Simplified Quality Assessment for Small-molecule Ligands in the Protein Data Bank

Chenghua Shao[1,2*], John D. Westbrook[1,2], Changpeng Lu[2], Charmi Bhikadiya[1], Ezra Peisach[1,2], Jasmine Y. Young[1,2], Jose M. Duarte[5], Robert Lowe[1], Sijian Wang[2,3], Yana Rose[5], Zukang Feng[1,2], and Stephen K. Burley[1,2,4,5,6,7*]

[1]Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB), 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]Department of Statistics and Biostatistics, Rutgers, The State University of New Jersey, New Brunswick, NJ, 08903, USA.

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

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

[6]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, CA 92093, USA.

[7]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) & Dr. Stephen K. Burley (stephen.burley@rcsb.org)

Lead Contact: Dr. Chenghua Shao (chenghua.shao@rcsb.org)

Mailing Address: Institute for Quantitative Biomedicine, Rutgers University, 174 Frelinghuysen Road, Piscataway, NJ 08854, USA.

SUMMARY

More than 70% of the experimentally determined macromolecular structures in the Protein Data Bank (PDB) contain small-molecule ligands. Quality indicators of ~643,000 ligands present in ~106,000 PDB X-ray crystal structures have been analyzed. Ligand quality varies greatly with regard to goodness-of-fit between ligand structure and experimental data, deviations in bond lengths and angles from known chemical structures, and inappropriate interatomic clashes between the ligand and its surroundings. Based on Principal Component Analysis, correlated quality indicators of ligand structure have been aggregated into two largely orthogonal composite indicators measuring goodness-of-fit to experimental data and deviation from ideal chemical structure. Ranking of the composite quality indicators across the PDB archive enabled construction of uniformly distributed composite ranking score. This score is implemented at RCSB.org to compare chemically identical ligands in distinct PDB structures with easy-to-interpret 2D ligand quality plots, allowing PDB users to quickly assess ligand structure quality and select best exemplars.

KEYWORDS

ligand structure quality, composite ranking score, ligand structure, small-molecule ligand, ligand quality indicator, multivariate analysis, Principal Component Analysis, Protein Data Bank, PDB, RCSB PDB

INTRODUCTION

Most pharmaceutical agents are small molecules that bind to target proteins (or nucleic acids) and modify biochemical function of their targets. Experimental studies of three-dimensional (3D) structures of ligands bound to proteins or nucleic acids have proven themselves useful for understanding binding strength and selectivity (Burley, 2021). Drug hunters in academia and the biopharmaceutical industry have come to rely on the open access Protein Data Bank (PDB) (Berman et al., 2000; wwPDB consortium, 2019) as a source of drug target information (Westbrook and Burley, 2019) and starting points for structure-guided drug discovery (Westbrook et al., 2020; Burley, 2021).

The PDB was established in 1971 as the first open-access digital data resource in biology (Protein Data Bank, 1971) with seven X-ray structures of proteins. During its first 50 years of continuous operations, the PDB has grown more than 24,000-fold to become the single global archive of 3D-structures of proteins, nucleic acid, and their complexes with one another and small-molecule ligands determined using macromolecular crystallography (MX), nuclear magnetic resonance (NMR) spectroscopy, electron microscopy (3DEM), and micro-electron diffraction (μED). Open access to expertly biocurated PDB structures enables advances in scientific advances across fundamental biology, biomedicine, bioenergy, and biotechnology/bioengineering (wwPDB consortium, 2019; Goodsell et al., 2020).

The Worldwide Protein Data Bank (wwPDB, wwpdb.org) (Berman et al., 2003; wwPDB consortium, 2019) manages the PDB archive according to the FACT principles of Fairness-Accuracy-Confidentiality-Transparency (van der Aalst et al., 2017) and the FAIR principles of Findable-Accessible-Interoperable-Reusable (Wilkinson et al., 2016). Current wwPDB members include RCSB Protein Data Bank (RCSB PDB), (Berman et al., 2000; Burley et al., 2019)), Protein Data Bank in Europe (PDBe) (Mir et al., 2018), Protein Data Bank Japan (PDBj) (Kinjo et al., 2018), the 3DEM data resource Electron Microscopy Data Bank (EMDB) (Abbott et al., 2018), and the NMR data resource of Biological Magnetic Resonance Bank (BMRB) (Ulrich et al., 2008). The wwPDB global OneDep system for 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 on every inhabited continent. In 2020, The OneDep data deposition and biocuration platform received 15,436 new structures, bringing the total number of PDB structures housed in the archive to over 173,000 by the end of the calendar year.

PDB structure data are described and defined by the PDBx/mmCIF (Westbrook and Fitzgerald, 2009) data dictionary. PDB structures are composed of amino acids or nucleotide building blocks that comprise biopolymers, and associated small molecules such as water molecules, solute molecules, ions, co-factors, enzyme inhibitors, drugs, etc. More than 70% of the PDB macromolecular structures contain small-molecule ligands (excluding water molecules). All small molecule constituents of PDB structures are defined in the wwPDB Chemical Component Dictionary (CCD) that contains detailed chemical description and identification (Westbrook et al., 2015).

Structure quality assessment and validation have been extensively discussed by the wwPDB X-ray Validation Task Force (Read et al., 2011) and researchers from both academia and industry (Adams et al., 2016). wwPDB Validation Reports (Gore et al., 2012; Feng et al., 2021) generated for every PDB structure provide comprehensive quality assessments calculated using community-standard software tools including Mogul (Bruno et al., 2004), MolProbity (Chen et al., 2010), Xtriage (Adams et al., 2010), DCC (Yang et al., 2016), and EDS (Kleywegt et al., 2004). Because the PDB is a core biological data archive serving many millions of users who are not structural biologists, distilled views of the wwPDB validation report are in order. For every PDB structure (identified with a PDB ID, e.g., 4HHB), a slider image appearing both in the validation report and on wwPDB member websites (RCSB.org, PDBj.org, and PDBe.org) provides an easy way for PDB data consumers to gauge the overall quality of each structure. Quality metrics depicted in the slider include agreement with experimental data, inappropriately close contacts between atoms, unlikely polypeptide chain backbone torsion angles, and unlikely sidechain conformations(Shao et al., 2017).

In addition to overall structure quality assessment, the wwPDB validation report summarizes individual ligand quality (Feng et al., 2021), including the local electron density goodness-of-fit indicators of Real space R factor (RSR) (Jones et al., 1991) and real space correlation coefficient (RSCC) (Brändén and Jones, 1990; Tickle, 2012) for X-ray structures calculated using EDS (Kleywegt et al., 2004); the chemical structure quality indicators of Root-Mean-Squared deviation Z-score of all bond lengths (RMSZ-bond-length) and all bond angles (RMSZ-bond-angle) provided by Mogul (Bruno et al., 2004) based on small-molecule structures in the Cambridge Structural Database (CSD) (Groom et al., 2016); and a measure of inappropriate interatomic clashes computed by MolProbity (Chen et al., 2010).

Continuous growth of X-ray co-crystal structures in the PDB over the past fifty years has provided an enormous body of open access data for biomedical research. It has also created considerable challenges for PDB data consumers, who may encounter difficulty when deciding which PDB structure to use and for what purpose. Because every ligand present in a PDB structure is the product of a particular experiment, ligand structure quality varies greatly across the archive (Warren et al., 2012; Tickle, 2012; Deller and Rupp, 2015; Smart et al., 2018). Lower quality ligand structures can mislead researchers, and waste precious time and resources. Ideally, researchers studying a particular ligand want to know, in advance of doing any work, which specific instances of the ligand in which PDB structures are well resolved, depending on (1) how well the atomic coordinates are supported by experimental data, and (2) how well the ligand 3D structure agrees with known chemical and geometric parameters (bond lengths, bond angles, etc.). Herein, we describe construction of a simplified ligand structure quality assessment metrics by (1) aggregating correlated quality indicators into a unidimensional indicator, and (2) establishing a uniformly distributed composite ranking score that simplifies interpretation for all users. The constructed ligand quality score has been implemented at RCSB.org, accessible from the structure summary pages of PDB structures with ligands.

RESULTS

Ligands in the PDB archive

(A) Numbers, types, and sizes of ligands

The number of unique ligands represented in the wwPDB Chemical Component Dictionary (CCD) has grown continuously over the past five decades, particularly in the last seven years during which the CCD doubled in size (Figure 1A). More than 3,000 new ligands were added in 2020 alone. CCD ligands include both constituent monomers within macromolecular polymers and individual small molecules associated with macromolecules in archival structures (Westbrook et al., 2015). Among the constituent monomer ligands, there are ~1,400 standard or modified amino acids from proteins, ~700 distinct nucleotides from nucleic acids. An amino acid or nucleotide may be either a residue within a polymeric sequence of a macromolecule, or an isolated non-polymeric entity associated with a macromolecule. Although all small molecules defined in the CCD are generally called ligands, herein we use the term ligand to refer to small molecules that are not part of a protein or nucleic acid sequence in the PDB structure being studied. These individual small-molecule ligands associated with macromolecules can be roughly classified as "functional" (i.e., ligands likely playing biological/biochemical roles) or "non-functional". Functional ligands include enzyme co-factors, activators, inhibitors, substrates, products, intermediates, and analogs thereof. Non-functional ligands include water molecules and other solvents, salts, and ions, and crystallization and cryoprotection agents (e.g., 2-methyl-2,4-pentanediol). Many functional ligands have identified by the structure depositor as "Ligands of Interest", indicating the focus of research or the subject of investigation. As of the end of 2020, ~6,000 PDB CCD ligands also occur in DrugBank (Wishart et al., 2018).

CCD ligands vary considerably in size (Figure 1B). The number of non-hydrogen atoms range from one (e.g., metal ions) to >100 (e.g., Sulfonated Quinoline-derived Foldamer, CCD ID L0T; Di-PEGylated

Sulfonatocalix[4]arene, CCD ID B4X). The median number of non-hydrogen atoms/ligand across the CCD is ~24. The right-skewed formula weight (FW) distribution for all CCD ligands depicted in blue in Figure 1B has mean, standard deviation, and median values of ~373 Da, ~193 Da, and ~352 Da, respectively, with an Interquartile Range (IQR) of ~216 Da. Compared to the FW distribution for all CCD ligands (blue in Figure 1B), increasing numbers of lower FW ligands have been added to the CCD over the past three years (yellow in Figure 1B). This trend was particularly striking in 2020 (red in Figure 1B), following deposition of >400 X-ray co-crystal structures of fragments (or chemical scaffolds, with a median FW of ~203 Da) bound to SARS-CoV-2 proteins (Newman, 2020; Schuller et al., 2020; Douangamath et al., 2020) that were determined using the Pan-Dataset Density Analysis method (Pearce et al., 2017).

(B) Ligand occurrences in PDB structures

Most CCD ligands occur in more than one PDB structure. For example, functional cyclic monosaccharide ligands N-acetyl-beta-D-glucosamine (CCD ID NAG), beta-D-mannose (CCD ID BMA), and alpha-D-mannose (CCD ID MAN) appear in thousands of PDB structures as components of core glycans covalently attached at protein glycosylation sites (Varki, 2017; Shao et al., 2021). Other functional ligands present in more than 1,000 PDB structures include Heme (CCD ID HEM), Adenosine diphosphate (CCD ID ADP), Adenosine triphosphate (CCD ID ATP), Flavin adenine dinucleotide (CCD ID FAD), Nicotinamide adenine dinucleotide (CCD ID NAD), Nicotinamide adenine dinucleotide phosphate (CCD ID NAP), Guanosine diphosphate (CCD ID GDP), and Flavin mononucleotide (CCD ID FMN). While multiple occurrences of CCD ligands are identical in chemical composition, they are rarely identical in 3D structure owing to conformational flexibility, experimental data quality, occasional errors during data interpretation, etc.

Multiple occurrences of the same CCD ligand can also be found in the same PDB structure. For example, PDB ID 5MR6 (Kugel et al., 2017) contains 24 instances of CCD ID FAD, with one bound to each of the 24 instances of the XiaF protein in the deposited structure. Another, PDB ID 5MCP (Buey et al., 2017),

contains 24 instances of CCD ID ATP bound to eight instances of Inosine-5'-monophosphate dehydrogenase (three/protomer). When multiple instances of the same ligand bound to the same protein in a structure are compared, their 3D structures typically adopt similar conformations.

Ligand structure quality across the PDB archive

(A) Ligand structure quality indicators

Approximately 643,000 individual ligand structures occurring in ~106,000 X-ray crystallographic PDB-ligand complex structures were initially evaluated for quality by examining distributions of RSR, RSCC, RMSZ-bond-length, RMSZ-bond-angle, inappropriately close interatomic distances or clash-per-atom between ligand and nearby molecules, and chirality outliers (see Methods). Ion ligands without bond length or angle parameters were not included in the analysis. Chirality outliers were considered separately because they occur in only 2% of ligand structures and reflect major errors in experimental data interpretation, which may require structure re-determination. Detailed quantitative analyses were performed with five primary quality indicators, including RSR, RSCC, RMSZ-bond-length, RMSZ-bond-angle, and clash-per-atom. Characteristics of the distributions for each of the five quality indicators are summarized in Table 1. Since ~49% ligands in PDB structures exhibit no steric clashes, the median value (at 50% percentile) for clash-per-atom is near zero.

While analyzing the primary quality indicators for identical CCD ligands occurring multiple times in a given PDB structure, we observed that their quality characteristics do not differ significantly. For example, among the 24 instances of CCD ID FAD in PDB ID 5MR6, computed standard deviations for RSR, RSCC, RMSZ-bond-length, and RMSZ-bond-angle were minuscule (0.008, 0.003, 0.05, and 0.05, respectively). Structure quality indicator variation for identical CCD ligands in a single PDB structure is typically five- to 20-fold lower than structure quality indicator variation for identical CCD ligands in different PDB structures

(Supplementary Table S1). We, therefore, averaged primary quality indicators for identical CCD ligands occurring within the same PDB structure to prevent undue bias in our analyses.

After "within-structure" averaging, ~197,000 ligand occurrences in PDB X-ray structures remained, each identified with unique combinations of PDB ID and CCD ID. This averaging process gave the same weight to each PDB structure for the subsequent investigation on the quality distribution of any unique CCD ligand, so that many instances in one PDB structure do not eclipse fewer instances in another structure. The five primary structural quality indicators were analyzed for their association with other characteristics specific to the ligand (e.g., FW) or the PDB complex structure (e.g., high resolution limit). Cases for which RSR and RSCC quality indicators are not available because experimental structure factors were not deposited, were omitted from subsequent analyses. Also excluded were cases with (1) incomplete ligands structures with missing non-hydrogen atomic coordinates, (2) unknown ligands, and (3) ligands with occupancy <0.9.

In total, ~159,000 ligands in PDB structures with complete data for all five quality indicators were used to examine relationships between indicators. Pairwise Pearson correlation coefficients were calculated (Figure 2). The two experimental data agreement indicators (RSR, RSCC) are negatively correlated (correlation coefficient~-0.67). The two chemical agreement indicators (RMSZ-bond-length, RMSZ-bond-angle) are positively correlated (correlation coefficient~0.63). No other quality indicator pairs show strong correlation.

(B) Principal Component Analysis

Initial Principal Component Analysis (PCA) was carried out on all five quality indicators to explore interrelationships and rigorously assess whether ligand quality measures can be reduced in dimensionality. Table 2 documents that the three most significant principal components collectively explain ~86% of total variance. Variances explained by the first three principal components are

comparable to each other, at 39%, 29%, and 18%, respectively. Hence, the 1st principal component alone cannot sufficiently represent all five input quality measures. The fractional of contributions from each original quality indicator indicates that the first principal component (PC1-overall) is dominated in approximately equal proportion by RSR and RSCC. The second principal component (PC2-overall) is dominated in approximately equal proportion RMSZ-bond-length and RMSZ-bond-angle. The third principal component (PC3-overall) is dominated by clash-per-atom alone. Thus, overall ligand structure quality in the PDB can be represented by the principal components of roughly three groups of: (1) RSR and RSCC, (2) RMSZ-bond-length and RMSZ-bond-angle, and (3) clash-per-atom. Mutual orthogonality between PC1, PC2, and PC3 is consistent with the Pearson correlation coefficient analyses (Figure 2) that demonstrated the relative independence between the three groups, allowing us to perform subsequent analyses within each group.

A secondary PCA was performed on the group of RSR and RSCC goodness-of-fit quality metrics (Table 2). The first principal component of this group, designated as PC1-fitting, accounts for ~84% of total variance (Table 2). Therefore, PC1-fitting can be used as the one-dimensional (1D) composite indicator to measure the goodness-of-fit between a ligand structure and corresponding local electron density. Another secondary PCA was computed on the group of RMSZ-bond-length and RMSZ-bond-angle (Table 2), yielding PC1-geometry that accounts for ~82% of total variance (Table 2). Similarly, PC1-geometry can be used as the 1D composite indicator to assess agreement between the ligand structure and known chemical parameters. Since PC1-fitting and PC1-geometry are relatively independent of each other (with a correlation coefficient of -0.138 for all PDB ligands), ligand structure quality should be separately assessed by PC1-fitting for agreement with experimental data and PC1-geometry for geometrical accuracy.

The five primary ligand quality indicators in Table 1 and 2 are not normally distributed. The relationship between any of the two indicators is not strictly linear. The impact of the data distribution and non-linear relationship have been assessed in the STAR methods and supplementary data (Supplementary Table

S2 and Figure S1). When a ligand is an outlier with exceptional value for any of the primary quality indicators, its quality may not be well represented by the composite quality indicators and should be separately marked to alert PDB data users (see STAR methods).

(C) Developing composite ranking score of ligand quality

Our goal was to allow any PDB data consumer, independent of structural biology expertise, to ask, "How does the quality of this ligand structure compare with other instances in the PDB archive?" For this purpose, composite ranking scores were developed to use composite indicators of PC1-fitting and PC1-geometry to compare and rank ligand structures by quality. For either composite indicator, the composite ranking score of a ligand structure is defined as the percentage of other PDB ligand structures with inferior quality versus the particular ligand, which is consistent with the ranking defined for overall structure quality in the wwPDB validation report (Gore et al., 2017; Feng et al., 2021). Since ranking is uniformly distributed, composite ranking score carries the simplest interpretation: 0% for the worst, 100% for the best, and 50% for median quality.

Figure 3 illustrates the relationship between local electron density map features and composite ranking scores of PC1-fitting quality indicator in representative PDB structures at a similar high resolution limit of around 1.5 Å. CCD ID FAD instance A-501 in PDB ID 5NAK (Hutchinson et al., 2017) ('A' indicates that they are associated with protein polymer Chain A, and the number 501 is the ligand instance identifier) depicted at the top of Figure 3 has a composite ranking score of ~99% revealing superior quality in terms of goodness-of-fit between the ligand structure and experimental data (electron density map). Proceeding vertically downwards in Figure 3, reveals examples of CCD ID FAD ligand structures with progressively inferior composite ranking scores and lower quality local electron density map features for PDB IDs 4U7H (Leung and Shilton, 2015) and 2QWX (Calamini et al., 2008). Depicted at the bottom of Figure 3 is FAD B-1202 in PDB ID 2CZ8, which has a very low PC1-fitting composite ranking score of ~2%. Visual inspection of the electron density map revealed reasonable signal for the flavin group and minimal signal

for the adenine diphosphate moiety. Thus, PDB data consumers can use the PC1-fitting composite ranking score to identify readily ligand structures that are well supported by experimental data, without the need for time-consuming review of electron density maps.

Constructing useful rankings with PC1-geometry proved to be more challenging, because RMSZ-bond-length and RMSZ-bond-angle are positively correlated with the FW. Larger ligands in the PDB tend to have greater RMSZ-bond length and RMSZ-bond-angle values, which most likely reflects the tradeoff made during X-ray structure refinement between optimizing ligand chemical geometry versus fit of the atomic coordinates to the experimental electron density map. To account the impact of ligand size, we investigated limiting PC1-geometry ranking to instances of identical compounds. Using this metric, FAD A-501 in PDB ID 5NAK has a ranking of ~59% for PC1-geometry when compared to all instances of CCD ID FAD, which is significantly different from its composite ranking score of ~19% when all PDB ligands are considered. FAD B-1202 in PDB ID 2CZ8 has a ranking of ~40% for PC1-geometry when compared to all instances of CCD ID FAD, versus ~15% when ranked against all ligands.

Another contributor to ligand structure quality is the identity of the protein target to which the ligand is bound. Mode of ligand binding can differ substantially depending on the macromolecular target of the ligand. We, therefore, established an additional ranking system limited to similar protein structures clustered at 95% sequence identity (computed using MMseqs2 (Steinegger and Soding, 2017)). For the PDB data consumer, this ranking system enables selection of the PDB structure of a given protein containing the desired ligand with the highest ligand structure quality. For example, the PDB contains two structures of CCD ID FAD bound to the XiaF protein (i.e., PDB IDs 5MR6 and 5LVW (Kugel et al., 2017)). PC1-fitting and PC1-geometry composite ranking scores for CCD ID FAD in PDB ID 5MR6 are significantly higher than those in PDB ID 5LVW. Ranking based on the identity of the protein to which the ligand is bound will ensure that users can select PDB ID 5MR6 as the best exemplar of CCD ID FAD bound to the XiaF protein.

Non-linearity between primary ligand quality indicators has been analyzed in the STAR methods and shown impact on the absolute value of the principal components, especially on PC1-geometry. Using ranking of the absolute value as the PDB ligand quality composite score provides the needed robustness to counter the non-linearity. However, the PDB ligand quality composite ranking score is limited for use as the relative comparison measure between the majority of the ligands within the PDB archive and should neither be used as an absolute measure nor beyond the PDB archive.

Review of the clash-per-atom quality indicator for each structure in the PDB revealed that nearly half of ligand structures in the PDB exhibit no interatomic clashes (Table 1). Consequently, percentile ranking for clash-per-atom is not informative because the value can never exceed 51%. Ranking for clash-per-atom was not pursued further.

Managing cases of incomplete ligand structures

Approximately 6% of X-ray structures in the PDB have ligand structures with missing non-hydrogen atomic coordinates. In most cases, these occurrences reflect paucity of signal corresponding to parts of the ligand in the experimental electron density map. RSR and RSCC values for partial ligand structures require adjustment to permit valid comparisons with all-atoms-included ligand structures when generating the comparative composite ranking scores. Many partial ligand structures represent lipid or detergent components (with long flexible carbon chains) or fragments of polyethylene glycol (PEG). PEG fragments are considered exceptions because they are intrinsically inhomogeneous. We analyzed three non-PEG CCD ligands with the most incomplete atomic structures and substantial numbers of complete structures (Table 3), including 1-oleoyl-R-glycerol (CCD ID oLC), (hydroxyethyloxy)tri(ethyloxy)octane (CCD ID C8E), and oleic acid (CCD ID oLA).

For each of these three CCD ligands, differences of RSR and RSCC (ΔRSR and ΔRSCC, respectively) were calculated between values of partial atomic structures and average values of full atomic structures

at comparable resolution limits. ΔRSR and ΔRSCC were then analyzed as a function of incompleteness (i.e., the fraction of missing non-hydrogen atoms (Supplementary Figure S2)). All 3691 incomplete instances for the three CCD ligands were pooled together to estimate ΔRSR and ΔRSCC for partial atomic structures to "adjust" for missing atoms. The adjustment was applied to ligands missing coordinates for more than one non-hydrogen atom. We justify applying this correction on the grounds that missing atoms reflect absence of signal in the electron density map. Adjusted RSR and RSCC were subsequently used to compute the adjusted PC1-fitting composite ranking scores, so that the ligand structure quality of both partial and full atomic structures can be compared directly. For example, the ligand atomic model of CCD ID FAD in PDB ID 5LVW is incomplete, because the Adenosine monophosphate moiety is missing. The PC1-fitting composite ranking score was ~10% without any adjustment, which falls to ~6% with adjustment for the missing portion.

Quality measures for ligand structures with only one non-hydrogen atom missing were not "adjusted", because in most cases the single missing atom is part of the leaving group that departs in the formation of a covalent bond with an adjacent compound. For example, >47,000 instances of CCD ID NAG have missing atoms, but almost all of them are only missing the reducing-end hemiacetal hydroxyl that departs on formation of a glycosidic bond with either a glycosylation-site amino acid sidechain or another monosaccharide (Shao et al., 2021).

RCSB PDB ligand plot for structure quality review and comparison

Based on our ligand structure quality analyses, we have designed and implemented at RCSBb.org a new graphical presentation schema for ligand structure quality to better meet the needs of all PDB users regardless of their structural biology expertise. Because the overwhelming majority of PDB data consumers are not structural biologists, the new schema was designed to allow any user to quickly review ligand structure quality and unambiguously select the ligand (or ligands) in a particular PDB structure that will best serve their research or teaching needs. Figure 4 exemplifies the two-dimensional (2D) PDB

ligand quality plot with multiple instances of Cholesterol Hemisuccinate (CCD ID Y01) bound to the G-protein coupled receptor (GPCR) Muscarinic acetylcholine receptor M1 occurring in PDB ID 6WJC (Maeda et al., 2020). PDB ID 6WJC contains four instances of CCD ID Y01, designated as A-502, A-503, A-504, and A-505. Three interactive two-dimensional (2D) ligand quality plots (Figure 4A, 4B, and 4C) enable at-a-glance graphical review of within-structure and between-structure ligand quality comparisons. An accompanying tabular report (Figure 4D) provides additional quantitative information. Figures 4A and 4D reveal that one of the four instances (Y01 A-502) has a significantly higher quality PC1-fitting composite ranking score when compared to the other three (~31% versus ~0-4%), while all four instances have comparable PC1-geometry composite ranking scores (~30%). Figure 4B allows the user compare Y01 A-502 in PDB ID 6WJC with the best quality alternative structure of CCD ID Y01 bound to the same protein. The 2D plot shows that PDB ID 5CXV (Thal et al., 2016) contains a higher quality structure for CCD ID Y01 (PC1-fitting ~37% versus ~31%; PC1-geometry ~43% versus ~36%). For users interested in analyzing CCD ID Y01 in different contexts (i.e., bound to other proteins), Figure 4C compares the quality of CCD ID Y01 occurring in PDB ID 6WJC with the top five best-fitted quality structures of the same ligand bound to any protein. PDB ID 2Y00 (Warne et al., 2011) contains the best-fitted example of CCD ID Y01 across the entire PDB archive (PC1-fitting ~51%). With the ligand quality plot, users may quickly select ligand structures from Figure 4A-C, and then consult the tabular report in Figure 4D to review the ligand quality indicators and other details such as chirality errors, intermolecular clashes, and ligand atomic coordinate completeness. The most extreme outliers (the worst 1%) for each of the original ligand quality indicators are highlighted in the tabular report in red font. PDB ligand quality plot and the composite ranking scores at RCSB.org are made interactively. Clicking on the best-fitted ligand instance symbol on the 2D plot or the identifier in the tabular report brings up a 3D display of electron density focused on the ligand structure viewed by Mol* web-native molecular graphics system (Sehnal et al., 2021) (Supplementary Figure S3). Ligand quality data are also available via the RCSB PDB data APIs supporting programmatic access and comprehensive search (Rose et al., 2021).

Ligand of Interest and likely functional ligands

To showcase the most interesting ligands in a PDB structure, the RCSB.org structure summary page of the entry highlights the ligand fitting quality to experimental electron density by 1D slider (i.e., the horizontal axis of the 2D plot) on the following ligands:

1. Ligand of Interest or LOI (i.e., focus of research or subject of investigation) as designated by the PDB data depositor(s), independent of FW.
2. Likely functional ligands with FW>150 Da that were not in an exclusion list of likely non-functional ligands (e.g., solvent molecules, ions, salts, buffers, crystallization precipitants, common cryo-protectants, and reducing agents), if LOI designation was not provided by data depositor(s).

Because the LOI designation was not introduced into the wwPDB OneDep system (Young et al., 2017) until 2017, a FW cutoff was required for PDB structures deposited prior to that date. The FW>150 cutoff was chosen because it corresponds to ~95% of all author-designated LOIs currently present in the PDB archive (Supplementary Figure S4). The non-functional exclusion list was assembled by expert RCSB PDB biocurators and is reviewed periodically.

Assessing quality for singleton ligands in the PDB

Examples provided above utilized CCD ligands that occur more than once in the PDB archive. For singleton CCD ligands such as the potent opioid Fentanyl (CCD ID 7V7) present only in PDB ID 5TZo (Bick et al., 2017), the best-fitted CCD ID 7V7 instance has the composite ranking scores of PC1-fitting and PC1-geometry at ~82% and ~51%, respectively, when all ligand structures in the PDB are used as references. But we would also like to know how good the quality of singleton ligand is versus “comparable” ligands in the PDB. To assess similarity of ligand structure quality across the PDB archive, we analyzed other characteristics. Among them, ligand size and structure resolution limit have the greatest impacts on ligand quality indicators. With the benefit of further analyses, we defined ligands to

be “comparable” if (1) FW falls within 15 Da of the singleton ligand, and (2) structure resolution limit falls within 0.2 Å of the resolution limit of the structure that contains the singleton ligand. Using this approach, the best-fitted CCD ID 7V7 instance in PDB ID 5TZo has a relative PC1-fitting and PC1-geometry at ~54% and ~89%, respectively, when only comparable ligand structures in the PDB are used as references. These results underscore the importance of choosing a proper PDB subset as reference when assessing ligand structure quality, versus relying on rankings based on the entire PDB archive. (N.B.: We are using the term similarity as it pertains to ligand FW and structure resolution limit, not chemical structure similarity.)

DISCUSSION

RCSB.org now offers easy-to-understand graphical guidance on PDB ligand structure quality based on the data analyses and processes described herein. The primary motivation for this work was to enable all PDB users, regardless of their structural biology expertise, to assess ligand structure quality quickly and readily and example(s) best suited to their research, experiment design, or teaching needs. By looking broadly across all ligands represented in the PDB and various quality metrics, we were able to use Principal Component Analyses to develop composite quality indicators with reduced dimensionality. The composite measures enable facile comparisons among instances of the same ligands occurring within the same PDB structure or in different structures. Additional comparisons can be made among ligands bound to the same or distinct biological macromolecules. Reducing available quality metrics to just two dimensions PC1-fitting and PC1-geometry, allowed us to develop a 2D graphical presentation for PDB data consumers to understand ligand structure quality at a glance. Two orthogonal axes display two percentile sliders from worst to best for ligand structure quality reflecting (1) how well the atomic coordinates describing the ligand structure are supported by experimental data, and (2) how well the

ligand structure conforms to known chemical geometry. Other important quality metrics, including clashes and chirality errors, are provided in a brief tabular report.

With the RCSB PDB ligand quality composite ranking scores we can access the impact on ligand quality by factors such as high resolution limit and deposition date (Supplementary Figure S5). Approximately, 99% of the PDB X-ray structures have high resolution limits between 1.0 Å and 3.5 Å. Within this range, high resolution limit significantly impacts the goodness-of-fit between a ligand atomic structure and experimental data (i.e., on average, the higher the resolution of the X-ray diffraction data the better the PC1-fitting ranking, demonstrated in Supplementary Figure S5(A)) but does not have strong effect on ligand geometry quality (i.e., PC1-geometry). Higher resolution does not, however, guarantee a more reliable ligand structure. The examples depicted in Figure 3 clearly show that ligand structure quality is not uniformly superior for PDB structures of higher resolution. Comparison of PC1-fitting values and experimental electron density for CCD ID FAD in PDB IDs 5NAK, 2QWX, and 2CZ8 (all obtained at 1.5 Å resolution) reveals considerable variation in quality. Therefore, resolution alone is not sufficient to accurately assess the ligand fitting quality, and we suggest PDB users to use the composite fitting score as the primary measure. Deposition date also appears to have some impact on ligand structure quality as well (Supplementary Figure S5(B)). PC1-geometry quality of ligand structures has improved modestly since 1999, which probably reflects the impact of improved X-ray structure refinement software. Average PC1-fitting quality has not changed significantly as a function of the deposition date. Both trends can be seen for selected individual chemical compounds as a function of time (Supplementary Figure S5(C)).

Our analyses of ligand structure quality were limited to PDB structures determined using only X-ray crystallography. For ligand structures determined using 3DEM, the PC1-geometry quality assessment can be used without modification. Assessing goodness-of-fit to experimental data will require construction of alternative composite scoring systems. Following the approach used in this work, both Q score that measures atom resolvability in cryo-EM maps (Pintilie et al., 2020) and atom inclusion within electric Coulomb potential maps criteria (Lawson et al., 2021) could be used to construct composite ranking

scores with which to assess the goodness-of-fit between ligand structures and 3DEM experimental data. One potential advantage of the Q score is that it can also be computed with X-ray experimental data. A composite scoring system based on Q scores could, at least in principle, be used to assess both X-ray and 3DEM derived ligand structures in the PDB together and enable direct quantitative comparisons of the quality of protein-ligand complex structures coming from the two methods. In addition, various 3DEM map-model correlation coefficients (Afonine et al., 2018) may be calculated on local ligand regions and be subsequently included as additional measures in assessing goodness-of-fit to experimental data.

In analyzing how well a given ligand fits in an experimentally determined electron density map, the wwPDB OneDep system for deposition, validation, and biocuration of incoming PDB structures assumes that the depositors have correctly chosen the chemical identity of ligands represented in their structures. While the overwhelming majority of structures have correct ligand identities, a very small proportion have been found to have issues (Cereto-Massague et al., 2013; Touw et al., 2016; Brzezinski et al., 2021). During the global OneDep process of deposition, validation, and biocuartion wwPDB Biocurators make best efforts to identify errors and inconsistencies, inform depositors, and explain recommended steps required for correction. However, the PDB is an archival resource and wwPDB biocurators are not empowered to require that depositors make the recommended corrections before the structure deposition is finalized and released publicly. The responsibility for the accuracy of PDB structures rightly rests with the depositors, just as it does for authors of publications in scientific journals. For X-ray structures visualized with our research-focused RCSB.org web portal, the Mol* web-native molecular graphics system (Sehnal et al., 2021) is used to view ligands in 3D with accompanying display of surrounding electron density for help with verification of ligand identity (provided the experimental data is of sufficiently good quality). If the electron density itself does not provide reliable information as to the identity of the ligand, and the ligand identity is a concern, PDB users are urged to consult the associated scientific publication, the structure depositors, and other tools (e.g., CheckMyBlob (Brzezinski et al., 2021)) for further insights into the ligand identity.

ACKNOWLEDGEMENTS

RCSB PDB is funded by the National Science Foundation (DBI-1832184, P.I.: S.K. Burley), the US Department of Energy (DE-SC0019749, P.I.: S.K. Burley), and the National Cancer Institute, National Institute of Allergy and Infectious Diseases, and National Institute of General Medical Sciences of the National Institutes of Health under grant R01GM133198 (P.I.: S.K. Burley). We thank Christine Zardecki for the help in editing the manuscript.

AUTHOR CONTRIBUTIONS

Conceptualization, S.K.B., C.S., and J.D.W.;

Methodology, C.S., J.D.W., and S.K.B.;

Software, C.S., J.D.W., Z.F., E.P., C.B., Y.V., J.D., and R.L.;

Validation, C.S. and J.D.W.;

Formal Analysis, C.S., S.W., and C.L.;

Investigation, C.S. and J.D.W.;

Resources, E.P. and J.D.;

Data Curation, C.S., J.Y.Y., and J.D.W.;

Writing, C.S., S.K.B., J.D.W., and J.Y.Y;

Visualization, C.S., J.D.W., C.B., and R.L.;

Supervision, S.K.B., J.Y.Y., and J.D.W.;

Project Administration, S.K.B. and J.Y.Y.;

Funding Acquisition, S.K.B.;

DECLARATION OF INTERESTS

The authors declare no competing interests

FIGURE LEGENDS

Figure 1. CCD growth and ligand sizes. (A) Growth of CCD versus time (height of a bar for the cumulative number of ligands in the CCD, and dark blue for new ligands added to the public archive during the calendar year). (B) Formula weight (FW) distribution of CCD ligands (all ligands: blue; new ligands in 2018-2020: yellow; new in 2020: red).

Figure 2. Pairwise correlations among ligand structure quality indicators. Each filled circle represents the relative pairwise correlation. Circle diameters correspond to absolute values of the correlation coefficients (positive-blue; negative-red; no correlation-blank).

Figure 3. Composite ranking scores for PC1-fitting to electron density maps. CCD ID FAD structures determined at 1.5 Å resolution, from best (top) to worst (bottom) along a colored vertical bar (blue: superior; red: inferior). All five figures show ligand omit maps (blue wireframe, generated based on experimental data and contoured at 1.0 σ) superimposed on the ligand models in stick representation colored by elements (gray: carbon; red: oxygen; blue: nitrogen; orange: phosphorus). PC1-fitting composite ranking scores are provided in parentheses with PDB ID, Chain ID, and instance number. N.B.: Two instances from PDB ID 2QWX were selected: residue #232 of chain A (2nd from the top) and residue #232 of chain B (3rd from the top).

Figure 4. PDB ligand quality plot and tabular report. Each 2D graph depicted in (A), (B), and (C) has color coded ranking scales from worst (0%, red) to best (100%, blue) for PC1-fitting (horizontal axis) and PC1-geometry (vertical axis). Each symbol represents a ligand instance of CCD ID Y01 with its horizontal location marked by PC1-fitting and its height by that of PC1-geometry. The diamond symbol in each plot indicates the best-fitted instance in the current PDB ID 6WJC, corresponding to the top row of the tabular report (D) that details ligand quality metrics. Other rows of the report (D) highlighted in green, yellow, and gray background correspond to the circle symbols in (A), (B), and (C), respectively, as within-structure

and between-structure comparisons indicated by the plot titles. In the table (D), the worst 1% outliers of each ligand quality indicators are highlighted in red font, and the identifiers in the first column are hyperlinks to Mol* 3D electron density view focused on the ligands. The plots and table have been implemented at https://www.rcsb.org/ligand-validation/6WJC/Y01.

TABLES

Table 1. Distribution of the primary ligand structure quality indicators for all ligands in PDB X-ray crystallographic structures.

	Mean	Standard Deviation	Median	IQR	Nature of Distribution
RSR	0.18	0.09	0.16	0.10	0≤RSR≤1, Right-skewed, lower value indicating better quality
RSCC	0.89	0.09	0.92	0.11	0≤RSCC≤1, Left-skewed, higher value indicating better quality
RMSZ-bond-length	1.12	1.03	0.81	1.06	0≤ RMSZ-bond-length, Right-skewed, lower value indicating better quality
RMSZ-bond-angle	1.21	0.98	1.03	1.21	0≤ RMSZ-bond-angle, Right-skewed, lower value indicating better quality
clash-per-atom	0.07	0.15	0.01	0.08	0≤clash-per-atom, Right-skewed, lower value indicating better quality

Table 2. Principal component analyses of ligand structure quality indicators

Principal Component Analysis for		All 5 indicators			RSR and RSCC	RMSZ-bond-length and RMSZ-bond angle
Principal Component (PC)		PC1-overall	PC2-overall	PC3-overall	PC1-fitting	PC1-geometry
% of total variance explained		39%	29%	18%	84%	82%
Fractional contributions of each indicator	RSR	0.53	0.42	0.22	0.71	n/a
	RSCC	-0.55	-0.37	-0.25	-0.71	n/a
	RMSZ-bond-length	-0.43	0.56	0	n/a	0.71
	RMSZ-bond-angle	-0.43	0.56	0	n/a	0.71
	clash-per-atom	0.23	0.25	-0.94	n/a	n/a

Table 3. CCD ligands with significant numbers of both complete and incomplete atomic structures.

CCD ID	Incomplete Instances	Complete Instances	PDB IDs
oLC	1698	626	256
C8E	999	371	105
oLA	994	286	157

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 use or generate any physical material.

Data and code availability

- The PDB structure and validation data indicated in this study are available through FTP at ftp.wwpdb.org and through HTTP at RCSB.org under the individual PDB IDs. The aggregated PDB ligand quality data and the data analyses results are available at Zenodo (zenodo.org) under DOI: 10.5281/zenodo.5525191 as well as GitHub (https://github.com/rcsb/PDB_ligand_quality_composite_score).
- 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 at ftp.wwpdb.org. Data were extracted from various data sources of the PDB archive as described in the Supplementary Table S3 and were then aggregated through data process.

Data process on atomic clashes

Atomic clashes on ligands were re-processed to exclude the clashes between atoms within the same ligand, and then were scaled against the number of non-hydrogen atoms to generate clash-per-atom that is defined as the clashes per atom between a ligand and its surrounding components. The number of observed non-hydrogen atoms were subsequently used to calculate ligand structure atomic coordinate completeness as described in the Results.

Remove within-structure redundancy for multivariate analysis

As described in the text, multiple instances of the same ligand within the same PDB structure were averaged to generate the dataset with unique PDB-CCD combination. After removing the within-structure redundancy, 642,625 ligands instances were reduced to 105,548 unique PDB-CCD representations for multivariate data analyses such as correlation study and PCA. However, when any specific PDB structure is investigated, each ligand structure instance was still studied individually.

Data process on incomplete ligand structures

All instances of oLC, oLA, and C8E in the PDB archive were included. For each instance, the RSR-difference is calculated between the RSR of the incomplete structures and the RSR of the complete structures in the reference set. The reference set is the structures of the same ligand in structures of similar resolution (±0.1Å from the resolution of the queried structure). If such resolution bin gives less than 10 complete ligand structure instances, the bin is expanded by another ±0.1Å until there are 10+ complete ligand structures found. The RSR-difference is then used to run linear regression against the ligand incompleteness (missing fraction), without intercept.

By choosing the reference in similar resolution, confounding from resolution was reduced on studying RSR/RSCC relationship with ligand incompleteness. For example, although RSR has a strong correlation of 0.43 with resolution, the RSR-difference constructed above has only a weak correlation of -0.05 with the resolution, much smaller than the correlation between RSR-difference and ligand incompleteness. Therefore, resolution was not necessary to be used as a variable for the linear regression, allowing a simpler linear relationship between RSR-difference and incompleteness. During the data process, resolution factor and its interaction term with incompleteness were also tried in the regression models, but the result showed that the resolution does not impact RSR-difference significantly, and the regression models do not have much improved regression fitting compared to the simpler one without resolution term.

Sequence cluster generation

Protein sequence clusters of the entire PDB achieve were generated by MMseqs2 (Steinegger and Soding, 2017) at different sequence identity. 95% sequence identity was used so that smaller number of mutations of the same protein can be grouped together.

LOI criteria for early deposited structures without author designation

Several known LOI groups were studied to determine the rough FW threshold (Supplementary Figure S4). Among them, 100+ drug-candidate ligands recently deposited into the PDB for the Drug Design Data Resource (D3R) project (Gathiaka et al., 2016; Gaieb et al., 2018; Gaieb et al., 2019), and 700+ fragment ligands from ~1000 PDB structures by PanDDA method (Pearce et al., 2017). The exclusion list of likely non-functional ligands (e.g., solvent molecules, ions, salts, buffers, crystallization precipitants, common cryo-protectants, and reducing agents) are also provided at Zenodo (zenodo.org) under DOI: 10.5281/zenodo.5525191 as well as GitHub (https://github.com/rcsb/PDB_ligand_quality_composite_score).

Computation and software

Data process, visualization, search, tabulation, and statistical calculation were performed primarily by a combination usage of Python and R. Computation was performed on in-house workstations at RCSB PDB.

QUANTIFICATION AND STATISTICAL ANALYSIS

Data summary

As of December 31, 2020, there were 32862 publicly released ligands in the CCD from which the CCD data were collected. The validation data on the ligand instances in PDB structures were collected earlier at the end of September 2020. As of September 30, 2020, among 149,931 PDB X-ray structures, there were 122,970 (82%) structures with small molecule ligands. Since single-atom ions do not have bond, they were not included in the ligand quality analysis, nor were included modified amino acids and nucleotides due to the scope of data analysis as described in the Results. Ligand with undefined chemical

identity (unknown ligand) were also not include. The filtering left 108,819 (73%) structures with non-ion ligands. 725,359 ligand instances were extracted from these 108,819 X-ray structures. These ligand instances are structures of 29130 chemically unique compounds defined in the CCD. For example, there are 2672 ATP structures in 1250 PDB entries. It is common that one PDB structure contains multiple ligands: 43% CCD ligand-containing entries have only one unique ligand; 30% have 2+ unique CCD ligands; 15% have 3+; 6% have 4+; and 3% have 5+. It is also common that one PDB structure contains multiple copies of the same CCD ligand (see Results).

Missing data

RSR and RSCC calculation requires structure factor data, but deposition of such data was not mandatory for PDB structures deposited before 2008 (http://www.wwpdb.org/news/news?year=2007#29-November-2007). Ligand structures without calculated RSR and RSCC were removed for multivariate data analyses. There were also other missing data of some specific ligands. Some ligands do not have bond length RMSZ or bond angle RMSZ data from Mogul because of the insufficient references in the CSD. For multivariate data analyses, ligand structures with any missing data were removed, resulting the final dataset of 158,866 unique ligand structures representing 23,921 unique CCD ligands in 90,330 X-ray structures. This final data set is presented as a big data frame in the supplementary data for ligand structures that meet the following conditions: (1) Non-ion, (2) No missing data for any column, (3) Ligand structure is complete with no more than two missing non-hydrogen atom, and (4) Average occupancy greater than 0.9.

Data exploration and visualization

The preliminary data exploration was carried out by running R on the dataset collected above. Tables and figures were all made through standard R and packages. The probability density distribution was calculated using Gaussian kernel density estimate. The overall distribution of ligand quality indicators in

Table 1 and 2 were calculated on the 158,866 unique ligand structure representations, whereas the within-structure quality variances were separately calculated on each PDB structures between instances of the same ligand.

Principal Component Analysis (PCA)

Each input variable was scaled for PCA that was then carried out on the quality metrics by computing a correlation matrix from which ranked eigenvalues and eigenvectors were extracted. To ensure no loss of information in calculating the correlation, we examined all 158,866 CCD unique ligand structural representations as describe in the dataset.

The initial PCA Initial Principal Component Analysis (PCA) was carried out on all five quality indicators using correlation matrix. The 2nd PCA analysis was performed on the group of goodness-of-fit quality metrics of RSR and RSCC only using correlation matrix. The 3rd PCA was run on the group of RMSZ-bond-length and RMSZ-bond-angle only using correlation matrix. The loadings calculated from the 2nd and 3rd PCA runs were subsequently applied to the any individual ligand instances to calculate the principal components PC1-fitting and PC1-geometry, respectively. The ranking scales of fitting and geometry were established by ordering PC1-fitting and PC1-geometry of the 158,866 unique ligand structure representations in the PDB.

Principal Component Analysis (PCA) may be impacted by the univariate data distribution and the non-linear relationship between variables. The impact of these factors has been assessed numerically. We further explicated that the construction of the PDB ligand quality ranking scores minimized such impact.

Impact of univariate data distribution on PCA

The five primary ligand quality variables in Table 1 are not normally distributed. Among them the variable of clash-per-atom has the most skewed data distribution because nearly half of the values of

clash-per-atom are zeros. Even after mean-removal scaling, majority of the clash-per-atom values concentrate on the same value, which may significantly impact the correlation matrix used for PCA. To investigate the impact, we performed separate PCA analyses on ligands with non-zero clash-per-atom only and demonstrated the results in the Supplementary Table S2.

Comparison between this table and the Table 2 in the main text shows minor difference for the "% of total variance explained" and the "factional contributions of each indicator" (i.e., PCA loadings) for PC1/PC2/PC3-overall. But the grouping of variables is still evident: (1) group of RSR and RSCC; (2) group of RMSZs of bond length and angle; (3) clash-per-atom alone as a group. The initial PCA on all five variables was an investigative analysis on the overall data dispersion, specifically the variance distribution on different eigen vectors. From this qualitative analysis we drew the conclusion that the overall variance cannot be presented by a single combination of the variables, and there were roughly three groups. For the subset of ligands with non-zero clash, the variance distribution is similar, and the grouping is the same.

For the ligands with non-zero clash, we also performed the secondary PCA on the group #1 of RSR and RSCC, and group #2 of the two RMSZs, with the result shown in the Supplementary Table S2 as well. Comparison between the Supplementary Table S2 and the Table 2 in the main text shows little difference on PC1-fitting and PC1-geometry, which means the skewed distribution of clash-per-atom do not significantly impact the two composite quality measures on ligand fitting and geometry. Only the secondary PCA's results were used for constructing the PDB ligand quality composite ranking scores, so the distribution of clash-per-atom has little impact on the final scores.

Overall, after the zero clashes were removed, the additional analysis demonstrated that the qualitative conclusion of the initial PCA (on five indicators with clash-per-atom) still holds, and the quantitative results of the secondary PCAs used for constructing ligand quality scores change little. Similar analyses were performed on RSR, RSCC, RMSZ-bond-length, and RMSZ-bond-angle by excluding extreme

values of their distribution, and the results shows no major deviation from that of Table 2 in the main text.

Impact of non-linear relationship between variables on PCA

PCA also relies on a linear model and may be impacted by the potential non-linear relationship between variables. To address this concern, we examined the relationship between variables. Only the results from the secondary PCAs on the two groups were used for PDB ligand quality composite score construction. Therefore, we examined the relationship between RSR and RSCC in group #1 and between the two RMSZs in group #2.

It is rare to observe pure linear or a specific type of non-linear relationship in large natural data. Usually, mixed types were observed, and individual non-linear relationship (e.g., polynomial) may be numerically explored in model building process, and the final model may be 'simplified' to absorb only the major type(s) of the relationships. An intuitive way to explore the relationship is to look at the data distribution plot. Because the full data is too big, 500 random samples were chosen to avoid overlapping in the plot. Both linear and non-linear fitting were explored, and results are displayed in the Supplementary Figure S1. The code for random sample selection and plotting was uploaded the GitHub repo (https://github.com/rcsb/PDB_ligand_quality_composite_score), with fixed seed setting so that the results can be reproduced. The algorithm used for the non-linear fitting is the default local polynomial regression fitting (*stats::loess* option in *geom_smooth* function of the *ggplot2* package in the statistical programming language *R)(Cleveland et al., 1992; Wickham, 2016)*.

Based on the sample plots, RSR and RSCC generally follow a statistical linear relationship except at the very extreme values. Therefore, except for outliers, the linear relationship is a proper approximation of the bivariate distribution. The RMSZs of bond length and angle also follow an approximately linear relationship with slightly greater non-linear deviation. Therefore, the non-linearity does have certain

level of impact to the absolute values of PC1-geometry that was constructed from a linear combination of the two RMSZs. However, the monotonic feature of the non-linear fitting implies that the two RMSZs change in the same direction, which means the ranking of PC1-geometry will not be impacted by the non-linearity.

RCSB PDB ligand quality composite scores as robust ranking statistics

The eventual PDB ligand quality composite ranking scores were constructed as ranking statistics of PC1-fitting and PC1-geometry. The absolute values of the composite ligand quality indicators, PC1-fitting and PC1-geometry are difficult to be interpreted directly. The ranking of the composite indicators is uniformly distributed with easy interpretation, which was the primary reason to be used as the eventual PDB ligand quality scores. The added benefit of using ranking statistics is their robustness. The non-normal univariate data distribution and non-linear relationship do have impact on the absolute values of PC1-fitting and PC1-geometry, but not on their rankings for the majority of the ligands. The composite ranking scores tell the relative quality standing of a specific ligand structure among other ligand structures in the PDB archive, and the standing is not significantly impacted by either the skewed distribution or the non-linearity.

Because nearly half of the ligand structures do not have inter-molecular clashes, ranking is not an efficient measure for clashes or clash-per-atom. Therefore, the clashes were only reported in the tabular report in Figure 4.

Handling exceptions and outliers to warn data users

Supplementary Figure S1 shows that exceptions and outliers in PDB ligand quality data may have significantly impact on the PCA results and the constructed ligand quality composite ranking scores. For example, for ligands with very large RMSZ-bond-length but near-zero RMSZ-bond-angle, the PC1-geometry underestimate the significant problems of the ligand bond issue. Since our goal was to

display multiple ligand instances on a 2D plot, it was necessary to reduce the dimension. But the presence of exceptions is a general problem for dimension reduction of any data, because when dimension is reduced to gain simplicity, there is information loss. To address this concern, we marked exceptional values in the tabular report of Figure 4 based on the univariate distribution alone. For example, row #4 of the table (under the identifier of 6WJC_Y01_A_505) has very high RSR that is among the worst 1% RSR in the entire PDB archive, so the value of RSR is highlighted as a warning to data users regardless the value of PC1-fitting. If a ligand has very large RMSZ-bond-length but small RMSZ-bond-angle, the extreme value of RMSZ-bond-length will be marked even though the value of PC1-geometry may be moderate.

Composite ranking scores construction

For any instance of a ligand in a PDB structure, the composite ranking scores are constructed by comparing its PC1-fitting and PC1-geometry composite measures to the 158,866 CCD unique ligand structure representations as reference. Composite ranking score is defined as the percent of ligand structures with inferior quality in the reference set. Therefore, composite ranking score is uniformly distributed.

Future update of composite ranking scores

Reference data will be updated annually adding new ligand structures deposited during the previous year. Hence, composite ranking scores of any given ligand instance may change slightly from year to year.

ADDITIONAL RESOURCES

The ligand quality composite ranking scores reported in this study are available at RCSB.org, under the newly implemented “Ligands” tab for PDB structures with ligands. Please refer to user instruction at

https://www.rcsb.org/docs/general-help/ligand-structure-quality-in-pdb-structures, and the example at https://www.rcsb.org/ligand-validation/6WJC/Y01.

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.

Adams, P. D., Aertgeerts, K., Bauer, C., Bell, J. A., Berman, H. M., Bhat, T. N., Blaney, J. M., Bolton, E., Bricogne, G., Brown, D., et al. (2016). Outcome of the First wwPDB/CCDC/D3R Ligand Validation Workshop. Structure *24***,** 502-8.

Adams, P. D., Afonine, P. V., Bunkoczi, G., Chen, V. B., Davis, I. W., Echols, N., Headd, J. J., Hung, L.-W., Kapral, G. J., Grosse-Kunstleve, R. W., et al. (2010). PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallographica Section D *66***,** 213-221.

Afonine, P. V., Klaholz, B. P., Moriarty, N. W., Poon, B. K., Sobolev, O. V., Terwilliger, T. C., Adams, P. D. & Urzhumtsev, A. (2018). New tools for the analysis and validation of cryo-EM maps and atomic models. Acta Crystallogr D Struct Biol *74***,** 814-840.

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***,** 235-42.

Bick, M. J., Greisen, P. J., Morey, K. J., Antunes, M. S., La, D., Sankaran, B., Reymond, L., Johnsson, K., Medford, J. I. & Baker, D. (2017). Computational design of environmental sensors for the potent opioid fentanyl. Elife *6*.

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

Bruno, I. J., Cole, J. C., Kessler, M., Luo, J., Motherwell, W. D., Purkis, L. H., Smith, B. R., Taylor, R., Cooper, R. I., Harris, S. E., et al. (2004). Retrieval of crystallographically-derived molecular geometry information. J Chem Inf Comput Sci *44***,** 2133-44.

Brzezinski, D., Porebski, P. J., Kowiel, M., Macnar, J. M. & Minor, W. (2021). Recognizing and validating ligands with CheckMyBlob. Nucleic Acids Res *49***,** W86-W92.

Buey, R. M., Fernandez-Justel, D., Marcos-Alcalde, I., Winter, G., Gomez-Puertas, P., De Pereda, J. M. & Luis Revuelta, J. (2017). A nucleotide-controlled conformational switch modulates the activity of eukaryotic IMP dehydrogenases. Sci Rep *7***,** 2648.

Burley, S. K. (2021). Impact of structural biologists and the Protein Data Bank on small-molecule drug discovery and development. J Biol Chem**,** 100559.

Burley, S. K., Berman, H. M., Bhikadiya, C., Bi, C., Chen, L., Di Costanzo, L., Christie, C., Dalenberg, K., Duarte, J. M., Dutta, S., et al. (2019). RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. Nucleic Acids Res *47***,** D464-D474.

Calamini, B., Santarsiero, B. D., Boutin, J. A. & Mesecar, A. D. (2008). Kinetic, thermodynamic and X-ray structural insights into the interaction of melatonin and analogues with quinone reductase 2. Biochem J *413***,** 81-91.

Cereto-Massague, A., Ojeda, M. J., Joosten, R. P., Valls, C., Mulero, M., Salvado, M. J., Arola-Arnal, A., Arola, L., Garcia-Vallve, S. & Pujadas, G. (2013). The good, the bad and the dubious: VHELIBS, a validation helper for ligands and binding sites. J Cheminform *5***,** 36.

Chen, V. B., Arendall, W. B., 3rd, Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S. & Richardson, D. C. (2010). MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr D Biol Crystallogr *66***,** 12-21.

Cleveland, W. S., Grosse, E. & Shyu, W. M. 1992. Local regression modells. *In:* CHAMBERS, J. M. & HASTIE, T. J. (eds.) *Statisticall Models in S.* 1st ed. Boca Raton: Chapman and Hall/CRC.

Deller, M. C. & Rupp, B. (2015). Models of protein-ligand crystal structures: trust, but verify. J Comput Aided Mol Des *29***,** 817-36.

Douangamath, A., Fearon, D., Gehrtz, P., Krojer, T., Lukacik, P., Owen, C. D., Resnick, E., Strain-Damerell, C., Aimon, A., Abranyi-Balogh, P., et al. (2020). Crystallographic and electrophilic fragment screening of the SARS-CoV-2 main protease. Nat Commun *11*, 5047.
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.
Gaieb, Z., Liu, S., Gathiaka, S., Chiu, M., Yang, H. W., Shao, C. H., Feher, V. A., Walters, W. P., Kuhn, B., Rudolph, M. G., et al. (2018). D3R Grand Challenge 2: blind prediction of protein-ligand poses, affinity rankings, and relative binding free energies. Journal of Computer-Aided Molecular Design *32*, 1-20.
Gaieb, Z., Parks, C. D., Chiu, M., Yang, H., Shao, C., Walters, W. P., Lambert, M. H., Nevins, N., Bembenek, S. D., Ameriks, M. K., et al. (2019). D3R Grand Challenge 3: blind prediction of protein-ligand poses and affinity rankings. J Comput Aided Mol Des.
Gathiaka, S., Liu, S., Chiu, M., Yang, H., Stuckey, J. A., Kang, Y. N., Delproposto, J., Kubish, G., Dunbar, J. B., Jr., Carlson, H. A., et al. (2016). D3R grand challenge 2015: Evaluation of protein-ligand pose and affinity predictions. J Comput Aided Mol Des *30*, 651-668.
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.
Gore, S., Velankar, S. & Kleywegt, G. J. (2012). Implementing an X-ray validation pipeline for the Protein Data Bank. Acta Crystallogr D Biol Crystallogr *68*, 478-83.
Groom, C. R., Bruno, I. J., Lightfoot, M. P. & Ward, S. C. (2016). The Cambridge Structural Database. Acta Crystallogr B *72*, 171-9.
Hutchinson, J. P., Rowland, P., Taylor, M. R. D., Christodoulou, E. M., Haslam, C., Hobbs, C. I., Holmes, D. S., Homes, P., Liddle, J., Mole, D. J., et al. (2017). Structural and mechanistic basis of differentiated inhibitors of the acute pancreatitis target kynurenine-3-monooxygenase. Nat Commun *8*, 15827.
Jones, T. A., Zou, J.-Y., Cowan, S. W. & Kjeldgaard, M. (1991). Improved methods for building protein models in electron density maps and the location of errors in these models. Acta Crystallogr. *A47*, 110-119.
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.
Kugel, S., Baunach, M., Baer, P., Ishida-Ito, M., Sundaram, S., Xu, Z., Groll, M. & Hertweck, C. (2017). Cryptic indole hydroxylation by a non-canonical terpenoid cyclase parallels bacterial xenobiotic detoxification. Nat Commun *8*, 15804.
Lawson, C. L., Kryshtafovych, A., Adams, P. D., Afonine, P. V., Baker, M. L., Barad, B. A., Bond, P., Burnley, T., Cao, R., Cheng, J., et al. (2021). Cryo-EM model validation recommendations based on outcomes of the 2019 EMDataResource challenge. Nat Methods *18*, 156-164.
Leung, K. K. & Shilton, B. H. (2015). Quinone reductase 2 is an adventitious target of protein kinase CK2 inhibitors TBBz (TBI) and DMAT. Biochemistry *54*, 47-59.
Maeda, S., Xu, J., Fm, N. K., Clark, M. J., Zhao, J., Tsutsumi, N., Aoki, J., Sunahara, R. K., Inoue, A., Garcia, K. C., et al. (2020). Structure and selectivity engineering of the M1 muscarinic receptor toxin complex. Science *369*, 161-167.

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.
Newman, J. (2020). Structure and X-ray Fragment screening of SARS-Cov-2 helicase (Nsp13). openlabnotebooks.org.
Pearce, N. M., Krojer, T., Bradley, A. R., Collins, P., Nowak, R. P., Talon, R., Marsden, B. D., Kelm, S., Shi, J., Deane, C. M., et al. (2017). A multi-crystal method for extracting obscured crystallographic states from conventionally uninterpretable electron density. Nat Commun *8*, 15123.
Pintilie, G., Zhang, K., Su, Z., Li, S., Schmid, M. F. & Chiu, W. (2020). Measurement of atom resolvability in cryo-EM maps with Q-scores. Nat Methods *17*, 328-334.
Protein Data Bank (1971). Crystallography: Protein Data Bank. Nature (London), New Biol. *233*, 223-223.
Read, R. J., Adams, P. D., Arendall, W. B., 3rd, Brunger, A. T., Emsley, P., Joosten, R. P., Kleywegt, G. J., Krissinel, E. B., Lutteke, T., Otwinowski, Z., et al. (2011). A new generation of crystallographic validation tools for the protein data bank. Structure *19*, 1395-412.
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.
Schuller, M., Correy, G. J., Gahbauer, S., Fearon, D., Wu, T., Diaz, R. E., Young, I. D., Martins, L. C., Smith, D. H., Schulze-Gahmen, U., et al. (2020). Fragment Binding to the Nsp3 Macrodomain of SARS-CoV-2 Identified Through Crystallographic Screening and Computational Docking. bioRxiv.
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., Feng, Z., Westbrook, J. D., Peisach, E., Berrisford, J., Ikegawa, Y., Kurisu, G., Velankar, S., Burley, S. K. & Young, J. Y. (2021). Modernized Uniform Representation of Carbohydrate Molecules in the Protein Data Bank. Glycobiology.
Shao, C., Yang, H., Westbrook, J. D., Young, J. Y., Zardecki, C. & Burley, S. K. (2017). Multivariate Analyses of Quality Metrics for Crystal Structures in the Protein Data Bank Archive. Structure *25*, 458-468.
Smart, O. S., Horsky, V., Gore, S., Svobodova Varekova, R., Bendova, V., Kleywegt, G. J. & Velankar, S. (2018). Validation of ligands in macromolecular structures determined by X-ray crystallography. Acta Crystallogr D Struct Biol *74*, 228-236.
Steinegger, M. & Soding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol *35*, 1026-1028.
Thal, D. M., Sun, B., Feng, D., Nawaratne, V., Leach, K., Felder, C. C., Bures, M. G., Evans, D. A., Weis, W. I., Bachhawat, P., et al. (2016). Crystal structures of the M1 and M4 muscarinic acetylcholine receptors. Nature *531*, 335-40.
Tickle, I. J. (2012). Statistical quality indicators for electron-density maps. Acta Crystallogr D Biol Crystallogr *68*, 454-67.
Touw, W. G., Van Beusekom, B., Evers, J. M., Vriend, G. & Joosten, R. P. (2016). Validation and correction of Zn-CysxHisy complexes. Acta Crystallogr D Struct Biol *72*, 1110-1118.
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*, D402-8.
Van Der Aalst, W. M. P., Bichler, M. & Heinzl, A. (2017). Responsible Data Science. Business & Information Systems Engineering *59*, 311-313.
Varki, A. 2017. *Essentials of glycobiology,* Cold Spring Harbor, New York, Cold Spring Harbor Laboratory Press.

Warne, T., Moukhametzianov, R., Baker, J. G., Nehme, R., Edwards, P. C., Leslie, A. G., Schertler, G. F. & Tate, C. G. (2011). The structural basis for agonist and partial agonist action on a beta(1)-adrenergic receptor. Nature *469***,** 241-4.
Warren, G. L., Do, T. D., Kelley, B. P., Nicholls, A. & Warren, S. D. (2012). Essential considerations for using protein-ligand structures in drug discovery. Drug Discov Today *17***,** 1270-81.
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. & Fitzgerald, P. M. D. 2009. Chapter 10 The PDB format, mmCIF formats, and other data formats. *In:* BOURNE, P. E. & GU, J. (eds.) *Structural Bioinformatics, Second Edition.* Hoboken, NJ: John Wiley & Sons, Inc.
Westbrook, J. D., Shao, C., Feng, Z., Zhuravleva, M., Velankar, S. & Young, J. (2015). The chemical component dictionary: complete descriptions of constituent molecules in experimentally determined 3D macromolecules in the Protein Data Bank. Bioinformatics *31***,** 1274-8.
Westbrook, J. D., Soskind, R., Hudson, B. P. & Burley, S. K. (2020). Impact of Protein Data Bank on Anti-neoplastic Approvals. Drug Discov Today *25***,** 837-850.
Wickham, H. 2016. *ggplot2: Elegant Graphics for Data Analysis*, Springer-Verlag New York.
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.
Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., et al. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res *46***,** D1074-D1082.
Wwpdb Consortium (2019). Protein Data Bank: the single global archive for 3D macromolecular structure data. Nucleic Acids Res *47***,** D520-D528.
Yang, H., Peisach, E., Westbrook, J. D., Young, J., Berman, H. M. & Burley, S. K. (2016). DCC: a Swiss army knife for structure factor analysis and validation. J. Appl. Cryst. *49***,** 1081-1084.
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

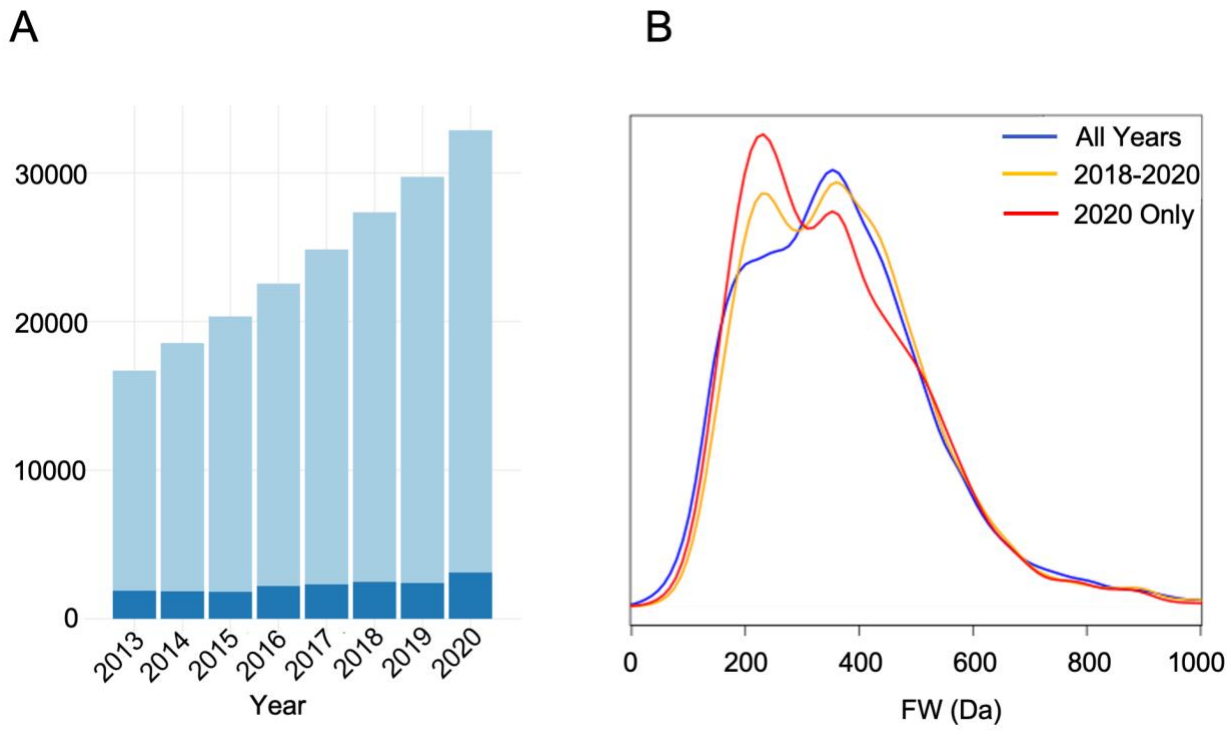

A
30000
20000
10000
0
2013
2014
2015
2016
2017
2018
2019
2020
Year
B
All Years
2018-2020
2020 Only
0
200
400
600
800
1000
FW (Da)

Figure 2

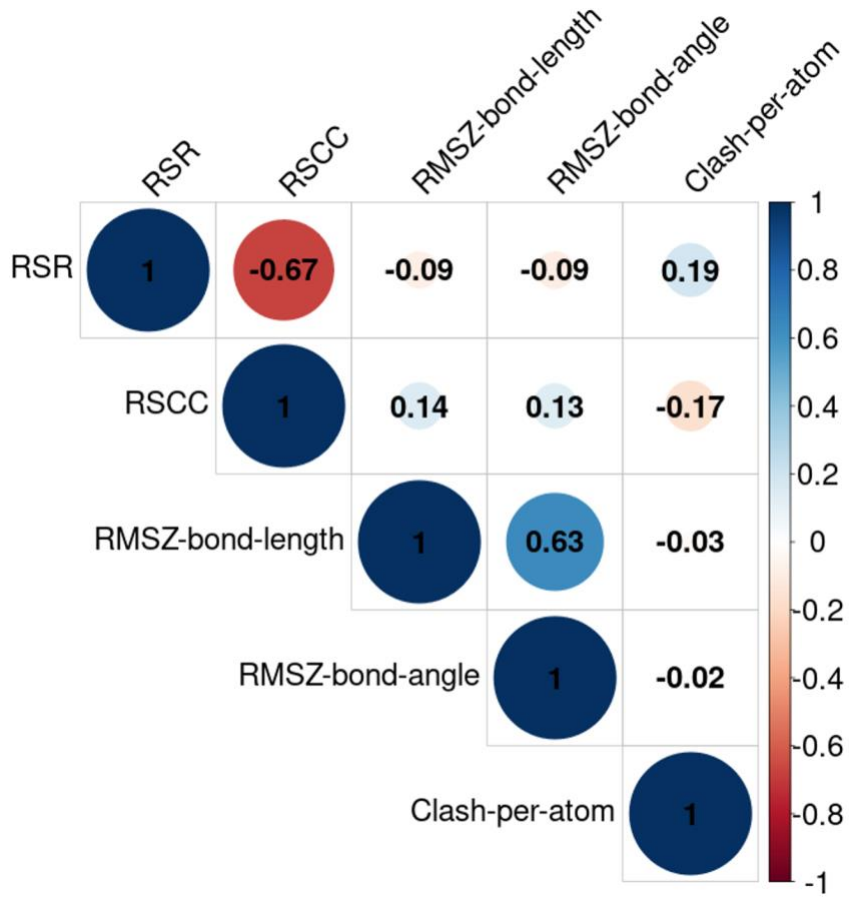

RSR
RSCC
RMSZ-bond-length
RMSZ-bond-angle
Clash-per-atom
RSR
RSCC
RMSZ-bond-length
RMSZ-bond-angle
Clash-per-atom
1
-0.67
-0.09
-0.09
0.19
1
0.14
0.13
-0.17
1
0.63
-0.03
1
-0.02
1
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

Figure 3

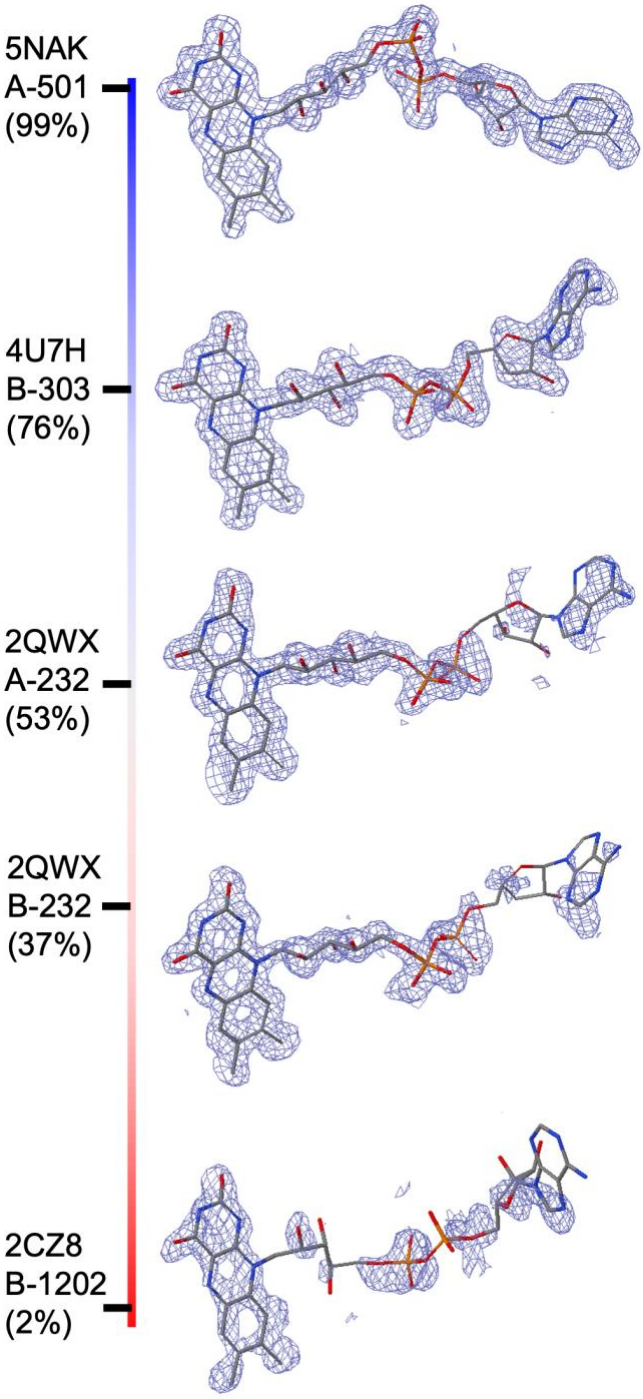
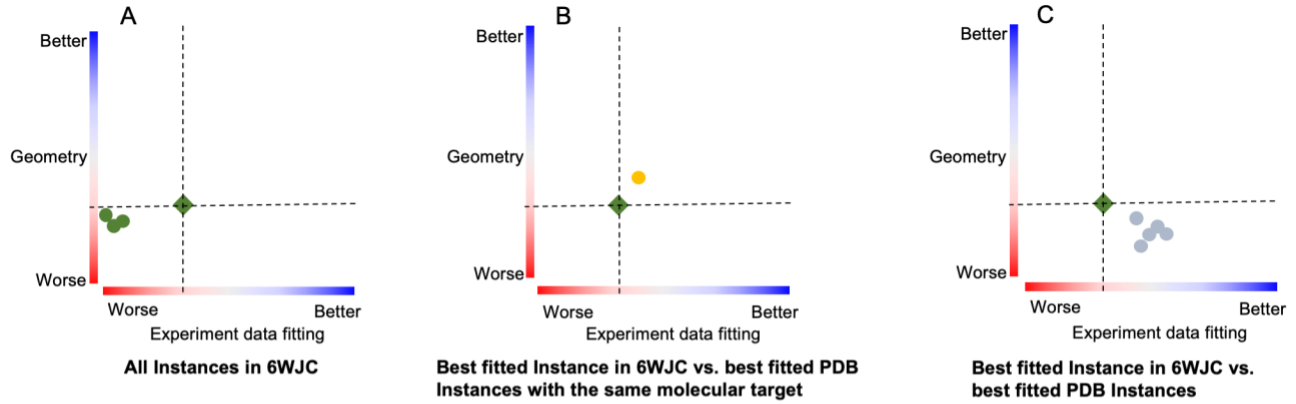

5NAK
A-501
(99%)
4U7H
B-303
(76%)
2QWX
A-232
(53%)
2QWX
B-232
(37%)
2CZ8
B-1202
(2%)

Figure 4



D

Identifier	Composite ranking of goodness-of-fit	Composite ranking of geometry	Real space R factor	Real space correlation coefficient	RMSZ-bond-length	RMSZ-bond-angle	Outliers of bond length	Outliers of bond angle	Atomic clashes	Stereo-chemical errors	Model completeness	Average occupancy
6WJC_Y01_A_502	**31.4%**	**35.7%**	**0.227**	**0.892**	**1.31**	**1.34**	**5**	**8**	**0**	**0**	**100%**	**1**
6WJC_Y01_A_503	3.5%	31.5%	0.518	0.871	1.28	1.58	4	12	0	0	100%	1
6WJC_Y01_A_504	2.3%	31.0%	0.422	0.714	1.27	1.62	4	9	3	0	100%	1
6WJC_Y01_A_505	0.4%	33.5%	0.727	0.761	1.32	1.44	4	10	3	0	100%	1
5CXV_Y01_A_502	37.1%	43.2%	0.207	0.899	1.02	1.24	1	5	0	0	100%	1
2Y00_Y01_B_401	51.2%	25.7%	0.195	0.946	1.24	1.94	3	13	1	0	100%	1
2Y01_Y01_A_401	50.3%	27.7%	0.208	0.956	1.23	1.84	3	15	1	0	100%	1
2Y03_Y01_B_401	49.3%	25.3%	0.213	0.957	1.21	1.99	3	13	2	0	100%	1
3ZPR_Y01_A_401	48.5%	17.8%	0.215	0.956	1.24	2.46	3	13	1	0	100%	1
4XNV_Y01_A_1103	47.4%	28.7%	0.175	0.911	1.9	1.15	9	3	0	0	100%	1