PDB MX structure is plotted in Supplementary Figure S1 (median value~0.41, range -0.48 to 0.95; Supplementary Data S1). RSCC values and pLDDT scores were also compared on a per-residue basis for various representative PDB MX structures.

Figure 3 illustrates our findings for full-length human RNA-binding protein Nova-1 (UniProt ID P51513), which consists of an N-terminal segment plus three globular K-homology or KH domains (KH1, KH2, and KH3), well separated from one another in the polypeptide chain sequence (Figure 3A). Two related MX structures are available from the PDB (PDB: 2ANR (KH1 and KH2) (Teplova et al., 2011); and PDB: 1DT4 (KH3) (Lewis et al., 1999)). To compare RSCC and pLDDT graphically, per-residue pLDDT scores were scaled by 1/100 (resulting metrics falling between 0 and 1) and plotted *versus* UniProt sequence numbering (Figure 3B). Figure 3B shows that per-residue RSCC values for PDB: 2ANR (KH1 and KH2) and per-residue pLDDT scores for the AlphaFoldDB CSM are well correlated (RSCC/pLDDT-CC~0.75 for common residues). Close inspection of the experimental-data-derived electron density for the KH1 and KH2 domains revealed that most residues were well resolved by the MX experiment (*i.e.*, they have high RSCC values).

Minor exceptions include the loops connecting the second and third β-strands in the three-stranded, anti-parallel β-sheet characteristic of KH domains (Lewis et al., 1999), and the C-terminus of KH1 (Figure 3C). Limited proteolysis studies of full-length human Nova-1 protein documented susceptibility to cleavage in these same regions (see Figure 2 of (Lewis et al., 1999)), suggesting that they are conformationally flexible in solution. The AlphaFoldDB CSM for human Nova-1 superposes well on PDB: 2ANR (KH1 and KH2) with Cα Root-Mean-Square-Deviation or RMSD~0.3 Å. A small number of residues falling within the inter-strand loops and the inter-domain region of the polypeptide chain differ in 3D structure between PDB: 2NAR (KH1 and KH2) and the AlphaFoldDB CSM, providing further evidence that these segments of the human Nova-1 polypeptide chain are flexible. Not surprisingly, the atomic coordinates of these residues have low RSCC values in PDB: 2NAR (KH1 and KH2), and low pLDDT scores in the AlphaFoldDB CSM.

Figure 3B also reveals that per-residue RSCC values for PDB: 1DT4 (KH3) and pLDDT scores for the corresponding residues in the AlphaFoldDB CSM are not as well correlated as seen for domains KH1

and KH2 (RSCC/pLDDT-CC~0.39). The AlphaFoldDB CSM and PDB: 1DT4 (KH3) superpose well (Cα RMSD~0.7 Å). Structural differences between PDB: 1DT4 (KH3) and the AlphaFoldDB CSM are restricted to an inter-strand loop (residues 461 to 466), the position of which appears to be stabilized by interactions with a neighboring protomer in the crystal lattice (Figure 3D). Again, these findings are consistent with KH domain inter-strand loops being conformationally flexible.

Comparison of Residues with High RSCC Values in PDB MX Structures and High pLDDT Scores in AlphaFoldDB CSMs

Backbone atoms in AlphaFoldDB CSMs with high per-residue pLDDT scores (median pLDDT>90) generally superpose very well on their corresponding PDB MX structures, even in cases when the RSCC/pLDDT-CC falls below the median value of ~0.41 (*i.e.*, 0.2 to 0.4). Substantial differences in 3D between high RSCC value human PDB MX structures and high pLDDT value AlphaFoldDB CSMs typically occur in flexible internal loop regions or N- and C-termini of MX structures (with per-residue RSCC values <0.8 and pLDDT scores <50). Accurate predictions of backbone atomic positions are, however, not always enough to understand mechanistic details of proteins that function as molecular machines dependent on precise arrangements of amino acid sidechains. To examine this issue, we compared representative high-resolution PDB MX structures of human proteins with their corresponding AlphaFoldDB CSMs.

Table 1 summarizes results obtained by comparing the AlphaFoldDB CSM for human hemoglobin α subunit (HbA-α, UniProt ID P69905; median per-residue pLDDT score~98.6) and high-resolution PDB MX structures of the same protein in different oxidation states. Three sets of atomic coordinates for HbA-α were extracted from among PDB MX structures of the human hemoglobin $\alpha_2\beta_2$ hetero-tetramer as follows: PDB: 2DN1 (oxy); PDB: 2DN2 (deoxy); and PDB: 2DN3 (carbonmonoxy). All three structures were experimentally determined at ~1.25 Å resolution by the same research group with well-resolved electron density for most non-hydrogen atoms and median per-residue RSCC values>0.94 (Park et al., 2006). The polypeptide chain backbone of the single AlphaFoldDB CSM for HbA-α superposes well on

those of all three experimental-data-derived structures (Cα RMSD~0.3 to ~0.5 Å) with RSCC/pLDDT-CC ranging from ~0.2 to ~0.6. For reference, the precision of the atomic coordinates for non-hydrogen atoms in an MX structure determined at 1.0 to 2.0 Å resolution is expected to be 0.1 to 0.2 Å. RSMD values between each of the three PDB MX structures and the AlphaFoldDB CSM calculated using all non-hydrogen atoms are higher than those obtained for Cα atoms alone (ranging from ~0.7 to ~1.5 Å). These results document that predicted sidechain atomic positions in the AlphaFoldDB CSM for human HbA-α are less reliable than those of main chain atoms.

Supplementary Figure S2(A) shows that His88 (responsible for binding to the heme group iron atom) is predicted to occur in the position observed by MX in PDB: 2DN2 (deoxy), which is different in the other two oxidation state structures. In the same view, the predicted position of Leu83 resembles that observed in PDB: 2DN1 (oxy), but not the other two oxidation states. Also in the same view, the predicted position of Trp14 differs dramatically from that observed in PDB: 2DN1 (oxy). Therefore, even AlphaFoldDB CSMs with very high median pLDDT scores should not be relied upon to reproduce the "ground-truth" of well-determined, high-resolution PDB MX structures, particularly when the macromolecules in question are structurally dynamic and knowledge of amino acid sidechain positions is critical. Because pLDDT scores are calibrated against lDDT-Cα (Jumper et al., 2021), they may not reflect confidence in sidechain atomic positions. For amino acids with pLDDT>90 in AlphaFold2 CSMs, 80% of χ^1 torsion angles about the Cα-Cβ bond fall within 40° of values in PDB reference structures (Tunyasuvunakool et al., 2021). For longer amino acid sidechains, even modest errors in χ^1 torsion angle predictions can translate into large position errors for sidechain atoms distal to Cβ.

Discordance between RSCC Values and pLDDT Scores for Individual Human Proteins

We also analyzed human proteins for which AlphaFoldDB CSMs exhibited very low pLDDT scores while corresponding PDB MX structures (containing the same polypeptide chain or segment thereof) had high RSCC values, and *vice versa*.

First, AlphaFoldDB CSMs for human proteins with low overall pLDDT scores were examined and compared to their PDB MX structure counterparts. 71 PDB MX structures differ substantially in 3D structure from corresponding AlphaFoldDB CSMs with median pLDDT<50 (Supplementary Data S1). For example, PDB: 3LK3 (Hernandez-Valladares et al., 2010), determined at 2.68 Å resolution, encompasses residues 971 to 1035 of UniProt ID Q5VZK9 (human F-actin-uncapping Leucine-rich repeat-containing protein 16A; entity ID 3; median RSCC~0.95). The corresponding AlphaFoldDB CSM has very low confidence (pLDDT<50 for every residue; median pLDDT~36). The AlphaFoldDB CSM and the PDB MX structure are too different to be superposed in 3D. Residues 971 to 1035 of human F-actin-uncapping Leucine-rich repeat-containing protein are well resolved in PDB: 3LK3, probably because they are interacting with a globular protein in the crystal and may well be disordered in the absence of a binding partner (Supplementary Figure S2(B)).

Low pLDDT score segments within AlphaFoldDB CSMs do not always correspond to IDRs of proteins. The PDB MX structure of human PR domain zinc finger protein 4 (UniProt ID Q9UKN5 residues 393 to 530; PDB: 3DB5; DOI: 10.2210/pdb3DB5/pdb, determined at 2.15 Å resolution, median RSCC~0.96) reveals a compact, largely β-strand domain (Supplementary Figure S2(C)). The corresponding AlphaFoldDB CSM has median per-residue pLDDT score~37, with a very different topological arrangement of the polypeptide chain in 3D *versus* PDB: 3DB5 (Supplementary Figure S2(C)). Because PDB: 3DB5 is a high-quality structure with atomic coordinates of most of its residues well resolved, the AlphaFoldDB CSM is likely to be unreliable as indicated by its low pLDDT scores. Alternatively, some AlphaFoldDB CSMs with low pLDDT scores can be partially superposed on their corresponding PDB MX structures (see Supplementary Figure S2(D) comparing the CSM of UniProt ID P41182 and PDB: 7LWE).

Individual residues with pLDDT<50 in AlphaFoldDB CSMs were also examined. Particular attention was paid to residues with pLDDT<50 also present in 288 high-resolution (1 to 1.1 Å) PDB MX structures of human proteins. Among these PDB MX structures, we identified 70 markedly “discordant” residues with

RSCC>0.95 and pLDDT<50. For example, PDB: 4FKA (Prugovecki et al., 2012) is an MX structure of human insulin determined at 1.08 Å resolution (UniProt ID P01308). It does not superpose well on its corresponding AlphaFoldDB CSM. Thornton and co-workers previously observed for this case that "the AlphaFold model bears no resemblance to the PDB structure, possibly because it has missed the disulfide bonds that hold the protein together" (Thornton et al., 2021). Like the α subunit of human hemoglobin, the lesson from human insulin is that well-resolved PDB structures should be used preferentially (*versus* CSMs) whenever they are available. Neither AlphaFoldDB nor the ModelArchive should not be relied on as sole sources of 3D protein structure information.

Second, the obverse scenario of high pLDDT *versus* low RSCC was also examined for representative cases. For example, PDB: 7E5M (Sun et al., 2021) is an MX structure of residues 33 to 268 of human tumor-associated calcium signal transducer 2 determined at 3.2 Å resolution (UniProt ID P09758), with median RSCC~0.67 *versus* median pLDDT~94 for the corresponding AlphaFoldDB CSM. The experimental-data-derived structure and the CSM are very similar in 3D, with Cα RMSD~0.5 Å when loop residues 82 to 103 are excluded from the comparison (Supplementary Figure S2(E)). Close inspection of the experimental-data-derived electron density revealed that the lower median RSCC value likely stems from paucity of diffraction data. For resolutions worse than ~3.5 Å (*i.e.*, higher number), MX structures are not as well resolved (particularly amino acid sidechains), because the number of experimental observations (diffraction measurements) per atom may be insufficient for the method to precisely determine atomic positions.

RSCC-based Confidence Criteria and Color Scheme for PDB MX Structure Display

Atomic coordinates of most PDB MX structures are well resolved in the experimental-data-derived electron density. Within individual MX structures, however, atomic coordinates for individual residues or short segments of the polypeptide chain(s) may not be as accurate. Statistically rigorous outlier detection

of RSCC values can provide readily interpretable measures of local structure quality for PDB data consumers who are not experts in structural biology.

Similar to the AlphaFoldDB per-residue pLDDT display color scheme, residues in PDB MX structures can be assigned an RSCC-based confidence and colored-coded so that RSCC outliers are readily apparent in ribbon representation 3D graphical displays. The vast majority of residues that are very well resolved by the MX method (Very well resolved - RSCC ordinal ranking between 25% and 100%, *i.e*., the most probable RSCC range) can be colored blue. Well-resolved residues with RSCC ordinal ranking between 5% and 25% can be colored cyan. Outlier residues that are not well resolved by the MX method can be colored either yellow (Low confidence, RSCC ordinal ranking between 1% and 5%) or orange (Very low confidence, lowest 1% of RSCC values).

Figure 4 illustrates application of this RSCC probabilities-based color-coding scheme for PDB: 1DTJ (third domain (KH3) of human RNA-binding protein Nova-2 determined at 2.0 Å resolution, UniProt ID Q9UNW9 (Lewis et al., 1999)). Most human Nova-2 KH3 residues are colored blue or cyan both in 1D (Figure 4A) and 3D (Figure 4B) representations of the MX structure, reflecting the fact that they were Well resolved or Very well resolved by the structure-determination method. Some residues occurring in the inter-strand loop, a short segment between the first and second α-helices, and the C-terminus of the domain are colored yellow or orange, because their RSCC values were deemed to be statistical outliers (falling within the lower 1% to 5% or lowest 1% of the probability distributions for those particular amino acids at 2.0 Å resolution). The similarity of the suggested color scheme with that of Alphafold2 pLDDT confidence scores, buttressed by the comparability of the distributions depicted in Figures 1 and 2, could help users of PDB data make informed assessments of experimentally-determined structure quality without having to delve into the details of the wwPDB validation report. The color scheme choice was intended to facilitate direct comparison of PDB MX structures with CSMs of proteins generated *via* deep learning methods. The RSCC-based quality classification and color scheme has been implemented on

the RCSB PDB research-focused web portal RCSB.org. Supplementary Figure S3 shows the Mol* 3D structure display of PDB: 1DTJ, reproducing the ribbon representation view of the Nova-2 KH3 domain illustrated in Figure 4. (N.B.: A comparable graphical representation of nucleotide quality was not implemented on RCSB.org because there are currently insufficient data with which to define reliable probability rankings of per-residue RSCC values.)

Discussion

Analyses of the distribution of RSCC values for PDB MX structures demonstrated the utility of RSCC as a residue-level MX structure quality indicator that can be used to assess the goodness of fit of atomic coordinates to experimental-data-derived electron density. Outlier residues can be objectively identified as those with lowest 1% or lowest 5% probability ranking. Typically, such outliers have little, if any, corresponding experimental-data-derived electron density signal. Low confidence and very low confidence portions of PDB MX structures should be treated with caution. In many, perhaps most, cases, low RSCC values indicate that the corresponding residue(s) or short polypeptide chain segments are not well ordered in the crystal, and do not, therefore, contribute to the Bragg diffraction signal measured in the MX experiment. They may be statically or dynamically (*i.e*., flexible) disordered. The conventional MX method cannot distinguish these possibilities.

Many structural biologists are taking a “glass half full” view of CSMs generated using AlphaFold2, RoseTTAFold, *etc*. They use the CSMs of full-length eukaryotic proteins to design protein expression constructs that exclude Low confidence (50≤pLDDT<70, color coded yellow) and Very low confidence (pLDDT<50, color coded orange) segments of longer polypeptide chains to generate samples of truncated proteins suitable for structure/function studies using MX, NMR, or 3DEM. They also scrutinize segments of polypeptide chains with Low and/or Very low confidence predictions for potentially globular segments that have not been previously characterized (within the dashed line circle in Figure 3A, for example). For the other 99% of PDB data consumers, poorly resolved residues in PDB MX structures

can be viewed in two ways. They can be seen as an inconvenience because they are unreliable. Alternatively, they can be viewed as a source of opportunities for designing experiments using methods other than MX to probe the biological or biochemical function of poorly resolved residues.

To explore the question of disorder in MX structures, RSCC values were compared and contrasted to AlphaFoldDB CSM pLDDT scores for all human protein structures in the PDB. Per-residue RSCC values in PDB structures and AlphaFoldDB CSM pLDDT scores for the same human proteins are correlated, suggesting that both metrics can be used to assess polypeptide chain flexibility or disorder. Cases wherein RSCC values and AlphaFoldDB CSM pLDDT scores are not correlated may serve as useful case studies for those seeking to improve *de novo* protein structure prediction methods.

Acknowledgements

The authors thank the tens of thousands of structural biologists who deposited structures to the PDB since 1971 and the many millions of researchers, educators, and students around the world who consume PDB data. We also gratefully acknowledge contributions to the success of the PDB archive made by past members of RCSB PDB and our Worldwide Protein Data Bank partners (PDBe, PDBj, EMDB, and BMRB). RCSB PDB core operations are jointly funded by the National Science Foundation (NSF, DBI-1832184, PI: S.K. Burley), the US Department of Energy (DE-SC0019749, PI: S.K. Burley), 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 (R01GM133198, PI: S.K. Burley). Other funding awards to RCSB PDB by the NSF and to PDBe by the UK Biotechnology and Biological Research Council are jointly supporting development of a Next Generation PDB archive (DBI-2019297, PI: S.K. Burley; BB/V004247/1, PI: Sameer Velankar) and new Mol* features (DBI-2129634, PI: S.K. Burley; BB/W017970/1, PI: Sameer Velankar). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author Contributions

Conceptualization: S.K.B., C.S., and S.W.; Methodology: C.S., S.B., S.W., and S.K.B.; Statistical Analysis: C.S. and S.W.; Writing: C.S., S.K.B., S.B., and S.W.; Supervision: S.K.B.; Funding Acquisition: S.K.B.

Declaration of Interests

Author Stephen K. Burley is a member of the Structure's advisory board. Otherwise, the authors declare no competing interests.

Figure Titles and Legends

Figure 1. RSCC value distribution of standard residues in the PDB archive. Probability density plots of RSCC values for (A) all standard protein and nucleic acid residues, (B) amino acid residues *versus* nucleotides, (C) individual amino acid types ordered by number of atoms (plus MSE or seleno-methionine), and (D) individual amino acid types *versus* resolution. Vertical lines denote the median value, and the 5% and 1% percentiles, respectively (*i.e*., cumulative percentages of data values to the left of each line). 25% (1^{st} quartile) line is shown in (A).

Figure 2. Per-residue pLDDT score distributions of AlphaFoldDB CSMs of human proteins. Dashed Line: All human protein residues in AlphaFoldDB (major peak~95, minor peak~35). Dotted Line: Residues of human protein sequences present in PDB structures determined by MX, 3DEM, and NMR. A non-redundant sequence is chosen by the maximal overlap with its UniProt reference sequence. Solid Line: Human protein residues observed in PDB MX structures. Y-axis is probability density using the same scale for all three distributions. Prediction confidence color coding: blue-Very high confidence pLDDT≥90; cyan-Confidence 70≤pLDDT<90; yellow-Low confidence 50≤pLDDT<70; orange-Very low confidence pLDDT<50.

Figure 3. Human RNA-binding protein Nova-1: Comparison of per-residue RSCC values for two PDB structures and AlphaFoldDB CSM per-residue pLDDT scores. (A) Mol* (Sehnal et al., 2021) ribbon representation graphical display of the AlphaFoldDB CSM of the entire polypeptide chain, color coded by pLDDT scores as for Figure 2. PDB: 2ANR (KH1 and KH2) and PDB: 1DTJ (KH3) superposed on the CSM, drawn with semi-transparent gray shading. Dashed line circle denotes a portion of the full-length human Nova-1 protein not previously recognized as potentially globular. (B) Per-residue overlay of pLDDT/100 (magenta, full-length UniProt ID P51513) and RSCC (black, observed residues in PDB: 2ANR (KH1 and KH2) and PDB: 1DT4 (KH3)). The gap in the KH2 overlay corresponds to an unobserved polypeptide chain segment in PDB: 2ANR. (C) Experimental-data-derived electron density overlayed on Very low confidence PDB: 2ANR atomic coordinates for residues Pro124 to Gln126 of domain KH1 ($2|F_{observed}|-|F_{calculated}|$ map, contoured at 1.0 σ). (D) Experimental-data-derived electron density overlayed on mostly Well resolved PDB: 1DT4 (KH3) atomic coordinates for residues Gly461-Gly466 ($2|F_{observed}|-|F_{calculated}|$ map, contoured at 1.0 σ). In this particular MX structure, the position of the Gly461-Gly466 loop is stabilized by the crystal contacts with a neighboring protomer (see neighboring protomer electron density within black ellipse).

Figure 4. Human RNA-binding protein Nova-2 KH3 domain: RSCC-based confidence levels and corresponding color scheme for PDB MX structures. (A) Amino acid sequence of PDB: 1DTJ (KH3) serving as X-axis, each residue is denoted by a solid circle in the 2D graph based on RSCC value (Y-axis). For each residue, both its 1-letter amino acid code and circle are color coded by per-residue structure quality using RSCC probability distributions of same residue type at similar resolutions in PDB MX structures: Very well resolved (RSCC ranking >25%, high probability, blue); Well resolved (RSCC ranking 5% to 25%, intermediate probability, cyan); Low confidence (RSCC ranking 1% to 5%, low probability, yellow); Very low confidence (RSCC ranking <1%, lowest probability, orange). (B) Mol* ribbon drawing of PDB: 1DTJ (KH3) using the same quality classification and color scheme, implemented on the RCSB PDB research-focused web portal RCSB.org Mol* 3D structure view under “Quality

Assessment" (Supplementary Figure S3). The outlier is Leu residue number 449 in the UniProt reference sequence, corresponding to residue number 45 in PDB: 1DTJ (KH3).

Tables

Table 1: Comparison of three high-resolution PDB structures of human HbA-α and its corresponding AlphaFoldDB CSM.

PDB MX Structure	Resolution Limit	Oxidation State	Median RSCC	Overall RSCC/pLDDT-CC	Cα atom RMSD (Å)	Non-hydrogen atom RMSD (Å)
2DN1	1.25 Å	oxy	0.97	0.59	0.53	1.45
2DN2	1.25 Å	deoxy	0.94	0.16	0.29	0.71
2DN3	1.25 Å	carbonmonoxy	0.97	0.57	0.54	1.32

STAR Methods

RESOURCE AVAILABILITY

Lead Contact

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Dr. Chenghua Shao (chenghua.shao@rcsb.org).

Materials Availability

This study did not generate new unique reagents.

Data and Code Availability

- PDB structure and validation data utilized in this study are available through FTP at ftp.wwpdb.org and through HTTP at RCSB.org under individual PDB IDs.
- This paper does not report original code.
- Any additional information required to reanalyze the data reported in this paper is available from the Lead Contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

All data are generated from the datasets provided in the KRT.

METHOD DETAILS

Data Collection

All data used for this study were based on the publicly released PDB archive available at ftp.wwpdb.org. Data were extracted from both atomic coordinate files and wwPDB validation reports released before Mar 18, 2022 and then aggregated through data processing. Human protein AlphaFold2 CSMs were downloaded from AlphaFoldDB (Varadi et al., 2022) on Mar 18, 2022.

Sequence Alignments for AlphaFoldDB CSMs and PDB Structures

Amino acid sequences corresponding to human protein AlphaFoldDB CSMs (based on UniProt ID) were used to query and align with protein sequences represented in the PDB archive using the RCSB PDB 1D coordinate server API (https://1d-coordinates.rcsb.org/) (Segura et al., 2020). PDB structure sequences were then used for residue-level alignments to match RSCC values with pLDDT scores for the corresponding residues at identical locations in each polypeptide chain.

Pairing Per-residue RSCC values and pLDDT Scores for Human Protein PDB MX Structures

Only polypeptide chain segments >30 residues in length were included in this step because correlation coefficients calculated between RSCC values and pLDDT scores may not be reliable when matched sequence regions are too short. Sequence pairing between PDB MX structures and AlphaFoldDB CSMs of human proteins also followed additional criteria enumerated below:

1. Alignment lengths between PDB structures and CSMs must be of equal length. In case of gaps or insertions, they were not paired.

2. Pairing must be on the same residue type. In case of mutations, they were not paired.

3. Paired residues must have valid RSCC values, otherwise they were not paired.

4. Pairing was performed on the first instance of a CSM, if multiple CSMs were identified.

5. Pairing was performed on the first instance of the protein in each PDB structure if multiple instances of the protein in PDB were identified.

6. Pairing was only performed on fully-resolved residues (*i.e.*, all atoms present in the PDB structure).

7. Pairing was only performed on residues with occupancy ≥ 0.9

Computation and Software

Data processing, visualization, search, tabulation, and statistical calculation were performed primarily using a combination of Python and R. Calculations were performed on in-house RCSB PDB workstations. To review RSCC-pLDDT correlations in 3D, PDB MX structures were randomly selected. Both PDB structures and corresponding AlphaFold2 models were truncated to partial models with overlapped sequence. Mol* and Pymol (DeLano, 2002) were used for 3D superpositions of the paired PDB MX structures and CSMs for RMSD calculations and for review of residues exhibiting significant differences between per-residue RSCC values and per-residue pLDDT scores.

QUANTIFICATION AND STATISTICAL ANALYSIS

Data Summary

As of Mar 18, 2022, there were 164,404 PDB MX structures in the PDB archive. Among these MX structures, 149,757 had RSCC calculated successfully on 108,286,678 standard residues in wwPDB structure validation reports. The remainder do not have experimental data or do not have standard residues (*e.g.*, carbohydrate only structures), or failed RSCC calculation (RSCC was not calculated when

author-reported R factors could not be verified). Among these standard residues, ~97% are standard amino acids and ~3% are standard nucleotides. Overall statistical characteristics of RSCC values are as follows: mean=0.935, median=0.955, standard deviation=0.065, IQR=0.047, 25th percentile (1st quartile)=0.924, 5th percentile=0.822, 1st percentile=0.666.

AlphaFoldDB human protein CSMs encompass 20,504 UniProt IDs and 23,391 CSMs. Only the 1st isoform of the UniProt ID was used for the alignment to PDB MX structures. 7,502 human protein UniProt IDs could be aligned to 78,088 polymer entities occurring in 53,714 PDB structures. During RSCC/pLDDT pairing, 5,340 UniProt IDs could be aligned to 41,306 PDB structures.

Data Exploration and Visualization

Preliminary data exploration was carried out by running R on the dataset collected above. Tables and figures were generated use R and standard software packages. Probability density distributions were calculated using Gaussian kernel density estimate.

Implementation and Annual Update

The RSCC-based structure quality classification and color scheme has been implemented within 3D Mol* visualization tools provided on the RCSB PDB research-focused web portal RSCB.org. Outlier criteria will be updated annually with newly deposited structures included. Hence, outlier classification of individual residues in PDB MX structures may change slightly from year to year.

Examination of Fisher Z-transformation on RSCC

When (X, Y) data have a bivariate normal distribution and the data pairs (X_i, Y_i) are independent and identically distributed, then their correlation coefficient can be reliably subjected to Fisher Z-transformation and the resulting Z-scores will also be normally distributed.

Correct use of the Fisher Z-transformation requires that the bivariate normality assumption to generate the correlation is met by the original data. When the original data follow a bivariate normal distribution, the Z-score in turn will follow a normal distribution. This condition is not met for experimental-data-derived or calculated electron density distributions used to compute RSCC values. Neither of these distributions are normal, because per-residue RSCC values are only calculated for electron density map regions with relatively strong signal *versus* solvent channels that make up about 50% of the volume of a typical protein crystal.

When the original data do not follow the bivariate normal distribution, the distribution of Z-scores resulting from Fisher Z-transformation of the correlation coefficient can differ significantly from a normal distribution. A simulation study described below substantiates this assertion. Three different (X, Y) data sets with bivariate normal, near normal, or non-normal distributions, respectively, were simulated, then their correlation coefficients were subject to Fisher Z-transformation, and the resulting Z-scores were analyzed using Quantile-Quantile or Q-Q Plots. The Q-Q Plot is a general graphical method for comparing probability distributions before and after Z-transformation. See Supplementary Figure S4(A) for the distribution of correlation coefficients for Data Sets 1, 2, and 3 before and after Fisher Z-transformation.

Data Set 1 was simulated with a bivariate normal distribution. The resulting Z-scores match the normal distribution very well, as evidenced by the diagonal straight-line relationship depicted in Supplementary Figure S4(B). Data Set 2 was simulated with a near-normal bivariate t-distribution (degree of freedom=3). The Data Set 2 Q-Q plot (Supplementary Figure S4(C)) does not recapitulate the diagonal straight line behavior seen for Data Set 1 (Supplementary Figure S4(B)), because the distribution of Z-scores is not normal. Incorrectly applying the 1.96(or 2.6)-sigma rule in this case would result in fewer than 5%(or 1%) of the data items being classified as outliers. Data Set 3 was simulated with a non-normal bivariate t-distribution (degree of freedom=1). The Data Set 3 Q-Q plot (Supplementary Figure S4(D)) does not

recapitulate the diagonal straight line behavior seen for Data Set 1, because the distribution of Z-scores is not normal. In fact, the deviations from a diagonal straight line are even more profound than for Data Set 2 (Supplementary Figure S4(C)), because the bivariate t-distribution at the degree of freedom of 1 is even further away from normal.

In conclusion, whenever the original bivariate data distribution is not normal, use of Fisher Z-transformation will yield Z-scores that are not themselves normally distributed and the 1.96(or 2.6)-sigma rule cannot be reliably used to classify outliers. Supplementary Figure S4(E) illustrates the Q-Q plot for the non-normal distribution of Fisher Z-transformed RSCC values for all 984,838 Leu residues present in 17,123 PDB MX structures with resolution between 2.0 and 2.1 Å. Residue type Leu was chosen because it is the most abundant residue occurring in PDB MX structures of proteins. The resolution range 2.0-2.1 Å was chosen because the median resolution for PDB MX structures is ~2.02 Å. A Shapiro-Wilk normality test (*shapiro.test* in R) was run on this set of Leu RSCC values to provide direct evidence that use of the Fisher Z-transformation of RSCC values is not appropriate. Because *shapiro.test* in R has a limit of 5,000 data points, 100 simple random subsamples of 5000 were selected from 984,838 Leu RSCC values and the calculation was run independently on each subset. For all 100 *shapiro.test* runs, *p* values ranged between 1.3×10^{-29} and 7.5×10^{-19} (average *p* value ~1.1×10^{-20}), documenting that the null hypothesis (*i.e.*, Fisher Z-transformed RSCC follows a normal distribution) was rejected for every run. The distribution of Fisher Z-transformed Leu RSCC values was also examined in detail, revealing that ~3.5% of the data fall below μ(mean)-2σ, and ~1.4% of the data fall below μ-2.6σ. Both percentages are substantially higher than the values of 2.5% and 0.5%, respectively, expected for a normal distribution.

Examination on RSRZ

RSRZ was defined by (Kleywegt et al., 2004) as $Z = (RSR - \langle RSR_{resolution} \rangle)/\sigma(RSR_{resolution})$. $\langle RSR_{resolution} \rangle$ is the average, and $\sigma(RSR_{resolution})$ is the standard deviation of the real-space R-factor or RSR values of

all residues of the same type in the resolution range wherein the structure lies. RSRZ>2 was used by (Kleywegt et al., 2004) to identify outliers. Supplementary Figure S4(F) solid line illustrates the probability density plot of RSR values for all Leu residues (~985,000) present in ~17,000 PDB MX structures with resolutions between 2.0 and 2.1 Å. To calculate RSRZ, this RSR distribution was transformed into the distribution illustrated by dotted line in Supplementary Figure S4(F) based on the mean and σ values, which is significantly different from that of the real data. The Q-Q plot in Supplementary Figure S4(G) also shows that the distribution of RSR values for the Leu 2.0-2.1Å data set is not normal.

A Shapiro-Wilk normality test (*shapiro.test* in R) was run on the same data set to provide additional evidence that the assumption on RSR normality is not appropriate. Because the *shapiro.test* in R has a limit of 5,000 data points, 100 simple random subsamples of 5000 were selected from residue type Leu RSCC values and the calculation was run independently on each subsample. For all 100 *shapiro.test* runs, *p* values ranged between 1.1×10^{-65} and 2.6×10^{-55} (average *p* value ~7.1×10^{-57}), documenting that the null hypothesis (*i.e.*, RSR follows a normal distribution) was rejected for every run. Utilization of the RSRZ>2 criterion for identifying outliers classifies ~3.8% of the observations as outliers, which is considerably higher than the value of 2.5% expected for a normal distribution.

Supplementary Excel Table Titles and Legends

Title: Data S1, Related to STAR Methods

Four data sheets are included in Data S1:

RSCC_by_residue_and_resolution: Per-residue RSCC value distribution for all standard amino acid residues in all PDB MX protein structures by residue type and by resolution. 1%, 5%, and 25% thresholds are the percentile values from the lowest.

RSCC_pLDDT_per_PDB: Per-PDB structure comparison between per-residue RSCC values in PDB MX human protein structures and per-residue pLDDT scores in the corresponding AlphaFoldDB CSMs. Each row corresponds one PDB structure that may have >1 corresponding UniProt IDs.

RSCC_pLDDT_by_residue: Comparison between per-residue RSCC values and pLDDT scores, grouped by residue type.

RSCC_pLDDT_by_resolution: Comparison between per-residue RSCC values and pLDDT scores, grouped by resolution.

References

Abbott, S., Iudin, A., Korir, P. K., Somasundharam, S. & Patwardhan, A. (2018). EMDB Web Resources. Curr Protoc Bioinformatics *61***,** 5.10.1-5.10.12.

Baek, M., Dimaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., Wang, J., Cong, Q., Kinch, L. N., Schaeffer, R. D., et al. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. Science *373***,** 871-876.

Berman, H., Henrick, K. & Nakamura, H. (2003). Announcing the worldwide Protein Data Bank. Nat Struct Biol *10***,** 980.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). The Protein Data Bank. Nucleic Acids Res *28***,** 235-242.

Brändén, C. & Jones, T. (1990). Between objectivity and subjectivity. Nature *343***,** 687-689.

Burley, S. K., Arap, W. & Pasqualini, R. (2021). Predicting Proteome-Scale Protein Structure with Artificial Intelligence. N Engl J Med *385***,** 2191-2194.

Burley, S. K. & Berman, H. M. (2021). Open-access data: A cornerstone for artificial intelligence approaches to protein structure prediction. Structure *29***,** 515-520.

Burley, S. K., Berman, H. M., Christie, C., Duarte, J. M., Feng, Z., Westbrook, J., Young, J. & Zardecki, C. (2018). RCSB Protein Data Bank: Sustaining a living digital data resource that enables breakthroughs in scientific research and biomedical education. Protein Sci *27***,** 316-330.

Burley, S. K., Bhikadiya, C., Bi, C., Bittrich, S., Chen, L., Crichlow, G. V., Duarte, J. M., Dutta, S., Fayazi, M., Feng, Z., et al. (2022). RCSB Protein Data Bank: Celebrating 50 years of the PDB with new tools for understanding and visualizing biological macromolecules in 3D. Protein Sci *31***,** 187-208.

Delano, W. L. 2002. The PyMOL molecular graphics system.

Feng, Z., Westbrook, J. D., Sala, R., Smart, O. S., Bricogne, G., Matsubara, M., Yamada, I., Tsuchiya, S., Aoki-Kinoshita, K. F., Hoch, J. C., et al. (2021). Enhanced validation of small-molecule ligands and carbohydrates in the protein databank. Structure *29***,** 393-400.e1.

Gao, X., Bain, K., Bonanno, J. B., Buchanan, M., Henderson, D., Lorimer, D., Marsh, C., Reynes, J. A., Sauder, J. M., Schwinn, K., et al. (2005). High-throughput limited proteolysis/mass spectrometry for protein domain elucidation. J Struct Funct Genomics *6***,** 129-134.

Goodsell, D. S. & Burley, S. K. (2022). RCSB Protein Data Bank Resources for Structure-facilitated Design of mRNA Vaccines for Existing and Emerging Viral Pathogens. Structure *30***,** 252-262.e4.

Goodsell, D. S., Zardecki, C., Di Costanzo, L., Duarte, J. M., Hudson, B. P., Persikova, I., Segura, J., Shao, C., Voigt, M., Westbrook, J. D., et al. (2020). RCSB Protein Data Bank: Enabling biomedical research and drug discovery. Protein Sci *29***,** 52-65.

Gore, S., Sanz Garcia, E., Hendrickx, P. M. S., Gutmanas, A., Westbrook, J. D., Yang, H., Feng, Z., Baskaran, K., Berrisford, J. M., Hudson, B. P., et al. (2017). Validation of Structures in the Protein Data Bank. Structure *25***,** 1916-1927.

Hawkins, D. M. 1980. *Identification of outliers,* London ; New York, Chapman and Hall.

Hernandez-Valladares, M., Kim, T., Kannan, B., Tung, A., Aguda, A. H., Larsson, M., Cooper, J. A. & Robinson, R. C. (2010). Structural characterization of a capping protein interaction motif defines a family of actin filament regulators. Nat Struct Mol Biol *17***,** 497-503.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Zidek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596***,** 583-589.

Kinjo, A. R., Bekker, G. J., Wako, H., Endo, S., Tsuchiya, Y., Sato, H., Nishi, H., Kinoshita, K., Suzuki, H., Kawabata, T., et al. (2018). New tools and functions in data-out activities at Protein Data Bank Japan (PDBj). Protein Sci *27***,** 95-102.

Kleywegt, G. J., Harris, M. R., Zou, J. Y., Taylor, T. C., Wahlby, A. & Jones, T. A. (2004). The Uppsala Electron-Density Server. Acta Crystallogr D Biol Crystallogr *60*, 2240-9.
Lewis, H. A., Chen, H., Edo, C., Buckanovich, R. J., Yang, Y. Y., Musunuru, K., Zhong, R., Darnell, R. B. & Burley, S. K. (1999). Crystal structures of Nova-1 and Nova-2 K-homology RNA-binding domains. Structure *7*, 191-203.
Mariani, V., Biasini, M., Barbato, A. & Schwede, T. (2013). lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. Bioinformatics *29*, 2722-8.
Mir, S., Alhroub, Y., Anyango, S., Armstrong, D. R., Berrisford, J. M., Clark, A. R., Conroy, M. J., Dana, J. M., Deshpande, M., Gupta, D., et al. (2018). PDBe: towards reusable data delivery infrastructure at protein data bank in Europe. Nucleic Acids Res *46*, D486-D492.
Necci, M., Piovesan, D., Predictors, C., Disprot, C. & Tosatto, S. C. E. (2021). Critical assessment of protein intrinsic disorder prediction. Nat Methods *18*, 472-481.
Park, S. Y., Yokoyama, T., Shibayama, N., Shiro, Y. & Tame, J. R. (2006). 1.25 A resolution crystal structures of human haemoglobin in the oxy, deoxy and carbonmonoxy forms. J Mol Biol *360*, 690-701.
Protein Data Bank (1971). Crystallography: Protein Data Bank. Nature (London), New Biol. *233*, 223-223.
Prugovecki, B., Pulic, I., Toth, M. & Matkovic-Calogovic, D. (2012). High Resolution Structure of the Manganese Derivative of Insulin. Croatica Chemica Acta *85*, 435-439.
Rose, Y., Duarte, J. M., Lowe, R., Segura, J., Bi, C., Bhikadiya, C., Chen, L., Rose, A. S., Bittrich, S., Burley, S. K., et al. (2021). RCSB Protein Data Bank: Architectural Advances Towards Integrated Searching and Efficient Access to Macromolecular Structure Data from the PDB Archive. J Mol Biol *443*, 166704.
Ruff, K. M. & Pappu, R. V. (2021). AlphaFold and Implications for Intrinsically Disordered Proteins. J Mol Biol *433*, 167208.
Schwede, T., Sali, A., Honig, B., Levitt, M., Berman, H. M., Jones, D., Brenner, S. E., Burley, S. K., Das, R., Dokholyan, N. V., et al. (2009). Outcome of a workshop on applications of protein models in biomedical research. Structure *17*, 151-159.
Segura, J., Rose, Y., Westbrook, J., Burley, S. K. & Duarte, J. M. (2020). RCSB Protein Data Bank 1D Tools and Services. Bioinformatics *36*, 5526-5527.
Sehnal, D., Bittrich, S., Deshpande, M., Svobodova, R., Berka, K., Bazgier, V., Velankar, S., Burley, S. K., Koca, J. & Rose, A. S. (2021). Mol* Viewer: modern web app for 3D visualization and analysis of large biomolecular structures. Nucleic Acids Res *49*, W431–W437.
Shao, C., Liu, Z., Yang, H., Wang, S. & Burley, S. K. (2018). Outlier analyses of the Protein Data Bank archive using a probability-density-ranking approach. Sci Data *5*, 180293.
Sun, M., Zhang, H. L., Jiang, M., Chai, Y., Qi, J. X., Gao, G. F. & Tan, S. G. (2021). Structural insights into the cis and trans assembly of human trophoblast cell surface antigen 2. Iscience *24*, 103190.
Teplova, M., Malinina, L., Darnell, J. C., Song, J., Lu, M., Abagyan, R., Musunuru, K., Teplov, A., Burley, S. K., Darnell, R. B., et al. (2011). Protein-RNA and protein-protein recognition by dual KH1/2 domains of the neuronal splicing factor Nova-1. Structure *19*, 930-944.
Thornton, J. M., Laskowski, R. A. & Borkakoti, N. (2021). AlphaFold heralds a data-driven revolution in biology and medicine. Nat Med *27*, 1666-1669.
Tickle, I. J. (2012). Statistical quality indicators for electron-density maps. Acta Crystallogr D Biol Crystallogr *68*, 454-467.
Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Zidek, A., Bridgland, A., Cowie, A., Meyer, C., Laydon, A., et al. (2021). Highly accurate protein structure prediction for the human proteome. Nature *596*, 590-596.
Ulrich, E. L., Akutsu, H., Doreleijers, J. F., Harano, Y., Ioannidis, Y. E., Lin, J., Livny, M., Mading, S., Maziuk, D., Miller, Z., et al. (2008). BioMagResBank. Nucleic Acids Res *36*, D402-408.
Van Der Aalst, W. M. P., Bichler, M. & Heinzl, A. (2017). Responsible Data Science. Business & Information Systems Engineering *59*, 311-313.

Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., et al. (2022). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res *50*, D439-D444.
Westbrook, J. D. & Burley, S. K. (2019). How Structural Biologists and the Protein Data Bank Contributed to Recent FDA New Drug Approvals. Structure *27*, 211-217.
Westbrook, J. D., Soskind, R., Hudson, B. P. & Burley, S. K. (2020). Impact of the Protein Data Bank on antineoplastic approvals. Drug Discov Today *25*, 837-850.
Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J. W., Da Silva Santos, L. B., Bourne, P. E., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Sci Data *3*, 1-9.
Wwpdb Consortium (2019). Protein Data Bank: the single global archive for 3D macromolecular structure data. Nucleic Acids Res *47*, D520-D528.
Young, J. Y., Westbrook, J. D., Feng, Z., Peisach, E., Persikova, I., Sala, R., Sen, S., Berrisford, J. M., Swaminathan, G. J., Oldfield, T. J., et al. (2018). Worldwide Protein Data Bank biocuration supporting open access to high-quality 3D structural biology data. Database *2018*, bay002.
Young, J. Y., Westbrook, J. D., Feng, Z., Sala, R., Peisach, E., Oldfield, T. J., Sen, S., Gutmanas, A., Armstrong, D. R., Berrisford, J. M., et al. (2017). OneDep: Unified wwPDB System for Deposition, Biocuration, and Validation of Macromolecular Structures in the PDB Archive. Structure *25*, 536-545.

Figure 1

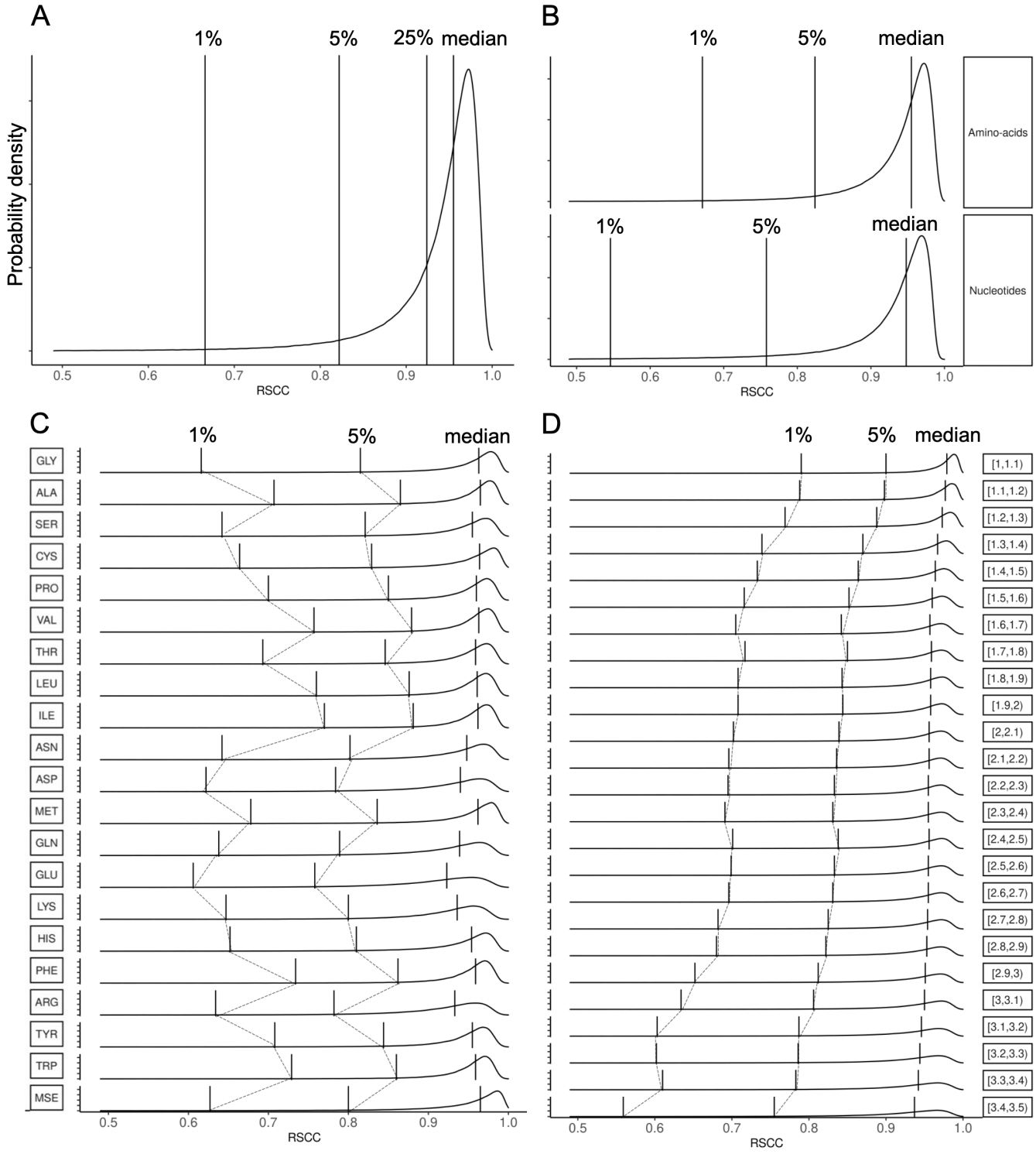

A
1%
5%
25%
median
Probability density
0.5
0.6
0.7
0.8
0.9
1.0
RSCC
B
1%
5%
median
Amino-acids
1%
5%
median
Nucleotides
0.5
0.6
0.7
0.8
0.9
1.0
RSCC
C
1%
5%
median
GLY
ALA
SER
CYS
PRO
VAL
THR
LEU
ILE
ASN
ASP
MET
GLN
GLU
LYS
HIS
PHE
ARG
TYR
TRP
MSE
0.5
0.6
0.7
0.8
0.9
1.0
RSCC
D
1%
5%
median
[1,1.1)
[1.1,1.2)
[1.2,1.3)
[1.3,1.4)
[1.4,1.5)
[1.5,1.6)
[1.6,1.7)
[1.7,1.8)
[1.8,1.9)
[1.9,2)
[2,2.1)
[2.1,2.2)
[2.2,2.3)
[2.3,2.4)
[2.4,2.5)
[2.5,2.6)
[2.6,2.7)
[2.7,2.8)
[2.8,2.9)
[2.9,3)
[3,3.1)
[3.1,3.2)
[3.2,3.3)
[3.3,3.4)
[3.4,3.5)
0.5
0.6
0.7
0.8
0.9
1.0
RSCC

Figure 2

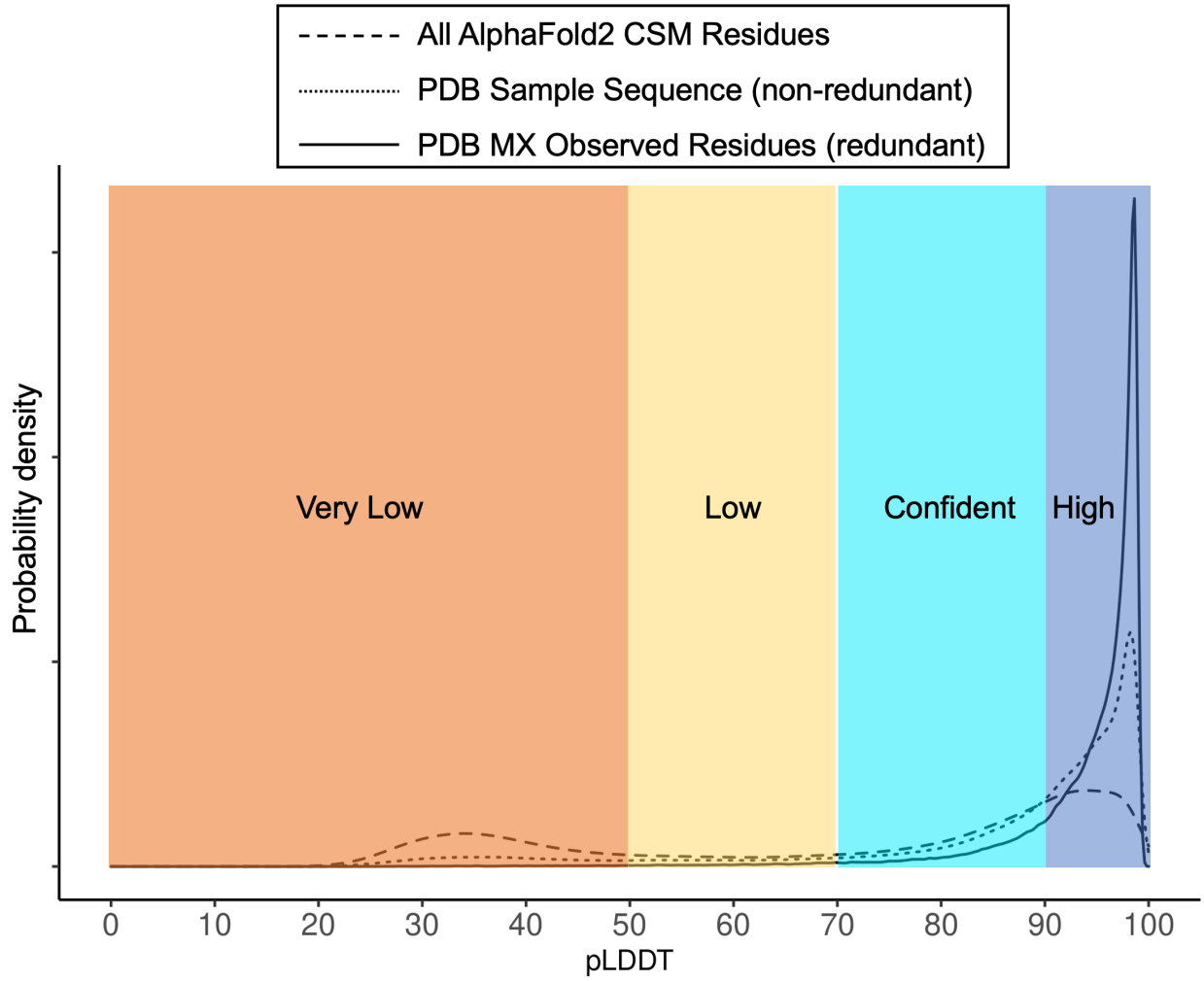

All AlphaFold2 CSM Residues
PDB Sample Sequence (non-redundant)
PDB MX Observed Residues (redundant)
Probability density
Very Low
Low
Confident
High
0
10
20
30
40
50
60
70
80
90
100
pLDDT

Figure 3

Figure 3

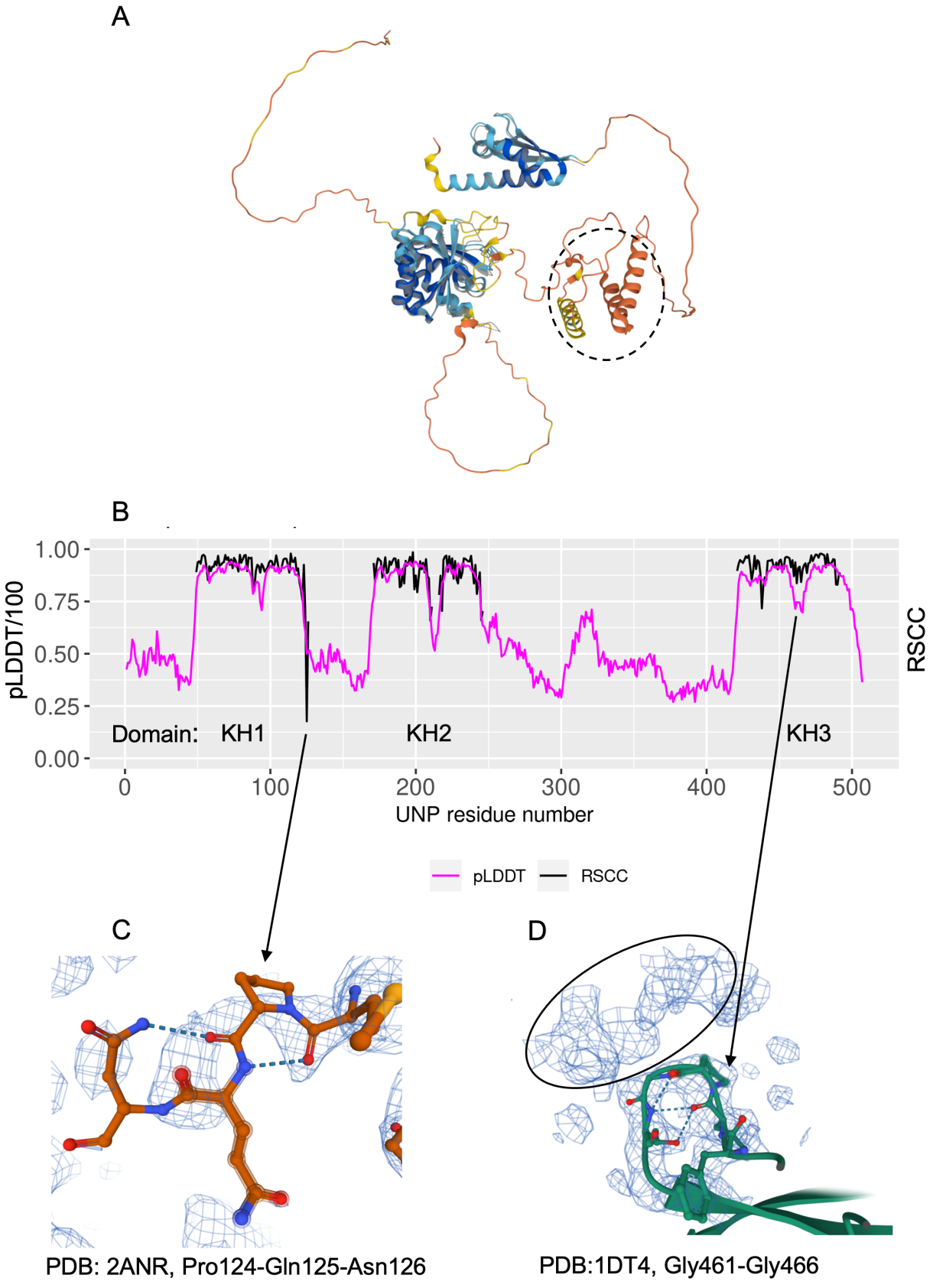
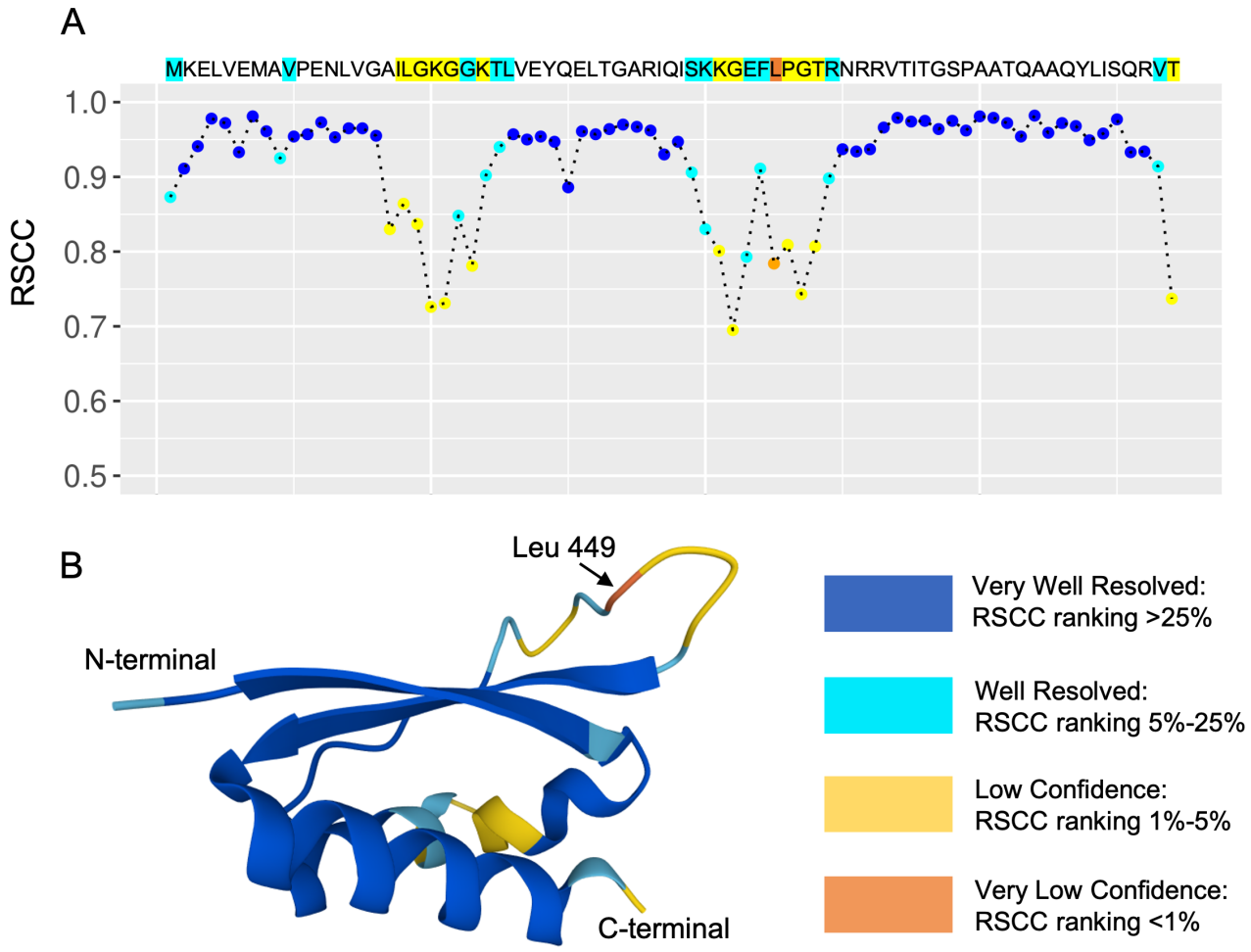

A
B
1.00
0.75
0.50
0.25
0.00
pLDDT/100
RSCC
Domain:
KH1
KH2
KH3
0
100
200
300
400
500
UNP residue number
pLDDT
RSCC
C
D
PDB: 2ANR, Pro124-Gln125-Asn126
PDB:1DT4, Gly461-Gly466

Figure 4


A
MKELVEMAVPENLVGAILGKGGKTLVEYQELTGARIQISKKGEFLPGTRNRRVTITGSPAATQAAQYLISQRVT
RSCC
1.0
0.9
0.8
0.7
0.6
0.5
B
Leu 449
N-terminal
C-terminal
Very Well Resolved:
RSCC ranking >25%
Well Resolved:
RSCC ranking 5%-25%
Low Confidence:
RSCC ranking 1%-5%
Very Low Confidence:
RSCC ranking <1%