RNAprecis: Prediction of full-detail RNA conformation from the experimentally best-observed sparse parameters

Henrik Wiechers[1◕], Christopher J. Williams[2◕], Benjamin Eltzner[1◕], Franziska Hoppe[1], Michael G. Prisant[2], Vincent B. Chen[2], Ezra Miller[3], Kanti V. Mardia[4], Jane S. Richardson[2*], Stephan F. Huckemann[1†]

March 5, 2025

1 Felix-Bernstein-Institute for Mathematical Statistics in the Biosciences, University of Göttingen, Göttingen, Germany
2 Department of Biochemistry, Duke University School of Medicine, Durham, North Carolina, USA
3 Mathematics Department, Duke University, Durham, North Carolina, USA
4 Department of Statistics, University of Leeds, Leeds, UK
◕ These authors contributed equally to this work.
* jsr@kinemage.biochem.duke.edu
† Stephan.Huckemann@mathematik.uni-goettingen.de

Abstract

We address the problem of predicting high detail RNA structure geometry from the information available in low resolution experimental maps of electron density. Here low resolution refers to $\gtrapprox 2.5$Å where the location of the phosphate groups and the glyocosidic bonds can be determined from electron density but all other backbone atom positions cannot. In contrast, high resolution determines all backbone atomic positions. To this end, we firstly create a gold standard data base for four groups of manually corrected suites, each reflecting one out of four sugar pucker-pair configurations. Secondly we develop and employ a modified version of the previously devised algorithm MINT-AGE to learn clusters that are in high correspondence with gold standard's conformational classes based on 3D RNA structure. Since some of the manually corrected classes are of very small size, the modified version of MINT-AGE is able to also identify very small clusters. Thirdly, the new algorithm RNAprecis assigns low resolution structures to newly designed 3D shape coordinates. Our improvements include: (i) learned classes augmented to cover also very low sample sizes and (ii) regularizing a key distance by introducing an adaptive Mahalanobis distance. On a test data containing many clashing and suites modeled as conformational outliers, RNA precis shows good results suggesting that our learning method generalizes well. In particular, our modified MINT-AGE clustering can be finer than the existing curated gold standard suite conformers. For example, the **0a** conformer has been separated into two clusters seen in different structural contexts. Such new distinctions can have implications for biochemical interpretation of RNA structure.

Code availability The code used to generate the analyses and plots presented in this paper can be found at `https://gitlab.gwdg.de/henrik.wiechers1/rnaprecis`.

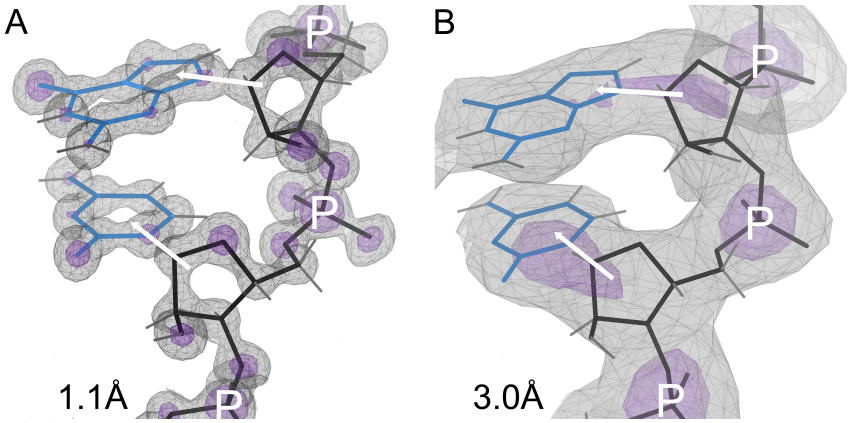

Figure 1: *The positions of individual atoms in the derived atomic model are experimentally determined by the electron density when at high resolution (Panel A), but not at low resolution (Panel B). In both panels electron density contours are shown at* 1.2σ *(gray) and* 3σ *(purple) where* σ *is the standard deviation of the corresponding electron density map. The derived models are shown as stick figures, backbone in black, bases in blue, phosphate positions (white P) and glycosidic bond direction (white line between sugar and base). Panel B: only the phosphate positions and glycosidic bond direction are well determined experimentally at low resolution (*$\sim 3\AA$*). Panel A: all individual atomic positions and their covalent bonds (blue or black) are well determined at high resolution* $\sim 1.1\AA$*). This RNAprecis project seeks to improve the process of deriving the full atomic model at low resolution.*

1 Introduction

Correctly modeling RNA backbone conformations in X-ray crystallography and cryoEM experimental 3D structures is a long standing challenge in structural biology. The RNA backbone has many degrees of freedom: 7 dihedral angles, plus the ribose sugar pucker. Additionally, even in structures solved at good resolution (≤ 2Å), the backbone is generally less well-resolved than its associated bases. At lower resolutions (> 2Å), the density map for the backbone gradually becomes more featureless and ambiguous (Figure 1), and a model's fit to that map becomes less able to distinguish between correct or incorrect local conformations.

Achieving an accurate model of the RNA backbone is important for understanding RNA structure and function. In contrast to proteins, where the sidechains are responsible for most binding contacts and chemistry, in RNA many of those significant contacts are with the backbone. Two such cases are the peptide transfer reaction of protein synthesis in ribosomes ([1]) and the specific base-binding control sites in messenger-RNA riboswitches ([2]). The latter sites control translation of the enzymes which synthesize that base, by changing the overall riboswitch tertiary structure drastically to prevent initiation of translation when that base is at high concentration.

The RNA backbone has the attractive property of rotamericity [3]. That means bond geometry and steric contacts generally constrain the backbone to distinct families of conformations. The traditional "residue", or "nucleotide", from phosphate to phosphate is a chemical definition, reasonable because that is the polymerization unit. From a conformational perspective, however, parsing the RNA backbone as the "suite" from ribose sugar to sugar is a more useful system, since it relates adjacent bases and its atoms have many more contacts with one another than inside a residue [3]. The suite includes atoms from two adjacent, bonded residues (see Section 2.1).

A full description of an RNA suite's conformation requires quite high-resolution experimental data to position all of the backbone atoms accurately. The ribose sugar pucker is especially difficult to model correctly at resolutions worse than about 2Å, as are the positions of the O5' and C5' atoms, which fall in an often-ambiguous tube of density across the $\alpha/\beta/\gamma$ dihedrals (Figure 2).

Resolving this problem is possible even at resolutions of 3-3.5Å, because the position of the phosphorus

atom and the direction between the base and the sugar (approximating the direction of the "glycosidic" bond that joins them) can usually be seen directly. The phosphorus is a much heavier atom than the rest with a strong map density signal, and is well located across this range of resolutions because it is the center of the symmetric PO_4 group (the purple contours in Figure 1). The nucleobase is large, planar, and frequently held in place by predictable base-pairing, and it usually produces a recognizable blob of density, while the sugar forms a blob where the base and backbone density join. It was discovered some time ago that the ribose sugar pucker, one of the hardest-to-observe details of RNA backbone structure, can be successfully predicted from those most reliable observations, even when the pucker itself has been modeled incorrectly (see Section 2.3).

Success of our pucker analysis made us hope that other combinations of parameters from well-observed features at low resolution might enable prediction of full-detail RNA backbone conformation as described by the multi-lab consensus suite conformers reported in 2008 [4]. We chose the name RNAprecis for this desired system, to indicate a very concise description that captures the important aspects. (Precis: *pray'-see*, in writing composition, a short formal summary of a much longer document.)

Standard clustering algorithms had earlier failed even at the easier high-resolution task, at least partly because 2/3 of the suites are in the single A-form helical conformation. Therefore those 2008 consensus conformers were defined by hand in a 7D graphics system developed for that purpose [4], plus testing whether close clusters should be separated or not by whether they were distinct or interchangeable in context; in their local structural motif, base interactions, or functional role. These suite conformers are defined in the space of the 7-dimensional dihedral angles along the backbone, and so at lower resolutions they are at the mercy of low-accuracy atomic models built into ambiguous experimental map density.

This current multidisciplinary collaboration between structural biochemistry and mathematical statistics is adopting more sophisticated treatments from non-Euclidean statistics, for automated use in a quality-curated training set from the much more numerous current RNA structures. We use a modified version of the mainly unsupervised MINT-AGE clustering algorithm introduced in [5]. Its mode hunting on the main one-dimensional nested stratified circle determined by torus PCA from [6] has been modified in [7] to identify also very low sized clusters, as are of concern here. Thus, modified MINT-AGE can accurately represent the toroidal relationships between pairs of 360° re-entrant angles or dihedral-angles (as observations on a 7-dimensional torus), and also provide better scaling among diverse distance, angle, and dihedral parameters and highly diverse cluster sizes (beginning with size 3), in up to 9 dimensions including the 2 already-determined puckers. The resulting automated conformational clusters mostly match one-to-one with the full-detail suite conformers. Where MINT-AGE divides one suite conformer into multiple clusters, the examples will be examined to see whether additional suite conformers should be defined. Where a MINT-AGE cluster combines two or more suite conformers, we will search for a single sparse parameter or combination of sparse parameters that can automatically separate the structural differences; for instance, the distance between base N1/9 atoms should be able to give a very good separation between the **1b** base-stacked conformer and the **1[** intercalated conformer.

The overall goal of RNAprecis is, at each backbone suite of a low-resolution RNA structure, to produce an automated analysis of which full-detail backbone conformer is compatible with the "precis" combination of parameters modeled at that suite's phosphorus atom and glycosidic bond directions. The answer will often be a unique conformer, but may be two or three possibilities, or sometimes none if the local map density and thus the model was very poor. The structural biologist would then rebuild and re-refine the model with the suggested new conformers and keep the changes shown to be improvements. The training stage of this process determined the overall transformation from the full-detail conformers of the quality-filtered training data set to the precis parameter space: specifically designed coordinates for 3D landmark based size-and-shape, that generalize [8] coordinates in [7]: The phosphorus group is used as the origin. This, along with the two neighboring C1' atoms forms a natural triangle, with respect to which the N1/9 atoms at the corresponding glycosidic bonds are considered in spherical coordinates. Among others, the transformation builds on an adaptive Mahalanobis distance ([9], see also [10]) to measure how far a given suite's multidimensional position lies from the centroid of the nearest cluster(s). The actual use case will perform that transformation in reverse for each suite in the structure being analyzed.

Our paper is structured as follows: In the following Section 2 we provide background for RNA structural biology and introduce our new quality-curated training high- and low-detail gold standard training data, as well as the low-detail test data. Also we describe previous work on the Pperp criterion that provides pucker discrimination for low-detail data. Then in Section 3 we continue with an introduction to non-Euclidean statistical methods used (MINT-AGE for high detail torus-valued data and wedge coordinates for size-and-shape analysis) and a detailed description of our new RNAprecis algorithm. Section 4 then shows our results for correspondence of RNA conformations with the clusters, found by MINT-AGE for the training data and by RNAprecis for the test data. We conclude with a detailed discussion and outlook in Section 5.

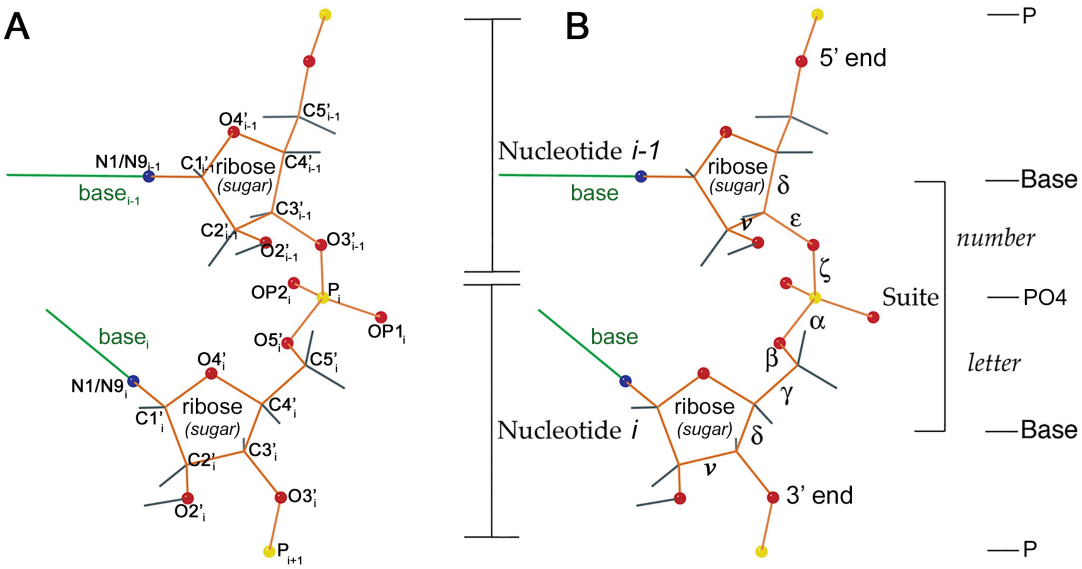

Figure 2: *Atomic structure and nomenclature for the 2 residues and the central sugar-to-sugar suite of an RNA dinucleotide. The bases are simplified to single green lines. Backbone atoms other than carbon are marked with balls, red for O, blue for N, yellow for P, and with gray bonds out to H. A. Names of heavy atoms used in describing suites. B. The 7 dihedral angles of the suite are labeled in order as δ, ϵ, ζ, α, β, γ, δ. Additionally we show the dihedral angle ν measuring the sugar pucker (see Fig. 4). Number and letter refer to suite conformer names as e.g. in Section 2.5 and Table 6 in the appendix.*

2 Structural biology of RNA: Background and data sets

2.1 RNA 3D structure

The relatively minor chemical differences between DNA and RNA, (the O2' atom in RNA and the methylated T base in DNA), produce surprisingly profound differences in overall conformation and behavior. The biological job of DNA is to preserve genetic information in base-paired B-form double-helix secondary structure as robustly as possible, and so those helices are quite flexible. DNA can also adopt A-form or Z-form helices, but seldom forms tertiary structure. In contrast, RNA cannot adopt B-form helix , and a large fraction of RNA molecules form well-ordered tertiary structures that have specific binding and even catalytic functions, more like protein structures.

Figure 2 diagrams the main features of full-detail local RNA 3D structure that we will describe and analyze in this work. The atomic model covers two phosphate-to-phosphate nucleotide residues and one central sugar-to-sugar suite. The suite backbone parsing is better suited than the nucleotide parsing for analyzing local atomic contacts along the backbone and relationships between adjacent bases [3], and will be used here. By convention, the suite is named by the residue number (i) of the second residue –

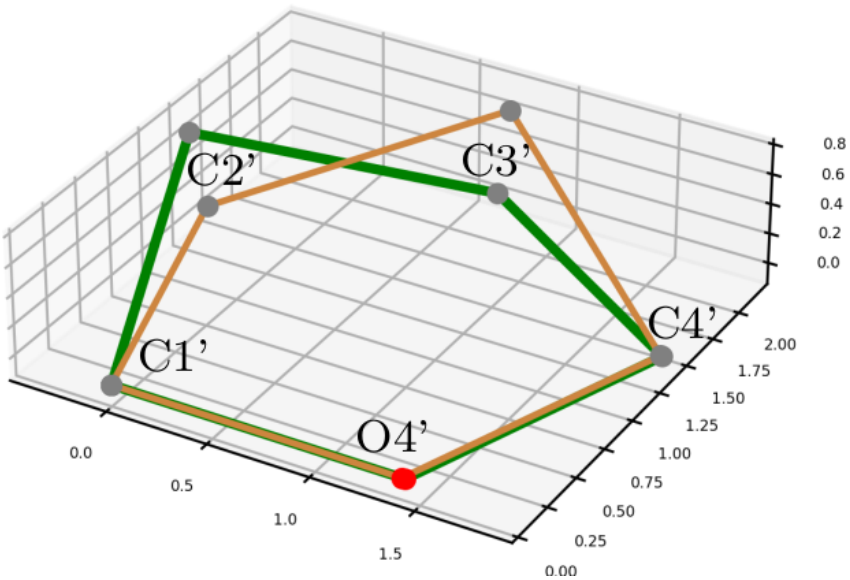

Figure 3: *The two sugar puckers: either C3' (brown, C3' endo) or C2' (green, C2' endo) is above (in positive orientation) the oriented plane spanned by $C4' - O4'$ and $C1' - O4'$. Actually all atoms but one are (nearly) co-planar.*

angle	atom bonds						
α	O3'	-	P	-	O5'	-	C5'
β	P	-	O5'	-	C5'	-	C4'
γ	O5'	-	C5'	-	C4'	-	C3'
δ	C5'	-	C4'	-	C3'	-	O3'
ϵ	C4'	-	C3'	-	O3'	-	P
ζ	C3'	-	O3'	-	P	-	O5'
ν	C4'	-	C3'	-	C2'	-	C1'

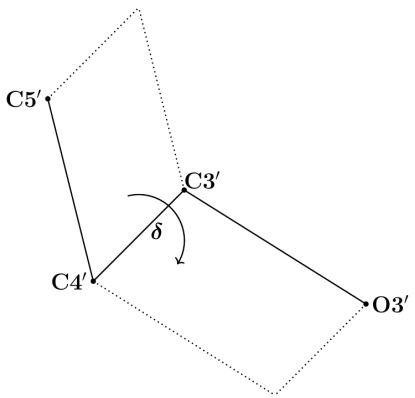

Figure 4: *Left: the dihedral angle (first column), gives the rotation from the plane spanned by the last three atoms (right column) with respect to the plane spanned by the first three atoms (right column), as depicted in the exemplary right panel (Figure and table adapted from [11]), see Figure 2 The first 7 measure backbone dihedral angles, distinguished from ν which measures the sugar pucker, see Figure 3.*

the same residue number as the P atom contained within the suite. The glycosidic bond joins the N1 or N9 atom of the base to the C1' atom of the ribose ring. At right, number and letter symbolize the names of the multi-lab consensus suite conformers, such as **1a**, **5z**, **6n**, etc., developed under the auspices of the RNA Society's RNA Ontology Consortium [4]. Suites with conformations that do not match any of the consensus suite conformers are considered outliers for validation purposes and given the **!!** designation.

2.2 Experimental methods of determining RNA atomic structures

X-ray crystallography and single-particle cryo Electron Microscopy (cryoEM) are the two predominant experimental methods for solving 3D structures of biological macromolecules, including RNA. Sample preparation, data collection, and data processing are completely different, but both methods produce

a quite similar-looking 3D density map, into which a model of the macromolecule is fitted or built from scratch, then cycled through refinement, validation, re-rebuilding, and re-refinement. CryoEM is sensitive to charge as well as mass, so for nucleic acids the positive bases are somewhat stronger and the negative phosphates somewhat weaker than in X-ray maps. Relatively small to mid-size macromolecules are easier for crystallography, and mid-size to very large ones are easier for cryoEM.

Resolution of the overall structure is defined in crystallography by the distance out in the reciprocal space of diffraction peaks to which the signal/noise stays acceptable. Overall resolution is defined in cryoEM by how far out in reciprocal Å the Fourier shell correlation (agreement between two half-maps calculated from random halves of the data) stays acceptable. In crystallography that overall value depends mostly on how well the crystal contacts line up the molecules between repeats across the crystal, and only somewhat on large changes within the molecule such as relative domain orientation, since if that changes very much it will require a different crystal form. In cryoEM the overall resolution depends on suitable diversity of particle alignment on the EM grid and on how well the different 2D images can be aligned in 3D, but distinct large-scale conformations in different particles can often be analyzed separately within one dataset. In most experimental RNA structures, the overall resolution applies to a large majority of the residues or suites.

Local resolution is even more important for determining local conformation. It depends on local mobility of various sorts: alternative conformations, local motion, and partial or complete disorder. In any macromolecule there is more mobility at the solvent-exposed surface than on the interior, especially at chain termini and for some loops. For cryoEM, uncertainty in image alignment sometimes produces an additional smooth gradient in local resolution from center to edge. As local mobility or uncertainty increases, the density becomes lower, then more and more patchy, and eventually disappears into noise. Across that spectrum, the RNAprecis assumption of fairly robust P position and glycosidic bond direction gradually becomes inapplicable.

2.3 Previous work on high-detail suite conformers and the "Pperp" ribose pucker criterion

The pucker state of the 5-membered ribose ring is a dominant influence on RNA conformation, because a pucker change greatly alters the relative directions of the three large ring substituents: the base, the preceding backbone, and the following backbone. In RNA the pucker is quite cleanly bimodal, as either C3'-endo (most common, and in A-form) or C2'-endo (frequent in irregular or non-helical regions), see Figure 3. Other puckers are extremely rare, but can occur. Like all examples of genuine (as opposed to wrong) outlier conformations, they are almost always at biologically important binding, catalytic, or structurally essential sites, such as the C1'-exo pucker at suite 14 for fluorophore binding in the 6e8u evolved aptamer [12]. The small changes in ring atom position between different puckers can only be seen directly at very high resolution (about 1.5Å or better).

Fortunately, however, the true ribose pucker can be determined quite successfully as C3'-endo or C2'-endo even if the resolution is relatively low and the pucker has been fit incorrectly in local ring detail. As noted in the introduction, the P atom and the direction of the glycosidic bond (between base and sugar) are the features most clearly observed in RNA at low resolution. The length of a perpendicular dropped from the next-in-sequence (3') phosphorus atom to the extended vector direction of the glycosidic bond that connects base to ribose (see Figure 5) is < 2.9Å for C2'-endo and ≥ 2.9Å for C3'-endo, as illustrated in Figure 6. This process was incorporated into MolProbity validation [13, 14], and the measure was named as the "Pperp" criterion in its full description [15]. As early as 2007, MolProbity validation showed that incorrect puckers were fairly common and produced many of the serious clash and geometry outliers in RNA. Correction of pucker outliers is seldom achieved by refinement, but once the problem is recognized it is fairly easy to fix in interactive rebuilding programs such as Coot [16].

2.4 Data sets used

Sugar pucker-pairs. Since the ribose puckers are bimodal, a good start on automatically identifying full-detail backbone conformers is to sort conformers into 4 categories by each suite's two puckers. We

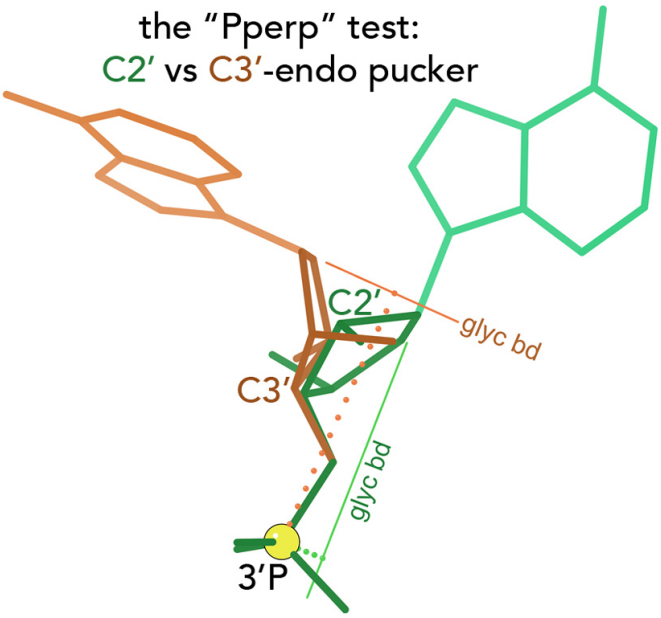

Figure 5: *Contrasted cases, superimposed at their phosphate ends, to illustrate the Pperp criterion for distinguishing between C3'-endo (brown) and C2'-endo (green) sugar puckers in RNA (same colors as in Figure 3). The atom puckered out of plane in each ring is labeled, and the vector extension of each glycosidic bond is drawn and labeled. The next P atom in sequence is marked by a large yellow ball. Dotted lines show the perpendiculars dropped from the P to each of the glycosidic bond vectors: long (and brown) for the C3'-endo puckered case; short (and green) for the C2'-endo case.*

call these the *pucker-pair* groups:

$$P33, \quad P32, \quad P23, \quad P22\,,$$

where C3'-endo and C2'-endo have been shortened to just 3 and 2. The P33 group is much larger than the others, because it includes the dominant A-form double-helical conformation (**1a**).

At high detail, sugar pucker is determined directly from atoms in the ribose ring, using the ν dihedrals (see Figures 3 and 4), where ν_1 refers to the first ribose in the suite and ν_2 to the second. At low detail, the ribose ring atoms are not assumed to be reliable, and sugar pucker is instead determined using the Pperp criterion (see Figure 5). Note that the P atom from the next sequential suite (with the index $i+1$) is required for determining the second sugar's Pperp distance and pucker conformation.

High detail gold standard. The first training dataset, denoted by HD, comprises complete atomic positions of RNA suites obtained by high resolution (< 2.0Å) X-ray crystallography and CryoEM. It is derived from a preliminary version of a carefully curated RNA residue dataset, RNA2023 [17]. The 3.150 release of BGSU RNA 3d Hub[1] was used to classify RNA chains by sequence and function, then select the best chain from each class based on resolution, RSR, RSCC, Rfree, clashes, and completeness [18]. Chains from two high-resolution CryoEM ribosome structures - 8b0x [19] and 8a3d [20] - were added to this set of chains.

Since even high-resolution structures may contain local disorder or local modeling errors [21], the selected RNA chains were then filtered at the suite level. Suites with strong indicators of modeling errors, such as Pperp/pucker mismatches, steric clashes, and covalent bond geometry outliers have been removed from the dataset. Likewise, suites with insufficient support from experimental data, as measured by local 2mFo-dFc map values, local real-space correlation coefficients, or alternate conformations were

[1] `http://rna.bgsu.edu/rna3dhub/nrlist`

removed. Additionally, suites involving a non-standard RNA base (i.e. other than A, U, C, or G), and nearly all suites identified as "!!" suite outliers by suitename validation were removed.

This quality-filtered data set serves as the *gold standard*. Since puckers are also recorded, HD comprises the following four datasets:

$$HDP33, \quad HDP32, \quad HDP23, \quad HDP22\,.$$

Low detail datasets comprise only positions of the large atomic agglomerations, namely the positions of the central P of the phosphate groups and the ends of the glycosidic bonds, i.e. the positions of the C1' and N1 or N9 atoms, respectively. For training purposes, from the *high resolution* dataset a low detail dataset denoted by LD has been obtained by taking the only corresponding 5 atom positions, yielding, four datasets featuring corresponding sugar puckers:

$$LDP33, \quad LDP32, \quad LDP23, \quad LDP22\,.$$

In Section A in the appendix, the PDB files of the high resolution dataset are listed with the corresponding resolution in Table 3 and the HD and LD representations are plotted in Figure 15 and Figure 16 in the appendix.

Low detail test data and homology conformer classes as probable answers. For testing how well our trained system can predict correct conformations from low detail parameters, a test set denoted LDT was created. The dataset comprises, for each member suite, measurements of the same five atomic positions as in LD (the central P and the two neighboring C1' and N1/N9), plus an additional sixth atom - P from the next residue following the suite. LDT is divided using Pperp criteria into the following pucker-pair datasets:

$$LDTP33, \quad LDTP32, \quad LDTP23, \quad LDTP22\,.$$

Suites for LDT were selected to provide examples of modeling errors with probably known solutions from related structures. RNA chains related by function and sequence homology were identified, according to the 3.328 release of `http://rna.bgsu.edu/rna3dhub/nrlist` [18]. To ensure that a good solution was possible for each chain, only homology classes that contributed a chain to the high-quality HD training set were considered for membership in LDT. Homology classes of low sequence variety (such as poly-U) were excluded, as the relatedness of their member chains' conformations is uncertain. Within each included homology class, a sequence alignment was performed on chains solved at 4.0Å or better. Chains that did not cleanly align on sequence, usually due to internal chain breaks, were excluded. From the remaining aligned chains, suites that were modeled as **!!** outlier conformations were selected as members of LDT if a homology-related chain had a recognized, non-**!!** conformer modeled at the same sequence location. The suite identities of the non-**!!** conformers were recorded as *homology-related conformer classes* and treated as probable correct conformations of the corresponding **!!** suites. In addition to providing probable answers for the test, selection of **!!** residues with non-**!!** counterparts selects against disordered residues and true outlier conformations, both of which may be expected to be modeled as **!!** across all members of a homology class.

In Section B in the appendix, the PDB files, the chain names, and the residue numbers of the elements that comprise the low detail test dataset are listed with their associated resolutions in Table 4 and Table 5 and the four LDT data sets are plotted in Figure 17.

The LDT test set addresses two related questions. First, are the model features used by RNAprecis reliably preserved in RNA modeling errors? Second, given reliable parameters, can RNAprecis predict probable solutions for RNA modeling errors?

2.5 Conformer classes, MINT-AGE clusters and RNAprecis

Conformer classes. For each suite in the high resolution training data HD, the suite conformer class was determined with `phenix.suitename`. `suitename` uses the complete 7 dihedral angles of the suite (Figure 3) to assign the suite to one of the manually defined rotameric conformers (e.g. **1a**, **1c**, **5z**,

etc.) or to !! if the suite does not match any recognized conformation. The conformer class assignments from `suitename` represent the current (best available to date) ground truth suite identities for the HD datasets [4].

Each suite conformer class corresponds to a single pucker-pair combination – identifiable from the δ dihedral angle at high resolution – and training data suites were assigned to pucker-pair sets accordingly. The HD and LD datasets comprise 4148 suites that could be read in with our python based code (of which 3529 are P33 - 3183 **1a** and 346 non-**1a**, 216 are P23, 265 are P32 and 138 are P22.)

Clustering of the four high resolution datasets has also been performed based on the purely statistical method of MINT-AGE developed by [11]. In order to accommodate for very low conformer class sizes, we have developed a new *mode hunting method for very small dataset sizes* from [7] described below in Section 3.2. Tuning parameters have been learned such that the *clusters* found are in high agreement with the above suite conformers, see Section 4.

MINT-AGE training was carried out on the HD datasets. The HD datasets are based on a preliminary version of the 'nosuiteout' RNA2023, as described above and available at `https://zenodo.org/doi/10.5281/zenodo.8103013` [17]. The files that have been used are listed in Appendix A. The source PDB files were downloaded from the RCSB PDB database [22]. MINT-AGE used the full 7 dihedral angles to automatically cluster suites into clusters similar to the manually-defined suitename conformers classes.

RNAprecis training was carried out on the LD datasets. The LD datasets comprise a minimal set of atom positions for the same files and suites as the HD datasets. The LD datasets are split into pucker-pair sets based on the Pperp ribose pucker criterion described above in Section 2.3, which requires only atom positions from this minimal set. Because the training data is filtered to remove any ambiguous or incorrect ribose puckers, the pucker-pairs sets are reliably identified by ν values.

RNAprecis testing was carried out on the LDT datasets. LDT datasets were based on the 3.328 version of the `http://rna.bgsu.edu/rna3dhub/nrlist` RNA homology list, dated 27.3.2024; files were downloaded from the RCSB PDB on 1.4.2024. The high-detail atom positions needed to determine ν angles were not considered reliable for test set suites, so the suites were separated into pucker-pucker sets based on the low-detail Pperp criteria (Figure 8). The test set comprises 849 suites - 452 P33, 123 P32, 176 P23, and 98 P22. The PDB files and residues used are listed in Section B in the appendix.

3 Non-Euclidean statistical learning methods for clustering and classification

We first explain how HD data is represented on the torus with its canonical distance and detail the modified *MINT-AGE* method to cluster such data.

Then we introduce the *simplex coordinates* representation for LD data and present the newly developed *RNAprecis* algorithm to assign LD data to HD clusters.

3.1 Dihedral angles on the torus

While individual bond lengths and angles between bonds at atomic centers are fairly fixed (with standard deviations on the order 10^{-2}Å for lengths and ~ 2 degrees for angles [23]), structural diversity is caused by rotation around bonds, i.e. dihedral angles measured in arc length on the unit circle

$$\mathbb{T} := [0, 2\pi)/\sim$$

where "$\sim$" indicates that 0 is identified with 2π.

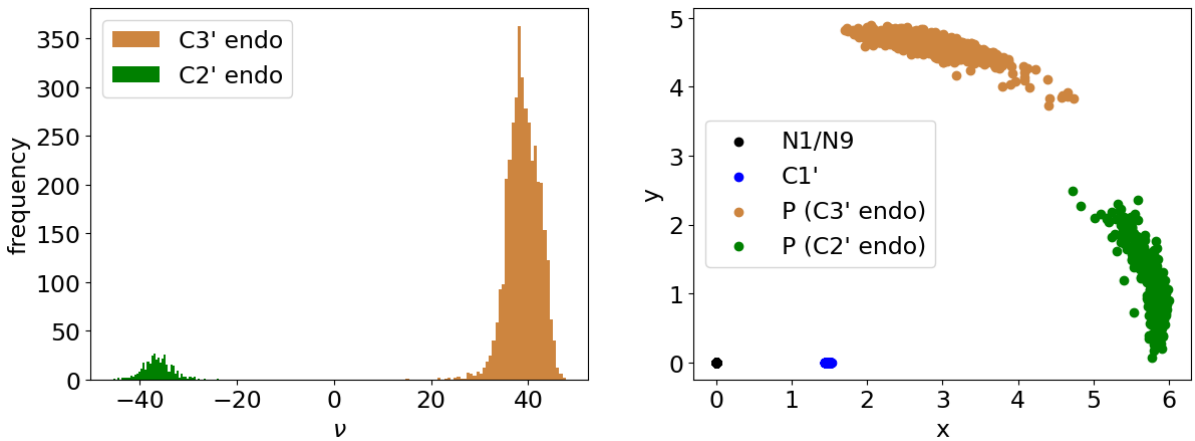

Figure 6: *For the high quality high resolution training data, the two different sugar puckers can be very well discriminated by the angle ν (left). Right: Using the coordinates described in Figure 5, namely mapping $N1/N9$ to the origin and $C1'$ to the first positive axis (x), the perpendicular distance (Pperp, positive vertical y axis) similarly perfectly separates the two pucker configurations (this 2D display reflects true atom position in Cartesian space).*

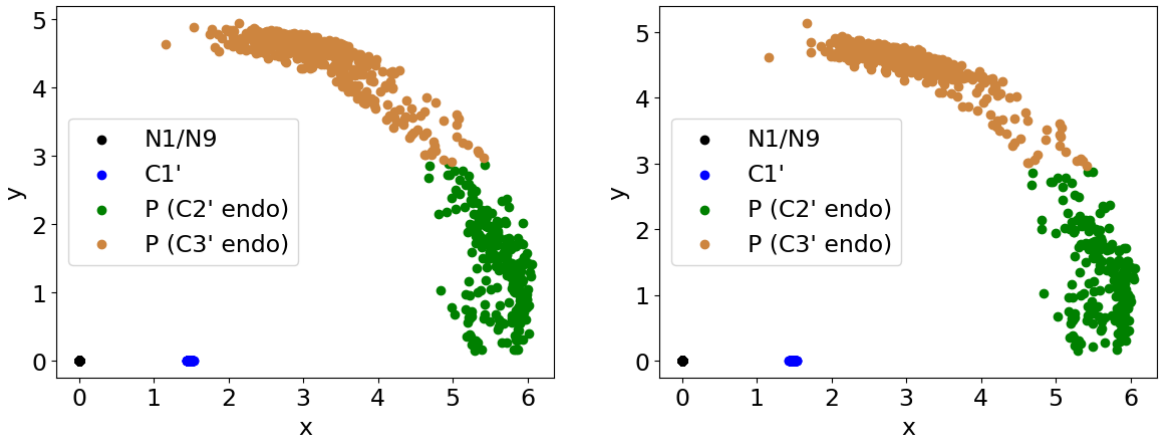

Figure 7: *Using the coordinates described in Figure 5, namely mapping $N1/N9$ to the origin and $C1'$ to the first positive axis, consider the perpendicular distance to the positive second (vertical) axis as in Figure 6 similarly for the low resolution test data. For all low resolution test suites depicting the perpendicular distance of the left sugar pucker (3' end of the suite's P, measured with respect to the preceding P) in the left panel and the perpendicular distance to the right sugar pucker (5' end of the suite's P). We predict the unknown true sugar pucker by vertically separating at 2.9, as learned from the high resolution data, depicted in Figure 6.*

The natural metric of $\mathbb{T}$ is given by arc length

$$d_{\mathbb{T}^7}(x, y) = \min\{|x - y|, 2\pi - |x - y|\} \text{ for } x, y \in \mathbb{T}\,.$$

As the shape of one suite, i.e. the RNA backbone from one sugar ring to the next is described by 7 dihedral angles, see Figure 2, high resolution structure is modeled by elements of the 7-dimensional torus

$\mathbb{T}^7$, equipped with the canonical distance for a direct product

$$d_{\mathbb{T}^7}(x,y) = \sqrt{\sum_{i=1}^{7} d_{\mathbb{T}}(x_i,y_i)^2}, \text{ for } x=(x_1,\ldots,x_7), y=(y_1,\ldots,y_7) \in \mathbb{T}^7 .$$

3.2 Statistical clustering of suites on the torus via MINT-AGE

MINT-AGE stands for *Mode huntINg after Torus pca on Adaptive iterative cutting Linkage trEes* and contains two building blocks. The first block, AGE, can be performed on any metric space and for the HD data we use the 7-dimensional torus with the canonical torus distance.

AGE first builds a linkage tree (e.g. average linkage), as introduced by [24] and then performs a specifically designed iterative adaptive clustering method inspired by [25, 26], identifying first the lowest density cluster, then iteratively higher density clusters. The method is detailed in [5] and our specific tuning parameters are listed in Table 1.

MINT is applied to every pre-cluster found by AGE. Each cluster is subjected to torus PCA from [6]. Here, the 7-dimensional torus is mapped to a 7-dimensional stratified sphere (where the stratification mimicks the torus' topology) and an adaption of the principal backward nested subsphere algorithm (PNS) from [27] (a variant of PCA for spheres), using torus distances rather than spherical distances, is applied to the data, which is then projected to its main one-dimensional small circular component.

These projected data are subjected to circular mode hunting where each statistically significant mode stands for a new subcluster. Originally MINT used nonparametric circular mode hunting ([6] building on the linear version by [28]), where cluster boundaries had been placed at antimodes of suitably kernel smoothed data utilizing a similar approach to the WiZer (Wrapped Gaussian kernel smoothed derivative's significant Zero crossings, from [29] building on [30]). In order to identify also very small subclusters, even of size 3 we use a parametric approach, namely a likelihood ratio test for the likelihoods of fitted unimodal and bimodal Gaussians recently developed in [7]. This test can be performed iteratively to find several subclusters on a main one-dimensional small circular component.

3.3 Multicentered polysphere coordinates for size-and-shape

LD data is given by $k=5$ *landmarks*

$$\mathbf{x_1},\ldots,\mathbf{x_k} \in \mathbb{R}^m$$

of atom center positions (P, the two adjacent C1' and the two N1/N9 of the corresponding glycosidic bonds) in dimension $m=3$. Shape and size information of such a configuration $(\mathbf{x_1},\ldots,\mathbf{x_k})$ is preserved by orientation preserving Euclidean motions: rotations and translations. Hence, we consider equivalence classes modulo the action of the Euclidean group,

$$[\mathbf{x_1},\ldots,\mathbf{x_k}] := \{R(\mathbf{x_1},\ldots,\mathbf{x_k}) + \mathbf{x}\,1_k^T : R \in SO(m), \mathbf{x} \in \mathbb{R}^m\}$$

where $SO(m)$ denotes the $m(m-1)/2$-dimensional special orthogonal group for $\mathbb{R}^m$ and $1_m \in \mathbb{R}^m$ has value 1 in every entry. While configurations have $15 = mk$ degrees of freedom, as the Euclidean group is of dimension $m + m(m-1)/2 = m(m+1)/2 = 6$, only $9 = 15-6$ degrees of freedom remain for *size-and-shape* (for more detail see e.g. [31]).

For statistical analysis of size-and-shape, two avenues open up. Under no additional information (on landmark prominence, say), these equivalence classes are often modeled in an abstract size-and-shape space obtained by first *Helmertizing*, i.e. removing translation by isometrically mapping (superimposing landmarks) to a configuration with $k-1$ landmarks (see, e.g. [31]). Alternatively a *frame* of prominent landmarks can be singled out and mapped by a unique group element to a standard frame. Then the images of the remaining landmarks serve as coordinates for the equivalence classes. This second avenue can be pursued for 2-dimensional shape ([32] coordinates, here, shape is obtained by additionally filtering

out size) or for arbitrary dimensional affine or projective shape ([33, 34]). For arbitrary dimensional size-and-shape this avenue is not as straightforward. However, we use *multicentered polysphere coordinates* as proposed by [7] (called multicentered polypolar coordinates there). These are specifically suited for our LD representation. Indeed the P atom center naturally serves as a reference point, i.e. its position is subtracted from all other atom centers and the two C1' atom center naturally define a planar reference triangle, with tilted coordinate system such that their vertical components agree. The two landmarks giving the N1/N9 atoms are then not centered with respect to the P' atom but with respect to their neighboring C1' atoms defining the glycosidic bond. Since these bonds lengths can be considered constant, the coordinates of the two N1/N9 atoms can thus be viewed to live on a unit sphere. Additionally, the opening angle at the P' atom of the reference triangle lives on a half circle.

Multicentered coordinates for 5 landmark configurations. Let the 5 landmark configuration $\mathbf{x_1}, \ldots, \mathbf{x_5}$ denote the estimated atom centers of $N1/N9, C1', P', C1'$ and $N1/N9$ in that order. Realistically we assume that the second, third and fourth landmark define a proper nondegenerate triangle. Then, in the following we define a the plane of the triangle spanned by $\mathbf{v}$ and $\mathbf{w}$ with respect to which the other two loci $\mathbf{y_1}$ and $\mathbf{y_2}$ are considered: Subtracting the phosphorus' coordinates from the two $C1'$ coordinates and the $C1'$ coordinates from the corresponding $N1/N9$, obtain

$$\mathbf{y_1} := \mathbf{x_1} - \mathbf{x_2},\ \mathbf{v} := \mathbf{x_2} - \mathbf{x_3},\ \mathbf{w} := \mathbf{x_4} - \mathbf{x_3},\ \mathbf{y_2} := \mathbf{x_5} - \mathbf{x_4}$$

and define orthogonal unit vectors

$$\left.\begin{array}{lcl} \mathbf{e_1} & := & \frac{\mathbf{w}-\mathbf{v}}{\|\mathbf{w}-\mathbf{v}\|}\,, \\ \mathbf{e_2} & := & \frac{\mathbf{v}-\mathbf{e_1}\mathbf{e_1}^T\mathbf{v}}{\|\mathbf{v}-\mathbf{e_1}\mathbf{e_1}^T\mathbf{v}\|}\,, \\ \mathbf{e_3} & := & \mathbf{e_1} \times \mathbf{e_2}\,, \end{array}\right\} \tag{1}$$

and the orthogonal matrix $R := (\mathbf{e_1}, \mathbf{e_2}, \mathbf{e_3})$. Then, the *multicentered simplex coordinates* (see [7]) for the above 5-point configuration are given by

$$(\|\mathbf{v}\|, \|\mathbf{w}\|, \arccos \frac{\mathbf{v}^T\mathbf{w}}{\|\mathbf{v}\| \cdot \|\mathbf{w}\|}, R^T\mathbf{y_1}, R^T\mathbf{y_2}) \in \mathbb{R}_+ \times \mathbb{R}_+ \times (0, \pi) \times \mathbb{R}^3 \times \mathbb{R}^3\,.$$

A polysphere representation Taking into account that the two glycosidic bonds have equal lengths that are the same for all configurations, we may assume $\|\mathbf{y_1}\| = 1 = \|\mathbf{y_2}\|$, to obtain the *polysphere representation*

$$W := (d_2, d_3, \alpha, \mathbf{s_1}, \mathbf{s_2}) \quad \in \quad \mathbb{R}+^2 \times \mathbb{S}^1_+ \times (\mathbb{S}^2)^2 \tag{2}$$

where

- $\mathbb{S}^1_+ = (0, \pi)$ is the unit half circle,
- $d_2 := \|\mathbf{v}\|$ and $d_3 := \|\mathbf{w}\|$ denote the distances between the atomic centers of P and the two C1', respectively,
- $\alpha := \arccos \frac{\mathbf{v}^T\mathbf{w}}{\|\mathbf{v}\| \cdot \|\mathbf{w}\|}$ gives the angle at P between the two pseudobonds to from P to the two C1', respectively and
- $\mathbf{s_i} := R^T\mathbf{y_i}$, $i = 1, 2$ give the directions of the two glycosidic bonds, respectively, with polar angles ϕ_1, θ_1 and ϕ_2, θ_2, respectively (see Figure 8).

Recalling that the above multicentered coordinates have 9 degrees of freedom, the polysphere representation ends up with 7 degrees of freedom. Notably, this is also the number of degrees of freedom for HD data modeled on the 7-dimensional torus.

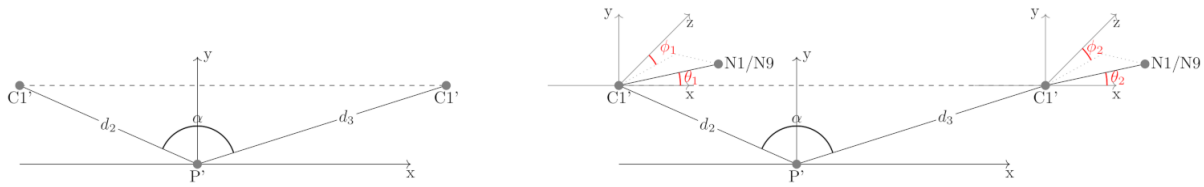

Figure 8: *Polysphere coordinates for LDT suite size-and-shape. Left: the triangle spanned by the phosphorus as foot point and the two neighboring C1' atoms. After normalization for rotation and height, three descriptors remain: the lenghts of the two pseudo bonds and the opening angle. Right: The two N1/9 atoms at the corresponding glycosidic bonds (which have fixed lengths) are additionally described by spherical coordinates with respect to the triangle.*

Remark 3.1. *The RNAprecis algorithm below computes a sample Fréchet mean. Due to [35], such sample means are unique on a complete connected Riemannian manifold if their distribution is absolutely continuous with respect to Riemannian measure. We can assume all of these conditions, except for completeness. The reason is that the polysphere representatives in (2) comprise a connected Riemannian manifold without the boundary points $\alpha = 0, \pi$. Since we exclude that data assume these boundary points and since our geodesic metric (induced by a product of Riemannian Euclidean and spherical metrics) in (Step 2 (b)) in the RNAprecis algorithm (Algorithm 3.2), none of the means will assume these points either. In consequence, we can assume uniqueness.*

3.4 RNAprecis

Certain atom positions are reliably observed even in low-resolution structures, and the underlying **assumption of RNAprecis** is that their positions can be used to expand on Pperp-based ribose pucker prediction and predict RNA backbone conformers.

Specifically, the phosphorus (P) is a much heavier atom than the rest with a strong map density signal, and is well located across this range of resolutions because it is the center of the symmetric PO_4 group (the purple contours in Figure 1). The nucleobase is large, planar, and frequently held in place by predictable base-pairing, and it usually produces a recognizable blob of density, while the sugar forms a blob where the base and backbone density join. The relationship between these two well-observed blobs can be represented with the glycosidic bond direction or the N1/N9 and C1' atom positions.

The following algorithm transforms LDT data to MINT-AGE clusters (introduced in Section 2.5) of HDT gold standard data (see Section 2.4) and via its correspondence with suite conformation classes to suite conformations (see Figure 1).

Algorithm 3.2 (RNAprecis)**.** ***Input:***

1. *Clusters obtained by MINT-AGE from HDT gold standard data,*
2. *test element*
$$W = (d_2, d_3, \alpha, \mathbf{s_1}, \mathbf{s_2}) \in \mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{S}^1_+ \times \mathbb{S}^2 \times \mathbb{S}^2$$
in polysphere representation from the LDT test data.

Algorithm:

1. *Determine the sugar pucker-pair of W via the perpendicular distances method (see Section 2.3) and let $C_1, \ldots, C_m$, $m \in \mathbb{N}$ be the MINT-AGE clusters of the HDT gold standard data set that have the same pucker configurations.*
2. *Determine the low detail distance dist_{LDT} between W and each cluster $C_1, \ldots, C_m$ as follows:*

(a) *Let*

$$W_j = (d_2^{(j)}, d_3^{(j)}, \alpha^{(j)}, \mathbf{s_1^{(j)}}, \mathbf{s_2^{(j)}}) \in \mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{S}_+^1 \times \mathbb{S}^2 \times \mathbb{S}^2\,, \quad j = 1\ldots, n_i$$

be the LDT polysphere representations of the suites in C_i where n_i is the number of elements in C_i, $1 \le i \le m$. Calculate a (Remark 3.1 ensures uniqueness under realistic conditions) ***Fréchet mean*** *for the canonical product distance:*

$$\mu^{(i)} \in \underset{\substack{\mu = (\mu_1, \mu_2, \mu_3, \mu_{\mathbf{4}}, \mu_{\mathbf{5}}) \\ \in \mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{S}_+^1 \times \mathbb{S}^2 \times \mathbb{S}^2}}{argmin} \frac{1}{n_i} \sum_{j=1}^{n_i} \Bigg((\mu_1 - d_2^{(j)})^2 + (\mu_2 - d_3^{(j)})^2 + (\mu_3 - \alpha^{(j)})^2 + \arccos^2({\mu_{\mathbf{4}}}^T \mathbf{s_1^{(j)}}) + \arccos^2({\mu_{\mathbf{5}}}^T \mathbf{s_2^{(j)}})^2 \Bigg).$$

(b) *Orthogonal projection of spherical coordinates $\mathbf{s_l^{(j)}}$ to the tangent spaces of their means $\mu_{\mathbf{l+3}}^{(\mathbf{i})}$, $l = 1, 2$: With the unit vectors from (1) select*

$$\mathbf{a_l} := \underset{\mathbf{e_k} \in \{\mathbf{e_1}, \mathbf{e_2}, \mathbf{e_3}\}}{\arg\min} |{\mathbf{e_k}}^T \mu_{\mathbf{l+3}}^{(\mathbf{i})}|$$

and set

$$\mathbf{b_l} := \frac{\mathbf{a_l} - \mu_{\mathbf{l+3}}^{(\mathbf{i})} {\mathbf{a_l}}^T \mu_{\mathbf{l+3}}^{(\mathbf{i})}}{\|\mathbf{a_l} - \mu_{\mathbf{l+3}}^{(\mathbf{i})} {\mathbf{a_l}}^T \mu_{\mathbf{l+3}}^{(\mathbf{i})}\|}\,, \quad \mathbf{c_l} := \mathbf{b_l} \times \mu_{\mathbf{l+3}}^{(\mathbf{i})}\,, Q_l := (\mathbf{b_l}|\mathbf{c_l})$$

Then

$$\tilde{\mathbf{s}}_{\mathbf{l}}^{(j)} = Q_l^T {\mathbf{s_l}}^{(j)} \in \mathbb{R}^2\,, \quad l = 1, 2\,, \quad j = 1, \ldots, n$$

yield the orthogonal projections of ${\mathbf{s_l}}^{(j)}$ to the tangent space of $\mathbb{S}^2$ at $\mu_{\mathbf{l+3}}^{(\mathbf{i})}$ in a coordinate system only depending on $\mu_{\mathbf{l+3}}^{(\mathbf{i})}$.

(c) *Define $\widetilde{W}_j := (d_2^{(j)}, d_3^{(j)}, \alpha^{(j)}, \tilde{\mathbf{s}}_{\mathbf{1}}^{(j)}, \tilde{\mathbf{s}}_{\mathbf{2}}^{(j)}) \in \mathbb{R}_+ \times \mathbb{R}_+ \times (0, \pi) \times \mathbb{R}^2 \times \mathbb{R}^2$ for all $j = 1\ldots, n_i$.*

(d) *Determine the* ***empirical covariance matrix***

$$S_i = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} \begin{pmatrix} d_2^{(j)} - \mu_1^{(i)} \\ d_3^{(j)} - \mu_2^{(i)} \\ \alpha^{(j)} - \mu_3^{(i)} \\ \tilde{\mathbf{s}}_{\mathbf{1}}^{(j)} \\ \tilde{\mathbf{s}}_{\mathbf{2}}^{(j)} \end{pmatrix} \begin{pmatrix} d_2^{(j)} - \mu_1^{(i)} \\ d_3^{(j)} - \mu_2^{(i)} \\ \alpha^{(j)} - \mu_3^{(i)} \\ \tilde{\mathbf{s}}_{\mathbf{1}}^{(j)} \\ \tilde{\mathbf{s}}_{\mathbf{2}}^{(j)} \end{pmatrix}^T \in \mathbb{R}^{7 \times 7}.$$

(e) *Calculate the empirical covariance matrix for Cluster 1 from the data set LDP33 as described in Steps 2(a) to 2(d) above and name the corresponding covariance matrix S_{target}.*

(f) *Calculate the regularized covariance matrix*

$$S_i^{(R)} = \lambda S_i + (1 - \lambda) S_{\text{target}},$$

where $\lambda = 0.5$.

(g) *The low detail distance is defined by the* **Mahanalobis distance**

$$\text{dist}_{LDT}(W, C_i) := \sqrt{\begin{pmatrix} d_2 - \mu_1^{(i)} \\ d_3 - \mu_2^{(i)} \\ \alpha - \mu_3^{(i)} \\ \tilde{\mathbf{s}}_{\mathbf{1}} \\ \tilde{\mathbf{s}}_{\mathbf{2}} \end{pmatrix}^T (S_i^{(R)})^{-1} \begin{pmatrix} d_2 - \mu_1^{(i)} \\ d_3 - \mu_2^{(i)} \\ \alpha - \mu_3^{(i)} \\ \tilde{\mathbf{s}}_{\mathbf{1}} \\ \tilde{\mathbf{s}}_{\mathbf{2}} \end{pmatrix}} \tag{3}$$

where $\tilde{\mathbf{s}}_{\mathbf{l}} = Q_l \mathbf{s_l}$ with Q_l from Step 2(b) above, $l = 1, 2$ for $i = 1, \ldots, m$.

3. *Return the conformation classes corresponding to the MINT-AGE cluster with the smallest low detail distance.*

Remark 3.3. *1) The spherical data analysis above is based on standard techniques in directional statistics, see e.g. [36].*

2) RNAprecis employs a Mahanalobis distance in order to ensure that variation of all coordinate representatives is taken equally into account.

*3) Regularization for the Mahanalobis distance with the overwhelming largest cluster (***1a** *has over 3,000, the second largest* **1c** *has only approx. 150 members) guarantees invertibility also for low size clusters whose variation does not cover the entire 7-dimensional space.*

4) The low detail distance amounts to the Mahalanobis distance of the point from each cluster. It can be interpreted as the distance of a point from a cluster in terms of number of standard deviations. However, since the covariances of smaller clusters are regularized by the main cluster, the distances cannot be immediately translated into statistical confidence statements.

4 Results

4.1 Clustering of HD and corresponding LD regions

In this section, the cluster results of the adapted MINT-AGE algorithm (see Section 3.2) for the individual data sets HDP33, HDP32, HDP23 and HDP22 (see Section 2.4) are reported and analyzed. The AGE algorithm requires 3 tuning parameters (see [5] and [11]) listed in Table 1 for the four different data sets. The resulting clusters are listed in Table 6 in Appendix C. Remarkably, as the confusion matrices in

dataset	q	κ	$d_{\max}$
HDP33	0.09	3	1.1%
HDP32	0.05	3	0.34%
HDP23	0.07	3	1.4%
HDP22	0.05	3	0.7%

Table 1: *Selection of tuning parameters required for the AGE algorithm from [5] for the four different pucker-pair HD data sets: relative branching distance q, minimal cluster size κ (set to $\kappa = 3$ for all four data sets in order to also identify clusters with very few elements) and the maximum outlier distance d_{max} (given as the percentage of elements that are in a cluster with less than κ elements).*

Figures 22, 23, 24 and 25 (all in Appendix C) teach,

8 out of 11 clusters in the HDP33 data set,
13 out of 15 clusters in the HDP32 data set,
10 out of 16 clusters in the HDP23 data set,
7 out of 10 clusters in the HDP22 data set

belong to a single suite conformer class. Figure 9 shows the dihedral angle combination of α and β for the data set HDP33 (by far the largest) in the left panel (see Figure 2 and Figure 4 for an explanation of the dihedral angles), in which the MINT-AGE clusters are best visually separated. The plots of all angle combinations for all data sets are shown in Appendix C. There is also a detailed comparison between the cluster results and the suite conformer classes identified with `suitename`, using confusion matrices.

MINT-AGE successfully identifies clusters with small numbers of individual suites for many suite conformers. **7d**, **5d**, **5j**, **4g**, **#a**, **4d**, **5p**, **5q**, **0b**, **2u**, and **2o** are all assigned unique clusters based on 4 or fewer examples. Nevertheless, some conformers remain unclustered: **7a**, **3d**, **5n**, **3g**, **5r**, **0k**, and **4s** could not be assigned to unique MINT-AGE clusters. Conformers **7a**, **3d**, **5n**, and **3g** each have enough individual suites for MINT-AGE clustering, but their data clouds are obviously too spread out to assign to a single cluster, while **0k**, **5r**, and **4s** comprise only one individual suite each.

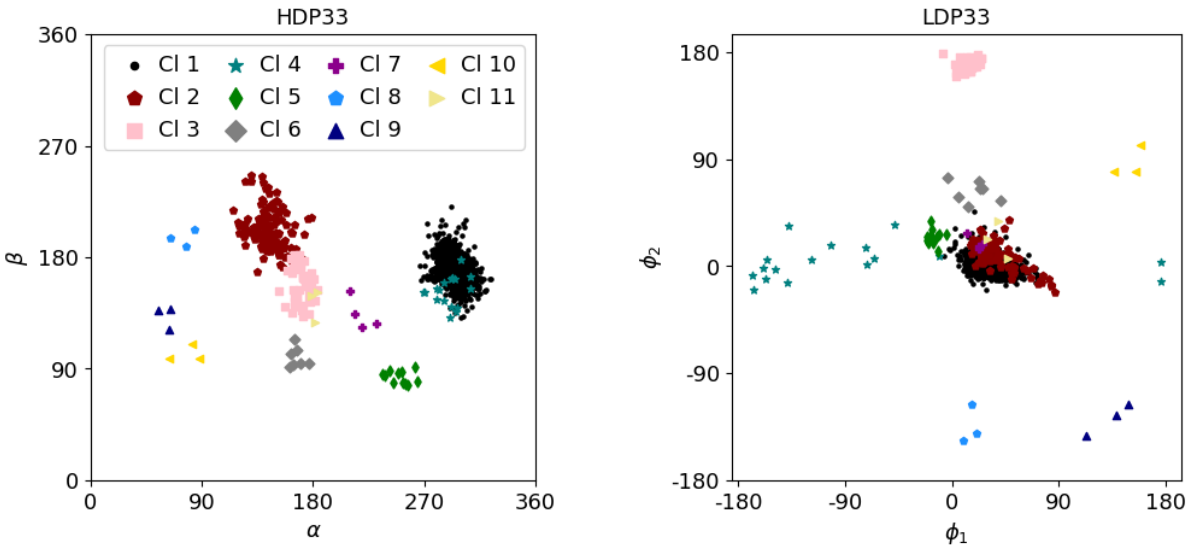

Figure 9: *Left panel: Illustration of HDP33 clusters in the α-β plane of the high resolution coordinates. Right panel: The low resolution coordinates corresponding to the same suites, representing the distribution of the clusters in the ϕ_1-ϕ_2-plane. One can see that many of the clusters in dihedral angles are mapped to distinct clusters also in the low resolution coordinates.*

4.2 MINT-AGE clusters comprising multiple conformer classes

In broad overview (see Figures 22, 23, 24 and 25 in Appendix C), HDP33 Clusters 1/4 and HDP32 Clusters 1/5 combine two or more different conformer classes into a single cluster. It was expected that the corresponding suites would be difficult to separate automatically, as their conformer classes boundaries were based on structural context in the previous work. We now go into detail.

HDP33 Cluster 1 combines **1a**, **&a**, **1L**, and **1m**. **1a** is the dominant conformer class (A-form helix) across all RNA and exerts a strong influence on clustering. The **&a**, **1L**, and **1m** conformer classes are shoulders on the main **1a** cluster, each with a distinct behavior, and are not cleanly separated from it. Similarly to the **1b**/**1[** case, these satellite conformers are distinguished from **1a** more by context than by conformation.

HDP33 Cluster 4 combines suite conformer Classes **9a**, **3a**, and possibly **7a**. These conformer classes differ only in the ζ dihedral. As in the other cases, the distribution is quite continuous, and the distinction is found in the suite's context. **9a** conformers start or end loops, while **3a** conformers enable stack switches. **7a** conformers, of which a single instance is included in cluster 4, are also stack switches.

HDP32 Cluster 1 combines suite conformer Classes **1b** and **1[**, as does Cluster 5. The **1b**/**1[** distinction is illustrative of the challenge that MINT-AGE faces in reproducing the manual conformer classes. In the `suitename` parameter space, **1b** and **1[** are closely related and share a continuous distribution with no clear conformational breakpoint between the behaviors. Instead the primary distinction comes from elsewhere in the structure - **1[** has a ligand or another base intercalated between the bases of the suite, while the bases of 1b are often stacked directly on each other. In defining the original suite conformer classes [4], the cutoff between **1b** and **1[** was manually set at β=206° to provide the cleanest possible separation between these behaviors.

Some clusters combine a main conformer with a small conformer that is otherwise unclustered. HDP33 Clusters 1 and 4 each included a single instance of **7a**. HDP23 Cluster 2 is majority **0a**, but also includes four **#a**. HDP23 Cluster 5 is majority **8d** and also contains two **4d**. HDP22 Cluster 6 does not contain a majority conformer, but combines three **2u** with two **2o**. These cases reflect the limitations of our relatively small training dataset; there are not enough examples of **#a**, **4d**, **2u**, and **2o** to create

independent clusters for these conformers.

Several clusters include one or two instances of a conformer that is otherwise clearly identified in its own cluster. HDP23 Cluster 2 includes two instances of **4a**, which is otherwise cleanly assigned to Cluster 7. HDP23 Cluster 3 includes two stray **01** and one stray **6j**. HDP23 Cluster 11 includes one stray **9d**. These cases are a reminder that RNA conformer space is complex and biological conformer boundary definitions are often somewhat ambiguous.

Finally, some conformers and clusters are particularly challenging to untangle. HDP23 Cluster 4 is majority **6g**, but contains two instances of **4g**, while **4g** is further split across Clusters 15 and 16. Cluster 16 also contains a single **4a**, which is otherwise assigned to Cluster 7. HDP22 Cluster 7 combines 4 instances of **4b** with two of **0b**, bother of which conformers themselves are heavily split across multiple clusters. Expansion of the training dataset should help resolve these ambiguous conformers. However, the splitting of suite conformers into multiple clusters is not necessarily due to the limitations of the training data and may, as discussed below, present an opportunity to revisit the existing suite conformer definitions.

4.3 Conformer classes split across multiple MINT-AGE clusters

Some suite conformer classes were split into multiple clusters by MINT-AGE, most notably **0a** in HDP23, but also **6d** in HDP23, **7p** in HDP32, **1f** in HDP33, and **0b** in HDP22 (see Figures 22, 23, 24 and 25 in Appendix C). For reproducing the suite conformer classes, it is easy to simply re-combine the clusters. However, the separation suggests that MINT-AGE has discovered real conformational divisions that were not visible during the previous work.

MINT-AGE strongly separates the **0a** conformer into two subclusters, with 16 suites in Cluster 2 and 12 suites in Cluster 6 (see Table 6 in Appendix C). These subclusters separate the two major structural motifs found within **0a**.

Cluster 2 contains stack switches. In standard **1a** conformation, the subsequent bases of the suite are near parallel and stack on top of each other. In these stack switches, the bases are still roughly parallel but change which helix or strand of a helix they are stacking with, rather than stacking on each other (Figure 10 A and C).

Cluster 6 contains kink-turns. The kink-turn motif [38] spans multiple suites – the most distinctive feature is a very tight turn with an outturned base that often interacts with proteins. The Cluster 6 suite comprises the two residues following the residue with the outturned base. The first base of the cluster 6 suite stacks with the backbone of the residue on the other side of the outturned base, stabilizing the sharp turn. The bases of a Cluster 6 suite are angled relative to each other due to this interaction with the preceding backbone (Figure 10 B and D).

0a was previously treated as a single conformer class because these suites have a continuous distribution along the ζ dihedral angle, best seen in the ζ/α panel of Figure 19 in the appendix. However, Clusters 2 and 6 are more clearly distinguishable by ϵ, which also separates the related conformer Class **4a**/Cluster 7, best seen in the ϵ/ζ panel of Figure 19 in the appendix.

The split of suite conformer Class **0a** provides the strongest evidence so far of MINT-AGE identifying a systematic functional/conformational distinction that was not previously recognizable. The **7p** split in HDP32 is also suggestive. MINT-AGE finds 4 clusters, combining 7 instances into Cluster 9, 5 into Cluster 10, 4 into Cluster 13, and 3 into Cluster 14. This set of clusters is complex, but Cluster 9 is clearly conformationally distinct from the rest (Figure 11). That this distinction is conformational rather than "just" contextual demonstrates the value in approaching suite conformer identification from both perspectives.

The **0a** and **7p** suite conformer classes should be split into new conformer classes, based on the MINT-AGE clustering. The success of an unsupervised method in recognizing new conformer classes not previously identified by manual work is an encouraging result.

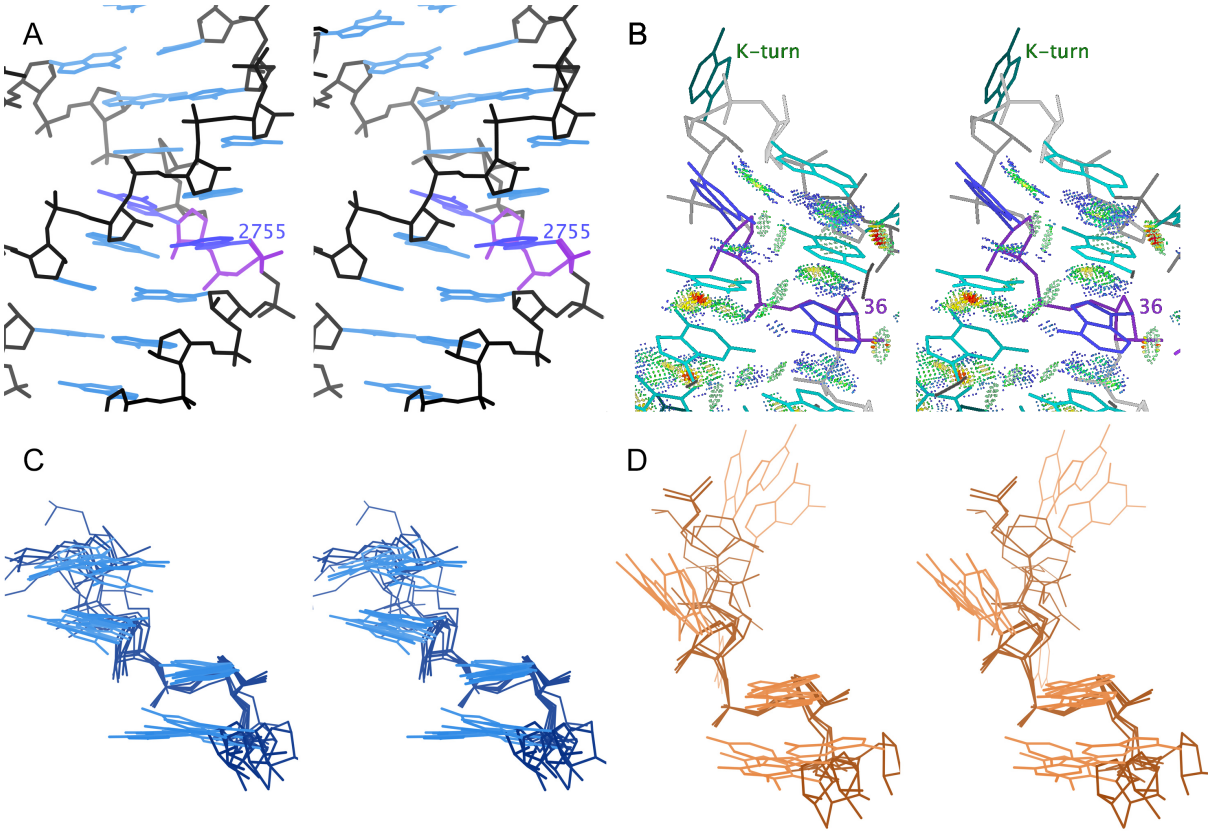

Figure 10: *Stereo images of* ***0a*** *conformers, as separated by MINT-AGE. A. Stack switch from 8a3d chain A ([20]). Suite 2755 is highlighted in purple, and its bases stack into two different helix strands. B. Kink-turn from 5fjc chain A ([37]). Suite 36 is highlighted in purple. Below the outturned base, the upper base of suite 36 shows blue and green dots for favorable van der Waals packing with the backbone of the strand on the right. C. Superposition of 6 representative MINT-AGE Cluster 2 suites. These are stack switchs and have bases close to parallel. D. Superposition of 4 representative MINT-AGE Cluster 6 suites. These are kink-turns and have bases at a high angle to each other.*

4.4 Statistical classification of conformations for low-detail data

The detailed results of the RNAprecis test are listed in Appendix E. The results are presented individually for the four pucker-pairs data sets LDTP33 (Section E.1), LDTP32 (Section E.2), LDTP23 (Section E.3) and LDTP22 (Section E.4) formed by the classification in step 1 of Algorithm 3.2. For each of the test elements, there are one or more probable answers of homology-related conformer classes determined by related experimental structures (see paragraph "Low detail test data and homology conformer classes as probable answers" in Section 2.4). There are three possibilities, whose colors correspond to tables and figures below:

Prediction matches: The cluster determined by RNAprecis (Algorithm 3.2) has at least one suite conformer class (see Table 6 in the appendix), which is in the list of probable answers for this element.

Prediction mismatches: The list of probable answers contains at least one suite conformer that can also be found in the pucker-pair group of the test element. However, the cluster determined by RNAprecis (Algorithm 3.2) has no suite conformer (see Table 6 in the appendix), which is in the list of probable answers for this element.

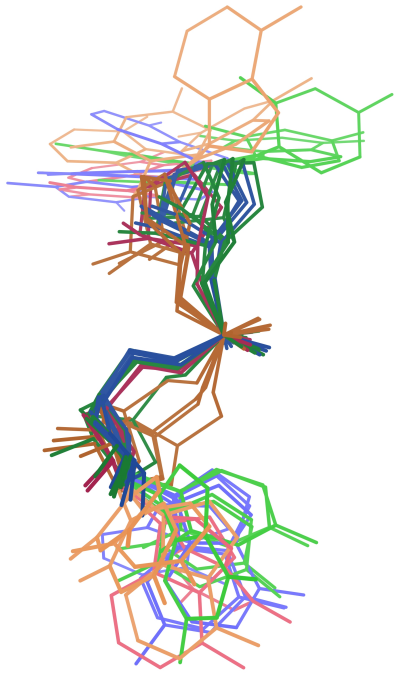

Figure 11: *Superposition of* ***7p*** *suites, colored by MINT-AGE cluster. Cluster 9 is orange, 10 is green, 13 is blue, 14 is red. Cluster 9 is clearly conformationally distinct from the other groupings within* ***7p****.*

Pucker mismatches: In the list of probable answers, there is not a single suite conformer that can be found in the pucker-pair group of the test element (determined by step 1 of Algorithm 3.2).

dataset	count	prediction matches	prediction mismatches	pucker mismatches
LDTP33	452	295 (65%)	143 (32%)	14 (3%)
LDTP32	123	80 (65%)	10 (8%)	33 (27%)
LDTP23	176	121 (69%)	38 (22%)	17 (10%)
LDTP22	98	73 (74%)	18 (18%)	7 (7%)

Table 2: *List of the four pucker-pairs test datasets in the first column formed by Step 1 in RNAprecis (Algorithm 3.2) with the respective total counts in the second column and counts and percentages in the third through fifth columns.*

For all four pucker-pairs sets, at least 2/3 of the elements are sorted into the prediction match group (blue), see Table 2 and top left in 30, Figure 31, Figure 33 and Figure 35 in the appendix). The results are robust to the resolution of the experiments and even relatively robust to local disorder, as indicated by whether the next and/or previous suite is a !!-suite conformer. Since adjacent suites share a common sugar, modeling errors can propagate to neighboring suites. As a result, suites which have at least one !!-suite adjacent are more likely to be misclassified. Whether the adjacent !! suite precedes or follows does not lead to a significantly different outcome, despite the following suite's P atom contributing to the Pperp criterion, see top center and bottom of 30, Figure 31, Figure 33 and Figure 35 in the appendix). If the smallest low detail distance is large (i.e. no cluster has a similar low detail representation), the proportion of prediction mismatch (orange) or pucker mismatch (red) elements increases for all data sets, see top right of 30, Figure 31, Figure 33 and Figure 35 in the appendix). The probable homology conformer class answers and the clusters determined with RNAprecis for all individual elements (sorted according to the four pucker-pairs sets and according to the blue, red and orange categories) are reported

in Table 7 and 8 (blue category of LDTP33), Table E.1 (red and orange category of LDTP33), Table 10 (blue, red and orange category of LDTP32), Table 11 (blue category of LDTP23), Table 12 (red and orange category of LDTP23) and Table 13 (blue, red and orange category of LDTP22). A comparison between the probable answers and the clusters determined by RNAprecis is shown in a classification matrix for the four data sets in Figure 12 below and in Figures 32,34 and 36 in the appendix. The classification results are separated into the four pucker pairs so that the **1a** conformer does not dominate the results. The classification results are good uniformly over all pucker pair classes.

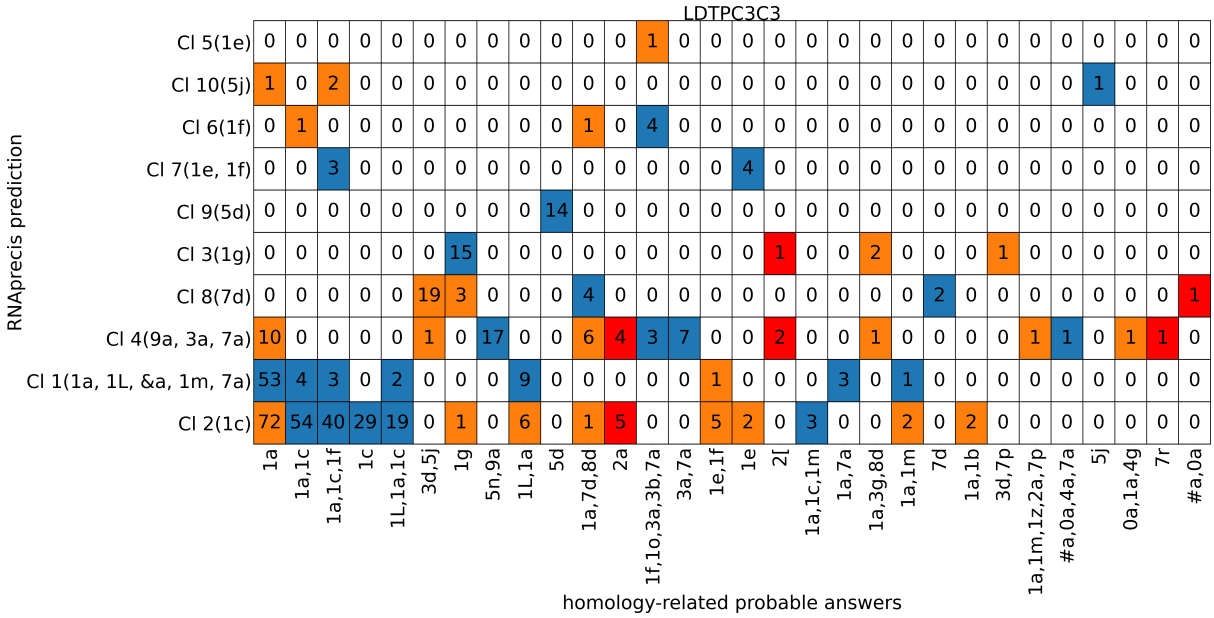

Figure 12: *Classification matrix listing counts of probable homology-related conformation classes (horizontal axis) versus MINT-AGE clusters (vertical, with conformer classes in parentheses from MINT-AGE training, see Figure 22 in the appendix) for the low detail test data set of sugar pucker-pairs LDTP33 with colors: blue for match, orange for mismatch and red for pucker-pair mismatch. Columns listing multiple suite conformers are from sites solved in different conformations across homology-related models, and any of the homology-related conformations were accepted as a matching predictions.*

The conformers **3d**, **5n**, **3g**, **0k**, **5r**, and **4s** were not clustered by MINT-AGE, probably due to having too few examples in the training dataset (between 1 and 4). Therefore, these conformers also could not be predicted by RNAprecis. Any test set suite where the homology-related conformation is among these unclustered conformers will necessarily result in a prediction mismatch or pucker mismatch. This inability to predict matching conformers affects 3 **4s** suites. It may affect some of the 20 **3d**,**5j** suites; the 3 **1a**,**3g**,**8d** suites; or the single **3d**,**7p** suite currently without a matching prediction. These latter cases have multiple homology-related conformations, and are affected only if the unclustered conformer is the correct one for a given test suite.

5 Discussion

5.1 MINT-AGE clustering

This clustering process is remarkably successful in both training and test stages. In the training stage with filtered, high-resolution data, 38 of the 51 suite conformers with ≥ 3 individual suites (75%) were clustered with one-to-one agreement by MINT-AGE and only 4 were not clustered at all. The rest either combineded 2 or more suite conformers into a single cluster or split one conformer into more than one cluster. Interesting and perhaps correct examples of those cases are described in Section 4 (Results).

For the testing stage, RNAprecis (which by design) uses the trained cluster relationships in reverse to predict the high-detail suite conformer for each **!!** suite that has a probable valid conformer (or several) in related structures. 67% were clear wins of prediction matches to a homology-related conformation, and 25% were prediction mismatches. The other 8% were mismatches of the Pperp-determined pucker-pair category (used by RNAprecis) to the pucker-pair of the probable answers, indicating cases where the modeling errors were too severe for this method. The details of the two mismatch categories have taught us several feasible ways to make the RNAprecis process significantly better, described below.

The most obvious way to improve MINT-AGE performance is use more training data. The RNA2023 dataset, from which the HD and LD training sets were adapted, is a general purpose dataset. One of its purposes is to reflect the relative frequency of different features in RNA structures. This means that several rare suite conformers (e.g. **3d**, **5n**, **4s**) are too rare (see above) in the dataset to form clusters with MINT-AGE. A new dataset, built to include more conformational diversity and more rare suite conformers (for instance by allowing two entries per homology class) would improve clustering and recognition of those conformers.

MINT-AGE clustering can also be improved through development of additional parameters. For instance, the N1/N9-to-N1/N9 distance across the suite may differentiate between stacking (shorter distance) and intercalation (longer distance) in **1b** and **1[**. Such parameters would be targeted to specific conformer families and invoked only when needed.

5.2 Pucker mismatches

The most concerning category of incorrect predictions are the pucker mismatches, although they account for only about 8% of the total predictions. In these cases, the ribose pucker combination predicted from a test suite's Pperp distance test does not match the pucker-pairs combination of any of the probable homology-related conformation class answers. For example, 1y6s [39], chain B, suite 7-8 has a Pperp-predicted sugar pucker combination of P32. The probable answers for this site have different pucker combinations, either P33 (**1a** and **3g**) or P23 (**8d**).

Quality-filtered residues from HD and LD are cleanly separated into pucker-pairs by Pperp distance (Figure 6). Because the separation is so clear, MINT-AGE and RNAprecis treat each pucker-pairs set separately, and RNAprecis prediction across pucker-pairs sets is not possible. However, residues from the LDT test set can have ambiguous Pperp values near the 2.9Å cutoff (Figure 7). The otherwise useful Pperp criterion loses predictive power near this boundary.

A poor electron density map - or a poor fit to map - is the primary observed cause of pucker mismatches. If the phosphate is not modeled near the center of a high-density peak, or if the base and ribose are not each in a blob of density with plausible local relationships, then the basic assumptions of RNAprecis are not satisfied. Figure 13 shows two such cases. Even without a pucker mismatch, RNAprecis is unlikely to find a correct answer in these cases, since the modeled suite does not necessarily bear any relationship to the correct conformation. Pucker mismatches thus represent a more severe category of modeling error than RNAprecis is intended to address, where the conformation of the model is very far from the intended structure.

The test set likely under-represents the actual prevalence of regions with electron density maps that cannot confidently support a model. All suites in the test set were solved as a recognized conformer in a related structure, so truly disordered regions were selected against. While Pperp distances in cases of pucker mismatch are significantly closer on average to the 2.9 boundary, than for correct matches, the variation is very high such that this property is not suited to predict pucker mismatches. Regularized Mahalanobis distances from the nearest MINT-AGE cluster are also not strong predictors, due to the large variation in size and density of the clusters. Other methods that directly assess map or fit-to-map quality will be necessary to identify suites that are likely to contain pucker mismatches. In regions with very poor density, it is usually better simply to build no model at all, rather than try to predict conformations from compromised data.

An interesting alternate case is Suite 8 of the HIV-1 DIS homology class. This homology class is somewhat overrepresented in the test set, due to a family of closely related structures bound to different ions [39]. The B chain contains an alternate conformation at this site. The B alternate is in the recognized

conformation **3g**, which is a member of the probable answer set. The A alternate is in a **!!** conformation, and is therefore part of the test set. Pucker mismatches arise because the B alternate already fulfills the expected match to the answer set, and the A alternate is free to adopt a different – possibly very different - conformation. Instances of this alternate conformation account for 8 of the pucker mismatches in the P32 group.

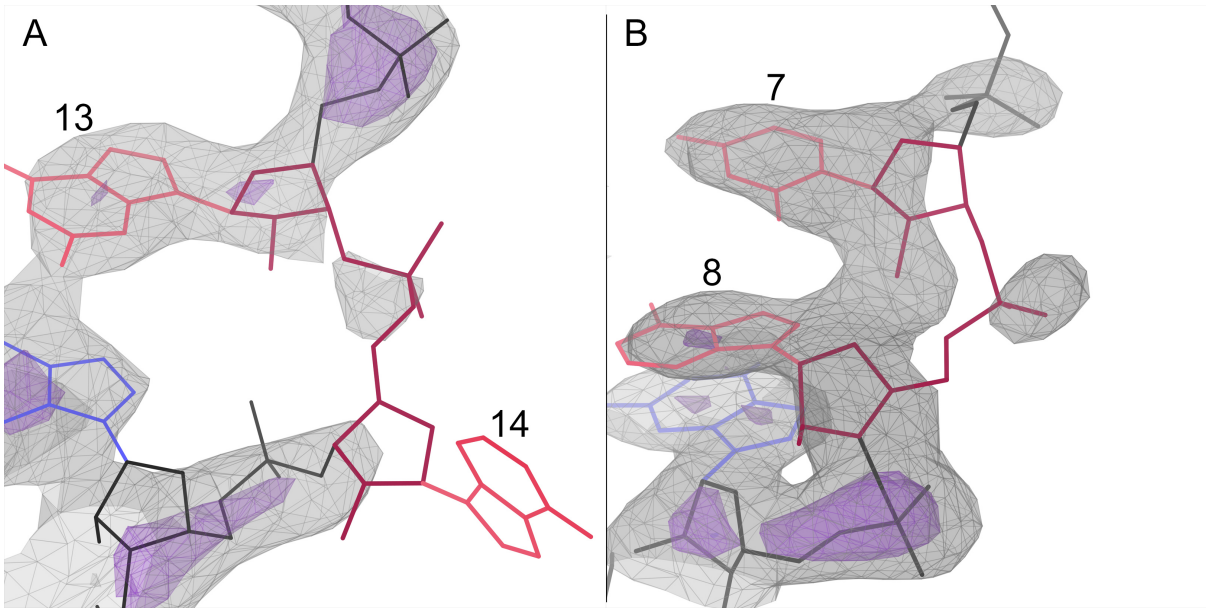

Figure 13: *Two examples where locally poor map density caused an RNAprecis pucker mismatch. A: Suite 14 of 5fk2, chain A([37]), from LDT. This !! outlier suite is modeled with a P22 sugar pucker-pair. The probable answer from related models is 1g, which has a P33 sugar pucker-psit. In this case the P of Residue 14 is at the edge of a quite low density peak, and the base and most of ribose of 14 are outside density. Density contours at* 1σ *(gray) and* 2σ *(purple). B: Suite 8 of 7kd1, chain A([40]) , from the LDT. This !! outlier suite is modeled with a P23 sugar pucker-pair. The probable answer from related models is 1a, which has a P33 sugar pucker-pair. In this case the backbone density is strong but unusual, and far offset from the modeled backbone. Density contours at* 1.2σ *(gray) and* 3σ *(purple).*

5.3 1a vs 1c prediction mismatches

The confusion matrix (Figure 12) for the LDTP33 portion of the test set shows that **1a** vs **1c** was one of the most difficult conformation pairs for RNAprecis to predict correctly. RNAprecis's prediction still matches the probable answer in the majority of cases – 58.6% for **1c** predictions, 77.2% for **1a**. **1a**/**1c** ambiguity is not unexpected.

A coordinated motion of α and γ allows $\approx 180°$ changes in those dihedral angles to occur with minimal effects on the parameters that RNAprecis uses (Figure 14). This motion resembles that of a physical "crankshaft", having opposite and equal changes in A and G that drive changes in C5'/O5' atom positions, while B is held unchanged. In the Figure 14 superpositions, the glycosidic bond vectors of the lower base are systematically different in **1a** versus **1c**, but the difference is subtle relative to the huge size and population of the **1a** conformer. RNAprecis therefore struggles to distinguish between these conformers.

We examined whether RNAprecis might be identifying the correct conformations in cases where the reference structure was mistaken. Examples from the reference structures of each sequence homology class that contained a **1a**/**1c** prediction mismatch were manually reviewed based on visual agreement with their electron density maps. Well-populated homology classes tended to have a clear consensus, so

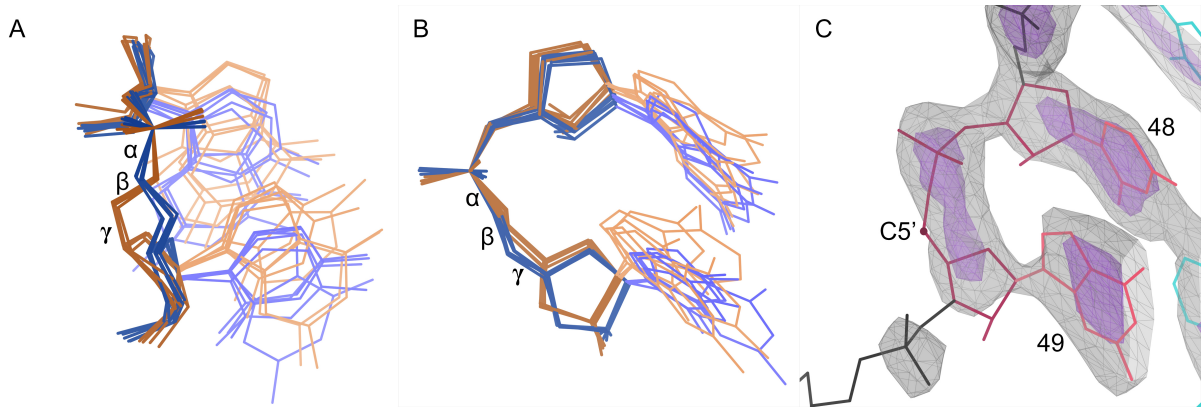

Figure 14: *Superposition of representative **1a** (orange) and **1c** (blue) suites. A. A view showing full backbone detail. **1a** and **1c** are easily distinguished by differences in the α and γ dihedrals. Therefore, MINT-AGE can easily distinguish these conformers. B. Another view of the same superposition, showing the alignment of the riboses and bases. Despite the large difference in α and γ, the related changes in glycosidic bond directions are relatively small. Therefore, RNAprecis cannot readily see the difference between **1a** and **1c**. C. A **1c** conformer modeled into density that suggests a **1a** at Suite 49 in 4kze chain R. The low-detail gray contour (1.2σ) shows how this error is possible; the position of the purple contour (2.2σ) indicates that the α/γ crankshaft should be reversed to move the C5' atom into the density peak.*

more attention was given to classes with few members.

Only one homology class's reference structures showed a clear modeling error (Figure 14), with a **1c** conformation where a **1a** should go (4kze [41] and 6b3k [42] at suite 49). All other examined reference structure suites showed fits to map ranging from plausible to definitely correct.

Manual review of the **1a**/**1c** prediction mismatches confirms that the homology-related probable answers for the test set are reliable enough to use as targets. Where MINT-AGE disagrees with the probable answers for this known difficult case, the homology-related conformation is generally correct.

The **1a**/**1c** ambiguity may be partially resolvable by increased reliance on the d_3 parameter in this case. d_3 is the distance across the relevant α, β, and γ dihedrals (see Figure 8). While there is some overlap between the Cluster 1/**1a** and Cluster 2/**1c** distributions in all parameters, d_3 provides the cleanest single-parameter distinction between the two, as seen in the d_2/d_3 panel of Figure 26 in the appendix.

5.4 Other prediction mismatches

The **2a** and **6n** conformers from LDTP23 also resulted in large numbers of prediction mismatches in the RNAprecis test. The **2a** and **6n** cases that resulted in prediction mismatches are largely drawn from two homology classes, and from sets of related riboswitch structures apparently from a single lab. As in the pucker mismatches at suite 8 of the HIV-1 DIS homology class discussed above, these cases are repetitions of related modeling challenges, rather than independent problems.

At suite 10 of the SAM riboswitch ([43]), where a **2a** conformer should be predicted, the first base of the suite has little to no density. This loss of information from the map breaks the underlying assumptions of RNAprecis and makes the RNAprecis parameters unreliable. At suite 64 of the guanine riboswitch ([44]), where a **6n** conformer should be predicted, the backbone P atom is out of the center of its density peak, and often shows nearby difference density suggesting that a rotation of the backbone is needed. This is a modeling error that damages the RNAprecis parameters. Indeed, misplacement of the P may be particularly damaging, since the P position is used as the origin of the RNAprecis parameter system.

The **3d**/**5j** conformers from LDTP33 result in a different type of prediction mismatch. These conformers come from the same SAM riboswitch homology class as the 2a conformers discussed above. In the highest resolution reference structure from this homology class, 5fjc (1.71Å, where the next highest,

3gx5, is 2.402Å), suite 61 is modeled in a **!!** outlier conformation. The modeled conformation is similar to the **3d** conformer, but differs by a crankshaft motion across the α, β, and γ dihedrals, similar to the difference between **1a** and **1c** conformers. However, the electron density clearly supports the modeled conformation. This is likely to be a genuine outlier conformation, and not an error to be corrected.

3d was one of the conformers that could not be clustered by MINT-AGE, due to too few examples in the training data. Therefore, to date, **3d** also cannot be predicted by RNAprecis. Because the above outlier conformation differs from **3d** in a detail that is difficult for RNAprecis to capture, it is likely that this outlier conformation would have been (incorrectly) predicted as **3d** if that were possible. This increases the importance of developing additional methods to resolve the α/γ crankshaft ambiguity.

The test set guards against inclusion of genuine outliers like this one by requiring that its member suite be solved in a non-!! conformation in a related structure. In this case, the lower-resolution related structures contained decoy **3d** solutions, because the low-resolution density envelope cannot easily distinguish the α/γ crankshaft. Adding residue-level quality checks, like those used in preparing the training data, to the test set would further enhance the confidence of the probable answers and better guard against decoy answers.

6 Outlook

Ultimately, we intend to develop RNAprecis predictions into an RNA structure validation and recommendation tool for use in Phenix [45], MolProbity, and similar platforms.

Challenges in RNAprecis prediction have identified additional targeted structural tools to be developed alongside RNAprecis. The **1a**/**1c** ambiguity may yet be resolvable with the RNAprecis parameters such as the d_3 distance. But such difficult cases might also be addressed with paired, comparative local refinements, once RNAprecis identifies a likely site. Automatically identifying suites where the critical RNAprecis atoms are not well-aligned with the density will be a more challenging problem, but would benefit both model building and RNAprecis predictions, especially avoiding pucker mismatches. RNAprecis predictions will be enhanced by development of a prediction confidence measure, likely using Mahalanobis distance to the predicted cluster, probability contours based on cluster distribution, or both.

MINT-AGE clustering provides an opportunity to revisit and expand previous suite definitions. Suite conformers such as **0a** and **7p** which contain multiple subclusters with distinct enough conformations to merit their own suite identities have been discussed above. Additionally, there may exist additional valid suite conformers that were too rare to be detected at all at the time of the previous work. In the current work, unrecognized suite conformers (**!!**s) were intentionally excluded from training. MINT-AGE clustering can be performed on the **!!**s as well, to identify any new clusters of sufficient population and consistency to merit recognition as suite conformers, though human inspection will be required to distinguish between real conformations and decoys derived from common modeling errors.

Acknowledgements

BE, SH and HW acknowledge support from DFG SFB 1456; SH acknowledges support from DFG HU-1575-7 and DFG GK 2088; CJW, MGP, VBC, and JSR acknowledge support from NIH P01-GM063210 and the Phenix Industrial Consortium.

Appendices

A High detail training data set (HD)

This section introduces in Table 3 the PDB files with the corresponding resolution of the high resolution training data set (introduced in Section 2.4), from which the HD and LD data sets (and the corresponding

four pucker-pairs sets) are composed. In Figure 15 the four pucker-pairs sets of HD are plotted and in Figure 16 the four sub-data sets of LD are plotted.

PDB	Res	PDB	Res	PDB	Res	PDB	Res
2jLv	1.90	3t5n	1.79	4z4d	1.60	6cmn	1.80
2v7r	1.20	3td0	1.60	4zLd	1.60	6d92	1.81
2xLk	1.80	3zp8	1.55	5L00	1.25	6db8	1.87
2xs7	1.45	4Lgt	1.30	5L2L	1.55	6dtd	1.65
2y8w	1.80	4Lvw	1.77	5Lr3	1.65	6du4	1.70
2y8y	1.44	4aLp	1.48	5Lr4	1.80	6e1w	1.69
2zko	1.70	4d25	1.90	5ay3	1.20	6e8u	1.55
2zy6	1.75	4fen	1.35	5ay4	1.70	6f4g	1.90
3La5	1.70	4gv6	1.98	5b2q	1.70	6gc5	1.90
3bso	1.74	4j39	1.70	5c5w	1.25	6gd2	1.90
3c3z	1.50	4j50	1.65	5da6	1.05	6h0r	1.73
3cgp	1.57	4jah	1.50	5dhb	1.80	6h9h	1.75
3cgs	1.65	4jrd	1.00	5eLh	1.80	6hau	1.86
3d2s	1.70	4k31	1.42	5eim	1.54	6hc5	1.41
3dd2	1.90	4mcf	1.90	5ev4	1.57	6i0v	1.85
3diL	1.90	4mdx	1.50	5ew4	1.47	6iv9	1.86
3gLp	1.23	4ms9	1.32	5fjc	1.71	6jjh	1.74
3gvn	1.20	4n0t	1.70	5gjb	1.70	6kwq	1.76
3h5y	1.77	4nLf	1.00	5h1k	1.90	6oon	1.90
3i62	1.95	4p3t	1.60	5hby	1.18	6ozn	1.90
3iev	1.90	4p97	1.86	5jaj	1.50	6ppq	1.81
3jxq	1.45	4pcj	1.90	5jrc	1.90	6q1h	1.45
3nd3	1.37	4pco	1.32	5kLa	1.14	6qit	1.50
3nd4	1.52	4qpx	1.86	5mwi	1.51	6sce	1.83
3nmr	1.85	4qvi	1.90	5nxt	1.38	6tqb	1.60
3oin	1.90	4rbq	1.05	5udi	1.58	6u8d	1.81
3pdr	1.85	4rby	1.19	5v3f	1.70	6ugg	1.95
3pey	1.40	4rc0	1.13	5vj9	1.57	6ugi	1.75
3qgc	1.90	4rne	1.01	5wLh	1.80	6uv4	1.70
3r1d	1.45	4u3L	1.48	5www	1.80	8a3d	1.67
3rer	1.70	4wan	1.80	6b14	1.64	8b0x	1.55
3sj2	1.36	4xw0	1.81	6cb3	1.89		
3syw	1.57	4xwf	1.80				

Table 3: PDB IDs (sorted in alphabetical order) and resolutions of the different measurements of the HD data set (introduced in Section 2.4 in the main text)

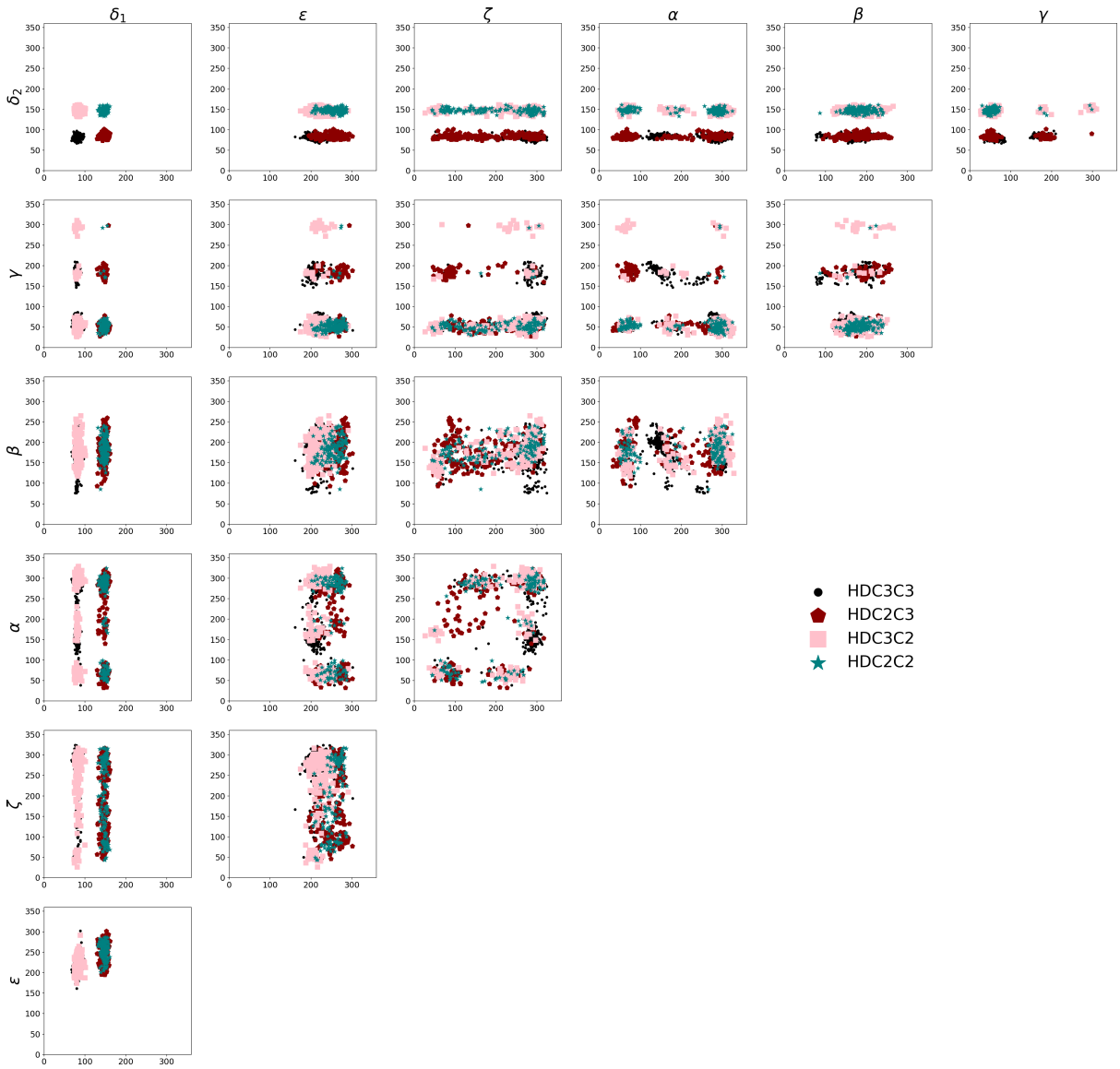

Figure 15: The HD representation of the high resolution training data set, resolved in pucker-pairs (introduced in Section 2.4 in the main text), represented by scatter plots of all two-dimensional dihedral angle pairs.

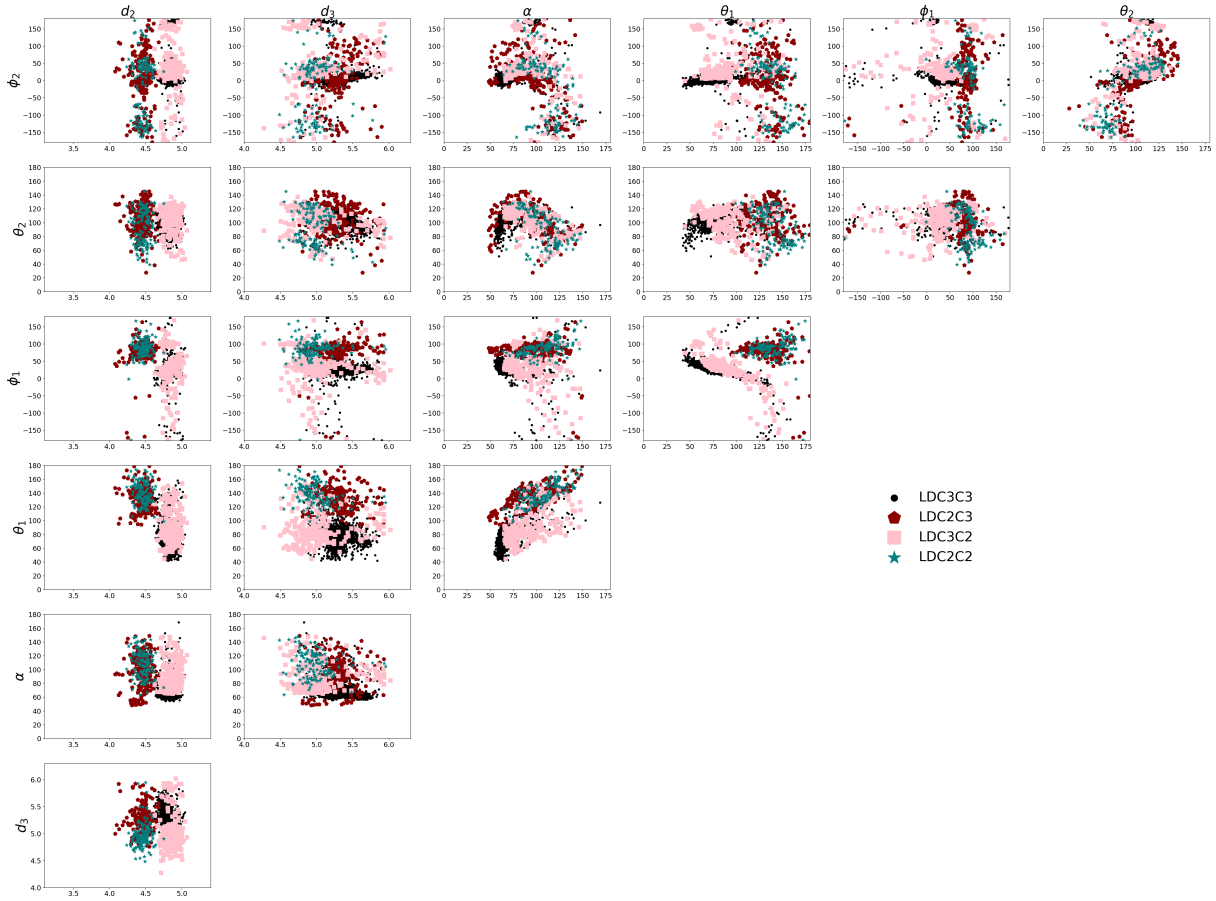

Figure 16: The LD representation of the high resolution training data set, resolved in pucker-pairs (introduced in Section 2.4 in the main text), represented by scatter plots of all combinations of parameters (described in Section 3.3 in the main text).

B Low detail test data set (LDT)

This section introduces in Table 4 and Table 5 the PDB files, the name of the chain and the residue number of the elements of the low resolution training data set (introduced in Section 2.4), from which the LDT data set (and the corresponding four pucker-pairs sets) is composed. In Figure 17 the four pucker-pairs sets of LDT are plotted.

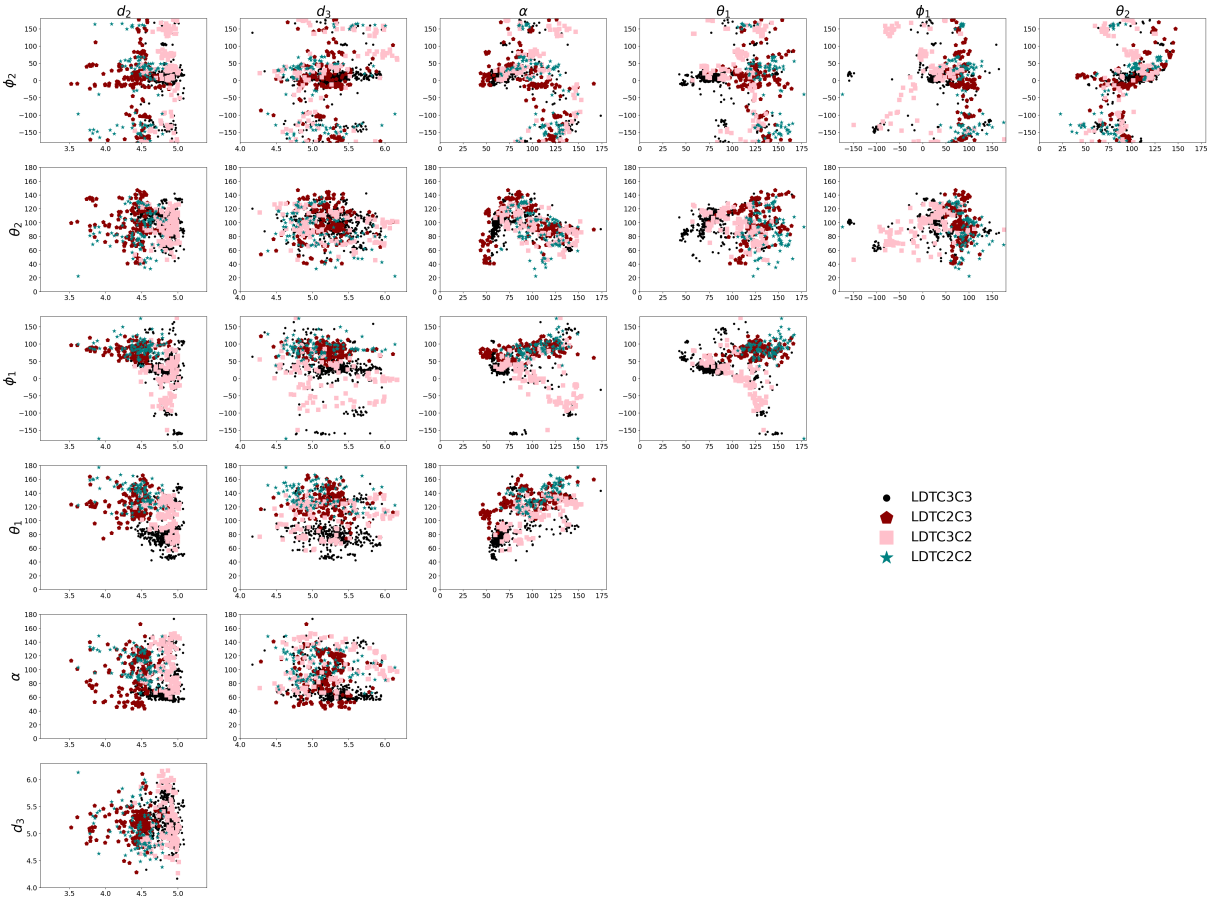

Figure 17: The LDT representation of the low resolution test data set, resolved in pucker-pairs (introduced in Section 2.4 in the main text), represented by scatter plots of all combinations of parameters (described in Section 3.3 in the main text).

PDB	chain name and residue number	Res
1c0a	B637	2.40
1nLc	A8, A9, B4, B8	1.85
1o3z	A8, A9, B8, B9	2.65
1q9a	A2654	1.04
1wvd	A8, A9, B4, B8, B9	2.93
1y6s	A8, B8	2.90
1y6t	A8, B2, B8	2.60
1y73	A8, A9, B8	2.90
1y90	A8, A9, B9	3.08
1y95	A8, A9, B8	2.80
1y99	A8	2.40
2ees	A16, A39, A40, A48, A50, A51	1.75
”	A54, A55, A64, A78, A79, A80	”
2eet	A16, A18, A20, A21, A22, A45	1.95
”	A54, A63, A64	”
2eeu	A18, A22, A23, A50, A52, A53	1.95
”	A54, A63, A64, A66, A67, A68	”
”	A69	”
2eev	A16, A22, A27, A54, A62, A63	1.95
”	A64, A80	”
”	A16, A39, A49, A54, A63, A64	”
2g9c	A22, A54	1.70
2gis	A10, A13, A14, A15, A16, A21	2.90
”	A24, A33, A34, A35, A47, A50	”
”	A52, A53, A57, A61, A64, A65	”
”	A69, A74, A9	”
2oij	A8, B8	2.31
2qek	A10, B8	1.80
2xLi	B17	2.33
2ydh	A10, A14, A15, A16, A2, A20	2.90
”	A24, A30, A33, A34, A35, A53	”
”	A61, A65, A69, A76, A77, A78	”
”	A83, A9	”
2ygh	A20, A28, A35, A54, A57, A61	2.60
”	A65, A9	”
3bsn	T6, T7	1.80
3c44	A8, B8	2.00
3cgq	B21	2.55
3diL	A128, A14, A39, A40, A49, A60	1.90
”	A72, A82	”
3dig	X114, X128, X14, X154, X40, X44	2.80
”	X46, X60, X63, X72, X82	”
3dim	A11, A113, A12, A13, A14, A154	2.90
”	A168, A2, A41, A43, A44, A46	”
”	A52, A60, A63, A72, A82	”
3dio	X128, X151, X154, X164, X35, X41	2.40
”	X61, X82	”
3diq	A128, A14, A154, A168, A41, A42	2.70
”	A43, A50, A52, A54, A60, A72	”
”	A82	”
”	A63, A81	”

PDB	chain name and residue number	Res
3dis	A11, A117, A12, A126, A13, A14	3.10
”	A154, A164, A165, A35, A39, A40	”
”	A41, A42, A43, A44, A50, A52	”
3dix	A114, A12, A128, A14, A42, A43	2.90
”	A52, A70, A72, A82	”
3diy	A11, A12, A128, A14, A150, A154	2.71
”	A34, A36, A39, A40, A52, A82	”
3diz	A126, A128, A14, A154, A164, A172	2.85
”	A39, A40, A41, A42, A43, A44	”
”	A46, A50, A52, A60, A70, A72	”
”	A82	”
3dj2	A128, A14, A154, A168, A40, A41	2.50
”	A52, A60, A72	”
3ds7	B223, B249, B250, B252	1.85
3dvv	A8, A9, B8	2.00
3dw4	A2671	0.97
3dw5	A2648	0.96
3far	B8	2.40
3fo6	A49, A54, A63, A64	1.90
3g4m	A17, A27, A39, A50, A54, A63	2.40
”	A64	”
”	A16, A36, A49, A50, A54, A64	”
3ger	A16, A49, A54, A63, A64	1.70
”	A27, A48, A49, A54, A63, A64	”
3got	A16, A18, A48, A49, A50, A54	1.95
”	A64, A80	”
3gx2	A10, A11, A13, A14, A15, A16	2.90
”	A17, A20, A21, A22, A33, A34	”
”	A35, A36, A37, A47, A53, A57	”
”	A61, A64, A65, A66, A69, A70	”
”	A71, A74, A9	”
3gx3	A10, A11, A14, A15, A16, A17	2.70
”	A20, A21, A22, A24, A28, A33	”
”	A34, A46, A47, A53, A57, A61	”
”	A63, A64, A68, A69, A7, A70	”
”	A78, A92	”
3gx5	A10, A14, A16, A20, A24, A30	2.40
”	A33, A34, A35, A36, A53, A61	”
”	A64, A65, A69	”
3gx6	A10, A13, A14, A15, A16, A20	2.80
”	A21, A24, A33, A34, A35, A36	”
”	A52, A53, A57, A61, A64, A65	”
”	A69, A74, A9	”
3gx7	A10, A11, A13, A14, A15, A16	2.95
”	A17, A18, A20, A21, A22, A24	”
”	A28, A33, A34, A35, A36, A37	”
”	A46, A47, A52, A55, A56, A57	”
”	A61, A64, A65, A66, A68, A69	”
”	A70, A71, A78, A9	”

Table 4: First part (from 1c0a to 3gx7) of the LD data set from Section 2.4 in the main text. The PDB ID, the name of the chain and the residue number of the second residue in the suite and the resolution of the respective measurement are listed in alphabetical order.

PDB	chain name and residue number	Res
3iqn	A10, A14, A15, A16, A17, A20	2.70
”	A21, A24, A25, A34, A35, A46	”
”	A47, A50, A51, A52, A54, A61	”
”	A64, A65, A68, A69, A70, A71	”
”	A72, A74, A75, A78, A83	”
3iqp	A10, A13, A14, A15, A20, A21	2.90
”	A23, A24, A33, A34, A35, A36	”
”	A50, A52, A54, A58, A61, A63	”
”	A64, A65, A69, A71, A73, A74	”
”	A78, A9	”
3iqr	A10, A13, A14, A20, A21, A24	2.55
”	A33, A34, A35, A47, A51, A52	”
”	A61, A63, A64, A65, A68, A69	”
”	A74, A9	”
3jxq	A4	1.45
3oij	C3	3.00
3syw	A11, A12, A13, A15, A16, B11	1.57
”	B12, B13, B15, B16	”
3t5q	L5	3.00
3u4m	B2116, B2148, B2159, B2160	2.00
3u56	B2159, B2183	2.10
3umy	B2116, B2145, B2182	1.90
480d	A2648, A2655	1.50
4Lvv	A54, A82	2.10
4Lvw	A43	1.77
4Lvx	A43, A54	1.90
4Lvy	A55	2.00
4Lvz	A43	1.77
4Lw0	A43	1.89
4b5r	A10, A15, A17, A18, A2, A24	2.95
”	A3, A33, A34, A35, A36, A39	”
”	A4, A41, A47, A49, A52, A54	”
”	A55, A60, A61, A65, A69, A78	”
”	A83, A9, A90, A92	”
4fe5	B16, B18, B22	1.32
4feL	B18, B22	1.60
4fej	B16, B22, B73	1.50
4fen	B22, B73	1.35
4feo	B16, B18	1.60
4fep	B50, B73	1.65
4fnj	A31, A4, A5	1.95
4kze	R28, R62	2.40
4mcf	E5	1.90
4qg3	B2126, B2148, B2159, B2173	2.00
4qvi	B2148, B2159, B2160	1.90
4u3L	B7	1.48
4u3p	B7	1.87
4xw7	A25, A28	2.50
4xwf	A27	1.80
4yn6	A10, A12, A13, A15, A3, B13	2.30
”	B2, B3	”
5b2o	B18	1.70
5c7u	B16, B17, B22, B37, B38, B39	3.05
”	B48, B50, B54, B55, B59, B60	”
”	B62	”
5d99	A2655	0.97
5fk1	A13, A14, A15, A17, A2, A21	2.50
”	A33, A47, A51, A54, A57, A61	”
”	A64, A65, A68, A78, A91	”

PDB	chain name and residue number	Res
5fk2	A10, A14, A15, A16, A17, A21	2.60
”	A35, A57, A61, A65, A77, A78	”
”	A83, A9, A92	”
5fk3	A10, A13, A14, A2, A21, A24	2.50
”	A34, A35, A51, A52, A57, A65	”
”	A77, A78, A86, A9	”
5fk5	A10, A13, A14, A15, A2, A21	3.31
”	A28, A33, A35, A46, A57, A61	”
”	A65, A78, A9, A92	”
5fk6	A14, A15, A21, A24, A51, A55	2.50
”	A57, A61, A65, A83, A9, A92	”
5fkd	A10, A13, A14, A15, A2, A21	3.00
”	A38, A46, A48, A55, A56, A61	”
”	A62, A65, A74, A76, A87, A9	”
5fke	A13, A14, A15, A2, A21, A33	2.80
”	A34, A54, A61, A62, A65, A78	”
”	A92	”
5fkf	A14, A15, A2, A21, A3, A34	2.80
”	A35, A44, A51, A54, A57, A65	”
”	A74, A78, A9	”
5fkg	A13, A14, A15, A2, A31, A33	2.95
”	A34, A61, A65, A78, A92	”
5fkh	A10, A13, A14, A15, A2, A21	2.65
”	A35, A53, A57, A61, A62, A65	”
”	A78, A9	”
5neo	A8	1.69
5npm	B15, B4, B55, B8	2.70
6b14	R28, R49	1.64
6b3k	R29, R58, R62	2.09
6d8a	H17	2.25
6d8f	H11, H17	2.15
6d92	H11, H12, H17	1.81
6d95	H12	1.85
6d9k	C12	2.00
6q57	A35, A43, A54, A83, A88	2.72
6qis	G5, G6	1.99
6qit	D6	1.50
6ubu	B16, B18, B22, B28	1.60
6uc7	B16, B22, B74, B75	1.80
6uc8	B16, B44	1.90
6uc9	B16, B22	1.94
6ugg	A12, A20, A22, A47, A48	1.95
6xh0	D36	3.10
6xh1	D34	2.60
6xh2	D34	1.71
6y0t	A2648, A2660	1.39
7eaf	A14, A15, A23, A24, A25, A34	2.85
”	A35, A57, A62, A77, A78, A80	”
”	A81, A9	”
7jje	A2648	1.25
7kd1	A21, A22, A35, A4, A43, A44	1.90
”	A45, A46, A79, A8	”
7m3v	A17, A19	2.19
7qp2	A2654	0.90
7ucr	A48	1.00

Table 5: The second part (from 3iqn to 7ucr) of the LD data set from Section 2.4 in the main text. The PDB ID, the name of the chain and the residue number of the second residue in the suite and the resolution of the respective measurement are listed in alphabetical order.

C Clustering results of HD and comparison with suitename

This section shows the clustering results of the four datasets HDP33, HDC2C3, HDP32 and HDP22 with the MINT-AGE clustering. The tuning parameters required for MINT-AGE are listed in Table 1 in the main text. The cluster results for the four different data sets are plotted in Figure 18, Figure 19, Figure 20 and Figure 21. In addition, the conformer class was determined for all elements of the four pucker-pairs sets with `suitename` software. A detailed comparison between the MINT-AGE cluster results and the conformer classification with `suitename` is provided in Table 6 and in the confusion matrices in Figure 22, Figure 24, Figure 23 and Figure 25.

HDP33

Cluster	nr elements	suite conformers
1	3214	1a (3179), 1L (14), &a (13), 1m (7), 7a (1)
2	146	1c (146)
3	43	1g (43)
4	17	9a (10), 3a (6), 7a (1)
5	11	1e (11)
6	7	1f (7)
7	4	1e (2), 1f (2)
8	3	7d (3)
9	3	5d (3)
10	3	5j (3)
11	3	1f (3)
outlier	75	

HDP23

Cluster	nr elements	suite conformers
1	57	2a (57)
2	22	0a (16), #a (4), 4a (2)
3	22	6n (19), 0i (2), 6j (1)
4	17	6g (15), 4g (2)
5	16	8d (14), 4d (2)
6	12	0a (12)
7	12	4a (12)
8	9	6d (9)
9	8	0i (8)
10	7	6j (7)
11	7	2g (6), 8d (1)
12	6	4n (6)
13	5	6d (5)
14	5	2h (5)
15	4	4g (4)
16	3	4g (2), 4a (1)
outlier	4	

HDP32

Cluster	nr elements	suite conformers
1	145	1b (121), 1[(24)
2	23	5z (23)
3	19	3b (19)
4	13	7r (13)
5	10	1[(6), 1b (3), !! (1)
6	9	1z (9)
7	8	1t (8)
8	7	1o (7)
9	7	7p (7)
10	5	7p (5)
11	4	5p (4)
12	4	5p (4)
13	4	7p (4)
14	3	7p (3)
15	3	5q (3)
outlier	1	

HDP22

Cluster	nr elements	suite conformers
1	41	2[(41)
2	33	6p (32), !! (1)
3	17	4p (17)
4	15	4b (15)
5	8	2z (8)
6	6	2u (3), 2o (2), 2[(1)
7	6	4b (4), 0b (2)
8	4	0b (4)
9	4	0b (4)
10	3	4b (3)
outlier	1	

Table 6: Left and middle column of the tables: MINT-AGE cluster numbers and outliers (left column) with size (middle column) for the different data HDP33, HDP32, HDP23 and HDP22 (introduced in Section 2.4 in the main text). Right column: the corresponding number of elements in the suitename two-digit conformer classes (the name of the class is a number and a letter or a bracket).

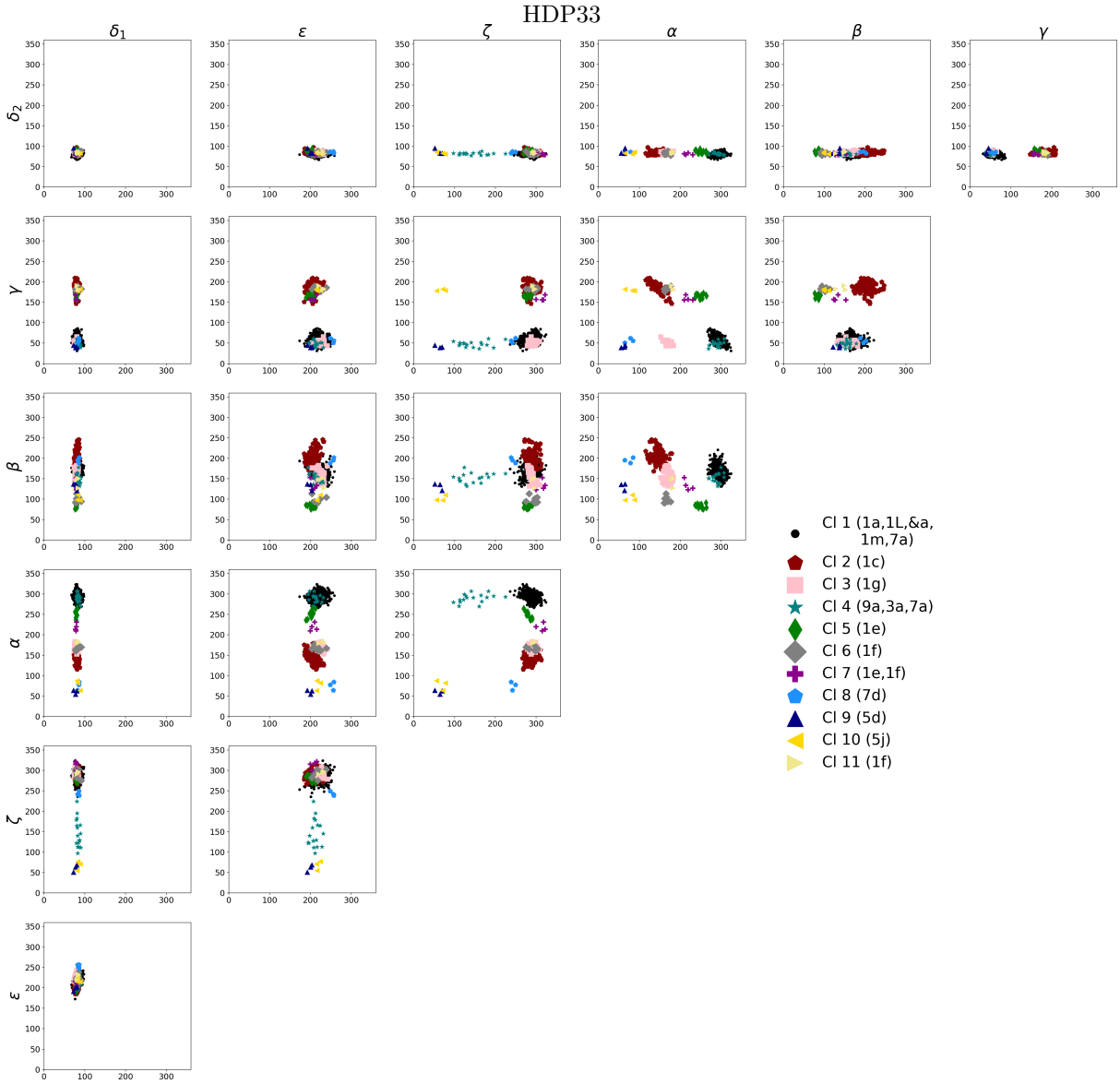

Figure 18: The MINT-AGE clusters (see Section 2.5 in the main text) from the data set HDP33 (introduced in Section 2.4 in the main text) represented by scatter plots of all two-dimensional dihedral angle pairs (given in degrees). See Table 6 for a detailed overview of the clusters and see Figure 22 for a detailed comparison with the conformer classification obtained by software `suitename`.

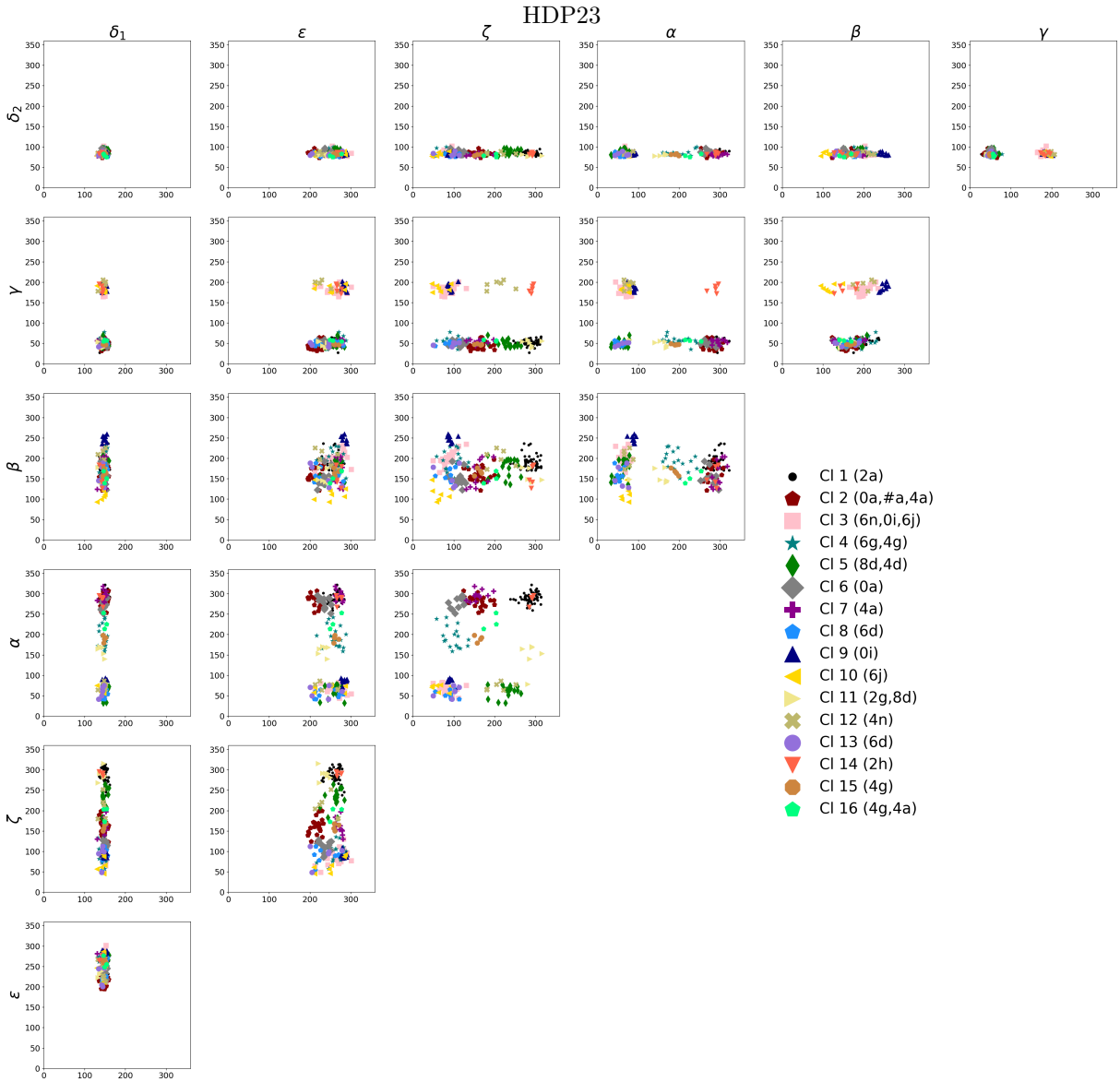

Figure 19: The MINT-AGE clusters (see Section 2.5 in the main text) from the data set HDP23 (introduced in Section 2.4 in the main text) represented by scatter plots of all two-dimensional dihedral angle pairs (given in degrees). See Table 6 for a detailed overview of the clusters and see Figure 24 for a detailed comparison with the conformer classification obtained by software `suitename`.

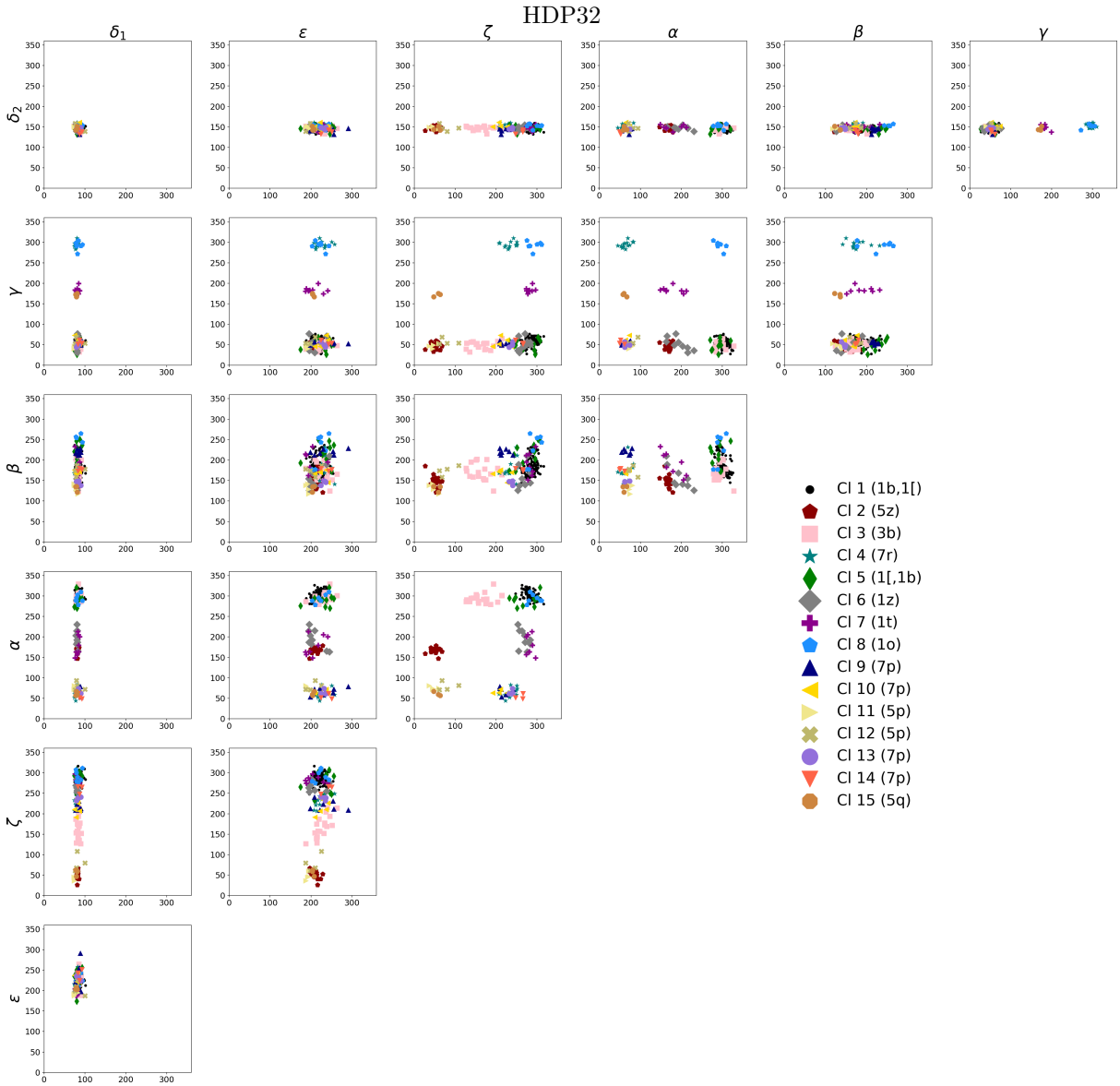

Figure 20: The MINT-AGE clusters (see Section 2.5 in the main text) from the data set HDP32 (introduced in Section 2.4 in the main text) represented by scatter plots of all two-dimensional dihedral angle pairs (given in degrees). See Table 6 for a detailed overview of the clusters and see Figure 23 for a detailed comparison with the conformer classification obtained by software `suitename`.

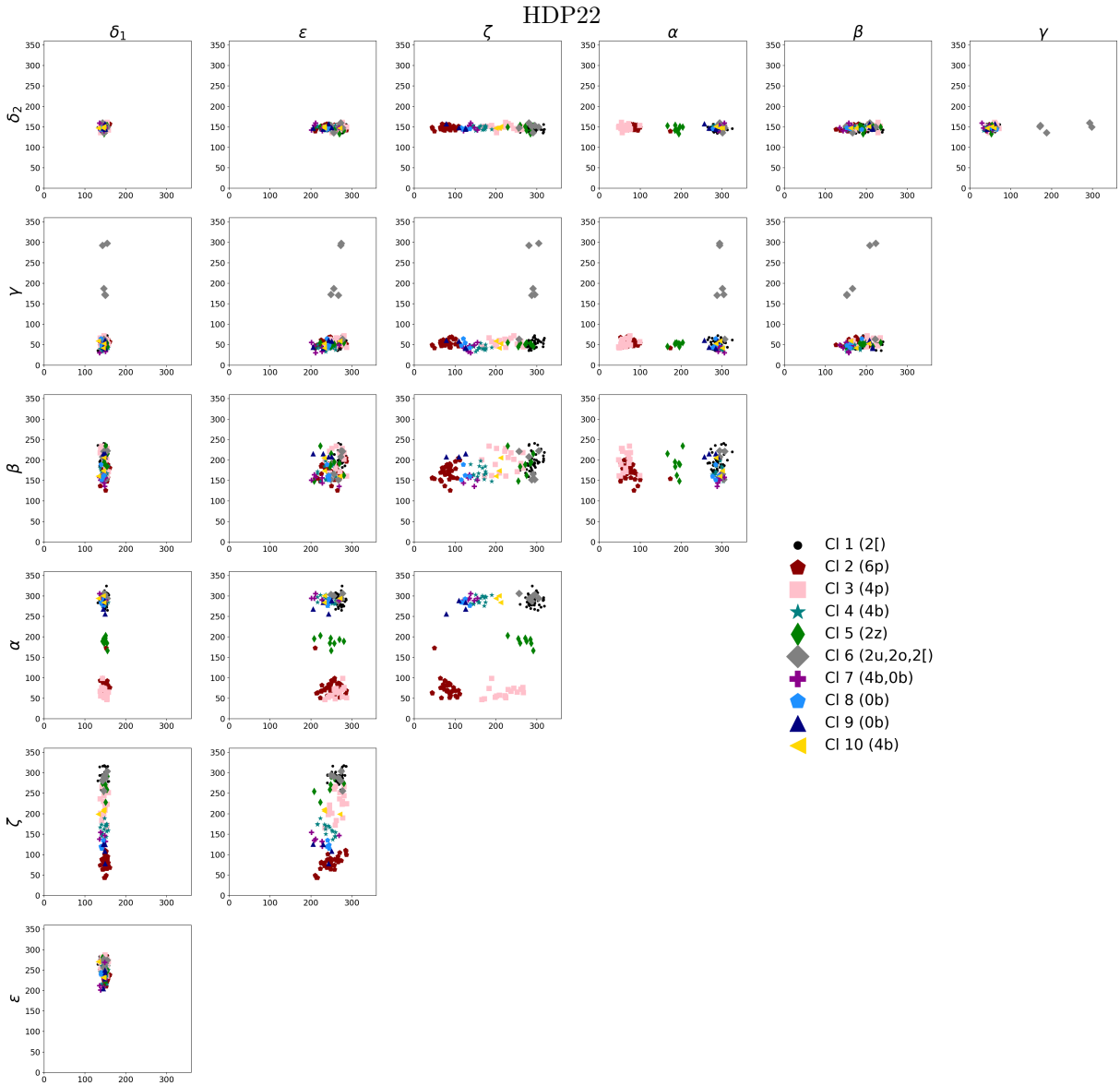

Figure 21: The MINT-AGE clusters (see Section 2.5 in the main text) from the data set HDP22 (introduced in Section 2.4 in the main text) represented by scatter plots of all two-dimensional dihedral angle pairs (given in degrees). See Table 6 for a detailed overview of the clusters and see Figure 24 for a detailed comparison with the conformer classification obtained by software `suitename`.

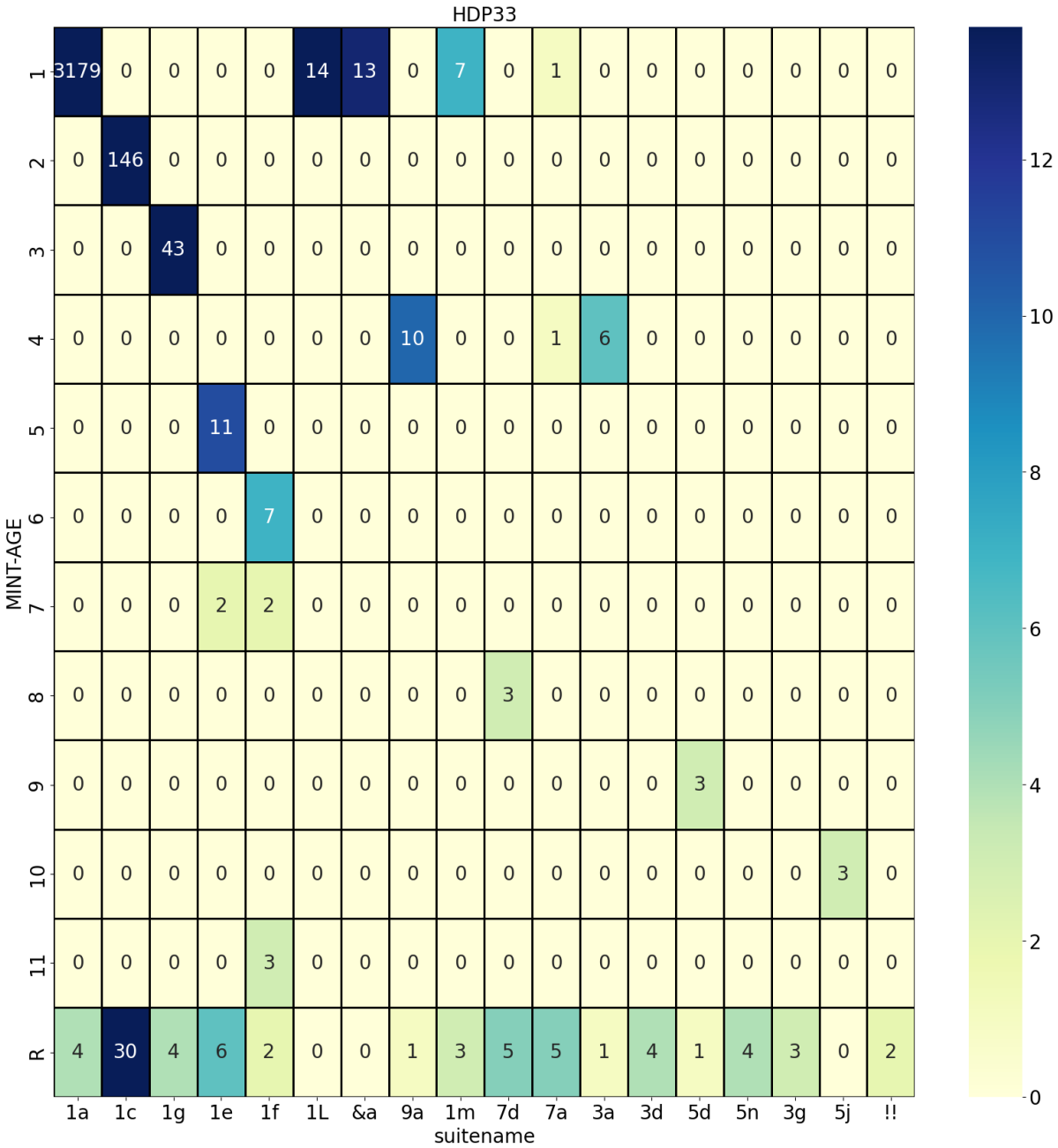

Figure 22: *Confusion matrix comparing the MINT-AGE cluster results with suite conformers for HDP33 training data. Every line in the matrix corresponds to a cluster and every column corresponds to a suite conformer. The last line (labeled R for rest) summarizes suites that were not assigned to any cluster by the MINT-AGE algorithm. One can see that most conformers are assigned to a single cluster with overwhelming majority. However, some clusters pool together several conformers.*

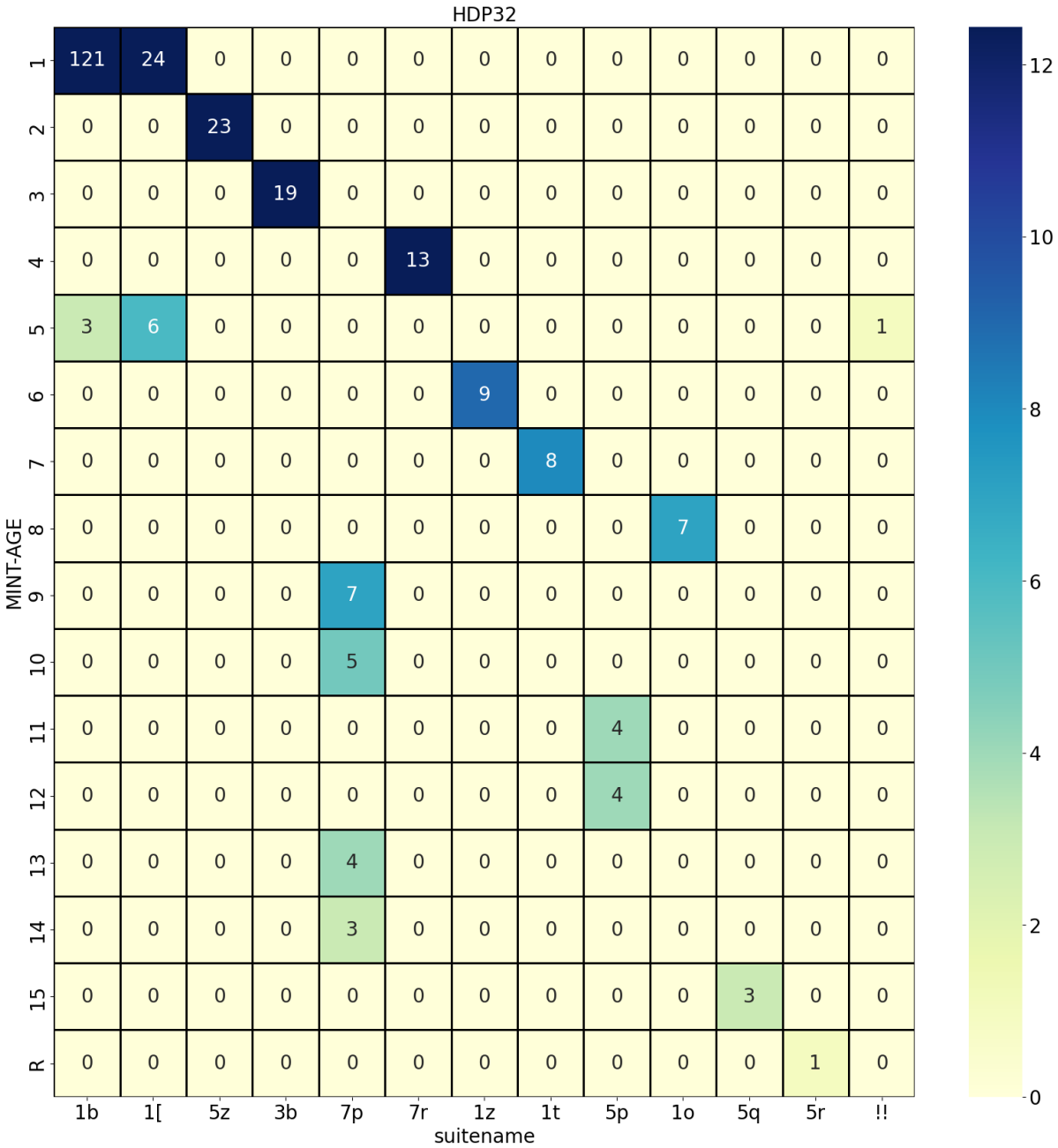

Figure 23: *Confusion matrix comparing the MINT-AGE cluster results with suite conformers for HDP32 training data. Every line in the matrix corresponds to a cluster and every column corresponds to a suite conformer. The last line summarizes outliers that were not assigned to any cluster by the MINT-AGE algorithm. Most clusters only comprise a single conformer, except for clusters 1 and 5 which contain both 1b and 1[conformers. Additionally, the 7p and 5p conformers are split up into several cluster.*

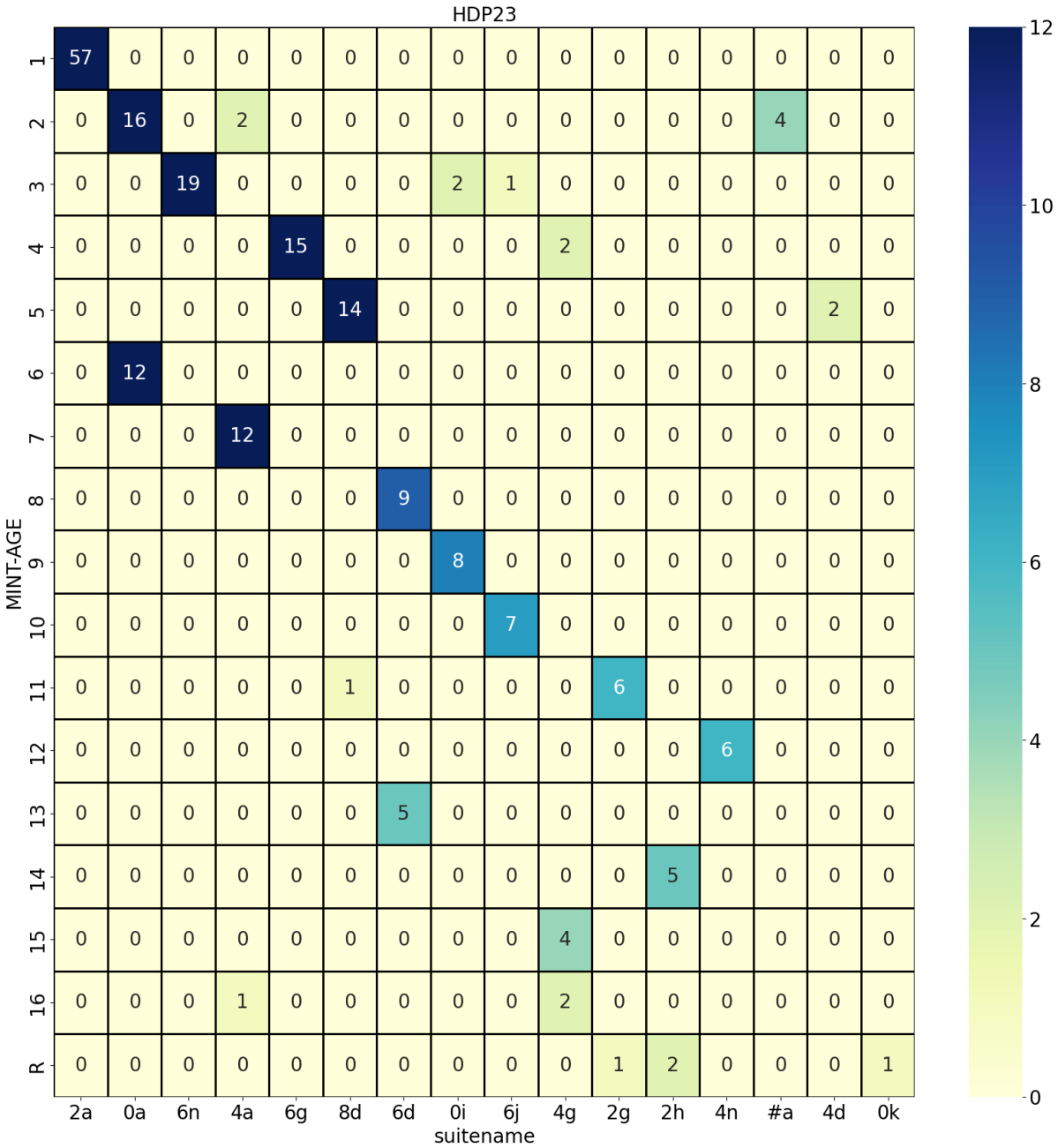

Figure 24: *Confusion matrix comparing the MINT-AGE cluster results with suite conformers for HDP23 training data. Every line in the matrix corresponds to a cluster and every column corresponds to a suite conformer. The last line summarizes outliers that were not assigned to any cluster by the MINT-AGE algorithm. Most clusters only comprise a single conformer, except for cluster 2 containing both 0a and #a conformers. Additionally, the 0a and 6d conformers are split up into several cluster.*

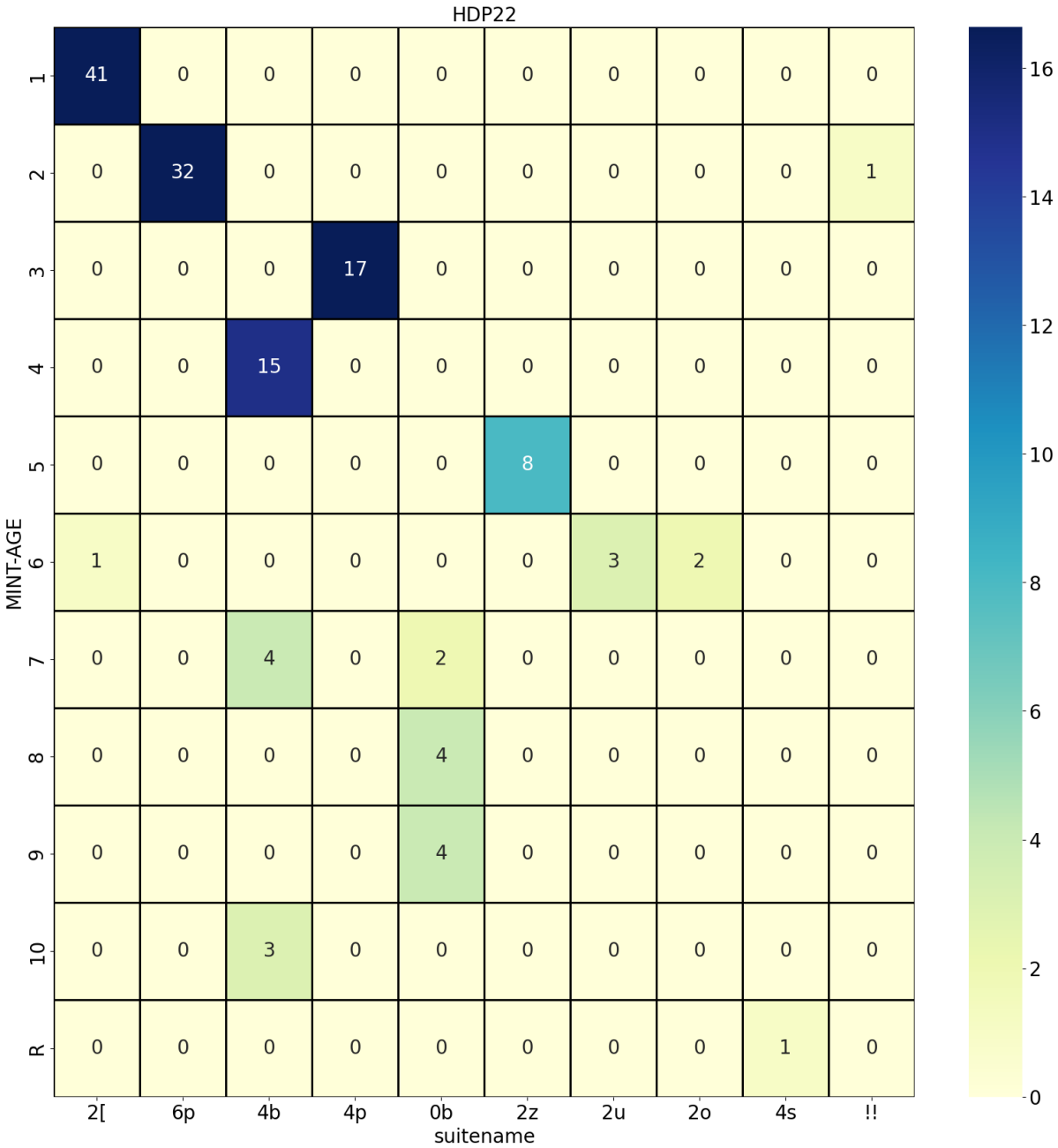

Figure 25: *Confusion matrix comparing the MINT-AGE cluster results with suite conformers for HDP22 training data. Every line in the matrix corresponds to a cluster and every column corresponds to a suite conformer. The last line summarizes outliers that were not assigned to any cluster by the MINT-AGE algorithm. While bigger clusters correspond pretty well with abundant conformers, the picture is less clear cut for smaller clusters and rare conformers.*

D Low detail (LD) parameter space

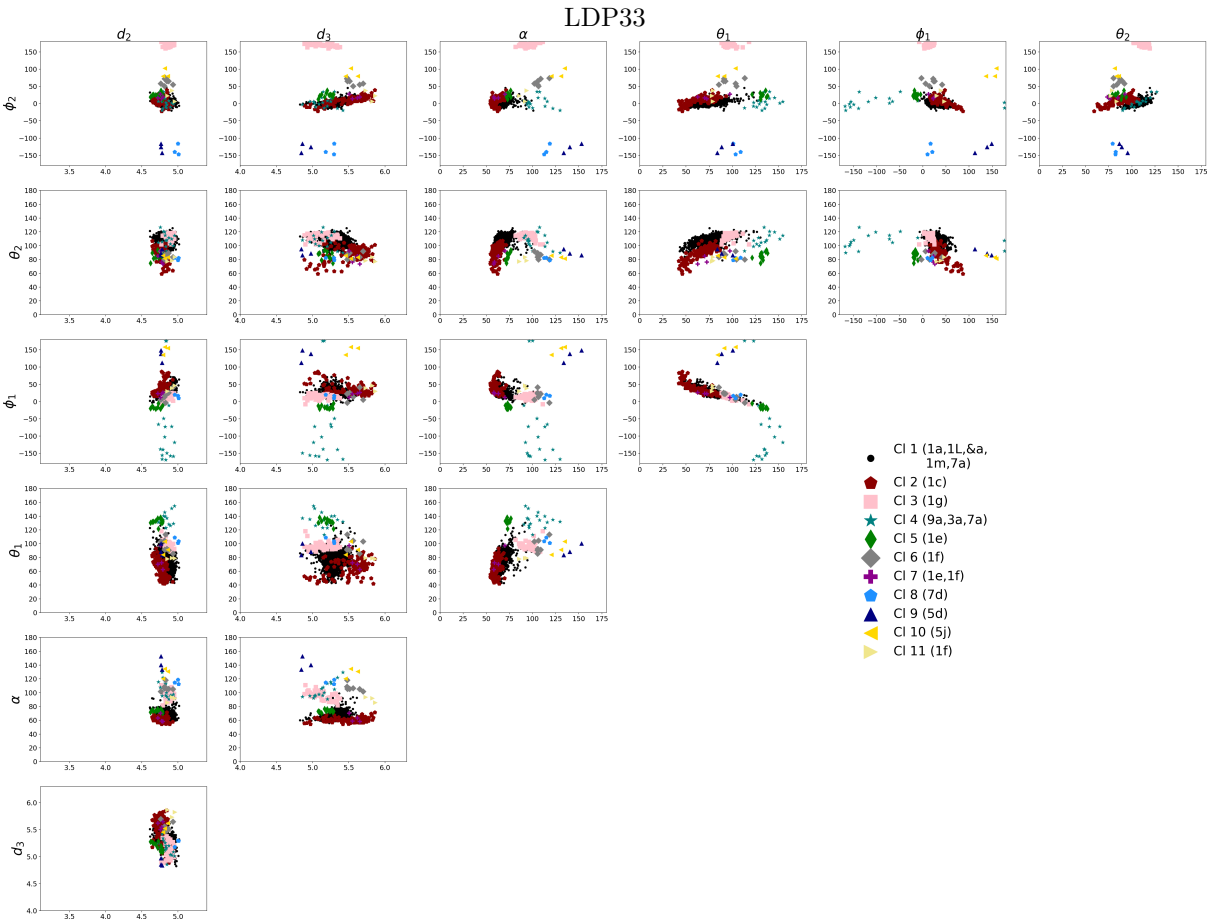

Figure 26: Scatter plot of LDP33 data in low detail parameter space introduced in Section 3.3 of the main text.

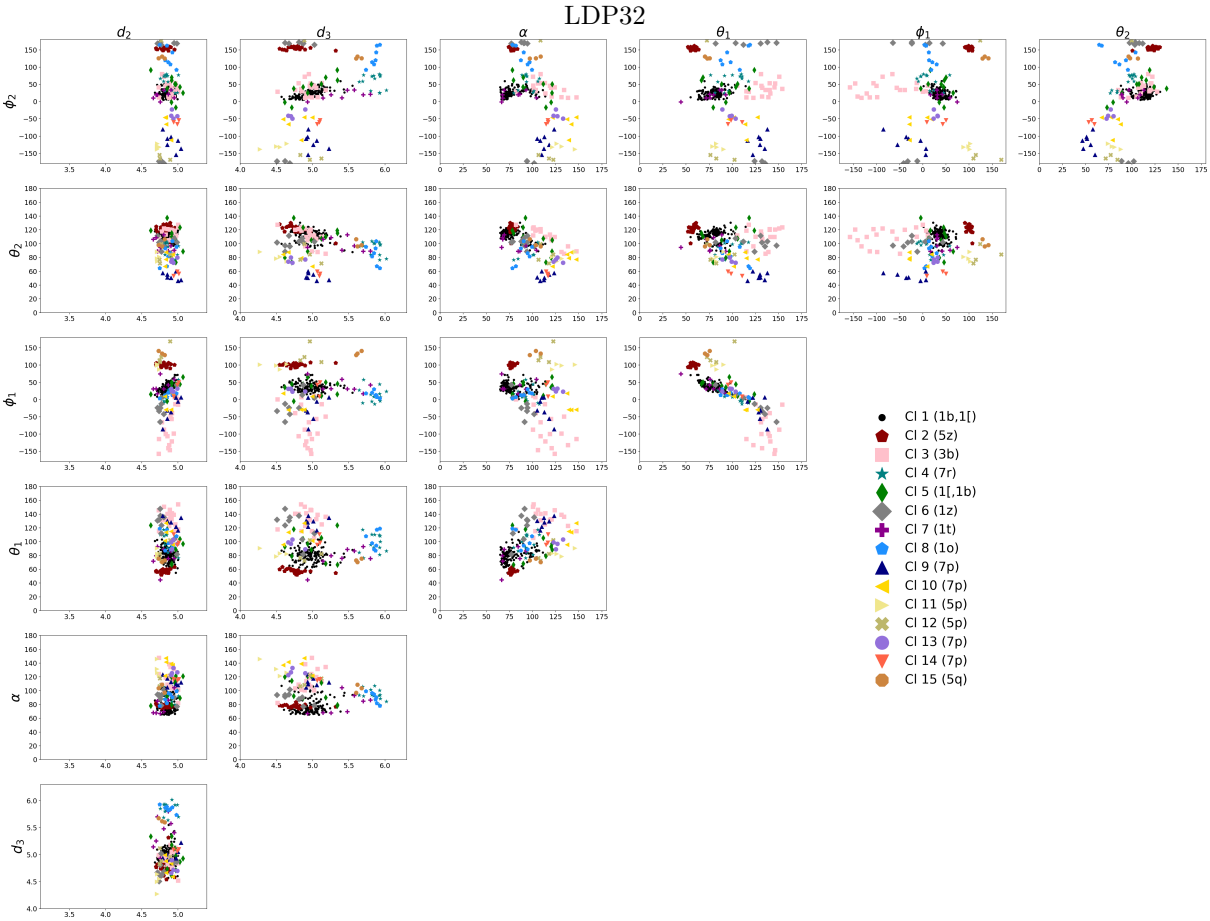

Figure 27: Scatter plot of LDP32 data in low detail parameter space introduced in Section 3.3 of the main text.

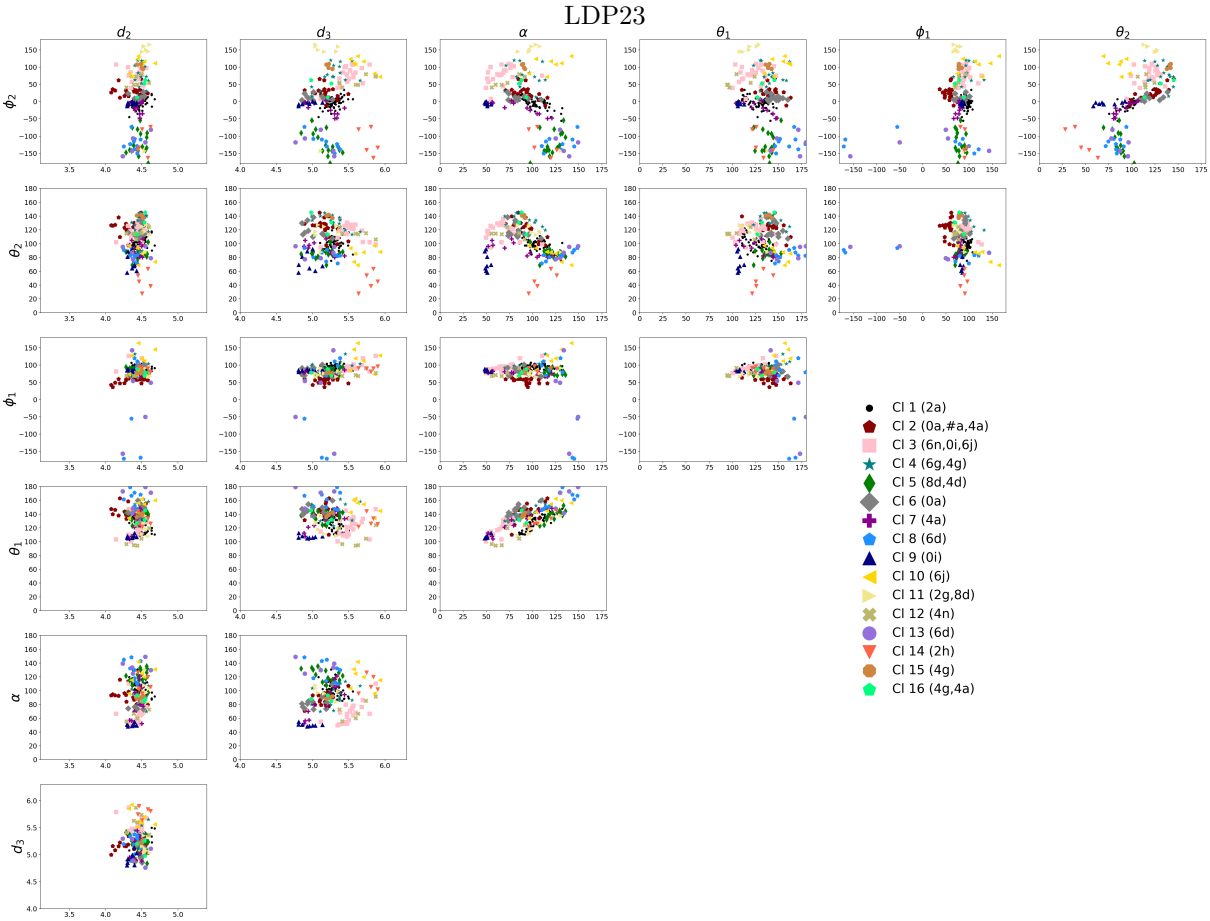

Figure 28: Scatter plot of LDP23 data in low detail parameter space introduced in Section 3.3 of the main text.

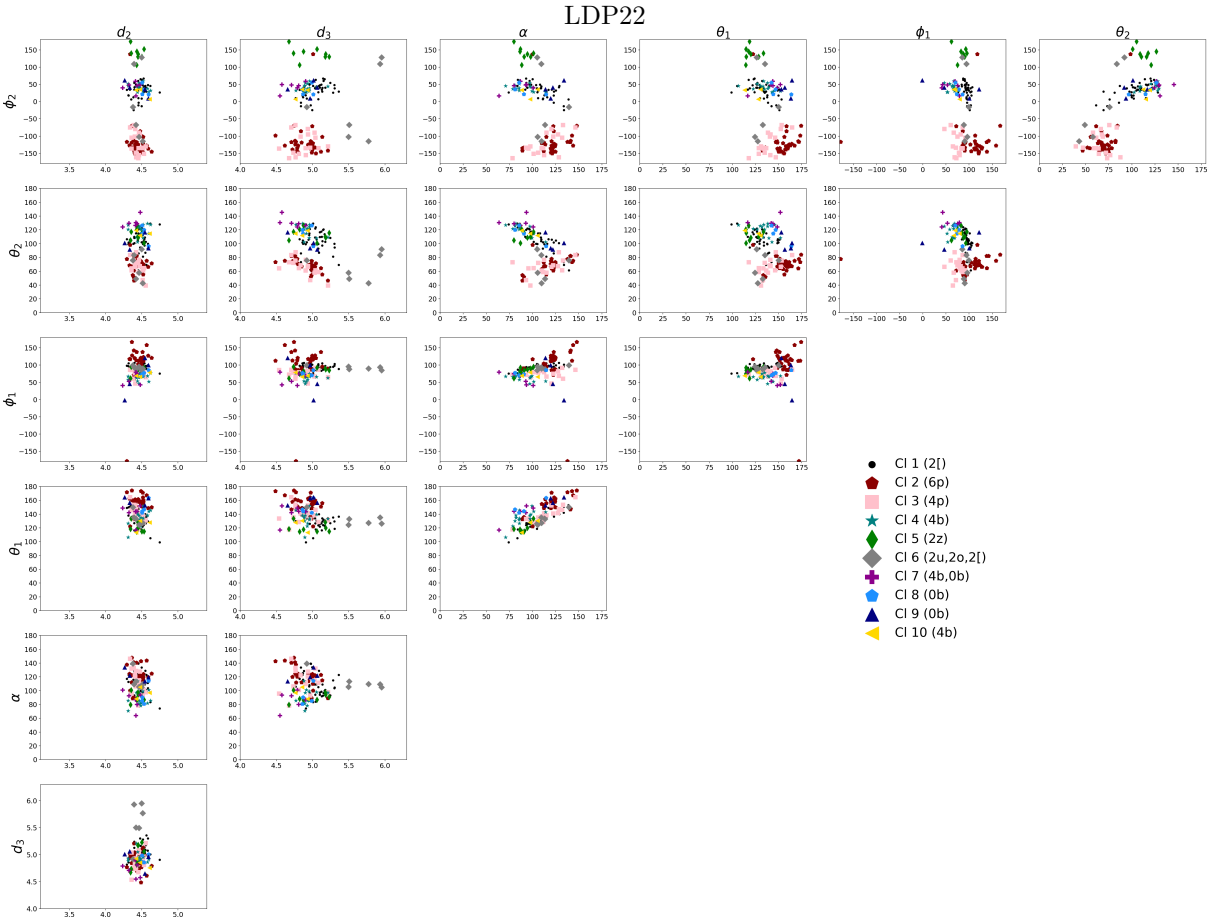

Figure 29: Scatter plot of LDP22 data in low detail parameter space introduced in Section 3.3 of the main text.

E Results RNAprecis

E.1 Results for the test data set LDTP33

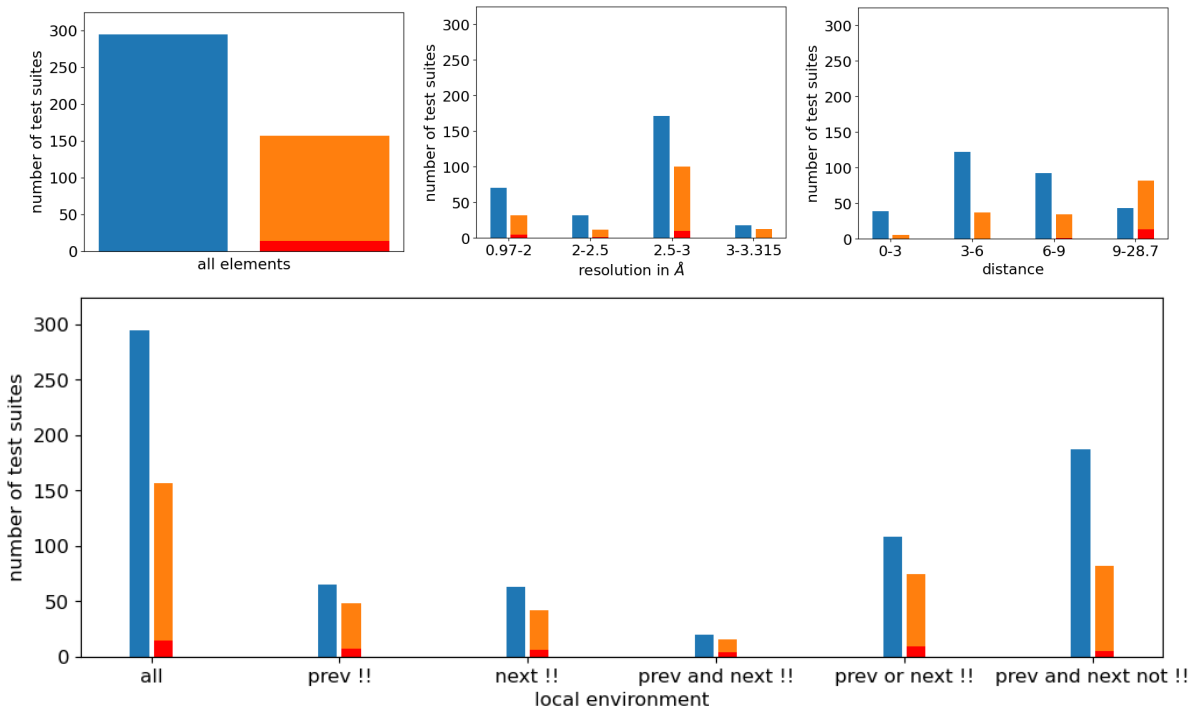

Figure 30: *Relative distribution of agreements and disagreements of MINT-AGE cluster prediction with homology related probable conformer classes for the test set LDT33 with color scheme (reflecting blue for match, orange for mismatch and red for pucker-pair mismatch) from Section 4.4 in the main text). Top left: overall histogram. Top middle: histogram reflecting varying resolutions of PDB suites. Top right: histogram reflecting varying Mahanalobis distances from the predicted cluster. One can see that with increasing distance the rate of mismatches increases. This indicates that despite its limitations discussed in Remark 3.3 in the main text the distance is useful as an empirical indicator of prediction confidence. Bottom: histogram also reflecting dependence on whether previous, next or both suites have been !! or not. While false predictions are more likely if at least one neighboring suite is a !! suite, roughly half of the total false predictions are from suites without a neighboring !! suite.*

LDTP33: matches

name	p. answer	RNAprecis	d
5fke_A_92	1a, 1c, 1f	**2** (1c)	1.23
5fk6_A_55	1a	**1** (1a, ...)	1.27
5fkg_A_92	1a, 1c, 1f	**2** (1c)	1.34
5fkf_A_3	1a	**1** (1a, ...)	1.66
2eet_A_18	1a, 1c	**2** (1c)	1.73
3gx3_A_17	1L, 1a	**1** (1a, ...)	1.78
4b5r_A_83	1a, 1c	**2** (1c)	1.99
3dw4_A_2671	1a, 1c	**2** (1c)	2.0
6q57_A_83	1a	**1** (1a, ...)	2.03
6uc8_B_44	1a, 1c	**1** (1a, ...)	2.03
7eaf_A_80	1a	**1** (1a, ...)	2.07
3diz_A_126	1g	**3** (1g)	2.12
3iqp_A_73	1a	**1** (1a, ...)	2.13
3got_A_80	1a, 1c, 1f	**1** (1a, ...)	2.16
3gx6_A_57	1a, 1c, 1f	**7** (1e, 1f)	2.22
5fk5_A_92	1a, 1c, 1f	**2** (1c)	2.28
4yn6_A_10	1a	**1** (1a, ...)	2.31
5fk2_A_92	1a, 1c, 1f	**2** (1c)	2.35
3dis_A_126	1g	**3** (1g)	2.35
3gx7_A_28	1a, 1c, 1f	**2** (1c)	2.4
6q57_A_88	1a	**1** (1a, ...)	2.4
3dj2_A_168	1a, 1c	**2** (1c)	2.44
3dis_A_117	1a	**1** (1a, ...)	2.51
5fkd_A_55	1a	**1** (1a, ...)	2.54
3oij_C_3	1c	**2** (1c)	2.58
4Lvv_A_82	1a, 1c	**2** (1c)	2.6
5fk6_A_92	1a, 1c, 1f	**2** (1c)	2.61
5fkf_A_74	1a	**1** (1a, ...)	2.63
3diy_A_150	1a	**1** (1a, ...)	2.66
3iqp_A_71	1a	**1** (1a, ...)	2.66
5fk6_A_83	1a, 1c	**2** (1c)	2.67
4yn6_A_3	1g	**3** (1g)	2.82
3iqn_A_75	1g	**3** (1g)	2.83
5fk1_A_54	1L, 1a, 1c	**2** (1c)	2.85
6q57_A_35	1a, 1c	**2** (1c)	2.87
5fk3_A_57	1a, 1c, 1f	**2** (1c)	2.93
4mcf_E_5	1c	**2** (1c)	2.95
3iqp_A_78	1a, 1c	**2** (1c)	2.97
3dim_A_168	1a, 1c	**2** (1c)	3.06
4yn6_B_3	1g	**3** (1g)	3.07
2ees_A_55	1a	**1** (1a, ...)	3.07
5fk1_A_57	1a, 1c, 1f	**2** (1c)	3.09
7eaf_A_23	1a	**1** (1a, ...)	3.13
3gx3_A_28	1a, 1c, 1f	**2** (1c)	3.13
3dim_A_46	1a	**1** (1a, ...)	3.14
5c7u_B_17	1a	**1** (1a, ...)	3.19
6ubu_B_18	1a, 1c	**2** (1c)	3.26
2ydh_A_30	1a, 1c, 1f	**1** (1a, ...)	3.29
3diq_A_168	1a, 1c	**2** (1c)	3.29
4b5r_A_78	1a, 1c	**2** (1c)	3.32
6b14_R_49	1c	**2** (1c)	3.33
5fk3_A_24	5d	**9** (5d)	3.4
3gx5_A_30	1a, 1c, 1f	**2** (1c)	3.4
3gx3_A_92	1a, 1c, 1f	**1** (1a, ...)	3.42
6ubu_B_28	1a, 1c	**2** (1c)	3.44
3u4m_B_2160	1a, 1c	**2** (1c)	3.49
2eet_A_20	1a	**1** (1a, ...)	3.5
2ygh_A_28	1a, 1c, 1f	**2** (1c)	3.5
5fk2_A_57	1a, 1c, 1f	**2** (1c)	3.53
5fk2_A_83	1a, 1c	**2** (1c)	3.53
5fk5_A_78	1a, 1c	**2** (1c)	3.54
7m3v_A_17	1c	**2** (1c)	3.54
2ees_A_79	1a, 1c	**1** (1a, ...)	3.56
2eeu_A_69	1a	**1** (1a, ...)	3.58
3diL_A_49	1c	**2** (1c)	3.6
5fkh_A_57	1a, 1c, 1f	**2** (1c)	3.6
3gx2_A_11	1L, 1a	**1** (1a, ...)	3.64
3bsn_T_7	1a	**1** (1a, ...)	3.69
3umy_B_2182	1a, 1c	**2** (1c)	3.7
5fkh_A_78	1a, 1c	**2** (1c)	3.73
3dig_X_46	1a	**1** (1a, ...)	3.74
2ygh_A_54	1L, 1a, 1c	**2** (1c)	3.74
2ygh_A_57	1a, 1c, 1f	**2** (1c)	3.76
4b5r_A_17	1L, 1a	**1** (1a, ...)	3.79
3iqp_A_54	1L, 1a, 1c	**2** (1c)	3.79
7eaf_A_78	1a, 1c	**2** (1c)	3.83
4b5r_A_90	1a	**1** (1a, ...)	3.84
4yn6_A_12	1c	**2** (1c)	3.86

LDTP33: matches

name	p. answer	RNAprecis	d
5fkd_A_87	1L, 1a	**1** (1a, ...)	3.92
3diz_A_172	1a	**1** (1a, ...)	3.99
5fk5_A_57	1a, 1c, 1f	**2** (1c)	4.0
5fkf_A_54	1L, 1a, 1c	**2** (1c)	4.01
4b5r_A_39	1a, 1c	**1** (1a, ...)	4.02
4b5r_A_92	1a, 1c, 1f	**2** (1c)	4.06
3gx3_A_22	1a	**1** (1a, ...)	4.1
7eaf_A_57	1a, 1c, 1f	**2** (1c)	4.13
5fk1_A_91	1a	**1** (1a, ...)	4.14
3gx2_A_17	1L, 1a	**1** (1a, ...)	4.21
5fkg_A_78	1a, 1c	**2** (1c)	4.21
3iqn_A_17	1L, 1a	**1** (1a, ...)	4.24
4b5r_A_3	1a	**1** (1a, ...)	4.24
5fkf_A_44	1a, 1m	**1** (1a, ...)	4.27
6d92_H_11	1a	**1** (1a, ...)	4.35
2ydh_A_76	1L, 1a	**1** (1a, ...)	4.36
3iqn_A_54	1L, 1a, 1c	**2** (1c)	4.4
5fk1_A_78	1a, 1c	**2** (1c)	4.45
3gx3_A_53	1a	**1** (1a, ...)	4.46
3iqn_A_78	1a, 1c	**2** (1c)	4.47
4u3p_B_7	1a, 1c	**2** (1c)	4.47
3got_A_18	1a, 1c	**2** (1c)	4.49
5fke_A_78	1a, 1c	**2** (1c)	4.49
5c7u_B_39	5n, 9a	**4** (9a, 3a, 7a)	4.49
5fk2_A_78	1a, 1c	**2** (1c)	4.51
3dix_A_70	1a	**1** (1a, ...)	4.51
3gx6_A_74	1a	**1** (1a, ...)	4.52
5fk3_A_51	1g	**3** (1g)	4.53
2eet_A_45	1a, 1c	**2** (1c)	4.53
4Lvv_A_54	1a, 1c	**2** (1c)	4.55
5fk3_A_78	1a, 1c	**2** (1c)	4.56
5fke_A_54	1L, 1a, 1c	**2** (1c)	4.56
4Lvx_A_54	1a, 1c	**2** (1c)	4.57
5c7u_B_38	1a	**1** (1a, ...)	4.6
5fkf_A_78	1a, 1c	**2** (1c)	4.61
3gx5_A_53	1a	**1** (1a, ...)	4.64
3syw_B_12	1c	**2** (1c)	4.67
3gx2_A_57	1a, 1c, 1f	**7** (1e, 1f)	4.68
3syw_A_12	1c	**2** (1c)	4.68
4b5r_A_4	1a	**1** (1a, ...)	4.75
5fkf_A_51	1g	**3** (1g)	4.76
2eev_A_80	1a, 1c, 1f	**2** (1c)	4.8
4b5r_A_54	1L, 1a, 1c	**2** (1c)	4.82
2eeu_A_18	1a, 1c	**2** (1c)	4.85
2ees_A_80	1a, 1c, 1f	**2** (1c)	4.85
3diq_A_50	1a	**1** (1a, ...)	4.88
4fnj_A_5	1c	**2** (1c)	4.89
3gx7_A_78	1a, 1c	**2** (1c)	4.95
3gx3_A_11	1L, 1a	**1** (1a, ...)	4.95
6d8f_H_11	1a	**1** (1a, ...)	4.95
4b5r_A_49	1a	**1** (1a, ...)	4.96
3gx5_A_16	1L, 1a, 1c	**1** (1a, ...)	4.97
2eeu_A_53	1a	**1** (1a, ...)	5.05
2ydh_A_53	1a	**1** (1a, ...)	5.1
3gx3_A_16	1L, 1a, 1c	**2** (1c)	5.16
5fkf_A_57	1a, 1c, 1f	**2** (1c)	5.16
5c7u_B_55	1a	**1** (1a, ...)	5.21
3gx6_A_53	1a	**1** (1a, ...)	5.36
4b5r_A_55	1a	**1** (1a, ...)	5.37
2ydh_A_78	1a, 1c	**2** (1c)	5.39
3iqn_A_83	1a, 1c	**2** (1c)	5.4
5fk6_A_51	1g	**3** (1g)	5.42
6d92_H_17	1c	**2** (1c)	5.43
3iqr_A_51	1g	**3** (1g)	5.45
3gx7_A_24	5d	**9** (5d)	5.45
5fkh_A_53	1a	**1** (1a, ...)	5.46
3gx7_A_57	1a, 1c, 1f	**7** (1e, 1f)	5.49
5fke_A_62	1L, 1a, 1c	**2** (1c)	5.49
2ees_A_39	5n, 9a	**4** (9a, 3a, 7a)	5.5
3iqr_A_24	5d	**9** (5d)	5.52

LDTP33: matches

name	p. answer	RNAprecis	d
2eew_A_39	5n, 9a	**4** (9a, 3a, 7a)	5.62
6d8a_H_17	1c	**2** (1c)	5.63
2ydh_A_83	1a, 1c	**2** (1c)	5.63
2ees_A_40	1a	**1** (1a, ...)	5.64
5fk1_A_51	1g	**3** (1g)	5.64
5fk1_A_2	1a	**1** (1a, ...)	5.77
3gx3_A_78	1a, 1c	**2** (1c)	5.78
4yn6_A_15	1c	**2** (1c)	5.81
4feL_B_18	1a, 1c	**2** (1c)	5.82
3g4m_A_27	1a, 7a	**1** (1a, ...)	5.84
4feo_B_18	1a, 1c	**2** (1c)	5.85
3gx3_A_24	5d	**9** (5d)	5.9
4fe5_B_18	1a, 1c	**2** (1c)	6.15
5fkd_A_13	1f, 1o, 3a, 3b, 7a	**6** (1f)	6.15
4u3L_B_7	1a, 1c	**2** (1c)	6.15
5fkd_A_74	1a	**1** (1a, ...)	6.16
3iqn_A_24	5d	**9** (5d)	6.17
2gis_A_57	1a, 1c, 1f	**2** (1c)	6.17
3ds7_B_252	1a	**1** (1a, ...)	6.2
6d8f_H_17	1c	**2** (1c)	6.2
3g4m_A_17	1a	**1** (1a, ...)	6.21
3iqn_A_74	1a	**1** (1a, ...)	6.22
5neo_A_8	5d	**9** (5d)	6.25
3gao_A_54	5n, 9a	**4** (9a, 3a, 7a)	6.26
5c7u_B_16	1a, 1c, 1f	**2** (1c)	6.34
3gx3_A_57	1a, 1c, 1f	**2** (1c)	6.4
3iqn_A_70	1a, 1c	**2** (1c)	6.41
5fk6_A_57	1a, 1c, 1f	**2** (1c)	6.48
2ees_A_78	1a, 1c	**1** (1a, ...)	6.52
2gis_A_50	1a, 1c, 1m	**2** (1c)	6.54
2ydh_A_24	5d	**9** (5d)	6.58
2eeu_A_54	5n, 9a	**4** (9a, 3a, 7a)	6.62
3ger_A_54	5n, 9a	**4** (9a, 3a, 7a)	6.65
2eeu_A_67	1e	**7** (1e, 1f)	6.68
3dix_A_114	7d	**8** (7d)	6.74
4b5r_A_18	1L, 1a, 1c	**2** (1c)	6.76
5fkd_A_38	1a	**1** (1a, ...)	6.76
3ges_A_54	5n, 9a	**4** (9a, 3a, 7a)	6.79
5fk5_A_28	1a, 1c, 1f	**2** (1c)	6.81
7eaf_A_62	1L, 1a, 1c	**2** (1c)	6.81
4b5r_A_24	5d	**9** (5d)	6.82
5fkh_A_13	1f, 1o, 3a, 3b, 7a	**6** (1f)	6.85
3ges_A_27	1a, 7a	**1** (1a, ...)	6.87
3gx7_A_11	1L, 1a	**1** (1a, ...)	6.89
3diz_A_70	1a	**1** (1a, ...)	6.9
3diq_A_82	1c	**2** (1c)	7.01
3gx2_A_53	1a	**1** (1a, ...)	7.02
3dig_X_114	7d	**8** (7d)	7.09
4qvi_B_2160	1a, 1c	**2** (1c)	7.12
6q57_A_54	1a, 1c	**2** (1c)	7.14
3fo6_A_54	5n, 9a	**4** (9a, 3a, 7a)	7.15
7eaf_A_25	1a	**1** (1a, ...)	7.17
3dim_A_82	1c	**2** (1c)	7.2
3t5q_L_5	1c	**2** (1c)	7.24
3iqp_A_13	1f, 1o, 3a, 3b, 7a	**6** (1f)	7.29
3diz_A_82	1c	**2** (1c)	7.3
4xw7_A_25	1c	**2** (1c)	7.31
2eev_A_27	1a, 7a	**1** (1a, ...)	7.31
3diq_A_72	3a, 7a	**4** (9a, 3a, 7a)	7.34
6y0t_A_2660	1g	**3** (1g)	7.37
3diL_A_82	1c	**2** (1c)	7.39
3iqp_A_50	1a, 1c, 1m	**2** (1c)	7.41
2eew_A_54	5n, 9a	**4** (9a, 3a, 7a)	7.42
3dim_A_72	3a, 7a	**4** (9a, 3a, 7a)	7.43
2xLi_B_17	1c	**2** (1c)	7.44
2eet_A_54	5n, 9a	**4** (9a, 3a, 7a)	7.44
3dix_A_72	3a, 7a	**4** (9a, 3a, 7a)	7.45
3iqp_A_36	#a, 0a, 4a, 7a	**4** (9a, 3a, 7a)	7.46
3gx5_A_24	5d	**9** (5d)	7.48

Table 7: *Suites for which the RNAprecis predicted cluster matches the homology related conformer (blue in the plots) sorted by distance.*

LDTP33: matches

name	p. answer	RNAprecis	d
5fk1_A_13	1f, 1o, 3a, 3b, 7a	**4** (9a, 3a, 7a)	7.48
3diy_A_82	1c	**2** (1c)	7.51
5fk3_A_52	1a, 1c	**2** (1c)	7.52
3dj2_A_72	3a, 7a	**4** (9a, 3a, 7a)	7.65
3dig_X_82	1c	**2** (1c)	7.65
3syw_A_15	1c	**2** (1c)	7.65
2gis_A_24	5d	**9** (5d)	7.66
3dix_A_82	1c	**2** (1c)	7.66
3syw_B_15	1c	**2** (1c)	7.67
5fkh_A_62	1L, 1a, 1c	**2** (1c)	7.72
3diz_A_50	1a	**1** (1a, . . .)	7.73
3dio_X_82	1c	**2** (1c)	7.74
3iqr_A_52	1a, 1c	**2** (1c)	7.8
2g9c_A_54	5n, 9a	**4** (9a, 3a, 7a)	7.92
3dig_X_72	3a, 7a	**4** (9a, 3a, 7a)	7.94
7eaf_A_24	5d	**9** (5d)	7.96
3gx3_A_70	1a, 1c	**2** (1c)	8.06
3gx7_A_52	1a, 1c	**2** (1c)	8.06
2eev_A_54	5n, 9a	**4** (9a, 3a, 7a)	8.12
3cgq_B_21	5j	**10** (5j)	8.14
3iqn_A_16	1L, 1a, 1c	**2** (1c)	8.14
3g4m_A_54	5n, 9a	**4** (9a, 3a, 7a)	8.14
3diL_A_72	3a, 7a	**4** (9a, 3a, 7a)	8.3
3gx6_A_16	1L, 1a, 1c	**1** (1a, . . .)	8.36
5fk5_A_14	1g	**3** (1g)	8.4
2ydh_A_16	1L, 1a, 1c	**2** (1c)	8.54
5fkd_A_62	1L, 1a, 1c	**2** (1c)	8.56
3iqn_A_50	1a, 1c, 1m	**2** (1c)	8.59
3gx6_A_24	5d	**9** (5d)	8.7
4b5r_A_52	1a, 1c	**2** (1c)	8.7
4qg3_B_2148	1c	**2** (1c)	8.71
3gx2_A_13	1f, 1o, 3a, 3b, 7a	**4** (9a, 3a, 7a)	8.72
3got_A_54	5n, 9a	**4** (9a, 3a, 7a)	8.85
3u4m_B_2148	1c	**2** (1c)	8.89
3g4m_A_39	5n, 9a	**4** (9a, 3a, 7a)	8.9
5fk2_A_16	1L, 1a, 1c	**2** (1c)	9.06
2gis_A_13	1f, 1o, 3a, 3b, 7a	**6** (1f)	9.09
4qvi_B_2148	1c	**2** (1c)	9.11
6ubu_B_16	1a, 1c, 1f	**2** (1c)	9.16
5npm_B_4	1a, 1c	**2** (1c)	9.16
3diz_A_72	3a, 7a	**4** (9a, 3a, 7a)	9.2
2gis_A_52	1a, 1c	**2** (1c)	9.25
6uc9_B_16	1a, 1c, 1f	**2** (1c)	9.31
3iqp_A_24	5d	**9** (5d)	9.37
3gx7_A_18	1L, 1a, 1c	**2** (1c)	9.38
6uc8_B_16	1a, 1c, 1f	**2** (1c)	9.54

LDTP33: matches

name	p. answer	RNAprecis	d
6uc7_B_16	1a, 1c, 1f	**2** (1c)	9.59
5fk5_A_13	1f, 1o, 3a, 3b, 7a	**4** (9a, 3a, 7a)	9.62
3iqp_A_52	1a, 1c	**2** (1c)	9.62
5fkd_A_14	1g	**3** (1g)	9.67
2eev_A_16	1a, 1c, 1f	**2** (1c)	9.67
2gis_A_53	1a	**1** (1a, . . .)	9.81
5fk6_A_14	1g	**3** (1g)	9.9
3gx6_A_52	1a, 1c	**2** (1c)	10.19
2ees_A_54	5n, 9a	**4** (9a, 3a, 7a)	10.21
5fk6_A_24	5d	**9** (5d)	10.52
6d95_H_12	1e	**7** (1e, 1f)	10.62
2gis_A_16	1L, 1a, 1c	**2** (1c)	10.72
2eew_A_16	1a, 1c, 1f	**2** (1c)	10.81
3gx7_A_70	1a, 1c	**2** (1c)	10.92
4feo_B_16	1a, 1c, 1f	**2** (1c)	11.04
7m3v_A_19	1e	**7** (1e, 1f)	11.44
2eet_A_16	1a, 1c, 1f	**2** (1c)	11.59
4fej_B_16	1a, 1c, 1f	**2** (1c)	11.79
6d9k_C_12	1e	**7** (1e, 1f)	11.8
3gao_A_16	1a, 1c, 1f	**2** (1c)	12.11
5fkh_A_14	1g	**3** (1g)	12.14
4fep_B_73	1a, 1c, 1f	**2** (1c)	13.01
4fen_B_73	1a, 1c, 1f	**2** (1c)	13.08
4fej_B_73	1a, 1c, 1f	**2** (1c)	13.36
4fe5_B_16	1a, 1c, 1f	**2** (1c)	13.46
3ger_A_16	1a, 1c, 1f	**2** (1c)	13.7
3gx7_A_16	1L, 1a, 1c	**2** (1c)	13.74
5c7u_B_54	5n, 9a	**4** (9a, 3a, 7a)	14.18
5fkf_A_15	1a, 7d, 8d	**8** (7d)	15.95
4b5r_A_15	1a, 7d, 8d	**8** (7d)	17.0
5fk6_A_15	1a, 7d, 8d	**8** (7d)	21.61
2gis_A_15	1a, 7d, 8d	**8** (7d)	25.79

Table 8: *Suites for which the RNAprecis predicted cluster matches the homology related conformer (blue in the plots) sorted by distance, continued from Table 7.*

LDTP33: pucker mismatches

7kd1_A_79	7r	4.63	4.55	4 (9a, 3a, 7a)	7.13
3iqr_A_69	2a	2.97	4.4	4 (9a, 3a, 7a)	13.03
3gx3_A_10	2a	3.83	4.42	2 (1c)	14.09
3gx3_A_69	2a	3.55	4.42	4 (9a, 3a, 7a)	14.65
2ydh_A_69	2a	3.56	4.52	2 (1c)	15.95
3ds7_B_250	#a, 0a	3.45	4.69	8 (7d)	16.29
7kd1_A_44	2a	4.25	3.45	4 (9a, 3a, 7a)	16.3
3gx5_A_69	2a	2.91	4.66	4 (9a, 3a, 7a)	18.33
3gx6_A_69	2a	3.07	4.62	2 (1c)	18.72
3ds7_B_249	2[	4.79	3.45	3 (1g)	18.73
3gx7_A_10	2a	3.31	4.26	2 (1c)	19.96
2gis_A_35	2[	4.33	4.03	4 (9a, 3a, 7a)	20.0
2gis_A_69	2a	2.94	4.52	2 (1c)	20.57
3iqp_A_35	2[	3.38	3.51	4 (9a, 3a, 7a)	28.69

LDTP33: conformer mismatch

name	p. answer	RNAprecis	d_1	first match	d_2
6ugg_A_12	1a	2 (1c)	1.58	1 (1a, ...)	6.43
3dim_A_63	1a	2 (1c)	2.33	1 (1a, ...)	5.82
5c7u_B_59	1a	2 (1c)	2.6	1 (1a, ...)	4.81
3dig_X_63	1a	2 (1c)	2.84	1 (1a, ...)	5.51
3dim_A_60	1e, 1f	2 (1c)	2.9	7 (1e, 1f)	5.91
3jxq_A_4	1a	2 (1c)	3.05	1 (1a, ...)	4.5
1wvd_B_4	1a	2 (1c)	3.07	1 (1a, ...)	5.0
1nLc_B_4	1a	2 (1c)	3.08	1 (1a, ...)	6.03
3dim_A_113	1a	2 (1c)	3.12	1 (1a, ...)	4.66
3diq_A_60	1e, 1f	2 (1c)	3.16	7 (1e, 1f)	6.66
3dig_X_60	1e, 1f	1 (1a, ...)	3.43	7 (1e, 1f)	7.79
7kd1_A_4	1a	2 (1c)	3.5	1 (1a, ...)	7.95
3syw_A_13	1a	2 (1c)	3.51	1 (1a, ...)	8.68
3syw_B_13	1a	2 (1c)	3.67	1 (1a, ...)	8.91
3diz_A_60	1e, 1f	2 (1c)	3.69	7 (1e, 1f)	8.0
3syw_A_11	1a	2 (1c)	3.82	1 (1a, ...)	5.33
3diL_A_60	1e, 1f	2 (1c)	3.86	7 (1e, 1f)	5.22
3syw_B_11	1a	2 (1c)	3.88	1 (1a, ...)	5.36
3dj2_A_60	1e, 1f	2 (1c)	4.06	7 (1e, 1f)	6.12
5c7u_B_60	1a	2 (1c)	4.11	1 (1a, ...)	4.18
7kd1_A_35	1a, 1c	11 (1f)	4.13	2 (1c)	11.57
3dis_A_63	1a	2 (1c)	4.14	1 (1a, ...)	5.33
3gx7_A_22	1a	2 (1c)	4.18	1 (1a, ...)	5.43
3diy_A_36	1L, 1a	2 (1c)	4.29	1 (1a, ...)	7.75
3gx7_A_66	1a	2 (1c)	4.34	1 (1a, ...)	5.09
3gx2_A_66	1a	2 (1c)	4.42	1 (1a, ...)	5.42
3iqp_A_23	1a	2 (1c)	4.51	1 (1a, ...)	4.98
5fk3_A_86	1a	2 (1c)	4.54	1 (1a, ...)	6.08
5fkg_A_31	1a	2 (1c)	4.64	1 (1a, ...)	6.61
3diz_A_46	1a	2 (1c)	4.85	1 (1a, ...)	8.89
5fkd_A_48	1a	2 (1c)	5.1	1 (1a, ...)	7.95
3gx7_A_55	1a	2 (1c)	5.3	1 (1a, ...)	6.27
4b5r_A_41	1a	2 (1c)	5.34	1 (1a, ...)	7.19
2ees_A_51	1a	2 (1c)	5.41	1 (1a, ...)	6.53
3diy_A_34	1a	2 (1c)	5.48	1 (1a, ...)	8.01
5fk3_A_2	1a	2 (1c)	5.52	1 (1a, ...)	7.73
3gx2_A_22	1a	2 (1c)	5.63	1 (1a, ...)	6.85
3dio_X_35	1a	2 (1c)	5.77	1 (1a, ...)	7.62
7kd1_A_46	1a	2 (1c)	5.8	1 (1a, ...)	6.67
3iqp_A_58	1a	2 (1c)	5.82	1 (1a, ...)	6.26
4b5r_A_2	1a	2 (1c)	5.86	1 (1a, ...)	9.36
3gx2_A_37	1a	2 (1c)	5.88	1 (1a, ...)	6.48
5fk1_A_17	1L, 1a	2 (1c)	6.14	1 (1a, ...)	6.81
3dim_A_2	1a	2 (1c)	6.17	1 (1a, ...)	6.56
3gx2_A_74	1a	2 (1c)	6.34	1 (1a, ...)	9.87
2ydh_A_77	1a	2 (1c)	6.46	1 (1a, ...)	6.93
5fkd_A_76	1L, 1a	2 (1c)	6.71	1 (1a, ...)	6.72
2gis_A_74	1a	2 (1c)	6.82	1 (1a, ...)	9.26
4b5r_A_60	1a	2 (1c)	6.83	1 (1a, ...)	8.3
4yn6_B_13	1a	2 (1c)	6.93	1 (1a, ...)	8.0
4yn6_A_13	1a	2 (1c)	6.94	1 (1a, ...)	9.95
5fk3_A_77	1a	2 (1c)	6.98	1 (1a, ...)	9.68
3u56_B_2183	1a	2 (1c)	7.17	1 (1a, ...)	10.34
5fk2_A_77	1a	2 (1c)	7.54	1 (1a, ...)	10.71
3iqn_A_25	1a	2 (1c)	7.67	1 (1a, ...)	8.24
1y6t_B_2	1a	2 (1c)	7.68	1 (1a, ...)	8.63
3gx7_A_71	1a	2 (1c)	7.68	1 (1a, ...)	9.33
3iqr_A_74	1a	2 (1c)	7.68	1 (1a, ...)	10.06
3iqp_A_15	1a, 7d, 8d	2 (1c)	7.72	1 (1a, ...)	9.36
5fk2_A_17	1L, 1a	2 (1c)	7.8	1 (1a, ...)	8.44
3gx3_A_7	1a	2 (1c)	7.87	1 (1a, ...)	9.95
3dis_A_50	1a	4 (9a, 3a, 7a)	7.88	1 (1a, ...)	10.01
3dis_A_35	1a	2 (1c)	7.88	1 (1a, ...)	11.08

LDTP33: conformer mismatch

name	p. answer	RNAprecis	d_1	first match	d_2
3bsn_T_6	1a	2 (1c)	7.92	1 (1a, ...)	8.39
3gx7_A_17	1L, 1a	2 (1c)	7.95	1 (1a, ...)	9.55
2eet_A_21	1a	2 (1c)	7.97	1 (1a, ...)	11.04
2ydh_A_2	1a	2 (1c)	8.0	1 (1a, ...)	8.0
4fnj_A_4	1a	2 (1c)	8.02	1 (1a, ...)	10.54
3iqp_A_74	1a	2 (1c)	8.03	1 (1a, ...)	14.75
2qek_B_8	1a, 3g, 8d	4 (9a, 3a, 7a)	8.03	1 (1a, ...)	20.69
7eaf_A_81	1a	2 (1c)	8.5	1 (1a, ...)	12.59
5fkd_A_56	1a, 1m	2 (1c)	8.78	1 (1a, ...)	9.58
4yn6_B_2	1a	2 (1c)	8.81	1 (1a, ...)	10.23
4fnj_A_31	1a	2 (1c)	8.82	1 (1a, ...)	10.93
7eaf_A_77	1a	2 (1c)	8.89	1 (1a, ...)	14.58
5fk5_A_2	1a	2 (1c)	9.18	1 (1a, ...)	11.79
3gx7_A_13	1f, 1o, 3a, 3b, 7a	5 (1e)	9.26	6 (1f)	12.91
5fkh_A_2	1a	2 (1c)	9.39	1 (1a, ...)	12.14
5fkf_A_2	1a	2 (1c)	9.58	1 (1a, ...)	12.65
3gx7_A_37	1a	2 (1c)	9.63	1 (1a, ...)	11.75
3gx3_A_68	1a, 1b	2 (1c)	9.69	1 (1a, ...)	12.58
5fkd_A_2	1a	2 (1c)	9.79	1 (1a, ...)	12.24
2gis_A_14	1g	8 (7d)	10.16	3 (1g)	24.12
5fke_A_2	1a	2 (1c)	10.17	1 (1a, ...)	12.28
5fkg_A_2	1a	2 (1c)	10.24	1 (1a, ...)	12.22
3dio_X_61	1L, 1a	2 (1c)	10.25	1 (1a, ...)	11.79
2eeu_A_52	1a	2 (1c)	10.37	1 (1a, ...)	13.46
3gx3_A_14	1g	8 (7d)	10.51	3 (1g)	18.62
5b2o_B_18	1e	2 (1c)	10.91	7 (1e, 1f)	11.04
3gx7_A_56	1a, 1m	2 (1c)	10.95	1 (1a, ...)	14.29
3gx2_A_71	1a	2 (1c)	11.13	1 (1a, ...)	13.88
3iqr_A_68	1a, 1b	2 (1c)	11.56	1 (1a, ...)	19.79
3got_A_16	1a, 1c, 1f	10 (5j)	11.95	2 (1c)	13.51
2eeu_A_68	1a	2 (1c)	12.0	1 (1a, ...)	17.18
3gx3_A_15	1a, 7d, 8d	4 (9a, 3a, 7a)	12.72	8 (7d)	20.02
5fke_A_15	1a, 7d, 8d	11 (1f)	12.82	8 (7d)	23.11
7kd1_A_45	1a	2 (1c)	12.92	1 (1a, ...)	21.73
3gx6_A_61	3d, 5j	8 (7d)	13.01	10 (5j)	36.59
3diq_A_128	1a	4 (9a, 3a, 7a)	13.11	1 (1a, ...)	19.46
3dio_X_128	1a	4 (9a, 3a, 7a)	13.18	1 (1a, ...)	17.73
6d92_H_12	1e	2 (1c)	13.18	7 (1e, 1f)	19.36
3dix_A_128	1a	4 (9a, 3a, 7a)	13.21	1 (1a, ...)	16.72
5fkf_A_14	1g	8 (7d)	13.24	3 (1g)	15.37
3diy_A_128	1a	4 (9a, 3a, 7a)	13.37	1 (1a, ...)	16.9
3iqn_A_71	1a	4 (9a, 3a, 7a)	13.49	1 (1a, ...)	21.58
3dj2_A_128	1a	4 (9a, 3a, 7a)	13.55	1 (1a, ...)	17.2
1y90_B_9	1a, 1m, 1z, 2a, 7p	4 (9a, 3a, 7a)	13.55	1 (1a, ...)	47.53
3diz_A_128	1a	4 (9a, 3a, 7a)	13.8	1 (1a, ...)	18.0
3dig_X_128	1a	4 (9a, 3a, 7a)	13.97	1 (1a, ...)	17.79
5fkd_A_15	1a, 7d, 8d	4 (9a, 3a, 7a)	14.03	1 (1a, ...)	22.73
3diL_A_128	1a	4 (9a, 3a, 7a)	14.12	1 (1a, ...)	16.7
5fk5_A_15	1a, 7d, 8d	4 (9a, 3a, 7a)	14.67	8 (7d)	23.37
3gx2_A_61	3d, 5j	8 (7d)	14.88	10 (5j)	33.04
2ydh_A_61	3d, 5j	8 (7d)	14.96	10 (5j)	28.37
3gx5_A_61	3d, 5j	8 (7d)	15.02	10 (5j)	30.81
3iqr_A_61	3d, 5j	8 (7d)	15.04	10 (5j)	31.75
3c44_A_8	1a, 3g, 8d	3 (1g)	15.1	1 (1a, ...)	54.82
3gx7_A_61	3d, 5j	8 (7d)	15.16	10 (5j)	28.15
2gis_A_61	3d, 5j	8 (7d)	15.22	10 (5j)	32.17
3dvv_B_8	1a, 3g, 8d	3 (1g)	15.28	1 (1a, ...)	64.64
5fke_A_61	3d, 5j	8 (7d)	15.56	10 (5j)	24.04
5fkg_A_61	3d, 5j	8 (7d)	15.64	10 (5j)	21.52
3syw_B_16	1a	2 (1c)	15.9	1 (1a, ...)	19.43
5fkh_A_61	3d, 5j	8 (7d)	15.91	10 (5j)	34.72
2ees_A_16	1a, 1c, 1f	10 (5j)	15.96	6 (1f)	16.89
3syw_A_16	1a	2 (1c)	16.1	1 (1a, ...)	19.61
3iqn_A_72	1a	10 (5j)	16.25	1 (1a, ...)	21.39
2ygh_A_61	3d, 5j	8 (7d)	16.27	10 (5j)	22.69
3iqn_A_61	3d, 5j	8 (7d)	16.29	10 (5j)	29.58
3iqp_A_61	3d, 5j	8 (7d)	16.33	10 (5j)	38.35
3gx3_A_61	3d, 5j	8 (7d)	16.42	10 (5j)	28.33
5fk6_A_61	3d, 5j	8 (7d)	16.79	10 (5j)	21.57
2gis_A_47	0a, 1a, 4g	4 (9a, 3a, 7a)	16.93	1 (1a, ...)	26.9
5fk2_A_61	3d, 5j	8 (7d)	17.06	10 (5j)	26.35
5fk1_A_61	3d, 5j	8 (7d)	17.22	10 (5j)	26.94
5fk5_A_61	3d, 5j	8 (7d)	17.52	10 (5j)	23.74
5fkd_A_61	3d, 5j	4 (9a, 3a, 7a)	17.63	10 (5j)	20.79
3iqp_A_14	1g	2 (1c)	17.87	3 (1g)	21.55
4b5r_A_61	3d, 5j	8 (7d)	17.93	10 (5j)	19.2
2ydh_A_15	1a, 7d, 8d	4 (9a, 3a, 7a)	18.58	8 (7d)	21.58
5fkh_A_15	1a, 7d, 8d	4 (9a, 3a, 7a)	18.81	8 (7d)	21.76
3gx7_A_9	3d, 7p	3 (1g)	20.1	-	-
5fkg_A_15	1a, 7d, 8d	4 (9a, 3a, 7a)	21.42	8 (7d)	22.52

Table 9: *Suites for which the sugar pucker-pairs determined by the Pperp criterion do not match the sugar pucker-pairs of the homology related conformers (red in the plots) or the RNAprecis predicted cluster does not match the homology related conformer (orange in the plots) sorted by distance.*

E.2 Results for the test data set LDTP32

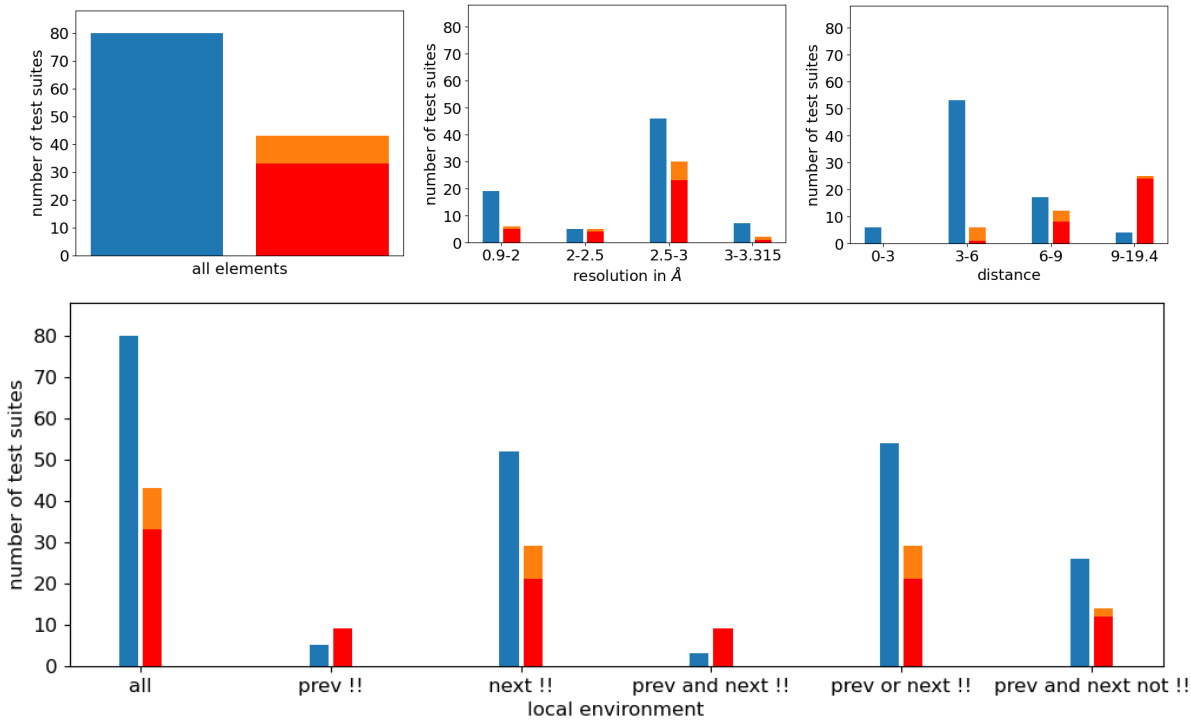

Figure 31: *Relative distribution of agreements and disagreements of MINT-AGE cluster prediction with homology related probable conformer classes for the test set LDT32 with color scheme (reflecting blue for match, orange for mismatch and red for pucker-pair mismatch) from Section 4.4 in the main text). Top left: overall histogram. Top middle: histogram reflecting varying resolutions of PDB suites. Top right: histogram reflecting varying Mahanalobis distances from the predicted cluster. One can see that with increasing distance the rate of mismatches increases. This indicates that despite its limitations discussed in Remark 3.3 in the main text the distance is useful as an empirical indicator of prediction confidence. Bottom: histogram also reflecting dependence on whether previous, next or both suites have been !! or not.*

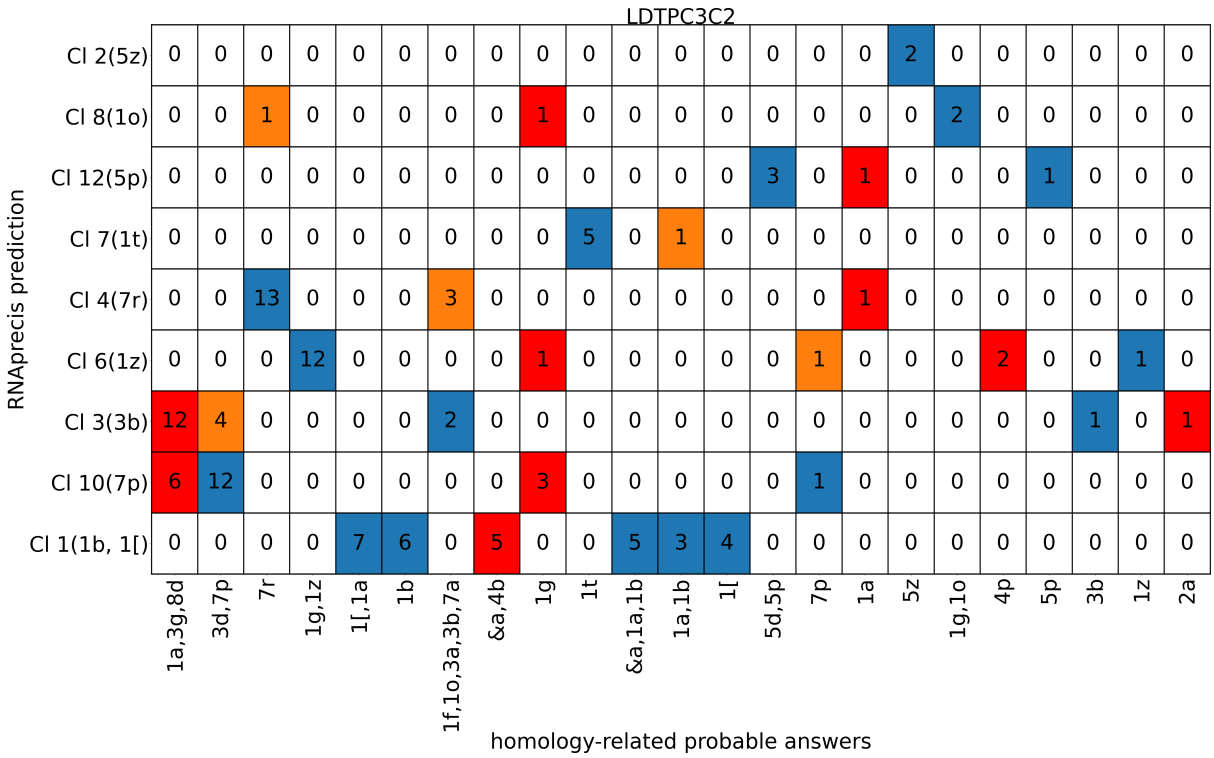

Figure 32: *Classification matrix listing counts of probable homology related conformation classes (horizontal axis) versus MINT-AGE clusters (vertical, with conformer classes in parentheses from MINT-AGE training, see Figure 23) for the low detail test data set of sugar pucker-pairs LDTP32 with colors (from Section 4.4 in the main text): blue for match, orange for mismatch and red for pucker-pair mismatch. Columns listing multiple suite conformers are from sites solved in different conformations across homology-related models, and any of the homology-related conformations were accepted as a matching predictions. As pointed out in Section 4.4, the* ***1a,3g,8d*** *and* ***3d,7p*** *suites have multiple homology-related conformations. This means that the counts in these two columns are inflated.*

LDTP32: matches

name	p. answer	RNAprecis	d
3gx5_A_33	7r	**4** (7r)	2.08
2gis_A_33	7r	**4** (7r)	2.14
3iqr_A_33	7r	**4** (7r)	2.62
4b5r_A_33	7r	**4** (7r)	2.85
5fk5_A_33	7r	**4** (7r)	2.86
2eev_A_62	1t	**7** (1t)	2.97
3iqp_A_33	7r	**4** (7r)	3.01
3umy_B_2145	1z	**6** (1z)	3.11
3gx3_A_33	7r	**4** (7r)	3.27
4Lvy_A_55	1b	**5** (1[, 1b, !!)	3.56
3diy_A_11	1t	**7** (1t)	3.64
5fk1_A_33	7r	**4** (7r)	3.68
3gx7_A_33	7r	**4** (7r)	3.72
3gx6_A_33	7r	**4** (7r)	3.72
3gx7_A_46	&a, 1a, 1b	**1** (1b, 1[)	3.77
5c7u_B_62	1t	**7** (1t)	3.78
3dix_A_52	1[, 1a	**5** (1[, 1b, !!)	3.91
3iqn_A_68	1a, 1b	**1** (1b, 1[)	3.92
3gx7_A_68	1a, 1b	**1** (1b, 1[)	3.95
3gx2_A_33	7r	**4** (7r)	3.97
3diy_A_52	1[, 1a	**5** (1[, 1b, !!)	3.98
6uc7_B_22	1g, 1z	**6** (1z)	4.03
5fke_A_33	7r	**4** (7r)	4.05
4feL_B_22	1g, 1z	**6** (1z)	4.06
3diz_A_52	1[, 1a	**5** (1[, 1b, !!)	4.06
4fen_B_22	1g, 1z	**6** (1z)	4.09
6ubu_B_22	1g, 1z	**6** (1z)	4.18
3diq_A_52	1[, 1a	**1** (1b, 1[)	4.25
4fej_B_22	1g, 1z	**6** (1z)	4.25
4fe5_B_22	1g, 1z	**6** (1z)	4.26
6uc9_B_22	1g, 1z	**6** (1z)	4.27
2eet_A_22	1g, 1z	**6** (1z)	4.3
2g9c_A_22	1g, 1z	**6** (1z)	4.33
3dis_A_11	1t	**7** (1t)	4.39
4qg3_B_2126	3b	**3** (3b)	4.39
3diy_A_39	1b	**1** (1b, 1[)	4.5
2eev_A_22	1g, 1z	**6** (1z)	4.54
3dio_X_151	1b	**1** (1b, 1[)	4.54
3dj2_A_52	1[, 1a	**1** (1b, 1[)	4.54
5c7u_B_22	1g, 1z	**6** (1z)	4.6
3umy_B_2116	1g, 1o	**8** (1o)	4.66
3dim_A_52	1[, 1a	**5** (1[, 1b, !!)	4.67
5fk5_A_46	&a, 1a, 1b	**1** (1b, 1[)	4.7
2eeu_A_22	1g, 1z	**6** (1z)	4.77
5fkd_A_46	&a, 1a, 1b	**1** (1b, 1[)	5.05
3gx3_A_46	&a, 1a, 1b	**1** (1b, 1[)	5.06
3u4m_B_2116	1g, 1o	**8** (1o)	5.08
5fk1_A_68	1a, 1b	**5** (1[, 1b, !!)	5.09
3dim_A_11	1t	**7** (1t)	5.13
3diL_A_39	1b	**1** (1b, 1[)	5.21
3diz_A_42	1[	**5** (1[, 1b, !!)	5.31
3dis_A_52	1[, 1a	**5** (1[, 1b, !!)	5.37
3iqn_A_46	&a, 1a, 1b	**1** (1b, 1[)	5.65
5fkg_A_33	7r	**4** (7r)	5.66
3gx6_A_13	1f, 1o, 3a, 3b, 7a	**3** (3b)	5.85
3dix_A_42	1[	**5** (1[, 1b, !!)	5.88
7kd1_A_21	5p	**12** (5p)	5.91
3diz_A_39	1b	**5** (1[, 1b, !!)	5.92
2ygh_A_9	3d, 7p	**10** (7p)	5.97
3gx3_A_63	5d, 5p	**11** (5p)	6.02
5fkg_A_13	1f, 1o, 3a, 3b, 7a	**3** (3b)	6.07
2gis_A_9	3d, 7p	**10** (7p)	6.09
3iqp_A_9	3d, 7p	**10** (7p)	6.14
3diq_A_42	1[	**5** (1[, 1b, !!)	6.32
5fk5_A_9	3d, 7p	**10** (7p)	6.34
5fkf_A_9	3d, 7p	**10** (7p)	6.36
3iqr_A_9	3d, 7p	**10** (7p)	6.4
3iqr_A_63	5d, 5p	**11** (5p)	6.49
7qp2_A_2654	5z	**2** (5z)	6.55
1q9a_A_2654	5z	**2** (5z)	6.88
6b14_R_28	7p	**9** (7p)	7.33
3dis_A_42	1[	**5** (1[, 1b, !!)	7.93
3dis_A_39	1b	**5** (1[, 1b, !!)	8.17
3iqp_A_63	5d, 5p	**11** (5p)	8.27
7eaf_A_9	3d, 7p	**10** (7p)	8.78
5fkh_A_9	3d, 7p	**10** (7p)	8.83
3gx2_A_9	3d, 7p	**10** (7p)	9.57
3gx6_A_9	3d, 7p	**10** (7p)	10.07
2ydh_A_9	3d, 7p	**10** (7p)	11.71
4b5r_A_9	3d, 7p	**10** (7p)	17.83

LDTP32: pucker mismatches

name	p. answer	p_1	p_2	RNAprecis	d
3iqn_A_51	1g	4.94	2.14	6 (1z)	3.29
3iqr_A_20	&a, 4b	3.83	0.99	5 (1[, 1b, !!)	6.62
3gx7_A_20	&a, 4b	3.34	1.65	5 (1[, 1b, !!)	6.79
3iqn_A_20	&a, 4b	3.53	1.99	5 (1[, 1b, !!)	7.42
3iqp_A_20	&a, 4b	3.56	1.98	5 (1[, 1b, !!)	7.53
6uc7_B_74	1a	4.18	0.31	4 (7r)	8.01
3iqn_A_14	1g	4.53	1.11	10 (7p)	8.03
3gx2_A_20	&a, 4b	3.33	0.65	5 (1[, 1b, !!)	8.67
3gx2_A_14	1g	4.62	1.16	9 (7p)	8.86
4Lvw_A_43	4p	3.2	0.56	6 (1z)	9.21
4Lw0_A_43	4p	3.29	0.79	6 (1z)	9.43
3c44_B_8	1a, 3g, 8d	4.77	2.88	9 (7p)	10.94
1y90_A_8	1a, 3g, 8d	4.72	2.33	3 (3b)	11.33
3dvv_A_8	1a, 3g, 8d	4.77	1.5	9 (7p)	12.05
1y95_A_8	1a, 3g, 8d	4.65	2.87	3 (3b)	12.29
3gx2_A_69	2a	3.07	2.85	3 (3b)	12.6
1c0a_B_637	1a	4.45	1.1	12 (5p)	12.71
1y6s_B_8	1a, 3g, 8d	4.65	0.86	9 (7p)	12.87
1nLc_A_8	1a, 3g, 8d	4.73	1.29	3 (3b)	12.92
1wvd_A_8	1a, 3g, 8d	4.71	1.51	3 (3b)	13.38
1o3z_A_8	1a, 3g, 8d	4.69	1.42	3 (3b)	13.41
1y73_A_8	1a, 3g, 8d	4.68	1.23	3 (3b)	13.99
1wvd_B_8	1a, 3g, 8d	4.67	1.58	3 (3b)	14.33
1nLc_B_8	1a, 3g, 8d	4.63	1.4	3 (3b)	14.46
1y6t_A_8	1a, 3g, 8d	4.82	1.13	3 (3b)	14.88
5fk1_A_14	1g	3.87	0.98	8 (1o)	14.89
1y99_A_8	1a, 3g, 8d	4.75	0.86	10 (7p)	14.93
1y6s_A_8	1a, 3g, 8d	4.74	0.97	10 (7p)	15.2
1o3z_B_8	1a, 3g, 8d	4.6	1.51	3 (3b)	15.88
1y73_B_8	1a, 3g, 8d	4.65	1.37	3 (3b)	17.25
1y6t_B_8	1a, 3g, 8d	4.68	1.21	9 (7p)	17.63
1y95_B_8	1a, 3g, 8d	4.36	1.56	3 (3b)	18.28
3gx7_A_14	1g	3.27	2.14	10 (7p)	19.41

LDTP32 conformer mismatches

name	p. answer	RNAprecis	d_1	first match	d_2
3iqr_A_13	1f, 1o, 3a, 3b, 7a	4 (7r)	3.16	8 (1o)	5.7
5fke_A_13	1f, 1o, 3a, 3b, 7a	4 (7r)	4.41	8 (1o)	4.64
6qis_G_5	1a, 1b	7 (1t)	5.12	1 (1b, 1[)	5.26
5fk3_A_13	1f, 1o, 3a, 3b, 7a	4 (7r)	5.73	3 (3b)	5.98
2ydh_A_33	7r	8 (1o)	5.77	4 (7r)	6.57
5fk2_A_9	3d, 7p	3 (3b)	6.02	10 (7p)	10.43
5fkd_A_9	3d, 7p	3 (3b)	6.49	10 (7p)	6.83
4kze_R_28	7p	6 (1z)	6.79	9 (7p)	12.33
5fk6_A_9	3d, 7p	3 (3b)	6.93	10 (7p)	6.93
5fk3_A_9	3d, 7p	3 (3b)	9.83	10 (7p)	10.72

Table 10: Suites for which the RNAprecis predicted cluster matches the homology related conformer (blue in the plots), the sugar pucker-pairs determined by the Pperp criterion do not match the sugar pucker-pairs of the homology related conformers (red in the plots), or the RNAprecis predicted cluster does not match the homology related conformer (orange in the plots); each sorted by distance.

E.3 Results for the test data set LDTP23

LDTP23: matches

name	p. answer	RNAprecis	d
4qg3_B_2173	2a	**1** (2a)	1.8
3gx5_A_36	#a, 0a, 4a, 7a	**2** (0a, #a, 4a)	2.09
3got_A_50	#a, 0a	**2** (0a, #a, 4a)	2.21
3gao_A_50	#a, 0a	**2** (0a, #a, 4a)	2.23
3gx6_A_65	8d	**5** (8d, 4d)	2.24
2eeu_A_50	#a, 0a	**2** (0a, #a, 4a)	2.27
5fk1_A_47	0a, 1a, 4g	**4** (6g, 4g)	2.28
3gx7_A_65	8d	**5** (8d, 4d)	2.28
4b5r_A_47	0a, 1a, 4g	**4** (6g, 4g)	2.33
3gx2_A_65	8d	**5** (8d, 4d)	2.35
3dw5_A_2648	0a, 1a	**6** (0a)	2.4
2ees_A_50	#a, 0a	**2** (0a, #a, 4a)	2.4
5fkg_A_65	8d	**11** (2g, 8d)	2.45
4fep_B_50	#a, 0a	**2** (0a, #a, 4a)	2.47
5fkd_A_65	8d	**5** (8d, 4d)	2.77
5fkh_A_65	8d	**5** (8d, 4d)	2.82
3gx5_A_65	8d	**5** (8d, 4d)	2.86
3g4m_A_50	#a, 0a	**2** (0a, #a, 4a)	2.89
3gx3_A_47	0a, 1a, 4g	**4** (6g, 4g)	2.91
2ydh_A_65	8d	**5** (8d, 4d)	2.97
5fk5_A_65	8d	**5** (8d, 4d)	3.01
5fk2_A_65	8d	**5** (8d, 4d)	3.08
5fk1_A_65	8d	**5** (8d, 4d)	3.09
3gx7_A_36	#a, 0a, 4a, 7a	**2** (0a, #a, 4a)	3.16
3dis_A_44	2g	**11** (2g, 8d)	3.16
5c7u_B_50	#a, 0a	**2** (0a, #a, 4a)	3.18
5fk3_A_65	8d	**5** (8d, 4d)	3.24
4b5r_A_36	#a, 0a, 4a, 7a	**2** (0a, #a, 4a)	3.27
5fkf_A_65	8d	**5** (8d, 4d)	3.28
4qvi_B_2159	0i	**9** (0i)	3.31
5npm_B_8	8d	**5** (8d, 4d)	3.33
3diz_A_41	0i, 6n	**3** (6n, 0i, 6j)	3.39
3dim_A_44	2g	**11** (2g, 8d)	3.46
3u56_B_2159	0i	**9** (0i)	3.52
3dig_X_14	8d	**5** (8d, 4d)	3.54
6y0t_A_2648	0a, 1a	**6** (0a)	3.64
3gx2_A_36	#a, 0a, 4a, 7a	**2** (0a, #a, 4a)	3.66
3dig_X_44	2g	**11** (2g, 8d)	3.68
7ucr_A_48	0a, 1a	**6** (0a)	3.71
5fk6_A_65	8d	**5** (8d, 4d)	3.77
3dim_A_41	0i, 6n	**3** (6n, 0i, 6j)	3.81
3dj2_A_41	0i, 6n	**3** (6n, 0i, 6j)	3.82
3diy_A_14	8d	**5** (8d, 4d)	3.83
3dis_A_165	2a	**1** (2a)	3.87
3dim_A_14	8d	**5** (8d, 4d)	3.89
3dix_A_14	8d	**5** (8d, 4d)	3.91
2ygh_A_65	8d	**5** (8d, 4d)	3.96
7jje_A_2648	0a, 1a	**6** (0a)	3.98
6b3k_R_62	8d	**5** (8d, 4d)	4.01
3diq_A_14	8d	**5** (8d, 4d)	4.06
3diq_A_41	0i, 6n	**3** (6n, 0i, 6j)	4.06
1o3z_B_9	1a, 1m, 1z, 2a, 7p	**1** (2a)	4.11
3dim_A_154	0a	**2** (0a, #a, 4a)	4.13
5fke_A_65	8d	**5** (8d, 4d)	4.15
4b5r_A_10	2a	**1** (2a)	4.15
4kze_R_62	8d	**5** (8d, 4d)	4.18
3gx6_A_36	#a, 0a, 4a, 7a	**2** (0a, #a, 4a)	4.24
4b5r_A_65	8d	**5** (8d, 4d)	4.34
3dio_X_41	0i, 6n	**3** (6n, 0i, 6j)	4.4
3dis_A_154	0a	**2** (0a, #a, 4a)	4.4
3dis_A_14	8d	**5** (8d, 4d)	4.43

LDTP23: matches

name	p. answer	RNAprecis	d
3diL_A_14	8d	**5** (8d, 4d)	4.43
3diz_A_154	0a	**2** (0a, #a, 4a)	4.51
3dis_A_41	0i, 6n	**3** (6n, 0i, 6j)	4.51
3dj2_A_14	8d	**5** (8d, 4d)	4.58
3dj2_A_154	0a	**2** (0a, #a, 4a)	4.61
3iqr_A_65	8d	**5** (8d, 4d)	4.62
1wvd_B_9	1a, 1m, 1z, 2a, 7p	**1** (2a)	4.63
3diz_A_14	8d	**5** (8d, 4d)	4.67
4b5r_A_69	2a	**1** (2a)	4.74
3diz_A_44	2g	**11** (2g, 8d)	4.77
5fkd_A_10	2a	**1** (2a)	4.8
3iqn_A_65	8d	**5** (8d, 4d)	4.85
3dig_X_154	0a	**2** (0a, #a, 4a)	4.99
3iqp_A_65	8d	**5** (8d, 4d)	5.09
3iqn_A_47	0a, 1a, 4g	**2** (0a, #a, 4a)	5.19
3diy_A_154	0a	**2** (0a, #a, 4a)	5.39
2gis_A_65	8d	**5** (8d, 4d)	5.41
5fk2_A_10	2a	**1** (2a)	5.47
6qis_G_6	2a, 3d, 7d	**1** (2a)	5.47
6xh0_D_36	2a, 8d	**1** (2a)	5.52
3u4m_B_2159	0i	**9** (0i)	5.55
3gx2_A_10	2a	**1** (2a)	5.57
1nLc_A_9	1a, 1m, 1z, 2a, 7p	**1** (2a)	5.73
1o3z_A_9	1a, 1m, 1z, 2a, 7p	**1** (2a)	5.73
1wvd_A_9	1a, 1m, 1z, 2a, 7p	**1** (2a)	5.81
1y73_A_9	1a, 1m, 1z, 2a, 7p	**1** (2a)	5.84
3gao_A_64	6n	**3** (6n, 0i, 6j)	6.03
2oij_A_8	1a, 3g, 8d	**5** (8d, 4d)	6.05
3gx7_A_15	1a, 7d, 8d	**5** (8d, 4d)	6.19
7eaf_A_15	1a, 7d, 8d	**5** (8d, 4d)	6.27
480d_A_2648	0a, 1a	**6** (0a)	6.32
3dvv_A_9	1a, 1m, 1z, 2a, 7p	**1** (2a)	6.37
4qg3_B_2159	0i	**9** (0i)	6.41
6ugg_A_22	4n	**12** (4n)	6.6
6b3k_R_29	2a	**1** (2a)	6.66
3iqr_A_47	0a, 1a, 4g	**2** (0a, #a, 4a)	6.7
5fke_A_21	0i	**9** (0i)	6.7
2eeu_A_66	4n	**12** (4n)	6.93
5fk1_A_21	0i	**9** (0i)	7.04
5fkf_A_21	0i	**9** (0i)	7.16
1y90_A_9	1a, 1m, 1z, 2a, 7p	**1** (2a)	7.46
5fkd_A_21	0i	**9** (0i)	7.57
3iqn_A_15	1a, 7d, 8d	**5** (8d, 4d)	7.59
2gis_A_10	2a	**1** (2a)	7.62
3gx3_A_21	0i	**3** (6n, 0i, 6j)	7.74
3gx2_A_47	0a, 1a, 4g	**2** (0a, #a, 4a)	7.75
5npm_B_55	0i	**9** (0i)	7.83
3diq_A_54	2a	**1** (2a)	8.1
5fk6_A_21	0i	**9** (0i)	8.52
5fk2_A_21	0i	**9** (0i)	8.71
6qit_D_6	2a, 3d, 7d	**1** (2a)	9.12
5fk3_A_21	0i	**9** (0i)	9.19
5fkh_A_21	0i	**9** (0i)	9.21
3gx6_A_21	0i	**3** (6n, 0i, 6j)	9.37
3iqr_A_21	0i	**9** (0i)	10.57
5fk5_A_21	0i	**9** (0i)	10.64
3iqn_A_21	0i	**3** (6n, 0i, 6j)	11.27
3gx7_A_21	0i	**3** (6n, 0i, 6j)	11.6
3gx2_A_21	0i	**3** (6n, 0i, 6j)	11.61
1y95_A_9	1a, 1m, 1z, 2a, 7p	**1** (2a)	11.68

Table 11: Suites for which the RNAprecis predicted cluster matches the homology related conformer (blue in the plots), sorted by distance.

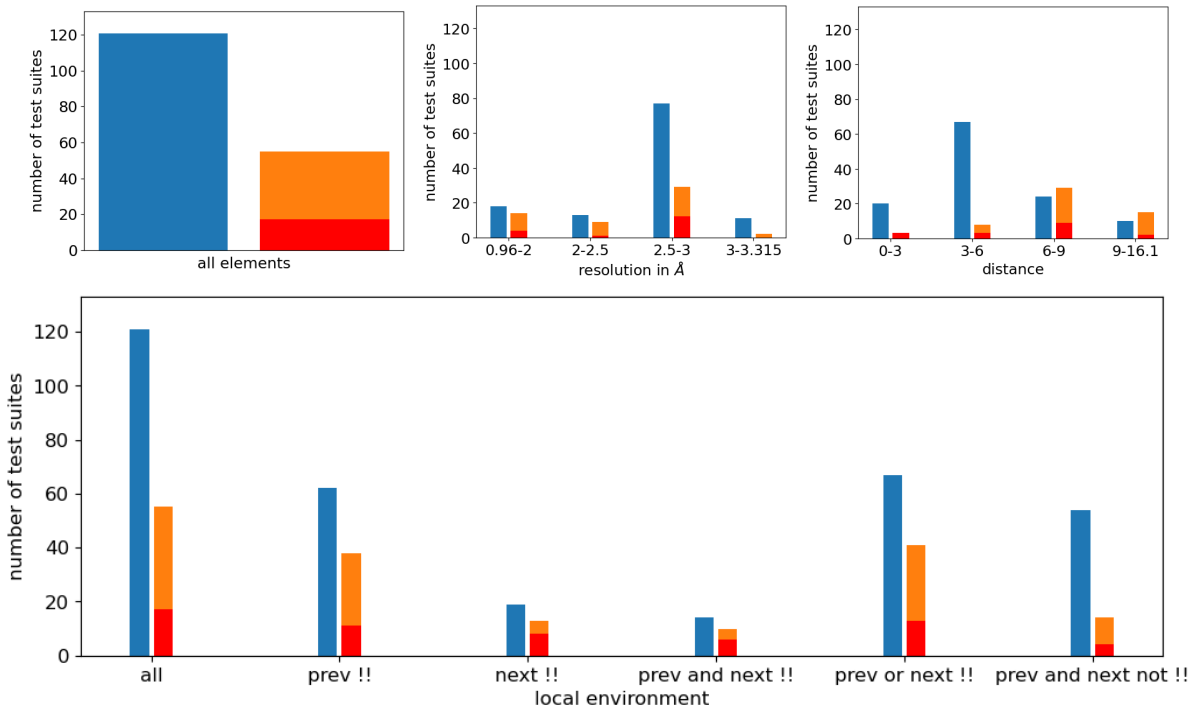

Figure 33: *Relative distribution of agreements and disagreements of MINT-AGE cluster prediction with homology related probable conformer classes for the test set LDT23 with color scheme (reflecting blue for match, orange for mismatch and red for pucker-pair mismatch) from Section 4.4 in the main text). Top left: overall histogram. Top middle: histogram reflecting varying resolutions of PDB suites. Top right: histogram reflecting varying Mahanalobis distances from the predicted cluster. One can see that with increasing distance the rate of mismatches increases. This indicates that despite its limitations discussed in Remark 3.3 in the main text the distance is useful as an empirical indicator of prediction confidence. Bottom: histogram also reflecting dependence on whether previous, next or both suites have been !! or not.*

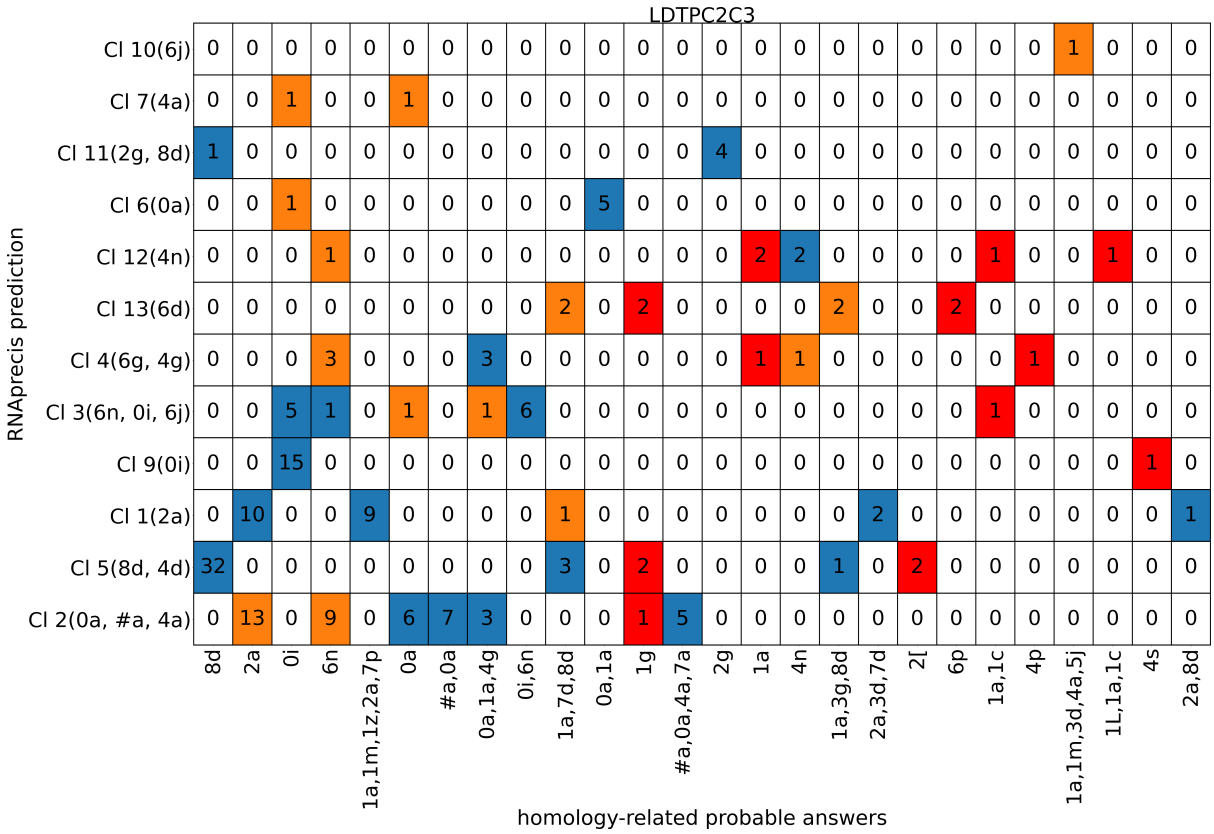

Figure 34: *Classification matrix listing counts of probable homology related conformation classes (horizontal axis) versus MINT-AGE clusters (vertical, with conformer classes in parentheses from MINT-AGE training, see Figure 24) for the low detail test data set of sugar pucker-pairs LDTP23 with colors (from Section 4.4 in the main text): blue for match, orange for mismatch and red for pucker-pair mismatch. Columns listing multiple suite conformers are from sites solved in different conformations across homology-related models, and any of the homology-related conformations were accepted as a matching predictions. Columns listing multiple suite conformers are from sites solved in different conformations across homology-related models, and any of the homology-related conformations were accepted as a matching predictions.*

LDTP23: pucker mismatches

name	p. answer	p_1	p_2	RNAprecis	d
2ydh_A_14	1g	0.3	4.41	8 (6d)	2.59
3gx5_A_14	1g	0.66	3.37	5 (8d, 4d)	2.74
6uc7_B_75	1a	0.31	4.65	4 (6g, 4g)	2.9
3gx2_A_70	1a, 1c	2.85	3.58	12 (4n)	4.4
5fke_A_14	1g	1.22	4.77	5 (8d, 4d)	4.66
7kd1_A_8	1a	2.21	4.77	12 (4n)	4.79
3gx2_A_16	1L, 1a, 1c	2.67	4.6	12 (4n)	6.71
5fkg_A_14	1g	1.28	3.87	8 (6d)	7.73
2gis_A_34	6p	0.68	4.33	13 (6d)	7.91
4xw7_A_28	1a	2.17	4.33	12 (4n)	7.91
3iqp_A_34	6p	0.35	3.38	13 (6d)	8.49
3iqn_A_35	2[	2.25	4.0	5 (8d, 4d)	8.68
3iqr_A_14	1g	1.97	3.98	2 (0a, #a, 4a)	8.69
7kd1_A_43	4p	1.18	4.25	4 (6g, 4g)	8.82
3iqn_A_52	1a, 1c	2.14	4.5	3 (6n, 0i, 6j)	8.98
4xwf_A_27	4s	0.54	4.59	9 (0i)	9.79
3iqr_A_35	2[	2.37	3.77	5 (8d, 4d)	10.32

LDTP23: conformer mismatches

name	p. answer	RNAprecis	d_1	first match	d_2
3ds7_B_223	6n	4 (6g, 4g)	3.16	3 (6n, 0i, 6j)	6.93
3diq_A_154	0a	7 (4a)	4.47	2 (0a, #a, 4a)	5.14
3dio_X_154	0a	3 (6n, 0i, 6j)	4.52	2 (0a, #a, 4a)	5.74
2ees_A_64	6n	4 (6g, 4g)	4.52	3 (6n, 0i, 6j)	7.15
3iqn_A_10	2a	2 (0a, #a, 4a)	5.17	1 (2a)	6.24
3dis_A_81	6n	12 (4n)	6.0	3 (6n, 0i, 6j)	8.17
3gx7_A_47	0a, 1a, 4g	3 (6n, 0i, 6j)	6.4	2 (0a, #a, 4a)	7.29
3got_A_64	6n	4 (6g, 4g)	6.41	3 (6n, 0i, 6j)	10.15
2eev_A_64	6n	2 (0a, #a, 4a)	6.42	3 (6n, 0i, 6j)	12.31
3far_B_8	1a, 3g, 8d	13 (6d)	7.1	5 (8d, 4d)	7.67
5fk5_A_10	2a	2 (0a, #a, 4a)	7.23	1 (2a)	8.68
3ges_A_64	6n	2 (0a, #a, 4a)	7.33	3 (6n, 0i, 6j)	21.17
5fk1_A_15	1a, 7d, 8d	13 (6d)	7.35	5 (8d, 4d)	7.83
3gx6_A_10	2a	2 (0a, #a, 4a)	7.5	1 (2a)	8.76
5c7u_B_37	2a	2 (0a, #a, 4a)	7.75	1 (2a)	8.06
5fk2_A_15	1a, 7d, 8d	1 (2a)	7.82	5 (8d, 4d)	8.14
3iqr_A_10	2a	2 (0a, #a, 4a)	7.86	1 (2a)	8.22
5fkh_A_10	2a	2 (0a, #a, 4a)	7.89	1 (2a)	8.0
2eeu_A_64	6n	2 (0a, #a, 4a)	7.93	3 (6n, 0i, 6j)	21.63
3ger_A_64	6n	2 (0a, #a, 4a)	8.05	3 (6n, 0i, 6j)	24.66
2eew_A_64	6n	2 (0a, #a, 4a)	8.09	3 (6n, 0i, 6j)	24.24
3fo6_A_64	6n	2 (0a, #a, 4a)	8.39	3 (6n, 0i, 6j)	19.74
3iqp_A_69	2a	2 (0a, #a, 4a)	8.72	1 (2a)	9.93
2qek_A_10	1a, 1m, 3d, 4a, 5j	10 (6j)	8.86	2 (0a, #a, 4a)	9.75
2eet_A_64	6n	2 (0a, #a, 4a)	8.92	3 (6n, 0i, 6j)	20.64
6b3k_R_58	4n	4 (6g, 4g)	9.09	12 (4n)	9.62
2ydh_A_10	2a	2 (0a, #a, 4a)	9.15	1 (2a)	11.37
3g4m_A_64	6n	2 (0a, #a, 4a)	9.53	3 (6n, 0i, 6j)	19.43
5fk3_A_10	2a	2 (0a, #a, 4a)	9.97	1 (2a)	11.51
2oij_B_8	1a, 3g, 8d	13 (6d)	10.69	11 (2g, 8d)	11.68
3iqp_A_21	0i	7 (4a)	10.79	3 (6n, 0i, 6j)	11.16
2eeu_A_23	6n	2 (0a, #a, 4a)	11.56	3 (6n, 0i, 6j)	15.17
3iqp_A_10	2a	2 (0a, #a, 4a)	11.71	1 (2a)	17.92
3gx7_A_69	2a	2 (0a, #a, 4a)	12.18	1 (2a)	12.98
2gis_A_21	0i	6 (0a)	12.2	3 (6n, 0i, 6j)	12.29
3gx5_A_10	2a	2 (0a, #a, 4a)	12.43	1 (2a)	15.86
3iqn_A_69	2a	2 (0a, #a, 4a)	13.6	1 (2a)	17.09
3gx6_A_15	1a, 7d, 8d	13 (6d)	16.14	5 (8d, 4d)	18.95

Table 12: Suites for which the sugar pucker-pairs determined by the Pperp criterion do not match the sugar pucker-pairs of the homology related conformers (red in the plots), or the RNAprecis predicted cluster does not match the homology related conformer (orange in the plots); each sorted by distance.

E.4 Results for the test data set LDTP22

LDTP22: matches

name	p. answer	RNAprecis	d
5npm_B_15	4p	**3** (4p)	1.64
3dix_A_43	4b	**10** (4b)	1.7
5fk3_A_34	6p	**2** (6p, !!)	1.91
3diz_A_43	4b	**10** (4b)	1.94
3diz_A_40	0b, 4b	**4** (4b)	2.4
3dim_A_43	4b	**10** (4b)	2.47
3gx6_A_35	2[	**1** (2[)	2.73
3diz_A_164	6p	**2** (6p, !!)	2.77
3dis_A_43	4b	**10** (4b)	2.84
3ger_A_49	2[	**1** (2[)	2.95
2eeu_A_63	0a, 0b, 4b	**8** (0b)	3.04
2eew_A_49	2[	**1** (2[)	3.06
3fo6_A_49	2[	**1** (2[)	3.11
3dig_X_40	0b, 4b	**4** (4b)	3.21
3gx7_A_35	2[	**1** (2[)	3.42
2ygh_A_20	&a, 4b	**4** (4b)	3.51
3dis_A_164	6p	**2** (6p, !!)	3.56
3g4m_A_63	0a, 0b, 4b	**7** (4b, 0b)	3.57
2eev_A_63	0a, 0b, 4b	**7** (4b, 0b)	3.59
7eaf_A_34	6p	**2** (6p, !!)	3.69
2eet_A_63	0a, 0b, 4b	**8** (0b)	3.7
2eew_A_63	0a, 0b, 4b	**7** (4b, 0b)	3.7
3ger_A_63	0a, 0b, 4b	**8** (0b)	3.75
5fkh_A_35	2[	**1** (2[)	3.85
5fke_A_34	6p	**2** (6p, !!)	3.87
7kd1_A_22	2[	**1** (2[)	4.02
6xh1_D_34	2[	**6** (2u, 2o, 2[)	4.08
3gx3_A_20	&a, 4b	**4** (4b)	4.13
3diq_A_43	4b	**10** (4b)	4.18
4b5r_A_35	2[	**1** (2[)	4.29
3dim_A_12	2[	**1** (2[)	4.39
3diy_A_12	2[	**1** (2[)	4.43
5fk3_A_35	2[	**1** (2[)	4.45
3gx2_A_34	6p	**2** (6p, !!)	4.46
3dio_X_164	6p	**2** (6p, !!)	4.52
3got_A_48	2h, 2u, 7a	**6** (2u, 2o, 2[)	4.66
3ges_A_63	0a, 0b, 4b	**7** (4b, 0b)	4.72
3dis_A_12	2[	**1** (2[)	4.78
3fo6_A_63	0a, 0b, 4b	**8** (0b)	4.79
3got_A_49	2[	**1** (2[)	4.82
4Lvz_A_43	4p	**3** (4p)	4.91
3diy_A_40	0b, 4b	**4** (4b)	5.0
5fk1_A_64	2o, 2z	**6** (2u, 2o, 2[)	5.2
3gx6_A_64	2o, 2z	**6** (2u, 2o, 2[)	5.26
3ges_A_49	2[	**1** (2[)	5.33
6xh2_D_34	2[	**6** (2u, 2o, 2[)	5.41
3dim_A_13	4p	**3** (4p)	5.43
3dis_A_13	4p	**3** (4p)	5.48
7eaf_A_35	2[	**1** (2[)	5.63
3gx6_A_20	&a, 4b	**7** (4b, 0b)	5.64
2ydh_A_35	2[	**1** (2[)	5.95
3gx3_A_34	6p	**2** (6p, !!)	6.04
3dj2_A_40	0b, 4b	**4** (4b)	6.05
5fk2_A_35	2[	**1** (2[)	6.14
3gx7_A_34	6p	**2** (6p, !!)	6.46
3gao_A_36	6p	**2** (6p, !!)	6.55
3gx5_A_64	2o, 2z	**6** (2u, 2o, 2[)	6.63
3diL_A_40	0b, 4b	**4** (4b)	6.63
3dis_A_40	0b, 4b	**4** (4b)	6.99
3gx5_A_35	2[	**1** (2[)	7.06
5fkf_A_35	2[	**1** (2[)	7.07
3iqn_A_64	2o, 2z	**6** (2u, 2o, 2[)	7.12
4b5r_A_34	6p	**2** (6p, !!)	7.15
3gao_A_49	2[	**1** (2[)	7.41
5fkf_A_34	6p	**2** (6p, !!)	7.61
3iqr_A_34	6p	**2** (6p, !!)	8.42
6q57_A_43	4p	**3** (4p)	8.45
2ydh_A_34	6p	**2** (6p, !!)	10.59
3gx5_A_34	6p	**2** (6p, !!)	10.72
3gx6_A_34	6p	**2** (6p, !!)	11.05
5fk5_A_35	2[	**1** (2[)	11.12
3iqn_A_34	6p	**2** (6p, !!)	11.15
5fkg_A_34	6p	**2** (6p, !!)	11.45

LDTP22: pucker mismatches

name	p. answer	p_1	p_2	RNAprecis	d
7eaf_A_14	1g	0.76	1.71	3 (4p)	5.06
5fk2_A_14	1g	0.9	1.05	3 (4p)	5.69
6ugg_A_47	1b	1.86	1.52	3 (4p)	5.8
6ugg_A_20	7p	1.58	2.2	1 (2[)	10.12
3gx2_A_15	1a, 7d, 8d	1.16	2.67	2 (6p, !!)	10.26
5fk3_A_14	1g	0.85	0.97	2 (6p, !!)	10.54
3gx6_A_14	1g	0.68	2.71	2 (6p, !!)	17.71

LDTP22 conformer mismatches

name	p. answer	RNAprecis	d_1	first match	d_2
2ees_A_48	2h, 2u, 7a	3 (4p)	4.09	6 (2u, 2o, 2[)	5.26
5c7u_B_48	2h, 2u, 7a	2 (6p, !!)	4.56	6 (2u, 2o, 2[)	9.75
3dix_A_12	2[	4 (4b)	4.65	1 (2[)	5.2
3gx2_A_35	2[	9 (0b)	4.86	1 (2[)	5.45
2ydh_A_20	&a, 4b	1 (2[)	5.03	4 (4b)	5.79
3gx5_A_20	&a, 4b	1 (2[)	5.07	4 (4b)	5.19
2ygh_A_35	2[	9 (0b)	5.49	1 (2[)	5.66
3ges_A_48	2h, 2u, 7a	3 (4p)	5.91	6 (2u, 2o, 2[)	5.93
4Lvx_A_43	4p	5 (2z)	6.12	3 (4p)	6.24
3gx3_A_64	2o, 2z	2 (6p, !!)	6.9	5 (2z)	8.63
3gx2_A_64	2o, 2z	2 (6p, !!)	6.91	5 (2z)	8.75
6ugg_A_48	0b	1 (2[)	8.13	9 (0b)	11.8
3gx7_A_64	2o, 2z	3 (4p)	9.04	5 (2z)	9.25
2gis_A_64	2o, 2z	2 (6p, !!)	9.13	6 (2u, 2o, 2[)	10.33
3iqr_A_64	2o, 2z	2 (6p, !!)	9.61	5 (2z)	13.63
3iqp_A_64	2o, 2z	2 (6p, !!)	11.17	5 (2z)	16.7
5d99_A_2655	4s	3 (4p)	12.94	-	-
480d_A_2655	4s	3 (4p)	13.28	-	-

Table 13: Suites for which the RNAprecis predicted cluster matches the homology related conformer (blue in the plots), the sugar pucker-pairs determined by the Pperp criterion do not match the sugar pucker-pairs of the homology related conformers (red in the plots), or the RNAprecis predicted cluster does not match the homology related conformer (orange in the plots); each sorted by distance.

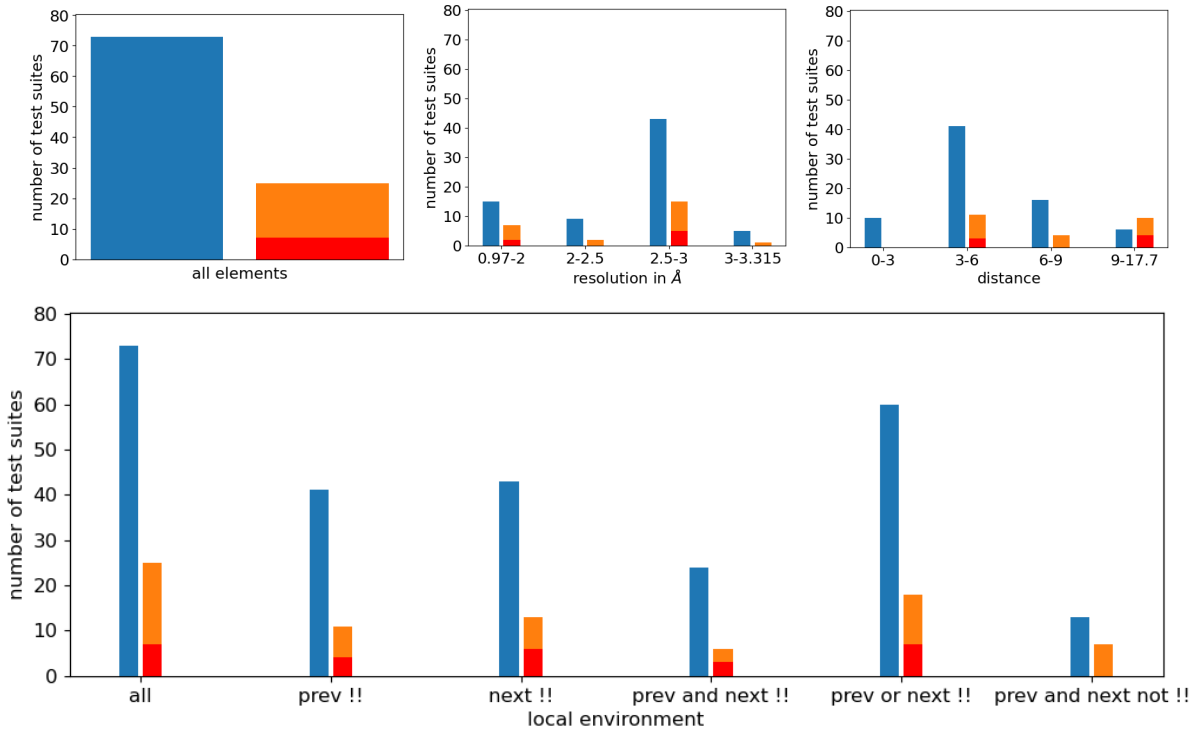

Figure 35: *Relative distribution of agreements and disagreements of MINT-AGE cluster prediction with homology related probable conformer classes for the test set LDT22 with color scheme (reflecting blue for match, orange for mismatch and red for pucker-pair mismatch) from Section 4.4 in the main text). Top left: overall histogram. Top middle: histogram reflecting varying resolutions of PDB suites. Top right: histogram reflecting varying Mahanalobis distances from the predicted cluster. One can see that with increasing distance the rate of mismatches increases. This indicates that despite its limitations discussed in Remark 3.3 in the main text the distance is useful as an empirical indicator of prediction confidence. Bottom: histogram also reflecting dependence on whether previous, next or both suites have been !! or not.*

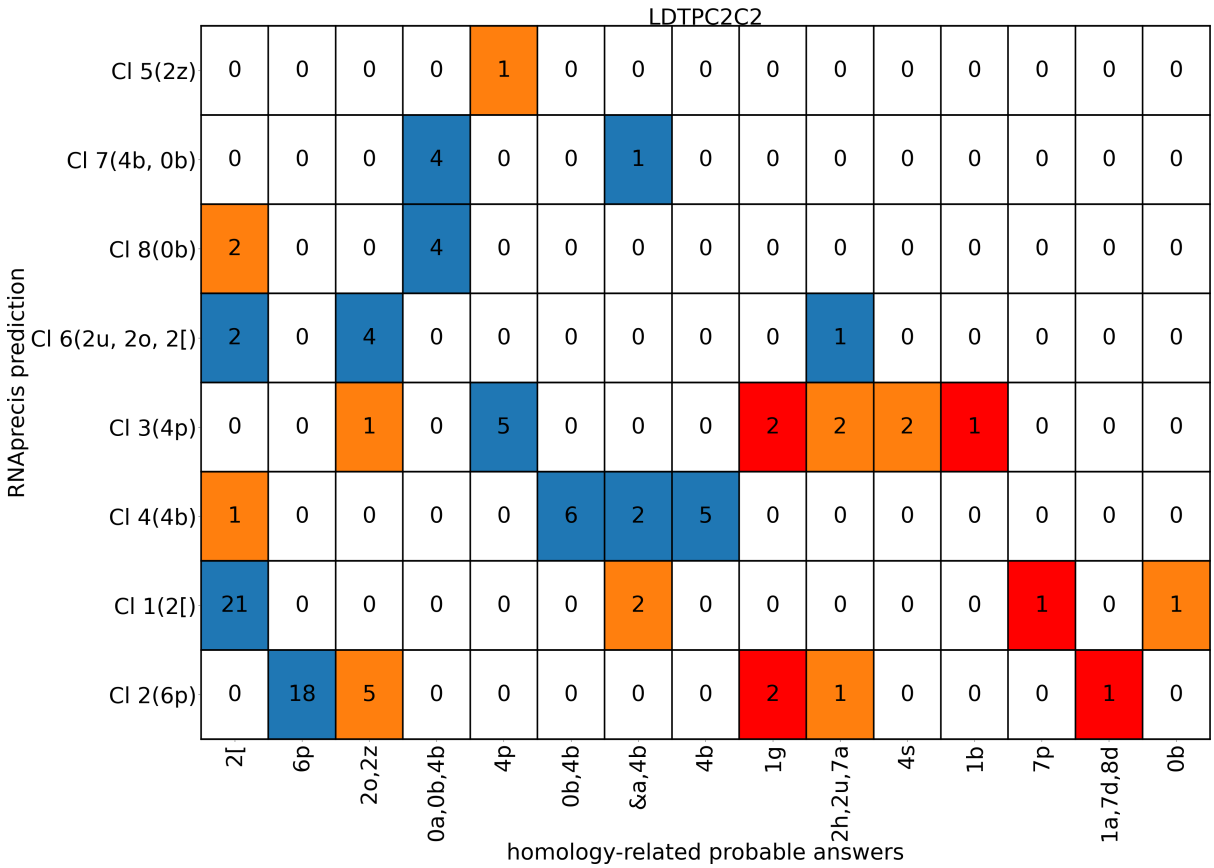

Figure 36: *Classification matrix listing counts of probable homology related conformation classes (horizontal axis) versus MINT-AGE clusters (vertical, with conformer classes in parentheses from MINT-AGE training, see Figure 25) for the low detail test data set of sugar pucker-pairs LDTP22 with colors (from Section 4.4 in the main text): blue for match, orange for mismatch and red for pucker-pair mismatch. Columns listing multiple suite conformers are from sites solved in different conformations across homology-related models, and any of the homology-related conformations were accepted as a matching predictions. Columns listing multiple suite conformers are from sites solved in different conformations across homology-related models, and any of the homology-related conformations were accepted as a matching predictions.*

References

[1] Steitz TA, Moore PB. RNA, the first macromolecular catalyst: the ribosome is a ribozyme. Trends in biochemical sciences. 2003;28(8):411–418.

[2] Batey RT, Gilbert SD, Montange RK. Structure of a natural guanine-responsive riboswitch complexed with the metabolite hypoxanthine. Nature. 2004;432(7015):411–415.

[3] Murray LJ, Arendall III WB, Richardson DC, Richardson JS. RNA backbone is rotameric. Proceedings of the National Academy of Sciences. 2003;100(24):13904–13909.

[4] Richardson JS, Schneider B, Murray LW, Kapral GJ, Immormino RM, Headd JJ, et al. RNA backbone: consensus all-angle conformers and modular string nomenclature (an RNA Ontology Consortium contribution). Rna. 2008;14(3):465–481.

[5] Mardia K, Wiechers H, Eltzner B, Huckemann SF. Principal component analysis and clustering on manifolds. Journal of Multivariate Analysis. 2021; p. 104862.

[6] Eltzner B, Huckemann S, Mardia KV. Torus principal component analysis with applications to RNA structure. Ann Appl Statist. 2018;12(2):1332–1359. doi:10.1214/17-AOAS1115.

[7] Mardia KV, Eltzner B, Huckemann SF. Constrained Shape Analysis with Applications to RNA Structure. arxiv:250216270. 2025;.

[8] Goodall CR, Mardia KV. Multivariate aspects of shape theory. The Annals of Statistics. 1993;21(2):848–866.

[9] Mahalanobis P. A note on the statistical and biometric writings of Karl Pearson. Sankhyā: The Indian Journal of Statistics (1933-1960). 1936;2(4):411–422.

[10] Mardia KV, Kent JT, Taylor CC. Multivariate analysis. vol. 88. John Wiley & Sons; 2024.

[11] Wiechers H, Eltzner B, Mardia KV, Huckemann SF. Learning torus PCA-based classification for multiscale RNA correction with application to SARS-CoV-2. Journal of the Royal Statistical Society Series C: Applied Statistics. 2023;72(2):271–293.

[12] Trachman III RJ, Autour A, Jeng SC, Abdolahzadeh A, Andreoni A, Cojocaru R, et al. Structure and functional reselection of the Mango-III fluorogenic RNA aptamer. Nature chemical biology. 2019;15(5):472–479.

[13] Davis IW, Leaver-Fay A, Chen VB, Block JN, Kapral GJ, Wang X, et al. MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. Nucleic acids research. 2007;35(suppl_2):W375–W383.

[14] Chen VB, Arendall WB, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, et al. MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallographica Section D: Biological Crystallography. 2010;66(1):12–21.

[15] Jain S, Richardson DC, Richardson JS. Computational methods for RNA structure validation and improvement. In: Methods in Enzymology. vol. 558. Elsevier; 2015. p. 181–212.

[16] Emsley P, Lohkamp B, Scott WG, Cowtan K. Features and development of Coot. Acta Crystallographica Section D: Biological Crystallography. 2010;66(4):486–501.

[17] Williams C, Richardson J. High-quality RNA residues: RNA2023; 2023. Available from: `https://doi.org/10.5281/zenodo.8103014`.

[18] Leontis NB, Zirbel CL. Nonredundant 3D structure datasets for RNA knowledge extraction and benchmarking. In: Leontis NB, Westhof E, editors. RNA 3D Structure Analysis and Prediction. Springer; 2012. p. 281–298.

[19] Fromm SA, O'Connor KM, Purdy M, Bhatt PR, Loughran G, Atkins JF, et al. The translating bacterial ribosome at 1.55 Å resolution generated by cryo-EM imaging services. Nature Communications. 2023;14(1):1095.

[20] Faille A, Dent KC, Pellegrino S, Jaako P, Warren AJ. The chemical landscape of the human ribosome at 1.67 Å resolution. bioRxiv. 2023; p. 2023–02.

[21] Williams CJ, Richardson DC, Richardson JS. The importance of residue-level filtering and the Top2018 best-parts dataset of high-quality protein residues. Protein Science. 2022;31(1):290–300.

[22] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Research. 2000;28:235–242.

[23] Parkinson G, Vojtechovsky J, Clowney L, Brünger AT, Berman HM. New parameters for the refinement of nucleic acid-containing structures. Acta Crystallographica Section D: Biological Crystallography. 1996;52(1):57–64.

[24] Florek K, Łukaszewicz J, Perkal J, Steinhaus H, Zubrzycki S. Sur la liaison et la division des points d'un ensemble fini. In: Colloquium mathematicum. vol. 2; 1951. p. 282–285.

[25] Langfelder P, Zhang B, Horvath S. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. Bioinformatics. 2008;24(5):719–720.

[26] Obulkasim A, Meijer GA, van de Wiel MA. Semi-supervised adaptive-height snipping of the hierarchical clustering tree. BMC bioinformatics. 2015;16:1–11.

[27] Jung S, Dryden IL, Marron JS. Analysis of principal nested spheres. Biometrika. 2012;99(3):551–568.

[28] Dümbgen L, Walther G. Multiscale inference about a density. Ann Statist. 2008;36(4):1758–1785. doi:10.1214/07-AOS521.

[29] Huckemann S, Kim KR, Munk A, Rehfeldt F, Sommerfeld M, Weickert J, et al. The circular SiZer, inferred persistence of shape parameters and application to early stem cell differentiation. Bernoulli. 2016;22(4):2113–2142.

[30] Chaudhuri P, Marron JS. Scale space view of curve estimation. The Annals of Statistics. 2000;28(2):408–428.

[31] Dryden IL, Mardia KV. Statistical shape analysis: with applications in R. 2nd ed. John Wiley & Sons; 2016.

[32] Bookstein FL. The Measurement of Biological Shape and Shape Change. vol. 24 of Lecture Notes in Biomathematics. 2nd ed. New York: Springer-Verlag; 1978.

[33] Mardia KV, Patrangenaru V. On affine and projective shape data analysis. Functional and Spatial Data Analysis, Proceedings of the 20th LASR Workshop (Eds: KV Mardia and RG Aykroyd). 2001; p. 39–45.

[34] Groisser D, Tagare HD. On the topology and geometry of spaces of affine shapes. Journal of Mathematical Imaging and Vision. 2009;34:222–233.

[35] Arnaudon M, Miclo L. Means in complete manifolds: uniqueness and approximation. ESAIM: Probability and Statistics. 2014;18:185–206.

[36] Mardia KV, Jupp PE. Directional Statistics. New York: Wiley; 2000.

[37] Huang L, Wang J, Lilley DM. A critical base pair in k-turns determines the conformational class adopted, and correlates with biological function. Nucleic acids research. 2016;44(11):5390–5398.

[38] Klein DJ, Schmeing TM, Moore PB, Steitz TA. The kink-turn: a new RNA secondary structure motif. The EMBO journal. 2001;20(15):4214–4221.

[39] Ennifar E, Walter P, Dumas P. A crystallographic study of the binding of 13 metal ions to two related RNA duplexes. Nucleic Acids Research. 2003;31(10):2671–2682.

[40] Wilt HM, Yu P, Tan K, Wang YX, Stagno JR. Tying the knot in the tetrahydrofolate (THF) riboswitch: a molecular basis for gene regulation. Journal of structural biology. 2021;213(1):107703.

[41] Huang H, Suslov NB, Li NS, Shelke SA, Evans ME, Koldobskaya Y, et al. A G-quadruplex–containing RNA activates fluorescence in a GFP-like fluorophore. Nature chemical biology. 2014;10(8):686–691.

[42] Koirala D, Shelke SA, Dupont M, Ruiz S, DasGupta S, Bailey LJ, et al. Affinity maturation of a portable Fab–RNA module for chaperone-assisted RNA crystallography. Nucleic Acids Research. 2018;46(5):2624–2635.

[43] Stoddard CD, Montange RK, Hennelly SP, Rambo RP, Sanbonmatsu KY, Batey RT. Free state conformational sampling of the SAM-I riboswitch aptamer domain. Structure. 2010;18(7):787–797.

[44] Gilbert SD, Love CE, Edwards AL, Batey RT. Mutational analysis of the purine riboswitch aptamer domain. Biochemistry. 2007;46(46):13297–13309.

[45] Liebschner D, Afonine PV, Baker ML, Bunkóczi G, Chen VB, Croll TI, et al. Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. Acta Crystallographica Section D: Structural Biology. 2019;75(10):861–877.