*Nucleic Acids Research*, 2026, **54**, gkaf1335
https://doi.org/10.1093/nar/gkaf1335
**Molecular and Structural Biology**

# New targets and procedures for validating the valence geometry of nucleic acid structures

**Jiří Černý[1,†], Robert A. Nicholls[2,†], Dariusz Brzezinski[3,†], Helen M. Berman[4,5], Miroslaw Gilski[6,7], Robbie P. Joosten[8], Marcin Kowiel[9], Catherine L. Lawson[10], Nigel W. Moriarty[11], Jane S. Richardson[12], Bohdan Schneider[1], Clemens Vonrhein[13], Christopher J. Williams[12], Mariusz Jaskólski[6,7,*], Martin Egli[14,*]**

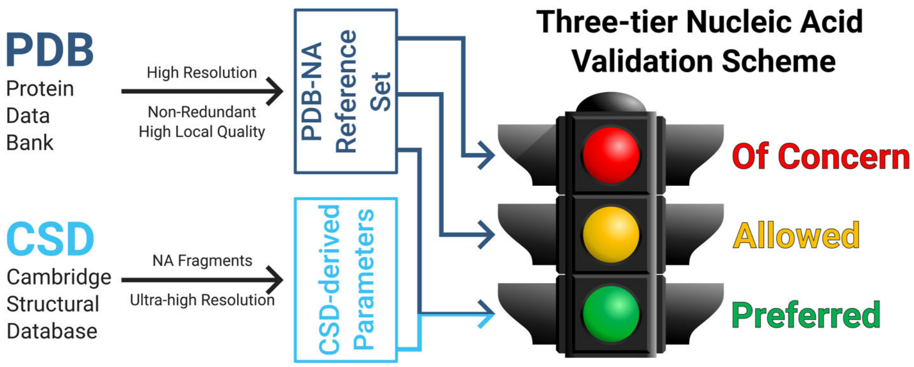
[1]Institute of Biotechnology, Czech Academy of Sciences, 252 50 Vestec, Czech Republic
[2]Scientific Computing Department, UKRI Science and Technology Facilities Council, Research Complex at Harwell, Rutherford Appleton Laboratory, Didcot OX11 0FA, United Kingdom
[3]Institute of Computing Science, Poznan University of Technology, Poznan 60-965, Poland
[4]Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, United States
[5]Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, CA 90089, United States
[6]Department of Crystallography, Faculty of Chemistry, Adam Mickiewicz University, Poznan 61-614, Poland
[7]Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan 61-704, Poland
[8]Oncode Institute and Division of Biochemistry, The Netherlands Cancer Institute, Amsterdam 1066CX, the Netherlands
[9]Ryvu Therapeutics, Kraków 30-394, Poland
[10]Institute for Quantitative Biomedicine, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, United States
[11]Molecular Biophysics and Integrated Bioimaging, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, United States
[12]Department of Biochemistry, Duke University, Durham, NC 27710, United States
[13]Global Phasing Ltd, Cambridge CB3 0AX, United Kingdom
[14]Department of Biochemistry, Vanderbilt University, School of Medicine, Nashville, TN 37232, United States

[*]To whom correspondence should be addressed. Email: martin.egli@vanderbilt.edu
Correspondence may also be addressed to Mariusz Jaskólski. Email: mariuszj@amu.edu.pl
[†]Equal contribution.

## Abstract

A Working Group consisting of the co-authors of this paper was established in 2020 to re-evaluate the standard valence geometry used for the validation of nucleic acid structure models in the Protein Data Bank (PDB). This Working Group re-examined the dependence of Cambridge Structural Database (CSD) derived targets on base and sugar type, sugar pucker, and phosphate and glycosidic conformation, before comparing those targets with the geometry of a quality-filtered reference set of nucleic acid crystal structural models held in the PDB. This revealed that the valence bond and angle mean values are close to the CSD targets, but many parameters have highly non-Gaussian or even multimodal distributions. One explanation is the inconsistency of restraints used over time and by different refinement programs. The Working Group recommends a new validation scheme for use by the PDB. For this purpose, we have developed a new three-tier scale for outlier detection—graded as Preferred, Allowed, and Of Concern intervals—based on a combination of quality-curated reference data from the CSD and the PDB. The proposed approach to validation should lead to improved nucleic acid models in (future) PDB-deposited macromolecular structures.

## Graphical abstract



**Received:** August 4, 2025. **Revised:** October 10, 2025. **Accepted:** November 3, 2025

## Introduction

Nearly 30 years ago, in 1996, ideal bond lengths and angles for nucleic acids were established by analysis of nucleic acid moieties in the small-molecule Cambridge Structural Database (CSD; [1]). Those values and error estimates are used to this day by the Protein Data Bank (PDB; [2]) for validation and in many crystallographic software packages.

Since then, the number of structural models containing nucleic acids has increased by more than an order of magnitude in the CSD and the PDB. In 1996, there were only ∼700 nucleic acid-containing entries in the PDB, with an additional 50 entries in the Nucleic Acid Database [3] (now the Nucleic Acid Knowledgebase [4]) and about 350 components of nucleic acids in the CSD. Currently, the PDB holds nearly 20 000 DNA and RNA structure models (by themselves or as part of complexes), and the CSD contains over 1600 nucleic acid components. They range from small duplexes to large, intricately folded structures, such as ribosomes, with diverse functions. These structures have been determined mostly by X-ray crystallography (∼12 000) but also by neutron diffraction (ND, 8), by cryo Electron Microscopy (cryoEM, ∼5600), and by Nuclear Magnetic Resonance spectroscopy (NMR, ∼1800).

A recent set of papers re-examined the much-expanded CSD to establish new bond length and angle targets in nucleic acids [5–7]. The revised targets were derived from highly accurate, full-matrix least-squares-refined small-molecule crystal structures that were additionally screened by a robust Z-score outlier elimination procedure. The analyses led to conformation-dependent targets for many bond lengths and angles, especially those related to sugar pucker, in the phosphodiesters, and around the glycosidic bond. The revised targets were made available through the RestraintLib Conformation-Dependent Library (CDL), referred to here as the NA RestraintLib to distinguish it from the CDL for protein geometry [8].

For all these reasons, a Working Group on Nucleic Acid Valence Geometry (NA-VAL) was convened by the ELIXIR Community on Structural Bioinformatics (3D-BioInfo) to analyze current data, report on findings, and make recommendations for geometric validation of nucleic acid structures [9]. The Working Group consists of developers of refinement packages and validation protocols, computational biologists and chemists who analyze nucleic acid structures, and representatives of the structure repositories, most of whom also solve and refine such structural models. The charge of the NA-VAL Working Group was to define, obtain agreement on, and implement an updated set of nucleic acid valence geometry parameters for use in validation.

### History of valence geometry standards for nucleic acids

In 1996, Clowney *et al.* [10] updated the standards for nitrogenous bases derived by Taylor and Kennard [11]. Structures of neutral T, A, U, G, and C, as well as protonated A and C were selected from the CSD, fulfilling quality criteria based on the R-factors (better than 6%), resolution (better than 1.0 Å), and C–C bond estimated standard deviation (e.s.d.; less than 0.01 Å). Additional criteria were that pyrimidines were substituted at the N1 atom, purines were substituted at N9, and entries with heavy atoms such as Br or metals were excluded. The final sample contained 225 structures. The distributions of each distance and angle were analyzed, and corresponding mean values and standard deviations were calculated. The idealized geometries of each of the bases were calculated from these data.

Following earlier analyses of β-nucleoside crystal structures [12], in which it was shown that the valence angles in sugars are conformationally dependent, Gelbin *et al.* [13] analyzed 108 structures from the CSD that contained a total of 127 sugar moieties. For that sample, the R-factors were below 8%, and the C–C bond e.s.d.s were less than 0.02 Å. Four conformational subgroups of structures were analyzed: ribose/C2’-endo, ribose/C3’-endo, deoxyribose/C2’-endo, and deoxyribose/C3’-endo. Statistically significant differences were found for some distances and angles between the ribose and deoxyribose samples, and for some cases between the C2’-endo and C3’-endo puckers.

A much smaller set of ∼10 crystal structures of dinucleoside phosphates that were refined with full-matrix least-squares with R-factors below 8% and did not contain modifications or ligands was used to analyze the phosphodiester valence geometry. The study hinted at possible conformation-dependence of the phosphate angles (“large” and “small”), but the sample was too small for a meaningful analysis of conformational dependence.

### Current Protein Data Bank validation for nucleic acids

The standard geometries that were derived from the CSD for the bases, sugars, and phosphodiester groups [10, 13], as described above, became the basis of nucleic acid geometry validation by the PDB. In the present implementation, this validation is performed as part of the geometric analysis of all standard amino acids and nucleotides, with results collected in the wwPDB validation reports available for each PDB entry [14]. The results include tables of summary statistics and individual outlier listings, with separate sections for bond lengths and bond angles.

In this validation procedure, a Z-score is calculated for each bond length and bond angle relative to the expected value and its standard deviation (a Z-score is the difference between the observed and expected values, divided by the standard deviation). Conformation-dependent differences in bond lengths and angles, as identified in the 1996 standards, are accounted for through separate evaluations of the two sugar puckers (C3’-endo versus C2’-endo). Bonds and angles with a Z-score greater than 5 or smaller than -5 are listed as outliers.

Following the determination of individual bond and angle Z-scores, root-mean-square Z-values (RMSZ) are calculated for bonds and angles at the level of each individual residue, each polymer chain, and the full structure. An RMSZ of 1 is consistent with the expected distribution based on the reference data, and higher values indicate a distribution that is wider than expected from the reference data. A tabulated summary reports these values, as well as percentages of bonds/angles with an absolute Z-score greater than 5.

### Refinement parameters

Implementation of the Clowney *et al.* and Gelbin *et al.* geometric parameters for the refinement of nucleic acids was undertaken within the X-PLOR refinement package [15] version 3.1, requiring expanded atom types [16]. The aim was to use the parameters to develop the appropriate weights for the refinement of nucleic acid-containing structures, specifically nucleic acid/protein complexes. The parameter file was populated with the target restraints (bond distances, angles, and dihedral angles) and with force constants. These constants were validated through the re-refinement of many DNA and protein/DNA complexes. This process resulted in a balanced dictionary for the refinement of macromolecular nucleic acid structures.

Most other refinement programs adopted some of the approaches summarized by Parkinson *et al.* [16]. CNS [17], for example, used the Parkinson DNA-based targets but added restraints for the 5′-phosphate based on Saenger [18]. Phenix [19] used the Parkinson targets but increased some of the e.s.d. values to a minimum value, initially pucker-agnostic. Since the introduction of the "Pperp" criterion to diagnose incorrect pucker choice [20], pucker-specific values have been used for RNA. The original distribution of the CCP4 Monomer Library (CCP4-ML) [21], which is used in Refmac5 [22] and by extension in PDB-REDO [23], also used the Parkinson values. Currently, the dictionary generator AceDRG [24] uses the small-molecule Crystallography Open Database [25] to generate restraints for all metal-free components in the CCP4-ML [26]. PDB-REDO has adopted conformation-independent restraints based on the CSD [5–7, 27]. BUSTER [28] makes use of the differentiation of main- and side-chains as implemented and described in the TNT package [29], using the 1996 targets in a pucker-agnostic way by default. In SHELXL [30], restraints are defined as part of the input file, and users usually create them from the Parkinson library.

## Rationale for new standards and validation procedures

It has been pointed out [31] that validation reports for nucleic acid structures in the PDB often list many more valence geometry outliers than are seen for proteins. The nucleic acid backbone has many conformational parameters, such as sugar puckers, that cannot be modeled reliably in poor electron density and cannot be fixed purely by coordinate refinement if incorrect. However, there are also multiple outliers arising from uniformly relaxing restraints at high resolution.

The current nucleic acid PDB validation procedures flag so many outliers partly because the validation targets were created almost three decades ago, when there were fewer high-quality structural data. The CSD validation targets have been recently updated [5–7] but have not yet been implemented in the PDB validation procedure. In addition, the e.s.d. values used for outlier detection of nucleic acid structures are derived differently from those for proteins. Estimated standard deviations and related Z-scores used to label outliers assume a Normal (Gaussian) distribution, which does not always reflect the physical reality of valence parameter distributions. Also, the weighting of geometric restraints in the refinement of nucleic acids, particularly in complexes with proteins, varies between refinement programs. Finally, the targets based mostly on small nucleic acid fragments from the CSD may not be fully representative of what can be found in macromolecular RNA and DNA structures in the PDB.

It is also worth reiterating that over the years, refinement programs have each used slightly different approaches to handle restraints of structures containing nucleic acids. Older versions of programs (CNS, Refmac5, and Phenix) used conformation-independent targets. Newer versions of Phenix first added pucker specificity in the ribose and now have the option to use conformation-dependent targets from the RestraintLib Conformation-Dependent Library (NA RestraintLib). Refmac5 allows the use of conformation-dependent targets, but this is not enabled by default and requires user intervention to prevent the automatic selection of incorrect conformers during the early stages of refinement.

## New validation system

In this paper, we propose a three-tier validation scale based on well-refined small-molecule CSD structures and quality-curated PDB structural models. Based on the context and rationale discussed above, the proposed validation approach aims to fulfill five goals: (i) modernize validation criteria by using up-to-date CSD and PDB data; (ii) anticipate skewed and small-separation multimodal valence bond and angle distributions, both genuine and artifactual; (iii) do not assume any particular distribution type; (iv) decouple validation from refinement targets; (v) and do not penalize individual users for which program they chose.

In the following sections, we explain how each of these goals was approached and discuss the details of the proposed three-tier validation scale.

# Materials and methods

## Selection of data from the Cambridge Structural Database

The reference dataset for defining standard bond length and angle values in nucleic acid structures was based on ultrahigh-resolution small-molecule fragments selected from the Cambridge Structural Database. We used the conformation-dependent data described by Kowiel *et al.* [5, 7] and Gilski *et al.* [6], which consisted of 238 phosphodiester groups, 970 bases, and 432 sugar fragments, retrieved using the CONQUEST software [32] and the CSD Python API [1]. The selection criteria used in those searches required structures to have an average estimated standard deviation of C–C bond lengths [$\sigma$(C–C)] $<$ 0.01 Å. The threshold of R $\leq$ 6% was set as a selector for the highest quality structures with nucleobases. The R criterion had to be slightly compromised to 7.5% for phosphates and 8.5% for sugars to guarantee a balance between the quality and sample size. Moreover, a robust Z-score test [33] was used to identify and reject outliers. Based on the statistical analyses described in the corresponding papers [5-7], some of the parameters were derived as single values (most notably for nucleobase geometry), while other parameters were grouped into distinct conformational clusters. For example, sugar bond distances were grouped by nucleic acid type (RNA/DNA), base type (A/C/G/T/U), and sugar pucker type (C2'-endo/C3'-endo/Other), and the phosphodiester parameters were divided into six possible conformational groups based on the values of the $\zeta$ and $\alpha$ torsion angles. For yet another group of parameters, a clear functional dependence on an appropriate conformational parameter was found. Thus, the endocyclic ribose angles show functional dependence on the sugar pucker amplitude $\tau_m$, while the glycosidic bond length and angles are functionally dependent on the glycosidic torsion angle $\chi$. To simplify validation, for the purposes of the proposed three-tier scheme, the parameters that showed functional dependencies were reduced to means and standard deviations within groups defined by nucleic acid and base type. The detailed counts, means, and standard de-

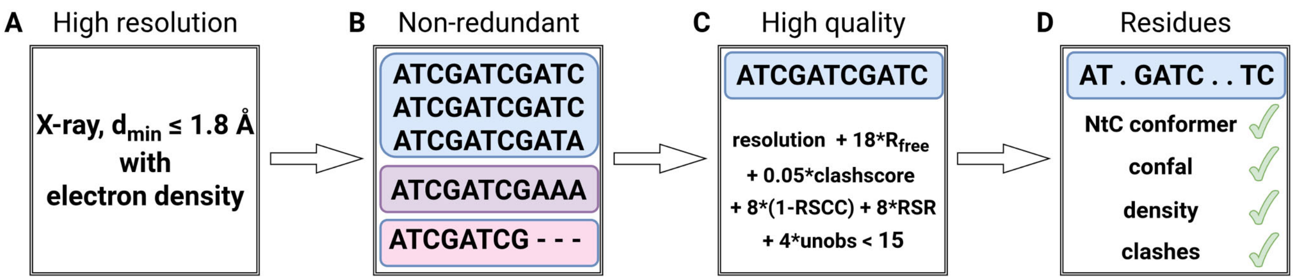


**Figure 1.** Scheme of the selection of the **PDB-NA Reference Set**. (**A**) X-ray structures of resolution 1.8 Å or better with available RCSB electron density information were selected. (**B**) These structures were clustered by sequence. (**C**) The structural models were tested for the best combination of resolution, $R_{free}$, clashscore, fraction of unobserved residues (unobs), and per-residue values of the real-space R-factor (RSR) and the real-space correlation coefficient (1-RSCC). All quantities under step (**C**) were taken from the PDB validation reports. (**D**) The final filtering was performed by testing per-nucleotide conformational agreement with the NtC classes, fit to the electron density map, and freedom from all-atom clashes or ribose pucker outliers. Steps (**A–D**) are described in detail in the text.

viations of the derived parameters, along with lists of CSD refcodes of the selected structures, can be found in the corresponding publications [5, 6, 7]. A complete listing of all the CSD-derived parameter values is available in the accompanying Zenodo repository [34]; see Data Availability section for details. The conformation-dependent means and standard deviations of bond lengths and angles within each conformation category were used as reference values for the **Preferred** tier, as described in the Results section.

## Curation of a high-quality PDB Nucleic Acid (PDB-NA) Reference Set

The preparation of a curated reference set of nucleic acid structures—which we will refer to as the **PDB-NA Reference Set** throughout the rest of this manuscript—involved four steps (Fig. 1): (A) identifying potentially suitable structural models from the PDB according to prescribed criteria; (B) defining clusters of redundant sequences in nucleic acid-containing PDB entries; (C) in each cluster of similar sequences, finding the chain deemed to be of highest structural quality; (D) filtering based on per-residue quality scores to remove residues with: incorrect nucleotide conformation, steric clashes, and poor fit to the electron density.

**Step A. Selection of X-ray structures containing nucleic acids.** A list of all PDB X-ray entries containing nucleic acids, initially having resolution better than 3.5 Å (later restricted to 1.8 Å, see Section Dependency of valence parameters on nucleotide identity) and reflection data available in RCSB was collected using an NAKB [4] query, returning 8783 PDB IDs (as of 16 October 2022, Fig. 1A). We note that—while the raw structure factor data are identical in the PDB archive—the derived electron density map coefficients and values exhibit slight variations across different wwPDB data centers. If not stated otherwise, in this study, we used data from the RCSB.

**Step B. Selection of sequence non-redundant structures.** The sequence information for each nucleic acid chain was obtained from the RCSB PDB using a GraphQL query for each PDB ID. All nucleic acid sequences were aligned using the *pairwise2.align.localds* function of BioPython (version 1.81), employing an extended nucleic acid substitution matrix. The aligned sequences were clustered separately for isolated DNA and RNA, and for DNA and RNA in complex with proteins. Chains containing both DNA and RNA nucleotides were omitted. Clusters of similar sequences were created separately for short chains of 6 to 24 nucleotides and for chains longer than 24 nucleotides. Short chains were put into one cluster if they had no more than two differences in sequence. Longer chains with a sequence identity of 90% or more were put into one cluster. At the end of the process, chains from isolated and complexed DNA were pooled into 3467 non-redundant clusters; the same was done for RNA chains, which were pooled into 1612 clusters (Fig. 1B). Further cleaning of the dataset was performed on these two sets of clusters. The NAKB query for performing Step A, and software tools needed in Step B, the GraphQL query, and the alignment code are available (see Data Availability).

**Step C. Selecting high-quality chains of nucleic acids.** In each sequence cluster, we identified the chain with the highest quality using an extended version of the Composite Quality Score (CQS) [35] (as modified for version 3.0 of the RNA non-redundant lists at https://rna.bgsu.edu/rna3dhub/nrlist). The CQS RNA quality score is a weighted combination of key quality indicators of each structure model, exactly as reported in the PDB validation reports. We modified the original CQS to also include DNA and modified the originally used weights as follows: resolution (Å) has weight 1, $R_{free}$ (as fraction) has weight 18, clashscore (as number of clashes per one thousand atoms) has weight 0.05, average per-residue value of the real space correlation coefficient (RSCC as a fraction) has weight 8, average per-residue real space R-factor (RSR, as fraction) has weight 8, and finally the fraction of unobserved residues has weight 4. It should be noted that, unlike for the other metrics, higher RSCC values are better; therefore, our extended CQS uses (1.0 − RSCC) in the scoring function. The weights of all these quality indicators were optimized such that each contributed roughly equally to the standard deviation of the composite quality score. For any structures for which the quality indicators were not available in the PDB validation reports, we used the following fallback values: 100 for resolution, 1 for $R_{free}$, 100 for clashscore, −1 for RSCC, and 40 for RSR. The CQS distributions for both DNA and RNA exhibited one major peak with values around 10 for higher-quality chains. The minor populations of lower-quality chains or structures with missing validation data typically cover scores above 20.

Cutoffs for CQS score and resolution were chosen as the highest for which the distribution did not show artifacts of bimodality from the choice of programs. The subset of the highest-quality non-redundant chains with a CQS score lower than 15 and crystallographic resolution better than 1.8 Å contained 539 DNA and 206 RNA chains. These chains comprised 6644 DNA and 4236 RNA nucleotides (Fig. 1C).

**Step D. Filtering based on residue quality.** Experience with the development of a similar high-quality reference set for proteins showed that chains of good overall quality almost always contain some extremely poor regions [36]. Therefore, we subjected the set of 539 DNA and 206 RNA non-redundant chains from high-resolution structures identified in Step C to two independent sets of quality filters:

(a) *Filter based on dinucleotide conformational classes, NtC* [37]. Nucleotides with incorrect local conformation may also deform valence geometry. Unlikely DNA and RNA conformers were filtered out using the concept of the NtC conformational classes as implemented in the dnatco.datmos.org web server [38]. This filtering step is based on the expectation that if all backbone torsion angles, sugar puckers, and the overall shape of a dinucleotide step are close to a known NtC class, then the covalent geometry of this residue also fits into the expected values. Dinucleotides assigned to one of the known NtC classes had to fulfill the following criteria: (1) the so-called *confal* quality score of dinucleotide conformations is larger than 60 (where 100 is the perfect score), (2) the harmonic mean of real space correlation coefficients between the experimental electron density and density calculated from the model (RSCC) is larger than 0.8; RSCC was calculated with *phenix.real_space_correlation* for the 18 atoms between C5' of the first nucleotide in the step and O3' of the second nucleotide (containing the C5'(n) to O3'(n + 1) backbone atoms and O4', C1', N1/N9, and C2/C4 atoms from both residues), and (3) RMSD between the same 18 atoms from the analyzed model and the structure of the representative NtC class is smaller than 0.5 Å. The choice of *confal* larger than 60 typically translates to deviations by no more than 20° from the corresponding NtC class average in each torsion, ensuring high similarity with the reference NtC. The 0.8 cutoff for the RSCC is generally used as an indicator of good model-to-density fit. The harmonic mean was used for stronger penalization of cases where a subset of atoms does not fit the electron density well, while the remaining atoms would fit nearly perfectly. The 0.5 Å RMSD cutoff in the Cartesian space supplements the *confal* (torsional space) filtering and was identified previously for dinucleotide geometries that can be, in most cases, successfully re-refined using torsional restraints closer to the representative Cartesian NtC geometry. This filtering procedure returned 4336 DNA and 3082 RNA residues.

(b) *MolProbity-based quality filters*. The second residue-level filtering system used MolProbity [39] and consisted of two main checks: model-to-map fit and model validation metrics, considering bases as well as backbone. For model-to-map fit, chains were assessed with *phenix.real_space_correlation detail = atom*, using .mtz reflection data files provided by the PDB, to obtain per-atom RSCC values and 2mFo-DFc map values at each atom position for all non-H atoms. It was observed that many otherwise well-supported residues had an atom (usually OP1 or an exocyclic base atom like thymine C7, Fig. 2) outside of any reasonable electron density contour. Therefore, for each residue, the average of the two lowest per-atom RSCC values and the average of the two lowest per-atom 2mFo-DFc map values were calculated. For a residue to be included in the **PDB-NA Reference Set**, it was required to have a worst-two-average RSCC $\geq 0.7$ and a worst-two-average 2mFo-DFc map value $\geq 1.2\sigma$. Additionally, the backbone P atom, which has about twice as many electrons as N/C/O atoms, was required to have a 2mFo-DFc map value $\geq 2.4\sigma$. The B-factor was not used as a filtering criterion, as its treatment was found to be too inconsistent across resolutions and refinement programs. Moreover, for a residue to be included, it was required to have no significant steric clashes [40]. Clashes < ($\Sigma$vdW radii - 0.5 Å) were defined as the "all-atom contacts" in MolProbity (they included hydrogen atoms, and were at the atom surfaces, not pairwise from center). A cutoff of 0.5 Å was used rather than the usual 0.4 Å to give more leeway than the region that is only sometimes wrong. Clashes with unsupported water molecules (2mFo-DFc map value $< 0.6\sigma$) were also considered insignificant. For RNA, residues with sugar pucker outliers [31] were also removed. Notably, because this **PDB-NA Reference Set** was prepared for assessing covalent bond geometry, bond length, and bond angle outliers were not used as explicit filtering criteria. Additionally, non-standard bases and residues with alternate conformations were not included in the **PDB-NA Reference Set**, as finding the correct traces through regions containing alternate positions is known to be prone to errors [41]. This filtering step returned 4561 DNA and 3175 RNA residues.

A total of 3202 DNA residues and 2544 RNA residues that passed both residue-level filtering protocols, as described in steps (a) and (b) above, were deemed of sufficiently high quality to allow further analysis with confidence (Fig. 1D).

## Dependency of valence parameters on nucleotide identity

Analysis of the bond length and angle distributions extracted from the **PDB-NA Reference Set** showed that, while the base parameters are residue-type specific, the sugar-phosphate backbone parameters could be pooled to obtain more extensive distributions. As an example, without pooling, the sample of uracil C5'–O5' bonds with a sharp X-ray resolution cutoff at 1.8 Å would consist of (just) 388 data points (Fig. 3, leftmost panel). One option to increase the amount of available data would be to extend the nominal resolution limit to worse than 1.8 Å. However, as shown in Fig. 3, increasing the amount of data by allowing lower-resolution curated structures in the **PDB-NA Reference Set** quickly introduces artifacts. The apparent bimodality of the bond length (Fig. 3, middle panels) originates from tighter refinement targets – the secondary peak corresponds to the value of 1.440 Å originally tabulated in the compilations of Parkinson *et al.* [16] and Gelbin *et al.* [13]. This effect is already visible when including data up to 2.2 Å, and continues to become more prominent when further extending the resolution to 2.4 Å. It was, therefore, decided that the resolution limit should remain at 1.8 Å and, instead, to pool sugar-phosphate backbone data from all four RNA nucleotide residues (and equivalently for all four DNA nucleotide residues).

The **PDB-NA Reference Set** for bond lengths and angles in the nucleobases contains between 648 (for residue U) and 1449 (for DG) observations per geometric feature. The pooled

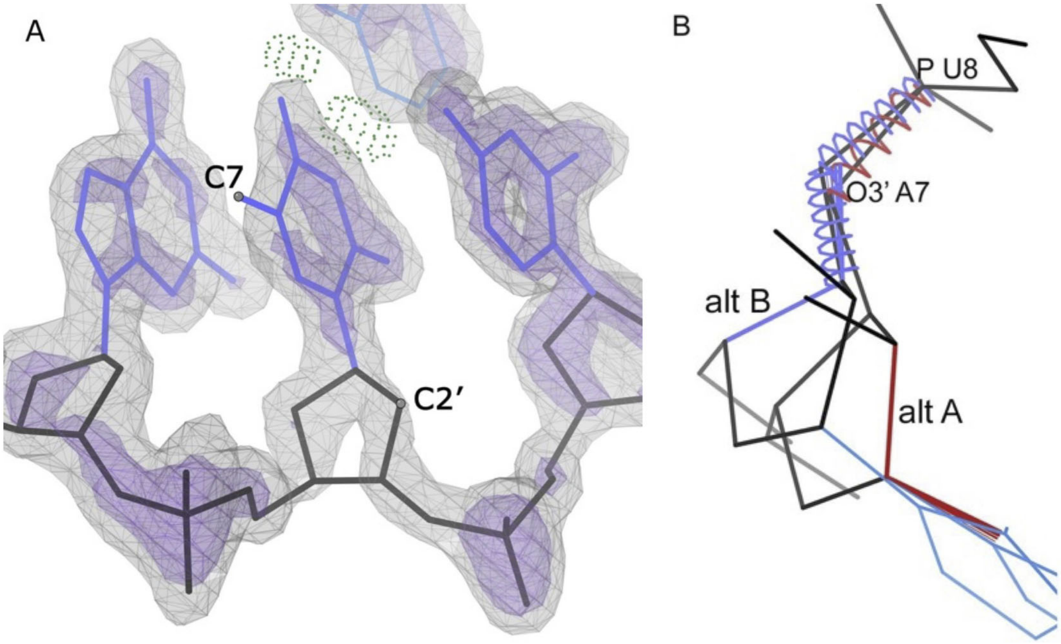


**Figure 2.** Background for selection of residue-level criteria in MolProbity data set curation. (**A**) Detail of the 1.52 Å resolution Human DNA polymerase eta - DNA ternary complex, PDB entry 4ed8 [42], in its 2mFo-DFc electron density. The gray contour is at 1.2$\sigma$, the purple contour at 2.4$\sigma$. The C7 atom (upper left) of residue DT 5 from chain P is outside the electron density envelope. However, the model of the base is clearly correct, and supported by hydrogen bonding (green dots), meaning that the atom out of density is not indicative of a modeling or conformation error. Residues like this pose a challenge for map-based residue curation, since a single density cutoff low enough to include this atom is no longer selective enough against real errors. Averaging strategies were therefore used to allow cases like this to pass residue-level filtering and be included in the **PDB-NA Reference Set**. The electron density measured at the C7 atom by *phenix.real_space_correlation* is 0.91 $\sigma$. The next lowest map value is 1.95$\sigma$ at C2′. Thus, the worst-two-average map value is 1.435$\sigma$, which exceeds the 1.2$\sigma$ filtering cutoff, indicating that the rest of the residue is experimentally supported with sufficient confidence to allow its inclusion in the **PDB-NA Reference Set**. (**B**) Serious geometrical problems common in alternate conformations, illustrated by a case in the 1.6 Å resolution PDB entry 6tqb of Roquin binding an AU-rich RNA hairpin [43]. The alternate conformation that should have been spread into the backbone causes steric strain at the discontinuity. At the residue junction, this produces O3′-P bond lengths that are too long for alt A (red spiral) or too short for alt B (blue spiral) by >0.5 Å, plus another bond-length and two bond-angle serious outliers further inside the alternate conformations.

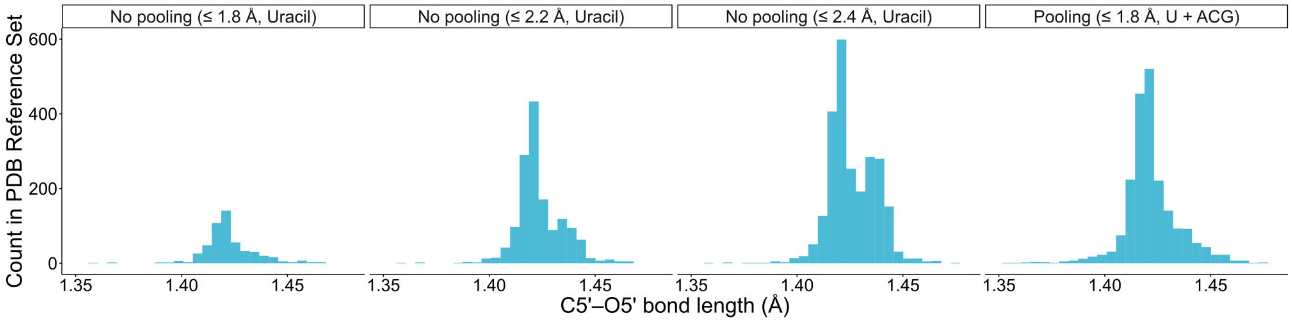


**Figure 3.** Parameter pooling and the effect of including lower resolution data in the **PDB-NA Reference Set**. Histograms of bond lengths between backbone atoms C5′ and O5′ are shown before data pooling for the uracil residue and after pooling the RNA data for A, U, C, and G residues.

data contain approximately 3300 and 2600 observations for DNA and RNA sugar-phosphate backbone parameters, respectively.

Importantly, some valence parameters were found to have multimodal distributions in the **PDB-NA Reference Set**. An example of such a case is provided in Fig. 4A by the distribution of the C1'-N9 glycosidic bond length for guanine in RNA vs in DNA. However, when comparing the PDB distributions with the corresponding CSD-derived NA fragments (Fig. 4B), the multimodality of the glycosidic bond length is not visible. For bond lengths in general, the empirical distributions for data from the PDB are a consequence of multiple factors—including varying refinement programs and protocols—and may or may not reflect the intrinsic variability of nucleic acid stereochemistry. Here, for the glycosidic bond of G in DNA, the histogram maxima for all programs agree closely with each other, as well as with the DNA-based most frequent C2'-endo values from 1996 (Fig. 4C). For the same bond in RNA, distributional bimodality is observed; this is primarily a result of combining CNS-refined deposits (left peak) and Phenix-refined deposits (right peak). The CNS software is still using the Parkinson DNA-based values for RNA, whereas Phenix now uses pucker-specific values most frequently invoked for the RNA-dominant C3'-endo pucker (Fig. 4D).

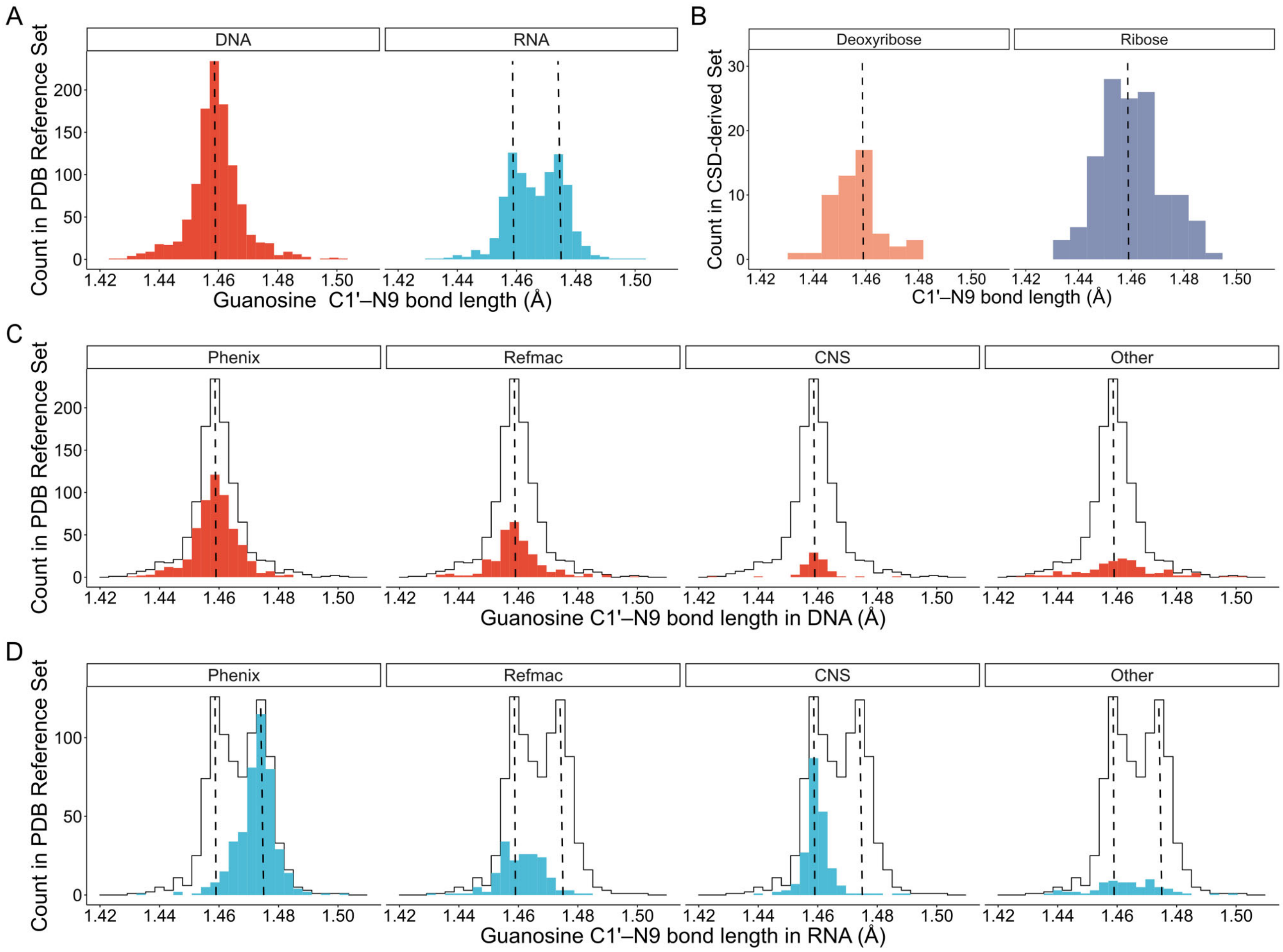


**Figure 4.** Distributions of guanosine C1′–N9 bond distances in the **PDB-NA Reference Set**. (**A**) Comparison of the **PDB-NA Reference Set** distributions for DNA and RNA. (**B**) Corresponding deoxyribose and ribose C1′–N9 bond length distributions from the CSD-derived set. (**C**) Distributions for guanosine C1′–N9 bond distances in DNA from structures refined by *Phenix*, *Refmac5, CNS*, and other (*BUSTER, X-PLOR*, *SHELX/L*, etc.) programs. (**D**) Distributions for guanosine C1′–N9 bond distances in RNA from structures refined by *Phenix, Refmac5, CNS*, and other (*BUSTER*, *X-PLOR*, *SHELX/L,* etc.) programs. In panels C and D, the hollow histogram represents the overall distribution, whereas the filled histograms show the bond lengths from structures refined using particular software packages. The dashed lines mark the maxima of distributions of the RNA and DNA histograms.

However, these are small differences, and our proposed validation system covers the range of both peaks within its **Preferred** tier.

## Expert curation of chemically questionable cases in the PDB-NA Reference Set

After performing all the quality checks, the distributions of parameter values that contained isolated data points were additionally inspected visually. Thus, 50 suspicious bond lengths and angles that strayed from the distributions were manually inspected in Coot [44] or KiNG [45]. Most of the examined residues were ambiguous. As a result, we removed 20 residues from the initially pooled data, based on detailed visual inspection. The final curated **PDB-NA Reference Set** and the list of manually excluded residues are available in the Zenodo repository (see Data Availability).

## Definition of probability percentile scores (ProSco)

To anticipate multimodal valence bond and angle distributions without assuming any particular distribution type, the three-tier validation score presented below in the Results Section largely relies on the empirical distributions observed in the curated **PDB-NA Reference Set.** More precisely, we propose a *probability percentile score (ProSco)* to numerically assess how common (i.e. "popular") a given bond length or angle is on a scale from 0 to 100. For this purpose, we perform probability density estimation through kernel smoothing [46] on the valence parameter distributions. By performing kernel smoothing, we estimate the relative frequency of particular bond lengths and angles. This non-parametric approach handles multimodal parameter distributions that may result from pooling data from different chemical environments (such as different charges on the phosphate groups [47]) or by historically inconsistent restraint handling procedures across different refinement programs. The ProSco metric is calculated as follows. For a given nucleic acid valence bond or angle data in the **PDB-NA Reference Set** (Fig. 5A), we perform kernel density estimation (KDE) on its value distribution (Fig. 5B), by smoothing using a Gaussian kernel with bandwidth determined through unbiased cross-validation multiplied by a 1.5 adjustment factor. The obtained probability density function assessing the frequency of different parameter values (Fig. 5C) is then divided into 512 bins of equal width, each bin being assigned its bond length/angle range and corresponding prob-

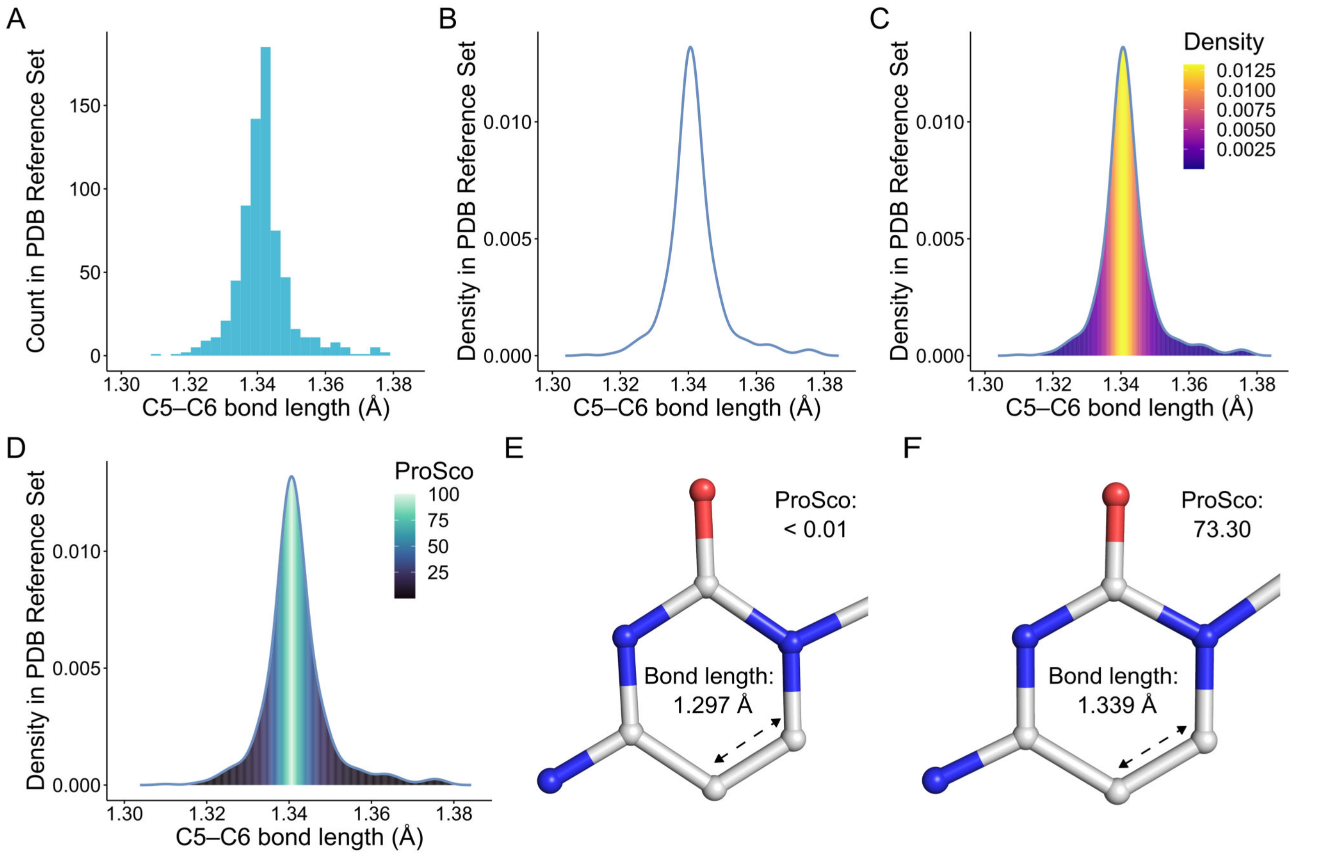


**Figure 5.** Statistical treatment of the **PDB-NA Reference Set** exemplified for the bond distances between atoms C5 and C6 of cytosine in RNA. (**A**) Histogram of bond length values in the **PDB-NA Reference Set**. (**B**) The KDE function was calculated for the parameter in question. (**C**) Relative frequencies (density-based probability estimates) of specific bond length values based on the KDE are divided into 512 bins. (**D**) ProSco scoring using percentile scores based on sorted KDE probabilities. (**E**) Example of scoring the C5–C6 bond length in residue C2610 of chain 14 in the 3.15 Å ribosome structure, PDB ID 5el5 [48]; the unusually short C5–C6 bond length of 1.297 Å translates to a ProSco score below 0.01. (**F**) Example of scoring the C5–C6 bond length in residue C34 of chain 13 in PDB structure 5el5; the bond length of 1.339 Å translates to a high ProSco score of 73.3.

ability. Finally, the bond length/angle bins are sorted according to probability, and each bin is assigned a ProSco value based on this ranking (Fig. 5D). The ProSco parameter can then be used to assess individual bond lengths or angles, yielding low scores for uncommon values (Fig. 5E) and high scores for values commonly found in the high-quality **PDB-NA Reference Set** (Fig. 5F). Computer code for calculating the probability percentile scores is available in the Zenodo repository (see Data Availability).

## A robust standard score to measure deviance from the typical distributions

The reason for any geometric feature being detected as spurious may stem from various sources. The local structural model may be conformationally incorrect or suboptimal in some way, the model may be suboptimally refined or restrained to an inappropriate target, or subject to incorrect assumptions regarding the chemical environment, or the deviation could be a genuine anomaly reflecting excessive strain from the structural environment, which in some cases might be of particular biological relevance. The **PDB-NA Reference Set** contains too few observations for reliable estimation of sufficiently extreme threshold levels directly from the empirical cumulative distribution functions. Consequently, it was necessary to explore other approaches to determining an appropriate protocol for defining what we will subsequently refer to as "**Of Concern**" cases.

The conventional approach to measuring the deviance of a particular value from a given distribution is to use a Z-score (or "standard score"), which measures how many standard deviations a data point is away from the mean of the distribution. However, such a parametric score is not statistically robust to outliers and has different interpretations for different distributions with varying forms. The conventional probabilistic interpretation of a Z-score is only valid if the underlying distribution can be considered to be Normal. Instead, we have developed a non-parametric analogue that is insensitive to distributional form, accounts for distributional asymmetry, is universally applicable to all feature types, and is robust to outliers.

Typical non-parametric analogues of a standard score measure how many interquartile ranges (IQR) or median absolute deviations (MAD) a given observation is away from the median of the distribution [49]. However, for the present application these measures had two major shortcomings: (i) they are symmetric, and thus would tend to systematically underestimate significance on one side of the distribution and overestimate significance on the other (noting that the **PDB-NA Reference Set** bond/angle distributions exhibit varying degrees of asymmetry); and (ii) they failed to sufficiently capture information about peripheral subpopulations, leading to a tendency to produce an inordinate number of outliers for some distributions. The IQR (the difference between the 25th and 75th percentiles) encapsulates information only about the dispersion of the central portion of the distribution, ignoring the peripheral regions. Similarly, although often favored as more robust than the IQR, the MAD was found to produce suboptimal results for the present application due to being overly influenced by the central portion of the distribution. Since our goal was to utilize information about smaller peripheral subpopulations, we instead consider the interdecile range, i.e. the difference between the 10th and 90th percentiles, which encapsulates information about the dispersion of a larger portion of the distribution while still retaining a reasonable degree of robustness to outliers. The interdecile measure has a breakdown point of 0.1, which means that 10% of the observations can be arbitrarily incorrect without affecting the statistic. We address the issue of requiring an asymmetric measure by considering different interquantile ranges for the upper and lower halves of the distribution. Specifically, we use the difference between the 90th and 50th percentiles (i.e. the $9^{\text{th}}$ decile and the median) as the standardization factor for values greater than the median, and the difference between the $50^{\text{th}}$ and $10^{\text{th}}$ percentiles (i.e. the median and the $1^{\text{st}}$ decile) as the standardization factor for values less than the median.

Since there are many cases where multiple instances of the same bond/angle type are present in a given PDB entry, it is also necessary to address the issue of sample bias. This is alleviated by using weighted statistics when estimating the median and deciles, in which the weight for a given bond/angle is equal to the reciprocal of the number of same-type features present in the same PDB entry.

This results in a *weighted asymmetric non-parametric standard score*, which we shall denote $Z'$:

$$Z'(x) = \begin{cases} \frac{x-\hat{M}}{\hat{D}_9-\hat{M}}\Phi^{-1}(0.9) & x \geq \hat{M} \\ \frac{x-\hat{M}}{\hat{D}_1-\hat{M}}\Phi^{-1}(0.1) & x < \hat{M} \end{cases}$$

where $\hat{M}$ is the weighted median, $\hat{D}_1$ and $\hat{D}_9$ are the weighted first and ninth deciles, and $\Phi^{-1}$ is the probit function. $Z$ A quantile function is the inverse of the Cumulative Distribution Function (CDF) of a random variable. We denote by $\Phi$ the CDF of the standard Normal distribution (mean 0, standard deviation 1), and $\Phi^{-1}$ its inverse, also known as the probit function. The name derives from 'probability unit' or 'bitwise probability', and the function is widely used in binary classification and other statistical applications. In the context of Z′(x), note that the scaling factor $\Phi^{-1}(0.9) = -\Phi^{-1}(0.1) \approx 1.28$ is used to place the $Z'$ score on the same scale as a unit standard deviation for Normal data. Indeed, in the case of normally distributed data, $Z'$ is asymptotically equivalent to a canonical $Z$-score. However, the exact probabilistic interpretation of $Z'$ is only strictly accurate for Normally distributed data; such probabilistic calculations are invalidated by violations of the Normality assumption. Consequently, since we know that we are dealing with non-Normally distributed data, any probabilistic interpretation associated with $Z'$ is inaccurate and thus should be neither overstated nor encouraged. Nevertheless, the overall scale and general interpretation of $Z'$ is consistent with the conventional standard $Z$-score, i.e. it is a unitless measure representing the number of standard deviations from the central tendency. In summary, $Z'$ is a flexible and robust measure of how far a given observation is from the distribution of values derived from the **PDB-NA Reference Set**.

Values of the weighted statistics

$$\hat{M}, \sigma_{upper} = \left(\hat{D}_9 - \hat{M}\right)/\Phi^{-1}(0.9) \text{ and}$$

$$\sigma_{lower} = \left(\hat{D}_1 - \hat{M}\right)/\Phi^{-1}(0.1)$$

for each of the **PDB-NA Reference Set** bond/angle distributions are tabulated and provided in the accompanying Zenodo repository (see Data Availability), allowing straightforward calculation of $Z'$ for observed bond/angle values as well as the associated thresholds for each bond/angle type.

## Results

### Three-tier nucleic acid validation scheme

In this report, we propose a new three-tier validation scheme for the quality assessment of nucleic acid structural model geometry. The three intervals of our system, classifying cases from commonly occurring to unusual (and thus suspicious), are labeled **Preferred**, **Allowed**, and **Of Concern**. We expect that "**Of Concern**" tags will suggest a need for careful inspection and possible remodeling. Any instance of geometry that is "**Of Concern**" is likely energetically unfavorable, and most of these are, therefore, incorrectly modeled, although occasionally some cases in this category may be genuine and could illuminate the demanding biological role there is to fill.

Each bond and angle parameter has a region of high probability, which we refer to as the **Preferred** tier. That region is surrounded on both sides by less probable parameter values: **Allowed** and then **Of Concern** (Fig. 6).

The inner **Preferred** interval is defined using both the CSD data and the curated **PDB-NA Reference Set** (see Materials and Methods). The CSD component of the **Preferred** tier is the CSD mean ± 3σ interval for the given bond or angle. The PDB component of the **Preferred** tier is the interval that contains 95% of the **PDB-NA Reference Set** (equivalent to ProSco $\geq 5$). The CSD and PDB intervals usually closely overlap but may occasionally be shifted or have different spreads. In our scheme, we always use whichever **Preferred** boundary is more permissive, i.e. the lower value for the left threshold, and the higher value for the right threshold of the interval. Our rationale for using both databases for the **Preferred** tier is as follows. The CSD data are of high quality and free of restraint bias, but they are derived from small molecular fragments that may not fully represent the stereochemical variability and macromolecular context of long nucleic acid chains. The PDB structural data are generally of lower resolution and are biased by the influence of previous restraints, especially for bond lengths (Fig. 4C). Still, they are undeniably our best representation of the structural variability and distributions found in biological nucleic acids.

The **Allowed** interval starts on either side of the **Preferred** tier boundary. It is defined as not **Preferred** and having a weighted asymmetric non-parametric standard score within the $|Z'| \leq 5$ threshold, leaving the more extreme observations with $|Z'| > 5$ classified as **Of Concern**. The *ad hoc* heuristic value of 5 was determined by manual inspection of all **PDB-NA Reference Set** distributions. We reiterate that the $|Z'| = 5$ threshold in this context does not have the same probabilistic interpretation as a canonical $Z$-score – we observe around 0.3% of bond and 0.2% of angle values more extreme than

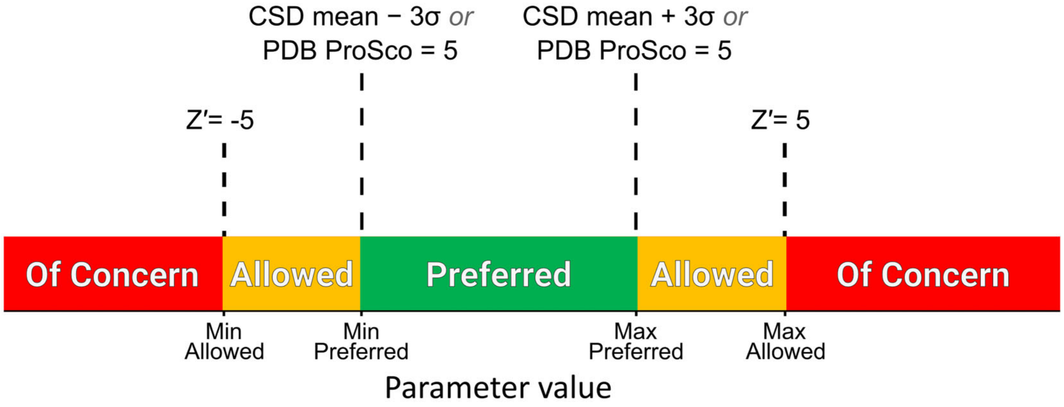


**Figure 6.** The three-tier validation scheme. The horizontal axis represents the value of the parameter being validated (bond length or angle). The lower bound for the **Preferred** tier is the lower ProSco = 5 limit from the **PDB-NA Reference Set** or the CSD mean − 3σ value, whichever is lowest. Similarly, the upper bound for the **Preferred** tier is the upper ProSco = 5 limit or the CSD mean + 3σ, whichever is highest.

this threshold in the **PDB-NA Reference Set**, which is around three-to-four orders of magnitude higher than the 0.00006% that would be usually expected for a conventional $Z$-score at the $5\sigma$ level. In addition to presenting the three-tier classification, we recommend providing the actual $Z'$-score and ProSco values as part of the validation of each individual geometric feature.

## Assessing overall geometric quality of a nucleic acid model

While detailed inspection of individual geometric features is essential in structural model validation, it is also important to consider global metrics that summarize overall geometric quality. One widely used approach is to compute the root-mean-square deviation (RMSD) of Z-scores—termed RMSZ—which quantifies how much a model's geometry deviates from expected "ideal" values:

$$\mathrm{RMSZ} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} Z_i^2}$$

where $Z_i$ are the Z-scores associated with individual geometric features, and $n$ is the number of features contributing to the RMSZ summary statistic.

RMSZ offers a robust, global measure of how well a structural model conforms to standard geometric restraints. It can be computed separately for individual geometric features (e.g. bond lengths, bond angles [50]), pooled across feature types, or extended to derived metrics such as Rama-Z [51]. Importantly, RMSZ is applicable across data obtained from diverse experimental methods, including X-ray crystallography [52] and cryoEM [53]. This versatility makes RMSZ a widely applicable and transferable tool for validating model geometry across a broad range of structural contexts.

In this work, we apply the same principle using the $Z'$ score, a variant tailored for nucleic acid geometries, and define the corresponding summary metric as RMSZ′. This serves as a single, global indicator of how closely a nucleic acid model conforms to its ideal geometry. The RMSZ′ score may be computed separately for bonds/angles, and for DNA/RNA, or combined into a single measure.

In the simple case that the $Z'$ scores are independent and identically distributed (i.i.d.) standard Normal variates, the expectation (mean) of RMSZ′ can be approximated as:

$$\mathbb{E}\left(RMSZ'\right) \approx 1 - \frac{1}{4n} + \mathcal{O}\left(n^{-2}\right)$$

This arises due to the relationship between RMSZ (thus also RMSZ′ in the case of i.i.d. standard Normal Z-scores) and the $\chi$ distribution. If $X \sim \chi(n)$, then we can express RMSZ as: $n^{-1/2}X$. The expectation of an $X$ variable is $\mathbb{E}(X) = 2^{1/2}\Gamma([n + 1]/2)/\Gamma(n/2)$, and Stirling's expansion for large $n$ gives: $\mathbb{E}(X) = n^{1/2}(1 - [4n]^{-1} + \mathcal{O}[n^{-2}])$, where $\Gamma$ is the gamma function, and $\mathcal{O}$ describes the asymptotic order of convergence. In the limit of large $n$:

$$\lim_{n\to\infty} \mathbb{E}\left(RMSZ'\right) = 1$$

Therefore, if $n$ is large—for instance, when computing RMSZ′ over all bonds or angles in a nucleic acid model—then it might be reasonable to expect the mean RMSZ′ to be approximately equal to 1. However, if $n$ is comparatively small—for instance, if there are only a few nucleotides in the model—then the mean RMSZ′ will be systematically less than 1. In such cases, it may be relevant to interpret RMSZ′ contextually, given the associated number of features $n$.

In some cases, the assumption of independence may be unreasonable, for example, if RMSZ′ is computed over multiple nucleotide types, then there may be internal correlations between the $Z'$ scores of same-type features, between features that are sequentially or spatially proximal, and/or dependent on similarity of atomic B-factors. As the average correlation between features increases, the mean RMSZ′ will decrease. The lower bound is reached in the extreme limit of all $Z'$ scores being perfectly positively correlated (i.e. all equal), in which case the expectation becomes:

$$\mathbb{E}\left(RMSZ'|Z'_i = Z'\right) = \mathbb{E}\left(|Z'|\right) = \sqrt{\frac{2}{\pi}} \approx 0.798$$

Since we know that Z′ is generally not normally distributed (see Materials and Methods), we should also contemplate the situation where the Z′ variates follow other distributions. While the mean RMSZ′ is approximately 1 when the underlying distribution is Normal, it will be systematically higher for distributions with high kurtosis (heavy tails) and lower for those with low kurtosis (light tails).

The degree to which the mean RMSZ′ deviates from that of the Normal case will depend on higher-order moments (i.e. the

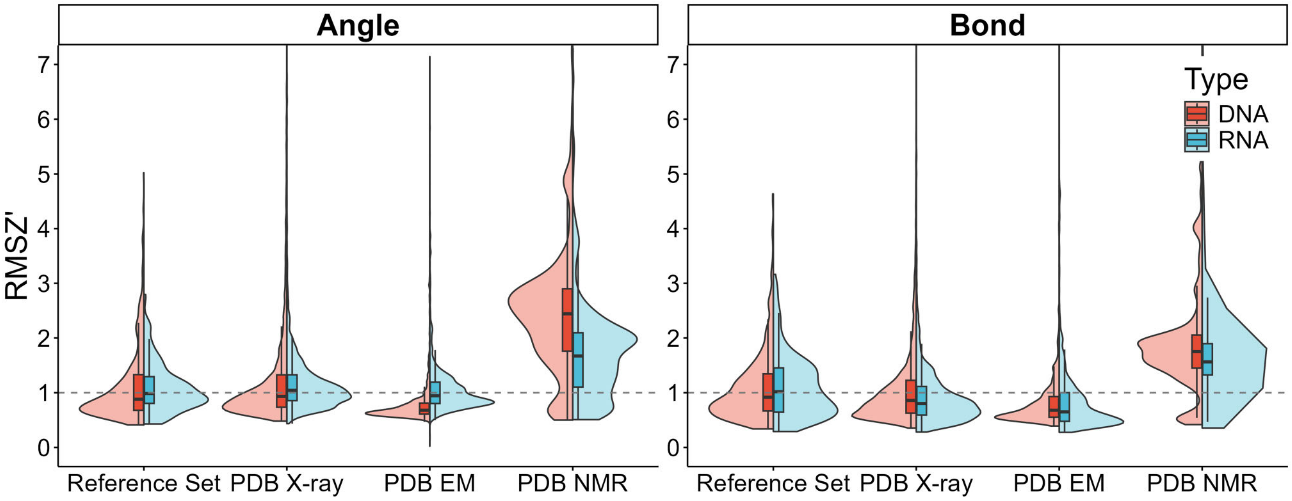


**Figure 7.** Violin plots representing the distributions of RMSZ′ scores across nucleic acid-containing models in the PDB. Separate distributions are shown for angles and bonds. The distributions of scores for models in the **PDB-NA Reference Set** are shown, as well as for all models in the PDB grouped by experiment type (X-ray, EM, and NMR). For each violin plot, data are further grouped by nucleic acid type—with DNA shown on the left (red) and RNA on the right (blue). The violin plots are capped at RMSZ′ = 7, with actual maximum RMSZ′ values for structures in the PDB exceeding 17 for X-ray, EM, and NMR.

exact shape) of the Z′ distribution. For example, the heavy-tailed t-distribution with 5 degrees of freedom—5 being the lowest integer for which finite kurtosis exists—yields an expected RMSZ′ of around 1.12. In contrast, the light-tailed Uniform distribution produces an expected RMSZ′ of approximately 0.92 (based on simulations).

Indeed, the "ideal" value of RMSZ′ is not precisely 1.0 but instead depends on the underlying distributional structure, and it is further reduced in the presence of internal correlations. Moreover, the variance of RMSZ′ is correlated with its expectation: distributions with heavier tails will result in greater uncertainty in the RMSZ′ estimate, a problem exacerbated when *n* is small.

A more detailed investigation of how data properties affect RMSZ′ calculations is beyond the scope of the present work. Nevertheless, appropriate caution should be exercised when directly comparing RMSZ′ scores between models with different compositions and feature counts (e.g. different combinations of bond/angle and nucleotide types). These variations imply a mix of underlying distributions with differing kurtoses and higher moments.

Pragmatically, a general rule of thumb is that RMSZ′ values around 1.0 are typical for well-refined models with geometries comparable to those in the **PDB-NA Reference Set.** Values significantly below 1.0 may indicate over-tightening of geometric restraints during refinement—potentially masking genuine deviations—whereas values above 1.0 may suggest overfitting, under-restrained refinement, modeling errors, or authentic structural deviations from idealized reference geometry. That said, high resolution does not guarantee model accuracy: even high-resolution structures may contain errors. Any individual **Of Concern** flag with elevated Z′ values should be scrutinized, regardless of the nominal resolution.

When examined in context, RMSZ′ distributions for X-ray models are broadly similar irrespective of whether they are part of the **PDB-NA Reference Set** (Fig. 7). In contrast, models derived from cryoEM experiments tend to show lower RMSZ′ values, possibly reflecting overly tight refinement protocols or sample bias. NMR ensembles, on the other hand, typically exhibit higher RMSZ′ distributions, which may be due to intrinsic variability or differences in refinement methodology.

There is also a notable temporal trend in RMSZ′ scores, particularly with a shift in the relative behavior of RNA versus DNA models around 2012 (Fig. 8). This may reflect changes in software defaults, refinement strategies, or shifts in the types of macromolecular systems being studied. Indeed, the number of nucleic acid-containing models deposited in the PDB has steadily increased in recent years (Fig. 9).

Interestingly, both bond and angle RMSZ′ values remain relatively consistent across resolution ranges (Fig. 10). Perhaps counterintuitively, high-resolution models often exhibit higher RMSZ′ values on average. This is likely because such models are of higher quality overall—capturing real deviations from ideal geometry—thereby increasing their apparent divergence from generalized averages. Medium-resolution models typically validate well, as this resolution range is the primary target of most refinement procedures. In contrast, very low-resolution models tend to perform poorly in validation metrics, likely due to inherent modeling inaccuracies.

## Implementation and analysis of the validation scheme

The validation scheme presented above has been implemented as a C++ library using the CSD data and **PDB-NA Reference Set** discussed in the Methods section. An example of full implementation is also available at https://dnatco.datmos.org/app, allowing detailed and interactive validation of nucleic acid geometry. The implemented system has been analyzed in terms of the agreement between the CSD and PDB **Preferred** ranges and the properties of the **Allowed** and **Of Concern** regions.

In most cases, the **Preferred** ranges defined by the CSD and PDB closely overlap (Fig. 11A). However, for some parameters, the CSD or PDB interval allows for a wider or shifted range of **Preferred** values (Fig. 11B). The most extreme example is the OP1–P–O3' angle, for which the CSD data considers multiple conformation groups defined by the $\zeta$ and $\alpha$

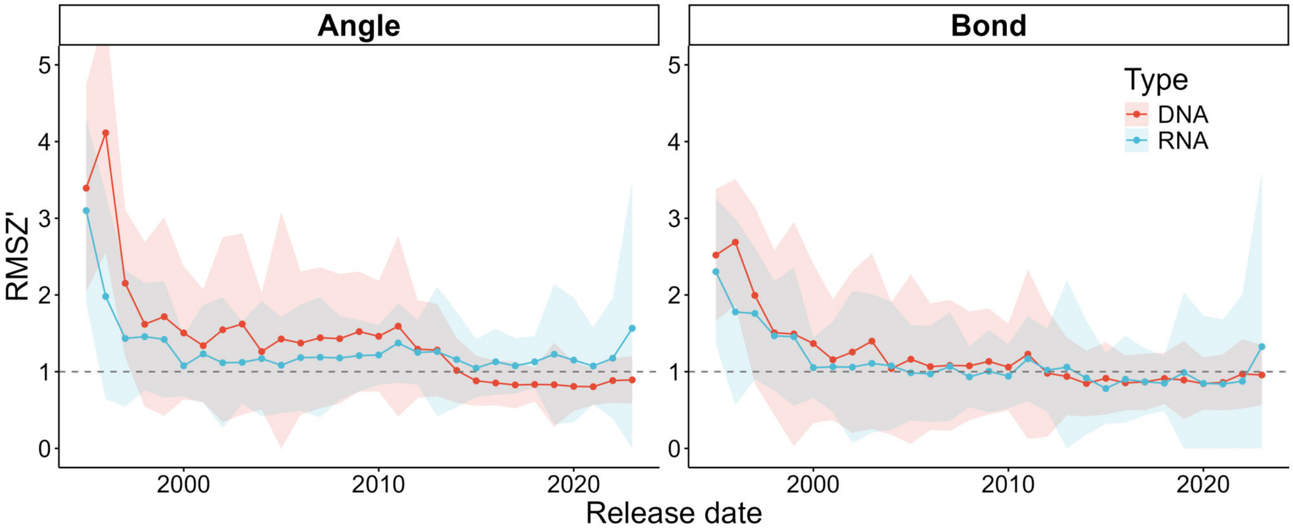


**Figure 8.** The evolution of RMSZ′ scores over time for all nucleic acid-containing models in the PDB derived from X-ray crystallographic data. Separate plots are shown for angles and bonds, and data are grouped by DNA (red) and RNA (blue). Connected points show the mean RMSZ′ score per year (according to release date), and the associated shaded areas represent ±1 standard deviation from the mean.

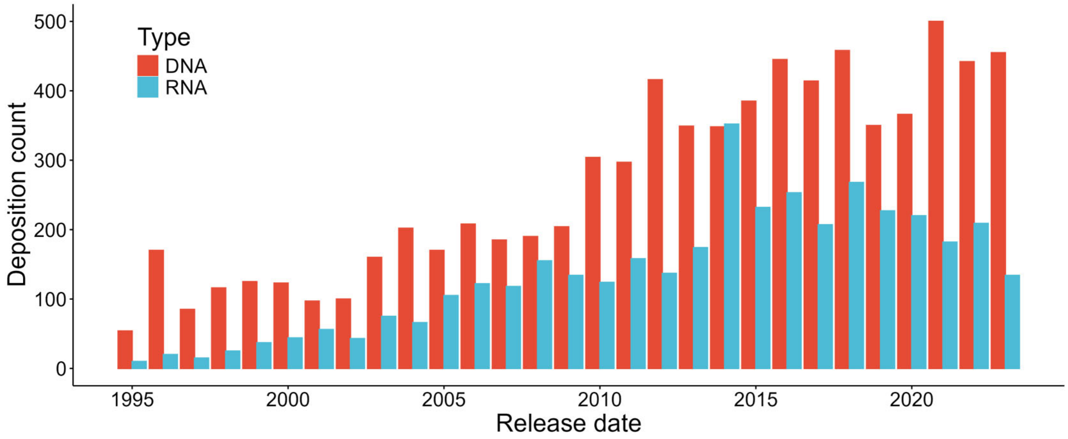


**Figure 9.** The number of nucleic acid-containing models in the PDB derived from X-ray crystallographic data per year (according to release date), grouped by DNA (red) and RNA (blue).

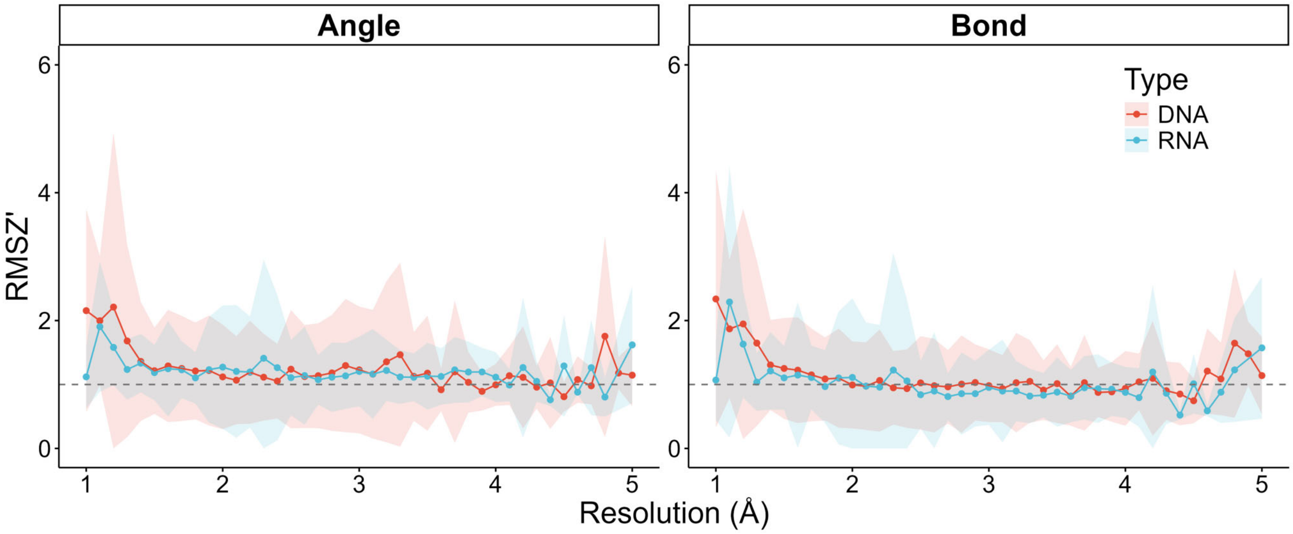


**Figure 10.** The relationship between RMSZ′ and nominal resolution for all nucleic acid-containing models in the PDB derived from X-ray crystallographic data. Separate plots are shown for angles and bonds, and data are grouped by DNA (red) and RNA (blue). Connected points show the mean RMSZ′ score per resolution bin (in 0.1 Å bins), and the associated shaded areas represent ± 1 standard deviation from the mean.

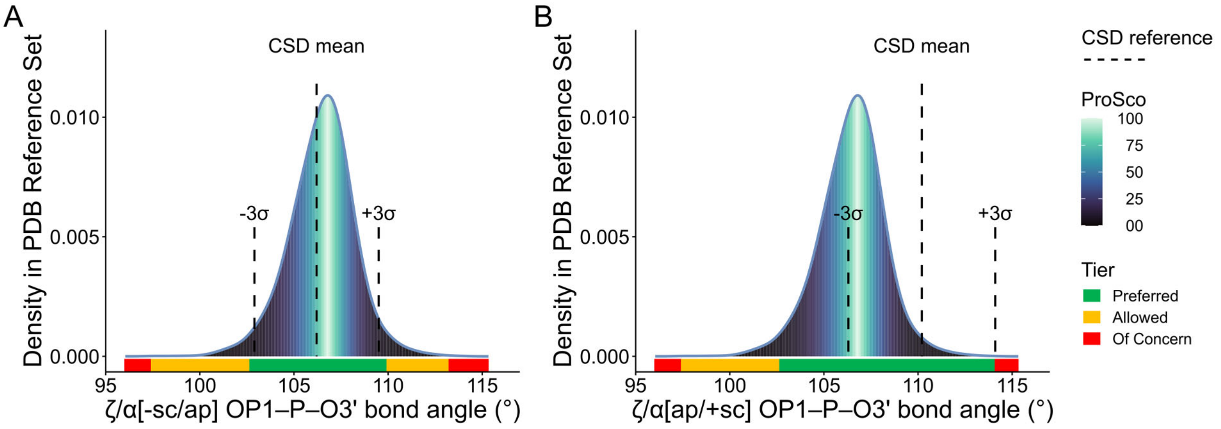


**Figure 11.** The three-tier validation scheme for the OP1–P–O3’ RNA backbone angle in two conformations of the phosphodiester group. Angle values falling into the **Preferred**, **Allowed**, and **Of Concern** tiers are denoted using green, yellow, and red bars above the horizontal axis. The density plot represents the probability percentile score (ProSco) of different values in the **PDB-NA Reference Set**. Ranges defined by the CSD data are marked by dashed lines. The threshold between **Preferred** and **Allowed** tiers utilizes information from both the conformation-dependent CSD data and the high-quality **PDB-NA Reference Set**. The definition of the threshold between the **Allowed** and **Of Concern** tiers is based solely on the **PDB-NA Reference Set**. (**A**) Example of a situation where 95% of the **PDB-NA Reference Set** and the CSD mean ± 3σ intervals almost perfectly overlap, for the ζ torsion angle –*synclinal* (–*sc*) and α torsion angle *antiperiplanar* (*ap*). (**B**) Example of a situation where the $\mu_{CSD} \pm 3\sigma_{CSD}$ span is shifted from the core 95% of data from the **PDB-NA Reference Set** for the ζ torsion angle *antiperiplanar* (*ap*) and α torsion angle +*synclinal* (+*sc*).

phosphate torsion angles. The CSD data corresponding to the *antiperiplanar* ζ and +*synclinal* α conformation extends the OP1–P–O3’ **Preferred** tier strongly to the right (Fig. 11B) compared to the –*synclinal* ζ and *antiperiplanar* α conformation commonly found in the PDB (Fig. 11A). In this case, the conformation-dependent CSD data allow angle values that, to date, have been rare in the PDB but are found in small-molecule structures. There are also cases where the CSD-defined range is tighter than the PDB-Preferred range. Overall, the Preferred ranges defined by the CSD and PDB show substantial agreement in the central portion of the distributions (Fig. 11). While some parameters exhibit differences in the tails or wider intervals in one dataset, the main bulk of the data overlaps, supporting the consistency of the two sources for validation purposes.

Although the conformation-dependence of many parameters is often apparent when looking at the CSD data [5, 7], it is usually not directly noticeable in the **PDB-NA Reference Set**. As discussed in Materials and Methods, the different refinement programs and the associated different restraint libraries are often responsible for the distribution of parameter values in PDB models (Fig. 4), especially for bond lengths. Even for bond angles, differences between conformations are rarely as large as 4° (Figs 11 and 12). As a result, the conformation-dependent subpopulations are hidden in the global value distribution of a given parameter, although they can be quite clear when plotted separately, as in Fig. 12. Nevertheless, the combination of CSD and PDB data ensures that the **Preferred** region encompasses data corresponding to different conformations as well as generated by different software packages.

## Demonstration of how the new validation scheme works for the whole PDB

To assess the impact of the proposed three-tier validation system, we applied it to all nucleic acids deposited to the PDB as of April 1, 2024. Table 1 compares the percentage of bond lengths and angles in the PDB that would be assigned to the **Of Concern** tier in the proposed validation scheme to the percentage of Outliers marked by the current PDB validation scheme. Notably, the proposed validation scheme assigns, on average, 1.9 times fewer bonds and angles to the worst category. The decrease is significant for both RNA and DNA, regardless of the experimental method (X-ray, cryoEM, NMR). It is worth remembering that bond and angle statistical distributions directly reflect how tightly software or users choose to restrain them. Since they are, therefore, a poor measure of structural accuracy, our goal was to minimize false alarms by setting more conservative thresholds, which we consider an improvement of the validation scheme.

The relative distributions of these statistics show informative patterns. Tabulation of the frequency of bonds and angles that would be assigned to each of the three tiers shows that, depending on the experimental method, 90–99% of parameter values would fall into the **Preferred** range of the presented scheme (Table 2). A clear distinction is visible between NMR and the other methods, with NMR structures having far fewer **Preferred** values (90–94%) than X-ray and cryoEM (97–99%) structures. We also note that, regardless of the experimental method, there is generally a higher percentage of rare (**Allowed** or **Of Concern**) bond angles than bond lengths.

We have also analyzed whether the resolution of deposited X-ray structures affects the proportions of bonds and angles falling into particular tiers. As can be seen from Table 3, there are slightly more bond and angle values classified as **Of Concern** in the highest-resolution bin ($d_{min}$ < 1.8 Å) compared to the lower-resolution groups. This is likely from a mixture of causes. Restraints are applied with less strength during refinement against high-resolution data, increasing the probability of seeing more “exotic” bond lengths and angles. Alternate conformations may be apparent, but difficult to model correctly, and uniformly loose restraints can cause clusters of enormous outliers at more conformationally flexible loops, termini, and side chains (see Discussion). Nevertheless, even in the set of high-resolution X-ray structures, over 95% of the parameter values fall into the **Preferred** tier.

Finally, we have studied the effect of the refinement software in intervals of the deposition date on the validation re-

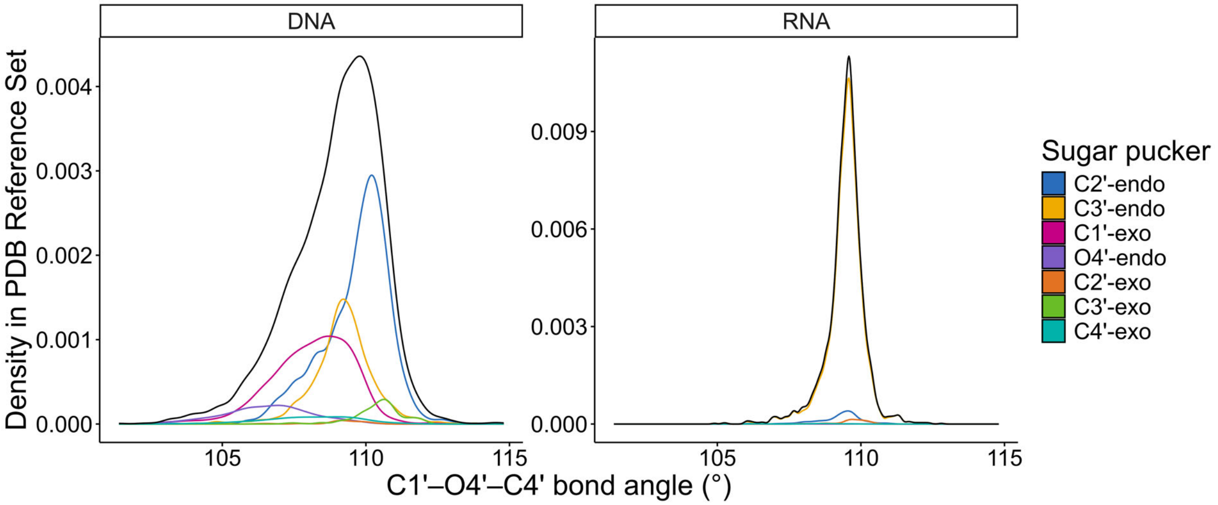


**Figure 12.** Distribution of the C1′-O4′-C4′ sugar bond angle in DNA and RNA structures in the **PDB-NA Reference Set**. The black line represents the overall distribution, whereas the coloured lines represent different sugar puckers. The distribution of the C1′-O4′-C4′ angle in DNA is pucker-dependent. However, the mode values of two dominant puckers, C2′-endo (blue) and C3′-endo (yellow), do not differ enough to produce distinct peaks in the overall distribution. In addition, a small number of sugars with rarer puckers, such as O4′-endo, with significantly different mode values, do not have reliable probability distributions. Numbers of other-than C3′-endo puckers in RNA are small and not sufficiently different from C3′-endo distributions.

**Table 1.** Comparison of validation results for the entire PDB, based on the validation protocol currently used by the PDB and the proposed three-tier system

| | | Bonds | | | Angles | | |
|---|---|---|---|---|---|---|---|
| NA type | Method | Count | Of Concern (%) | Current PDB outliers (%) | Count | Of Concern (%) | Current PDB outliers (%) |
| DNA | X-ray | 8.1M | 0.5 | 0.9 | 12.5M | 0.7 | 2.0 |
| | cryoEM | 9.0M | 0.5 | 1.3 | 13.9M | 0.8 | 1.8 |
| | NMR | 5.0M | 2.2 | 3.4 | 7.6M | 4.7 | 7.1 |
| RNA | X-ray | 92.0M | 0.2 | 0.3 | 143.5M | 0.5 | 1.2 |
| | cryoEM | 132.5M | 0.6 | 0.6 | 206.6M | 0.6 | 1.1 |
| | NMR | 6.2M | 0.9 | 0.9 | 9.6M | 2.3 | 4.3 |

**Table 2.** Percentages of PDB bond lengths and angles falling into the **Preferred**, **Allowed**, and **Of Concern** tiers of the proposed validation system, depending on the nucleic acid type (DNA, RNA) and experimental method (X-ray, cryoEM, NMR)

| | | Bonds | | | | Angles | | | |
|---|---|---|---|---|---|---|---|---|---|
| NA type | Method | Count | Preferred (%) | Allowed (%) | Of Concern (%) | Count | Preferred (%) | Allowed (%) | Of Concern (%) |
| DNA | X-ray | 8.1M | 98.3 | 1.2 | 0.5 | 12.5M | 97.8 | 1.5 | 0.7 |
| | cryoEM | 9.0M | 97.6 | 1.9 | 0.5 | 13.9M | 98.2 | 1.0 | 0.8 |
| | NMR | 5.0M | 92.3 | 5.5 | 2.2 | 7.6M | 89.9 | 5.4 | 4.7 |
| RNA | X-ray | 92.0M | 99.0 | 0.8 | 0.2 | 143.6M | 97.1 | 2.4 | 0.5 |
| | cryoEM | 132.5M | 98.2 | 1.2 | 0.6 | 206.6M | 97.6 | 1.7 | 0.6 |
| | NMR | 6.2M | 94.0 | 5.1 | 0.9 | 9.6M | 91.8 | 5.9 | 2.3 |

sults for bonds and angles in PDB X-ray structures. As a cutoff date for comparison, we chose 2010, the year the PDB introduced Validation Reports. As Table 4 shows, in most cases, structures deposited in 2010 or later have fewer **Of Concern** bond lengths and angles. The only exceptions are RNA bonds refined using Phenix and Other software. After 2010, Phenix was used with many very large RNA structures. Note that amongst the software in the "Other" group, CNS has had a small profile since 2010, and SHELXL is only used at high resolution.

It should be noted that the data reported in Table 4 may suffer from varying levels of statistical bias, as different software packages have used different restraint targets and weights, which may be more or less similar to the targets used in the validation system proposed herein. Consequently, the varying percentages reported for different software tools should not be misinterpreted as reflecting systematic differences in model quality. Such issues are always a concern when model geometry is restrained to the same values as are used for subsequent model validation, or when there are concerns regarding selection bias. Rather, Table 4 should be interpreted as simply giving an indication of how validation performance may reasonably vary contingent on such factors. Indeed, a high proportion of **Preferred** bonds/angles does not in itself imply that

**Table 3.** Percentages of PDB bonds and angles falling into the **Preferred**, **Allowed**, and **Of Concern** tiers of the proposed validation system, divided into nucleic acid type and resolution ranges of the deposited X-ray structural models

| | | Bonds | | | | Angles | | | |
|---|---|---|---|---|---|---|---|---|---|
| NA type | X-ray resol. (Å) | Count | Preferred (%) | Allowed (%) | Of Concern (%) | Count | Preferred (%) | Allowed (%) | Of Concern (%) |
| DNA | <1.8 | 747k | 96.1 | 2.6 | 1.4 | 1.1M | 95.9 | 2.6 | 1.5 |
| | 1.8–2.4 | 2.3M | 98.2 | 1.2 | 0.5 | 3.5M | 97.6 | 1.6 | 0.8 |
| | 2.4–3.0 | 3.0M | 98.7 | 1.0 | 0.3 | 4.6M | 98.0 | 1.4 | 0.6 |
| | 3.0–3.5 | 1.2M | 98.9 | 0.9 | 0.2 | 1.9M | 98.4 | 1.2 | 0.4 |
| | >3.5 | 868k | 98.4 | 1.2 | 0.4 | 1.3M | 98.8 | 0.9 | 0.3 |
| RNA | <1.8 | 322k | 97.6 | 1.8 | 0.6 | 499k | 96.7 | 2.8 | 0.5 |
| | 1.8–2.4 | 3.8M | 99.6 | 0.3 | 0.1 | 6.0M | 98.7 | 1.2 | 0.2 |
| | 2.4–3.0 | 32.1M | 99.1 | 0.7 | 0.2 | 50.1M | 97.4 | 2.2 | 0.4 |
| | 3.0–3.5 | 39.6M | 98.9 | 0.9 | 0.2 | 61.8M | 96.7 | 2.7 | 0.6 |
| | >3.5 | 16.2M | 98.9 | 0.8 | 0.3 | 25.3M | 97.1 | 2.3 | 0.6 |

**Table 4.** Percentages of PDB bonds and angles falling into the **Preferred**, **Allowed**, and **Of Concern** tiers for DNA and RNA X-ray structures determined by different refinement programs before and after 2010. Software in the "Other" category includes CNS, BUSTER, X-PLOR, SHELX/L, TNT, PROLSQ/NUCLSQ, and situations where more than one program was used

| | | | Bonds | | | | Angles | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| NA type | Software | Deposition date | Count | Preferred (%) | Allowed (%) | Of Concern (%) | Count | Preferred (%) | Allowed (%) | Of Concern (%) |
| DNA | Phenix | <2010 | 49k | 99.1 | 0.7 | 0.1 | 76k | 96.4 | 2.9 | 0.7 |
| | | ≥2010 | 2.7M | 99.0 | 0.8 | 0.1 | 4.2M | 99.2 | 0.7 | 0.1 |
| | Refmac5 | <2010 | 338k | 96.1 | 2.8 | 1.1 | 519k | 93.2 | 4.6 | 2.2 |
| | | ≥2010 | 1.3M | 98.8 | 0.9 | 0.3 | 1.9M | 97.9 | 1.6 | 0.5 |
| | Other | <2010 | 1.3M | 96.0 | 2.4 | 1.6 | 2.0M | 94.5 | 3.2 | 2.3 |
| | | ≥2010 | 456k | 99.0 | 0.8 | 0.2 | 702k | 98.8 | 0.9 | 0.3 |
| RNA | Phenix | <2010 | 280k | 99.4 | 0.5 | 0.1 | 437k | 97.1 | 2.6 | 0.3 |
| | | ≥2010 | 28.6M | 99.2 | 0.7 | 0.2 | 44.7M | 97.2 | 2.4 | 0.4 |
| | Refmac5 | <2010 | 330k | 97.8 | 1.6 | 0.6 | 514k | 95.4 | 3.6 | 1.0 |
| | | ≥2010 | 590k | 98.8 | 0.9 | 0.3 | 917k | 97.0 | 2.5 | 0.5 |
| | Other | <2010 | 5.5M | 99.6 | 0.3 | 0.1 | 8.6M | 99.0 | 0.9 | 0.2 |
| | | ≥2010 | 1.2M | 97.9 | 1.5 | 0.5 | 1.9M | 99.3 | 0.7 | 0.1 |

a model is "good"; such behavior can be artificially enforced by the use of overly tight restraints during refinement (which may even enforce an incorrect conformer) and thus simultaneous consideration of other validation metrics is required in order to support or contradict a model hypothesis (e.g. global agreement with data, local fit-to-map, etc.). We emphasize that the purpose of model validation is generally to draw the attention of the depositor or downstream interpreter to unusual features that are worthy of closer inspection.

## Discussion

The Nucleic Acid Valence Geometry Working Group has shown that there are many places where nucleic acid bond and angle target values, as well as their allowable variations, can now be improved over the older values of Gelbin *et al.* [13], Clowney *et al.* [10], and Parkinson *et al.* [16]. This is enabled by the availability of prolific new quality-curated high-resolution data, both in the CSD [5–7] and in the PDB.

There are two main applications for these revised targets: structure model refinement and validation, which have somewhat different requirements. In general, model refinement converges better with more accurate restraint targets, but in practice, this is less straightforward. Bond and angle restraints interact with other restraints in model refinement, and their effect varies with resolution. Modeling with conformation-dependent parameters is often not feasible due to limitations in the software and, more importantly, because initially the correct conformation may not be known. Given that appropriate use of conformation-dependence in refinement will vary, recommendations for restraints to be used in refinement were kept outside the scope of the Working Group.

The main focus of the Working Group has been on setting improved standards for validation of nucleic acid structural models, primarily for use by the wwPDB [14]. For most model geometry measures, the current PDB validation assessments are *preferred*, *allowed*, or *outlier*. They are applied both overall and per-residue, with *outlier* assigned only to cases that are deemed likely to have incorrect stereochemistry and thus require closer inspection. Most residues flagged as outliers are indeed incorrectly modeled, but the few genuine ones are nearly always chemically or biologically important [54]. The presumption that a structural model with only a few outliers is good in general has never been true for bonds and angles because their values are explicitly restrained in refinement [55]. Currently, this is also the case for some other measures, now also frequently restrained, such as Ramachandran angles, leading to artificial masking of persistent stereochemical problems [51, 56]. In line with the Principle of Parsimony, valence geometry (especially bond lengths) needs to be restrained tightly at low resolution, where the experimental evidence is too limited to support any strong deviations. However, even at

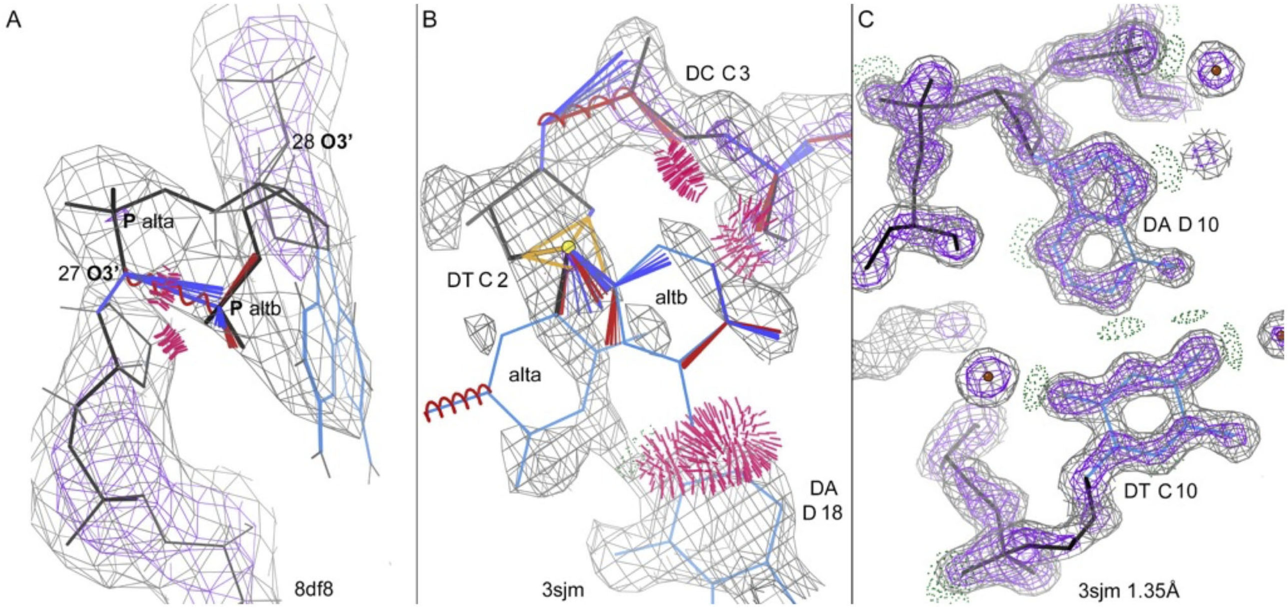


**Figure 13.** Extreme **Of Concern** clusters, most common at high resolution. (**A**) An RNA alternate conformation in PDB ID 8df8 topoisomerase at 2.92 Å [60]. It has clashes, four bad bond angle outliers and an absurd bond length of 2.56 Å (for O3′ 27 to P 28 alt b); it would seem that no restraint was applied to bond and angle parameters that involve the O3′-P bond in alternate b. (**B**) A DNA chain terminus in PDB ID 3sjm [61] with very poor and patchy electron density (gray 2mFo-DFc map contoured at 0.8σ), perhaps initially fit with plausible conformation but refined with no or very weak restraints, resulting in clashes, a chiral outlier, bond lengths up to 2.42 Å (O3′-P), and bond angles as low as 69° (O4′-C1-N1 alt a). (**C**) A typical, excellent part of the same 1.35 Å 3sjm structure model, with 2mFo-DFc map contoured at 1.2σ (gray) and 3σ (purple). Bad bonds are marked with spirals and bad angles with fans, red if too big and blue if too small. All-atom clashes are groups of hot-pink spikes, assumed hydrogen bonds are pillows of green dots, and the chiral outlier is in dark yellow.

higher resolution, it is possible to keep the geometry so tight that outliers are nearly impossible and the underlying problems are just pushed into other features, such as atomic B-factors, conformation, lack of favorable interactions, and/or fit-to-map [57].

Valence geometric validation is, therefore, of limited use in finding incorrect conformations. However, the PDB should flag suspicious residues when they do occur, typically due to the use of overly loose restraints in patchy or even absent electron density or from other inappropriate procedural choices. Conversely, validation should not penalize individual structures for reasonable differences between validation and restraint targets due to the use of different refinement software. With increasing requirements for, and availability of, automation, software defaults should be evaluated even more carefully.

Even though the **PDB-NA Reference Set** was designed to comprise only high-quality structural models, the resultant bond and angle distributions are heterogeneous in nature. Rather than representing the natural geometric variations for a particular chemical environment, with associated random errors, many of the empirical distributions comprise multiple subpopulations. There are a variety of potential reasons for such behavior, including some that are relevant to nucleic acid structure (e.g. differences in conformation, protonation, biologically relevant induced strain, etc.) and some that are simply a consequence of specific crystallographic software implementations (i.e. program packages or sources of prior information used to refine the models) rather than fundamental to the crystal structure itself. The presence of multiple subpopulations has a substantial effect on distributional skewness and kurtosis, and in some cases results in clear multimodality of the empirical distributions. Manual inspection of some of the most extreme cases revealed that it was not possible to straightforwardly identify any problems with these models, nor to figure out how such models might be adjusted to conform to more typical geometry while still retaining good fit to the electron density. It is thus necessary to validate such geometric features using an approach that is insensitive to distributional form and robust to outliers.

Our new validation standards were designed to achieve a good balance, and especially to minimize false alerts, by using the following strategy: 1) Setting the **Preferred/Allowed** boundary as the combined spread of ± 3σ around the CSD mean plus the range that includes 95% of the **PDB-NA Reference Set** data. 2) Setting the **Allowed/Of Concern** boundary to a weighted asymmetric non-parametric standard score of $Z' = \pm 5$ (see Methods and Results Sections). This strategy will allow for differences between extant (and perhaps also future) software target values, including whether ribose-pucker and/or phosphodiester-conformation dependence are used. Relatively few demonstrably correct cases should thus fall into the **Of Concern** category.

There are two major issues regarding the applicability of validating conformation-dependent differences in otherwise equivalent bond lengths or angles. The first concerns the CSD-established conformation-dependence of backbone geometry around the phosphate group, taken by necessity from a variety of small phosphodiester molecules [7]. We do not see such dependence in our polymeric **PDB-NA Reference Set**, but it may turn out to be visible in the future for the most affected, relatively rare conformers, given more high-resolution data. That dependence is present in the phosphate parameter CSD data used for the doubly defined **Preferred/Allowed** boundary, but it will not penalize software without that feature.

The other issue involves the effect of incorrectly modeled RNA ribose pucker on pucker-specific valence geometry in and around the ribose moiety. From the CSD pucker-specific geometry, targets were derived for RNA [7], but applying them to lower-resolution PDB structures is not straightforward because they need to be chosen according to the correct pucker, not the modeled pucker. In Phenix, the "Pperp" criterion (the perpendicular distance from the following P atom to the extended glycosidic bond direction) [58] can provide robust diagnosis of the correct C3'-endo vs C2'-endo ribose pucker because it relies only on features that are seen well up to ∼3.5 Å resolution. This feature enables pucker-specific targets to be used in refinement [19], which improves the behavior of RNA valence geometry to levels comparable with those typical for proteins [59]. However, this is not currently implemented in other software and does not apply to DNA.

For DNA, puckers are more variable due to the increased conformational freedom of the deoxyribose compared to the ribose ring (see Fig. 12). In the future, if and when Pperp-corrected ribose-pucker restraints have become a standard practice, or if more generalized corrected-pucker restraints (e.g. those available at dnatco.datmos.org for the DNA and RNA NtC conformational classes) are in regular use [37, 38], this should be reconsidered for validation as well.

Suboptimal refinement procedures can produce local clusters of large clashes and enormous geometry outliers (sometimes bond length > 0.5 Å or angle > 15° from expectation) – rather unintuitively, nearly always at high resolution. For nucleic acid bonds and angles in experimental structures (as well as for proteins), there are two common causes of these undesirable clusters [54]: 1) ending alternate conformations too soon for legitimate geometry to be possible (as in Fig. 2B), or with an erroneously missing restraint (as for alternate b in Fig. 13A). The residue boundary between O3' and P is the usual default but is seldom a good choice; 2) allowing unreasonable models in poor electron density to develop during refinement, by greatly down-weighting restraints uniformly everywhere (as in Fig. 13B). These rare but very extreme cases presumably force the more frequent **Of Concern** values seen at high resolution (Table 3), even though nearly all other parts of those structures are excellent, as exemplified in Fig. 13C. We considered classifying such extreme outliers as a $4^{th}$ tier in our proposed new system but decided to stay with the simpler three-tier scale, as such outliers should be obvious even without a separate tier.

As recommended by the respective PDB Validation Task Forces [62–64], and as is the case for current PDB model validation, the newly proposed validation criteria for nucleic acids would be applicable not only to X-ray, neutron, or microED crystal structures, but also to NMR and cryoEM structural models.

Even now, there are not enough high-quality datasets in the PDB to answer all our questions, especially for rare conformations. Despite that, we have used the much more numerous and higher-quality data from the CSD alongside data from the PDB to construct a new valence-geometry validation system. After 30 years since the last standard was introduced, we present a suitably improved approach to validation of the structural details of nucleic acid models, comprising more robust validation scores, ProSco and *Z*', as well as a three-tier classification scheme. The Working Group recommends this scheme for use as part of PDB model validation.

The curated **PDB-NA Reference Set** of high-quality nucleic acid-containing structural models from the PDB, created for this project (see Materials and Methods), has been made available for use in other projects.

## Conclusions and Recommendations

In conclusion, the Working Group on Nucleic Acid Valence Geometry proposes the following recommendations for consideration by the wwPDB when implementing a revised structure validation system for RNA and DNA bond lengths and angles.

**R1.** Use a three-tier scale with intervals for **Preferred**, **Allowed**, and **Of Concern** values of each specific bond or angle (see Section *Three-tier nucleic acid validation scheme*).

**R2.** The values of the ProSco and *Z*' scores should be computed and provided for each relevant geometric feature (see Sections *Definition of probability percentile scores* and *A robust standard score to measure deviance from typical distributions*) to avoid information loss, facilitate downstream analysis, and provide maximal feedback.

**R3.** The RMSZ' score should be computed and provided for each nucleic acid-containing structural model (see Section *Assessing overall nucleic acid model geometric quality*).

**R4.** In the case of observing an **Of Concern** bond or angle, links to appropriate resources should be provided to facilitate subsequent diagnostic analysis. Specifically, references to: (i) **Of Concern** bonds and angles observed in the **PDB-NA Reference Set** (see Data Availability); and (ii) the https://dnatco.datmos.org/app web server should be included.

## Acknowledgements

Dedicated to the memory of Dr. Olga Kennard. The authors would like to acknowledge and thank Paul Emsley, Garib Murshudov, and Gary Parkinson for discussions during the earlier stages of this project.

*Author contributions*: J.Č. (Conceptualization [equal], Data curation [lead], Formal Analysis [lead], Funding acquisition, Investigation [equal], Methodology [equal], Project administration [equal], Resources, Software, Validation [equal], Visualization [equal], Writing – original draft [equal], Writing – review & editing [equal]), R.A.N. (Conceptualization [equal], Formal Analysis [lead], Investigation [equal], Methodology [lead], Project administration [equal], Validation [equal], Visualization [equal], Writing – original draft [equal], Writing – review & editing [equal]), D.B. (Conceptualization [equal], Data curation [lead], Formal Analysis [lead], Investigation [equal], Methodology [equal], Project administration [equal], Software [equal], Validation [equal], Visualization [equal], Writing – original draft [equal], Writing – review & editing [equal]), H.M.B. (Conceptualization [equal], Formal Analysis [equal], Investigation [equal], Methodology [equal], Project administration [equal], Supervision [equal], Writing – original draft [equal], Writing – review & editing [equal]), M.G. (Conceptualization [equal], Investigation [equal]), R.P.J. (Conceptualization [equal], Formal Analysis [equal], Investigation [equal], Methodology [equal], Software [equal], Validation [equal], Writing – original draft [equal], Writing – review & editing [equal]), M.K. (Conceptualization [equal], Data curation [equal], Formal Analysis [equal], Investigation [equal], Validation [equal], Visualization [equal], Writing – original draft [equal]), C.L.L. (Conceptualization [equal], Data curation [equal], Formal Analysis [equal], Funding acquisition [equal], Investigation [equal], Methodology [equal], Software [equal], Validation [equal], Visualization [equal], Writing – original draft [equal], Writing – review & editing [equal]), N.W.M. (Conceptualization [equal], Data curation [equal], Formal Analysis [equal], Investigation [equal], Methodology [equal], Resources [equal], Software [equal], Validation [equal], Visualization [equal], Writing – original draft [equal]), J.S.R. (Conceptualization [equal], Data curation [equal], Formal Analysis [equal], Investigation [equal], Methodology [equal], Software [equal], Validation [equal], Visualization [equal], Writing – original draft [equal], Writing – review & editing [equal]), B.S. (Conceptualization [equal], Data curation [equal], Formal Analysis [equal], Funding acquisition [equal], Investigation [equal], Methodology [equal], Software [equal], Supervision [equal], Validation [equal], Visualization [equal], Writing – original draft [equal]), C.V. (Conceptualization [equal], Formal Analysis [equal], Investigation [equal], Methodology [equal], Software [equal], Validation [equal], Visualization [equal], Writing – original draft [equal], Writing – review & editing [equal]), C.J.W. (Conceptualization [equal], Formal Analysis [equal], Investigation [equal], Methodology [equal], Validation [equal], Visualization [equal], Writing – original draft [equal]), M.J. (Conceptualization [equal], Data curation [equal], Formal Analysis [equal], Investigation [equal], Methodology [equal], Project administration [equal], Software [equal], Supervision [equal], Validation [equal], Visualization [equal], Writing – original draft [equal], Writing – review & editing [equal]), and M.E. (Conceptualization [equal], Project administration [equal], Validation [equal], Visualization [equal], Writing – original draft [equal], Writing – review & editing [equal])

## Conflict of interest

None declared.

## Funding

R.P.J. was supported by the Horizon 2020 research and innovation programme [grant no. 871037] (iNEXT-Discovery) and by an institutional grant of the Dutch Cancer Society and of the Dutch Ministry of Health, Welfare and Sport. B.S. and J.C. were supported by the grant to the Institute of Biotechnology of the Czech Academy of Sciences [RVO 86652036]. This work was initiated during discussions of the steering committee of the 3D-Bioinfo ELIXIR community. Part of this work was carried out with the support of ELIXIR CZ Research Infrastructure (ID LM2023055, MEYS CR). C.L.L. was supported by the US National Institutes of Health [R01 GM085328]. J.S.R. and C.J.W. were supported by the US National Institutes of Health grants [R01 GM063210] and [R35 GM131883], and by a share of the Phenix industrial royalties. R.A.N. is supported by the Ada Lovelace Centre (ALC), part of the Science and Technology Facilities Council (STFC), part of the UK Research and Innovation (UKRI), and was supported by BBSRC [grant No. BB/S007083/1]. Funding to pay the Open Access publication charges for this article was provided by: The corresponding author is a member of the NAR editorial board and has the option of one publication in the journal per year at no cost.

## Data availability

The data and code generated and analyzed during this study are openly available in the Zenodo repository at https://doi.org/10.5281/zenodo.15804875. The Zenodo entry includes:

The complete PDB-NA Reference Set

Thresholds for the weighted asymmetric non-parametric standard score (Z′)

A listing of geometrical restraints for nucleic acid bond lengths and angles found in the literature and refinement programs

The code for filtering the PDB-NA Reference Set and for calculating the probability percentile score (ProSco)

A list of residues excluded from the PDB-NA Reference Set after manual inspection

Table with PDB-wide summary of the fractions of the lower and upper boundaries between the Preferred and Allowed tiers, defined by the CSD $\mu\pm3\sigma$ criterion rather than by ProSco 5 values

Html report with the visualizations used to inspect the effect of different Z′ thresholds

ProSco values in JSON format for all analyzed bond lengths and angles

## References

1. Groom CR, Bruno IJ, Lightfoot MP *et al.* The Cambridge Structural Database. *Acta Crystallogr B Struct Sci Cryst Eng Mater* 2016;72:171–9. https://doi.org/10.1107/S2052520616003954
2. Berman HM. The Protein Data Bank. *Nucleic Acids Res* 2000;28:235–42. https://doi.org/10.1093/nar/28.1.235
3. Berman HM, Olson WK, Beveridge DL *et al.* The nucleic acid database. A comprehensive relational database of three-dimensional structures of nucleic acids. *Biophys J* 1992;63:751–9. https://doi.org/10.1016/S0006-3495(92)81649-1
4. Lawson CL, Berman HM, Chen L *et al.* The Nucleic Acid Knowledgebase: a new portal for 3D structural information about nucleic acids. *Nucleic Acids Res.* 2024;52:D245–54. https://doi.org/10.1093/nar/gkad957
5. Kowiel M, Brzezinski D, Jaskolski M. Conformation-dependent restraints for polynucleotides: I. Clustering of the geometry of the phosphodiester group. *Nucleic Acids Res* 2016;44:8479–89. https://doi.org/10.1093/nar/gkw717
6. Gilski M, Zhao J, Kowiel M *et al.* Accurate geometrical restraints for Watson–Crick base pairs. *Acta Crystallogr B Struct Sci Cryst Eng Mater* 2019;75:235–45. https://doi.org/10.1107/S2052520619002002
7. Kowiel M, Brzezinski D, Gilski M *et al.* Conformation-dependent restraints for polynucleotides: the sugar moiety. *Nucleic Acids Res* 2020;48:962–73. https://doi.org/10.1093/nar/gkz1122
8. Berkholz DS, Shapovalov MV, Dunbrack RL *et al.* Conformation dependence of backbone geometry in proteins. *Structure* 2009;17:1316–25. https://doi.org/10.1016/j.str.2009.08.012
9. Schneider B, Bruno I, Burley SK *et al.* Nucleic Acid Valence Geometry Working Group. *IUCr Newsletter* 2020;28:16. http://iucr.org/news/newsletter/etc/articles?issue=150473&result_138339_result_page=16
10. Clowney L, Jain SC, Srinivasan AR *et al.* Geometric parameters in nucleic acids: nitrogenous bases. *J Am Chem Soc* 1996;118:509–18. https://doi.org/10.1021/ja952883d
11. Taylor R, Kennard O. The molecular structures of nucleosides and nucleotides. *J Mol Struct* 1982;78:1–28. https://doi.org/10.1016/0022-2860(82)85306-4

12. Murray-Rust P, Motherwell S. Computer retrieval and analysis of molecular geometry. III. Geometry of the β-1’-aminofuranoside fragment. *Acta Crystallogr B Struct Sci* 1978;34:2534–46. https://doi.org/10.1107/S0567740878008559
13. Gelbin A, Schneider B, Clowney L *et al.* Geometric parameters in nucleic acids: sugar and phosphate constituents. *J Am Chem Soc* 1996;118:519–29. https://doi.org/10.1021/ja9528846
14. Gore S, Sanz García E, Hendrickx PMS *et al.* Validation of structures in the protein data bank. *Structure* 2017;25:1916–27. https://doi.org/10.1016/j.str.2017.10.009
15. Brunger AT. *X-PLOR version 3.1: A system for X-ray crystallography and NMR*. New Haven, CT: Yale University Press. 1993.
16. Parkinson G, Vojtechovsky J, Clowney L *et al.* New parameters for the refinement of nucleic acid-containing structures. *Acta Crystallogr D Biol Crystallogr* 1996;52:57–64. https://doi.org/10.1107/S0907444995011115
17. Brunger AT. Version 1.2 of the crystallography and NMR system. *Nat Protoc* 2007;2:2728–33. https://doi.org/10.1038/nprot.2007.406
18. Saenger W. *Principles of nucleic acid structure*. New York, NY: Springer New York. 1984.
19. Liebschner D, Afonine PV, Baker ML *et al.* Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Crystallogr D Struct Biol* 2019;75:861–77. https://doi.org/10.1107/S2059798319011471
20. Davis IW, Leaver-Fay A, Chen VB *et al.* MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res* 2007;35:W375–83. https://doi.org/10.1093/nar/gkm216
21. Vagin AA, Steiner RA, Lebedev AA *et al.* REFMAC5 dictionary: organization of prior chemical knowledge and guidelines for its use. *Acta Crystallogr D Biol Crystallogr* 2004;60:2184–95. https://doi.org/10.1107/S0907444904023510
22. Murshudov GN, Skubák P, Lebedev AA *et al.* REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr D Biol Crystallogr* 2011;67:355–67. https://doi.org/10.1107/S0907444911001314
23. Joosten RP, Salzemann J, Bloch V *et al.* PDB_REDO: automated re-refinement of X-ray structure models in the PDB. *J Appl Crystallogr* 2009;42:376–84. https://doi.org/10.1107/S0021889809008784
24. Long F, Nicholls RA, Emsley P *et al.* AceDRG: a stereochemical description generator for ligands. *Acta Crystallogr D Struct Biol* 2017;73:112–22. https://doi.org/10.1107/S2059798317000067
25. Gražulis S, Daškevič A, Merkys A *et al.* Crystallography Open Database (COD): an open-access collection of crystal structures and platform for world-wide collaboration. *Nucleic Acids Res* 2012;40:D420–7. https://doi.org/10.1093/nar/gkr900
26. Nicholls RA, Joosten RP, Long F *et al.* Modelling covalent linkages in CCP4. *Acta Crystallogr D Struct Biol* 2021;77:712–26. https://doi.org/10.1107/S2059798321001753
27. de Vries I, Kwakman T, Lu X-J *et al.* New restraints and validation approaches for nucleic acid structures in PDB-REDO. *Acta Crystallogr D Struct Biol* 2021;77:1127–41. https://doi.org/10.1107/S2059798321007610
28. Bricogne G, Blanc E, Brandl M *et al.* *BUSTER* 2017; Cambridge, UK: Global Phasing Ltd.
29. Tronrud DE. TNT refinement package. In: Carter CW Jr., Sweet RM, *Methods in Enzymology*. Cambridge, MA, USA: Academic Press, 1997, pp.306–19.
30. Sheldrick GM. Crystal structure refinement with SHELXL. *Acta Crystallogr C Struct Chem* 2015;71:3–8. https://doi.org/10.1107/S2053229614024218
31. Jain S, Richardson DC, Richardson JS. Computational methods for RNA structure validation and improvement. In: Woodson SA, Allain FHT (eds.), *Methods in Enzymology*, Vol. 558, Cambridge, MA: Academic Press, 2015, 181–212. ISSN 0076-6879, ISBN 9780128019344https://doi.org/10.1016/bs.mie.2015.01.007
32. Bruno IJ, Cole JC, Edgington PR *et al.* New software for searching the Cambridge Structural Database and visualizing crystal structures. *Acta Crystallogr B Struct Sci* 2002;58:389–97. https://doi.org/10.1107/S0108768102003324
33. Iglewicz B, Hoaglin DC. *Volume 16: How to detect and handle outliers*. Milwaukee, WI: American Society for Quality Control, Statistics Division. Quality Press. 1993.
34. Černý J. Supplementary data for ‘new targets and procedures for validating the valence geometry of nucleic acid structures’. 2025. https://doi.org/10.5281/zenodo.15804876
35. Roll J, Zirbel CL, Sweeney B *et al.* JAR3D Webserver: scoring and aligning RNA loop sequences to known 3D motifs. *Nucleic Acids Res* 2016;44:W320–7. https://doi.org/10.1093/nar/gkw453
36. Williams CJ, Richardson DC, Richardson JS. The importance of residue-level filtering and the Top2018 best-parts dataset of high-quality protein residues. *Protein Sci* 2022;31:290–300. https://doi.org/10.1002/pro.4239
37. Černý J, Božíková P, Svoboda J *et al.* A unified dinucleotide alphabet describing both RNA and DNA structures. *Nucleic Acids Res* 2020;48:6367–81. https://doi.org/10.1093/nar/gkaa383
38. Černý J, Malý M Božíková P *et al.* DNATCO v5.0: integrated web platform for 3D nucleic acid structure analysis. *Nucleic Acids Res* 2026; https://doi.org/10.1093/nar/gkaf1491
39. Williams CJ, Headd JJ, Moriarty NW *et al.* MolProbity: more and better reference data for improved all-atom structure validation. *Protein Sci* 2018;27:293–315. https://doi.org/10.1002/pro.3330
40. Word JM, Lovell SC, LaBean TH *et al.* Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms. *J Mol Biol* 1999;285:1711–33. https://doi.org/10.1006/jmbi.1998.2400
41. Richardson JS, Moriarty NW, Keedy DA. Alternate conformations always want to spread. *Computat Crystallogr Newslett* 2023;14:2–9.
42. Nakamura T, Zhao Y, Yamagata Y *et al.* Watching DNA polymerase η make a phosphodiester bond. *Nature* 2012;487:196–201. https://doi.org/10.1038/nature11181
43. Binas O, Tants J-N, Peter SA *et al.* Structural basis for the recognition of transiently structured AU-rich elements by Roquin. *Nucleic Acids Res* 2020;48:7385–403. https://doi.org/10.1093/nar/gkaa465
44. Emsley P, Lohkamp B, Scott WG *et al.* Features and development of coot. *Acta Crystallogr D Biol Crystallogr* 2010;66:486–501. https://doi.org/10.1107/S0907444910007493
45. Chen VB, Davis IW, Richardson DC. KING (Kinemage, Next Generation): a versatile interactive molecular and scientific visualization program. *Protein Sci* 2009;18:2403–9. https://doi.org/10.1002/pro.250
46. Parzen E. On estimation of a probability density function and mode. *Ann Math Statist* 1962;33:1065–76. https://doi.org/10.1214/aoms/1177704472
47. Schneider B, Kabeláč M, Hobza P. Geometry of the phosphate group and its interactions with metal cations in crystals and *ab initio* calculations. *J. Am. Chem. Soc.* 1996;118:12207–17. https://doi.org/10.1021/ja9621152
48. Rozov A, Demeshkina N, Khusainov I *et al.* Novel base-pairing interactions at the tRNA wobble position crucial for accurate reading of the genetic code. *Nat Commun* 2016;7:10457. https://doi.org/10.1038/ncomms10457
49. Leys C, Ley C, Klein O *et al.* Detecting outliers: do not use standard deviation around the mean, use absolute deviation around the median. *J Exp Soc Psychol* 2013;49:764–6. https://doi.org/10.1016/j.jesp.2013.03.013
50. Shao C, Yang H, Westbrook JD *et al.* Multivariate analyses of quality metrics for crystal structures in the PDB archive. *Structure* 2017;25:458–68. https://doi.org/10.1016/j.str.2017.01.013
51. Sobolev OV, Afonine PV, Moriarty NW *et al.* A global ramachandran score identifies protein structures with unlikely stereochemistry. *Structure* 2020;28:1249–58. https://doi.org/10.1016/j.str.2020.08.005

52. Smart OS, Horský V, Gore S *et al.* Validation of ligands in macromolecular structures determined by X-ray crystallography. *Acta Crystallogr D Struct Biol* 2018;74:228–36. https://doi.org/10.1107/S2059798318002541
53. Liebschner D, Afonine PV, Moriarty NW *et al.* CERES: a cryo-EM re-refinement system for continuous improvement of deposited models. *Acta Crystallogr D Struct Biol* 2021;77:48–61. https://doi.org/10.1107/S2059798320015879
54. Richardson JS, Williams CJ, Hintze BJ *et al.* Model validation: local diagnosis, correction and when to quit. *Acta Crystallogr D Struct Biol* 2018;74:132–42. https://doi.org/10.1107/S2059798317009834
55. Kleywegt GJ, Jones TA. Where freedom is given, liberties are taken. *Structure* 1995;3:535–40. https://doi.org/10.1016/S0969-2126(01)00187-3
56. Richardson JS, Williams CJ, Chen VB *et al.* The bad and the good of trends in model building and refinement for sparse-data regions: pernicious forms of overfitting versus good new tools and predictions. *Acta Crystallogr D Struct Biol* 2023;79:1071–8. https://doi.org/10.1107/S2059798323008847
57. Moriarty NW, Janowski PA, Swails JM *et al.* Improved chemistry restraints for crystallographic refinement by integrating the Amber force field into Phenix. *Acta Crystallogr D Struct Biol* 2020;76:51–62. https://doi.org/10.1107/S2059798319015134
58. Jain S, Kapral GJ, Richardson JS. Fitting Tips #7 - Getting the pucker right in RNA Structures. *Computat Crystallogr Newsletter* 2014;5:2–9.
59. Kapral G. *RNA backbone validation, correction, and implications for RNA-protein interfaces*. Dissertation, Durham, NC: Duke University. 2013.
60. Osterman A, Mondragón A. Structures of topoisomerase V in complex with DNA reveal unusual DNA-binding mode and novel relaxation mechanism. *eLife* 2022;11:e72702. https://doi.org/10.7554/eLife.72702
61. Nair SK, Sliverman SK, Chen JH *et al.* Crystal structure analysis of TRF2-Dbd-DNA complex. 2012. https://doi.org/10.2210/pdb3SJM/pdb
62. Read RJ, Adams PD, Arendall WB *et al.* A new generation of crystallographic validation tools for the protein data bank. *Structure* 2011;19:1395–412. https://doi.org/10.1016/j.str.2011.08.006
63. Montelione GT, Nilges M, Bax A *et al.* Recommendations of the wwPDB NMR Validation Task Force. *Structure* 2013;21:1563–70. https://doi.org/10.1016/j.str.2013.07.021
64. Kleywegt GJ, Adams PD, Butcher SJ *et al.* Community recommendations on cryo-EM data archiving and validation. *IUCrJ* 2024;11:140–51. https://doi.org/10.1107/S2052252524001246

**Received:** August 4, 2025. **Revised:** October 10, 2025. **Accepted:** November 3, 2025