Classification of RNA backbone conformations into rotamers using ^{13}C' chemical shifts: Exploring how far we can go.

Alejandro A Icazatti [Corresp., Equal first author, 1] , Juan M Loyola [Equal first author, 1] , Igal Szleifer [2, 3, 4] , Jorge A Vila [1] , Osvaldo A Martin [1]

[1] IMASL - CONICET, Universidad Nacional de San Luis, San Luis, Argentina

[2] Department of Biomedical Engineering, Northwestern University, Evanston, Illinois, United States

[3] Chemistry of Life Processes Institute, Northwestern University, Evanston, Illinois, United States

[4] Department of Chemistry, Northwestern University, Evanston, Illinois, United States

Corresponding Author: Alejandro A Icazatti
Email address: ale.icazatti@gmail.com

The conformational space of the ribose-phosphate backbone is very complex as it is defined in terms of six torsional angles. To help delimit the RNA backbone conformational preferences, 46 rotamers have been defined in terms of these torsional angles. In the present work, we use the ribose experimental and theoretical ^{13}C' chemical shifts data and machine learning methods to classify RNA backbone conformations into rotamers and families of rotamers. We show to what extent the experimental ^{13}C' chemical shifts can be used to identify rotamers and discuss some problem with the theoretical computations of ^{13}C' chemical shifts.

Classification of RNA backbone conformations into rotamers using ^{13}C′ chemical shifts: Exploring how far we can go.

A. A. Icazatti[1,*], J. M. Loyola[1], I. Szleifer[2,3,4], J. A. Vila[1], and O. A. Martin[1]

[1]**IMASL - CONICET, Universidad Nacional de San Luis, San Luis, Argentina**
[2]**Department of Biomedical Engineering**
[3]**Chemistry of Life Processes Institute**
[4]**Department of Chemistry, Northwestern University, Evanston, Illinois 60208, United States**

Corresponding author:
A. A. Icazatti*

Email address: ale.icazatti@gmail.com

ABSTRACT

The conformational space of the ribose–phosphate backbone is very complex as it is defined in terms of six torsional angles. To help delimit the RNA backbone conformational preferences, 46 rotamers have been defined in terms of these torsional angles. In the present work, we use the ribose experimental and theoretical ^{13}C′ chemical shifts data and machine learning methods to classify RNA backbone conformations into rotamers and families of rotamers. We show to what extent the experimental ^{13}C′ chemical shifts can be used to identify rotamers and discuss some problem with the theoretical computations of ^{13}C′ chemical shifts.

INTRODUCTION

Nucleic acids are central macromolecules for the storing, flow and regulation of genetic and epigenetic information in cellular organisms. RNA can adopt a wide variety of 3D structural conformations and this structural variability explains the multiplicity of roles that RNA performs on cells (Wan et al., 2011; Eddy, 2001). The classification of RNA backbone conformations into rotamers is a very useful way to delimit the conformational space of RNA structures. Rotamers are defined in terms of the backbone torsional angles namely α, β, γ, δ, ε and ζ (as shown in Figure 1). This classification was proposed by Richardson et al. (2008), and has been achieved after the attempts of different research groups to find a consensus RNA backbone structural classification. There are 55 backbone rotamers, from which 46 are rotamers with well defined torsional angles distributions, and the remaining 9 rotamers were proposed as *wannabe* rotamers. The 'suite' is the basic subunit used for rotamer classification. The suite is defined from sugar-to-sugar (or from the δ torsional angle of residue *i-1* to the δ torsional angle of residue *i*), and it is contained within the dinucleotide (DN) subunit (see Figure 1). ^{13}C′ chemical shifts (CS) have been successfully used by our and other groups for proteins and glycans structural determination, validation and refinement (Shen and Bax, 2010; Martin et al., 2013; Frank et al., 2015; Garay and Vila, 2018). ^{1}H CS have been successfully used by Sripakdeevong et al. (2014) for structure determination and prediction of noncanonical RNA motifs. Methods incorporating ^{13}C CS for RNA structure determination, validation and refinement are also available (Frank et al., 2013, 2014; Brown et al., 2015) but, to our knowledge, none of them include the explicit use of backbone rotamers. In this work, we study how to use ^{13}C′ CS to classify RNA backbone conformations into rotamers with machine learning models. Overall, a complete understanding of the molecular basis of the biological processes in which RNA molecules are involved entails an accurate knowledge of their 3D structure. In this regard, it is well known that

44 the computation of the ^{13}C′ chemical shifts (CS) for RNA, at DFT-level of theory, is very sensitive to
45 the backbone conformation of the molecule. Thus, among other potential application of our work is to
46 build, for any possible combination of RNA backbone torsional-angles conformations, a detailed ^{13}C CS
47 look-up table. Hence, given a ^{13}C CS value the corresponding set of RNA backbone torsional angles
48 can be quickly determined, and vice versa, making the look-up table a very valuable tool with which
    determine, validate and refine RNA conformations.
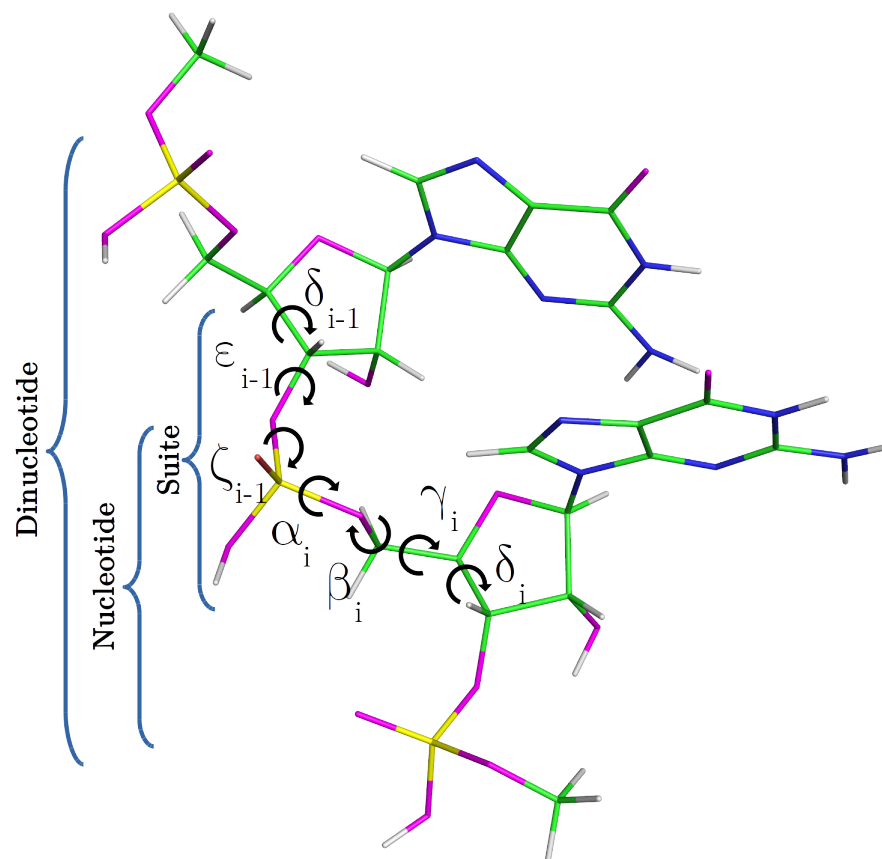
Figure 1. RNA DN template with sequence AA, obtained from a random PDB structure. C, H, O, N and
P nuclei are colored in green, white, magenta, blue and yellow, respectively. Torsional angles of RNA
backbone are named on Greek characters (α, β, γ, δ, ε, ζ). Suite (from δ_{i-1} to δ_i), DN and nucleotide
subunits are indicated.

49

METHODS

51 In order to provide a clear understanding of the methodology implemented in this work, a flowchart with
52 the overall work-flow is shown in Figure 2. A theoretical dataset of RNA backbone rotamers with ^{13}C′
53 CS values is necessary to train the machine learning models to classify RNA experimental suites into
54 rotamers. In the following two sections we explain how we obtained both datasets.

Experimental dataset

56 Experimental ^{13}C′ CS data for RNA molecules was retrieved from the BioMagResBank (BMRB;
57 www.bmrb.wisc.edu)(Ulrich et al., 2008), along with their corresponding structures from the Protein
58 Data Bank (PDB; https://www.rcsb.org/) (Berman et al., 2000). As it is fundamental to count on reliable
59 experimental ^{13}C′ CS values for an accurate structural analysis, data curation was carried out using
60 13Check_RNA (Icazatti et al., 2018) a Python module to correct RNA ^{13}C′ CS systematic errors, recently
61 developed in our group. The obtained dataset (see Supplementary Table S1) contains 26 RNA structures
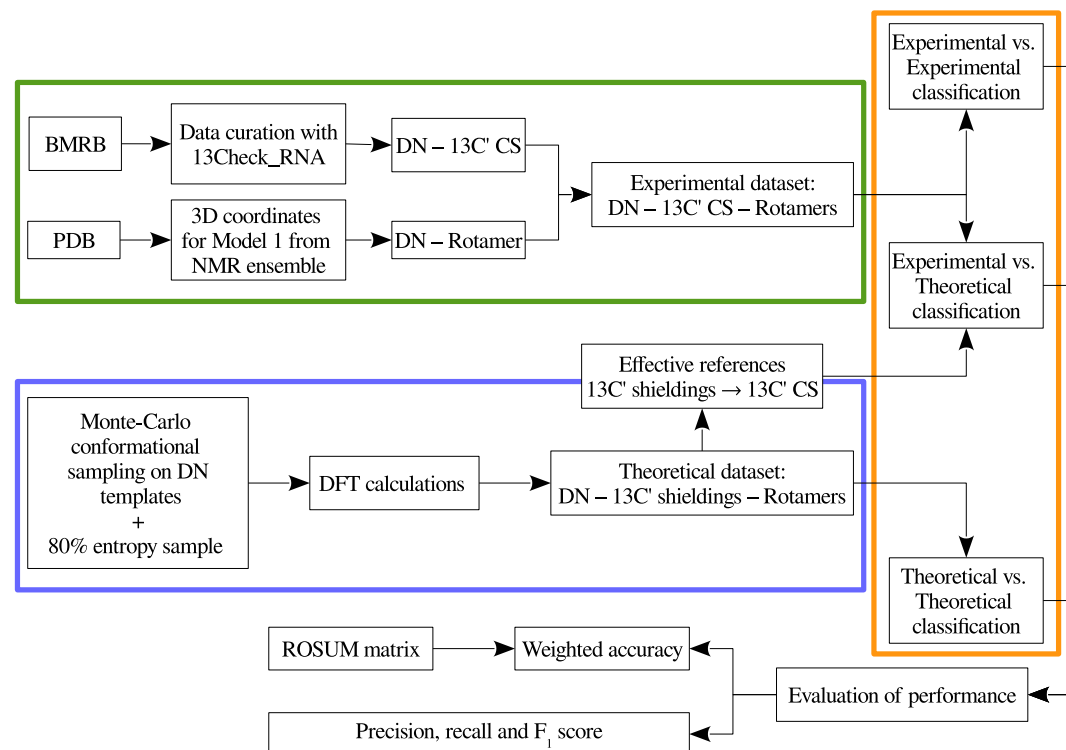
Figure 2. Flowchart of the general work-flow followed in this work. The experimental data retrieval process and the theoretical data generation steps are indicated inside the green and the blue boxes, respectively. The classification step using machine learning models is indicated with an orange box. The term *rotamers* could indicate the original backbone rotamers or redefined rotamer families.

62 with $^{13}C'$ CS for the five ribose carbon nuclei (C1′, C2′, C3′, C4′ and C5′), providing a total of 391 suite
63 subunits and 391 sets of $^{13}C'$ CS. As there are at least 8 models in the NMR ensembles for each RNA
64 molecule (up to 20 in some cases), the complete database contains 7612 conformations. Given that we
65 needed a one-to-one correspondence between the sets of CS and the rotamer suites, only the first structure
66 from each NMR ensemble was used, considering that the first model listed in the PDB files is usually
67 reported as the structure with the lowest energy scoring. This choice has a negligible average effect
68 on the results of our analysis (see Supplementary Figures S10 and S11). For every PDB entry, the 3D
69 coordinates of the first model were extracted in order to compute the backbone torsional angles (δ_{i-1},
70 ε_{i-1}, ζ_{i-1}, α_i, β_i, γ_i, δ_i) of the suites. Then, these torsional angles were used to assign the RNA suites to
71 their corresponding rotamer names. From the 46 original rotamers, only 38 are represented in the final
72 experimental dataset.

Theoretical dataset

74 In order to have a complete dataset with the 46 RNA backbone rotamers and their corresponding $^{13}C'$
75 CS, a theoretical dataset was also constructed. A template for each of the 16 possible combinations of
76 DN (A, C, U and G) sequences was obtained from RNA structures found in the PDB. A Monte-Carlo
77 conformational sampling was carried out by rotating the backbone torsional angles of the corresponding
78 suite contained in each DN, while keeping the bond-lengths and bond-angles fixed (rigid geometry
79 approximation). To perform such rotations, the torsional angle distributions for each of the 46 RNA
80 backbone rotamer suites (Richardson et al., 2008) were used. A function which eliminates conformers
81 with atom clashes was implemented as part of the routine. As a result, 10,340,852 conformations were
82 generated. Quantum-theory level computation of CS is very time-consuming. Therefore, to reduce the
83 number of calculations, a smaller number of conformations was selected. Aiming to keep most of the
84 variability of the originally generated conformations, we computed the Shannon entropy S (see Equation 1)
85 of the distribution of torsional angles. Here, P_i is the probability of the i conformation taken from a

86 histogram with a bin size of 5 degrees. The entropy was computed for different subsets of conformations
87 and sample sizes (from 5 to 100) (see Figure 3). We decided to use the 80% of the maximum entropy as
88 a cutoff, which implies a sample size of around 40 conformations per rotamer. As we also considered
89 the 16 combinations of DN sequences, the total number of conformations computed at the DFT level of
90 theory was 30,530.

$$S = -\sum_i P_i \ln P_i \qquad (1)$$
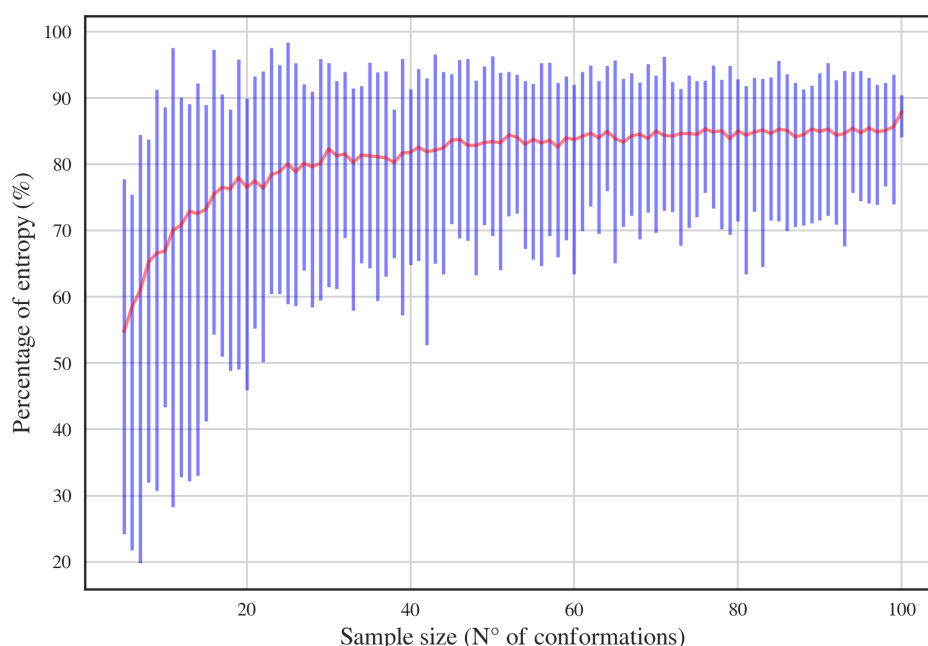
Figure 3. Percentage of entropy of the sample against sample size for a given DN sequence and rotamer,
UU and 1a, respectively, in this case. The red line and the blue bars represent the mean and the range of
percentage of entropy for a given sample size, respectively.

Details of the quantum-chemical calculations of the $^{13}C'$ shieldings

91
92 Previous to the DFT calculations of the obtained dataset, a test was performed over a subset of 41 rotamers
93 of sequence AA. A similar approach as described below for mononucleotides was used, except that the
94 templates were methyl-blocked DNs: $Me - O3'_{i-2} - A_{i-1} - A_i - O5'_{i+1} - Me$. Subsequent comparison
95 of the obtained $^{13}C'$ CS for these DNs with those obtained from the corresponding mononucleotides,
96 gave the same result within 10^{-2} ppm while the total computation time was approximately half the total
97 time for computing the complete DNs. Thus, the DN conformations from the final dataset were split in
98 their corresponding mononucleotide subunits. Nucleotide subunits were treated as terminally-blocked
99 mononucleotides with methyl groups (Me) in both termini $(Me - O3'_{i-1} - X_i - O5'_{i+1} - Me)$. Phosphate
100 groups of the backbone were treated as neutral, because we assume that all backbone charges are shielded
101 during the quantum-chemical calculations. Results based on the analysis of 139 conformations of ubiquitin
102 at pH 6.5 (Vila and Scheraga, 2008), indicate that use of neutral, rather than charged, aminoacids is a
103 significantly better approximation of the observed $^{13}C^{\alpha}$ CS in solution for the acidic groups, and a slightly
104 better representation, though significantly less expensive in computational time, for the basic groups.
105 Considering that the phosphate group in RNA is close to the nucleus of interest (as it happens with the
106 acidic groups) we can assume, without losing generality, that neutral rather than charged phosphate group

107 is a better approximation for the computation of the $^{13}C'$ CS in the RNA suites. This approach was
108 also adopted because under physiological conditions, the phosphate groups are completely ionized and
109 neutralized by positive charges (Lehninger A.L. and Cox, 2000). A 6–311+G(2d,p) locally dense basis
110 set (Chesnut and Moore, 1989) was used for calculation of backbone $^{13}C'$ CS and their nearest neighbor
111 nuclei, at the DFT level of approximation (see Figure 4 for details). The remaining nuclei were treated
112 with a 3-21G basis set. The OB98 density functional was used because good results were previously
113 observed for proteins and glycans in our group (Vila and Scheraga, 2009; Garay et al., 2014). All DFT
114 computations were done using the Gaussian package (Frisch et al., 2004). Summarizing, the adopted
115 strategies make the computed $^{13}C'$ CS from mononucleotides suitable for comparison with the $^{13}C'$ CS
116 observed from complete RNA molecules.

Figure 4. Example of a methyl blocked mononucleotide used for DFT calculations. The locally-dense basis-set approach is indicated by the different colors: the nuclei in magenta were treated with the extended 6-311+G(2d,p) basis set and the nuclei in green were treated with the smaller 3-21G basis set.

Families of rotamers

118 The original 46 RNA backbone rotamers were grouped in families based on their δ_{i-1}, δ_i, α and γ
119 torsional angles values. Only these 4 (out of 7) backbone torsional angles in the suite subunit were chosen
120 to group the rotamers because their distributions of observed values are bimodal (δ_{i-1} and δ_i) and trimodal
121 (γ and α), with clearly separated peaks (see Figure 5). This selection allowed us to group rotamers based
122 on the torsional angle values within the different peaks. As summarized in Table 1, 4 families were
123 found when both δ_{i-1} and δ_i torsional angles in the suite were used (see Table 2), 7 families for the $\alpha\gamma$
124 combination, 10 families for $\delta_{i-1}\delta_i\alpha$, and $\delta_{i-1}\delta_i\gamma$, and 22 families for $\delta_{i-1}\delta_i\alpha\gamma$. In order to evaluate
125 the classification performance of the RNA A–form helix conformations, the rotamers were also grouped
126 as: (i) A_noA families, where the 46 rotamers were separated in A–form helix (1a) vs no A–form helix
127 rotamers, and (ii) A*_noA* families, where the 46 rotamers were separated in rotamers related to A–form
128 helix (1a, 3d, 3b, 5d, 0a, 6b and 4b rotamers) vs the remaining rotamers.

Figure 5. Distribution plots for the six RNA backbone torsional angles α, β, γ, δ, ε and ζ in A), B), C), D), E) and F), respectively. Torsional angles values were obtained from the RNA09 database used in Laura Weston Murray (2007) "RNA Backbone Rotamers and Chiropraxis. Doctoral Dissertation; Dept. of Biochemistry; Duke University, 169 pages.

Classification

A series of machine learning methods were used to classify RNA suites as rotamers (or families of rotamers) based on their ribose ^{13}C' CS values. The following classification methods from the scikit-learn Python library (Pedregosa et al., 2011) were trained: K-Nearest Neighbors (NN), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM) and a class of neural network called Multi-Layer Perceptron (MLP). Different model parameters were tried out (see Supplementary Table S3). A random sampling algorithm was also used as a control, where suites were classified randomly. The sequence of the suite was considered for classification, because we found that the performance increased compared to a sequence–independent classification (see Supplementary Figure S10). The classification performance was assessed with four measures: weighted accuracy, precision, recall and F_1 score (van Rijsbergen, 1979). The weighted accuracy was used in order to recalibrate the contribution of the different rotamers, because the observed frequency of the rotamers is highly uneven (see Supplementary Figure S1).The weights used in the weighted accuracy were obtained from a substitution matrix (ROSUM, for ROtamers SUbstitution Matrix). The definition of the ROSUM matrix was inspired by the BLOSUM matrix used for protein sequence alignment (Henikoff and Henikoff, 1992). The matrix is used to weight the match or no match, between the true rotamer and the predicted rotamer, as a function of the euclidean distance between rotamers (in the seven-dimensional space of the suite backbone torsional angles) and the observed frequency of each rotamer. The torsional angles values and the observed frequencies are extracted from the rotamers table (Richardson et al., 2008). A ROSUM matrix was obtained for each of the rotamer families described in the previous section. Further details on the construction of the ROSUM matrices are provided in Supplementary Data Section 4. The precision and recall were used because they gave a general overview of the performance of the method. In particular, they allowed us to assess the fraction of

151 classified items that were correctly identified and the sensitivity of the method. The F_1 score was also
152 used as a performance measure because it is the harmonic mean of precision and recall and as such, it
153 gives more realistic measure of the classifier's performance.

154 ### *Experimental vs theoretical*
155 The classification models trained with theoretical data were used to classify the experimental suites.
156 The result of the theoretical calculations (described in a previous section) are theoretical NMR isotropic
157 shieldings (σ). The theoretical shieldings (σ_{comp}) must be subtracted from a reference shielding value
158 (σ_{ref}) to be transformed into theoretical CS (δ_{comp}) (see Equation 2) which can then be compared with
159 the experimental CS (δ_{exp}). A simple reference value of $\sigma_{ref} = 185.00$ ppm was used, which is very close
160 to the theoretical isotropic shielding for TMS ($\sigma_{TMS,th}$) (Vila and Scheraga, 2009), and it is consistent
161 with the reference value previously defined for proteins and glycans. Alternatively, a set of effective
162 references were obtained as a function of: (i) the nitrogenous base sequence, (ii) the combinations of
163 ribose puckering states in the four families of rotamers obtained from $\delta_{i-1}\delta_i$ torsional angles distributions,
164 (iii) the five carbon nuclei $^{13}C'$ CS mean values and (iv) a linear regression between theoretical and
165 experimental ribose $^{13}C'$ CS values for a set of suites (see Supplementary Table S2).

$$\delta_{comp} = \sigma_{ref} - \sigma_{comp} \tag{2}$$

166 ### *Theoretical vs theoretical*
167 The classification models trained with theoretical data were also used to classify the theoretical suites. In
168 this case, classification was assessed through a leave-one-out cross-validation (LOO-CV). In LOO-CV, the
169 dataset is split into a test set and training set in a one-folded manner, which means that at every iteration a
170 unique suite is taken apart from the dataset and the remaining suites are used for training. This process
171 continues until every suite from the theoretical dataset is evaluated.

172 ### *Experimental vs experimental*
173 A LOO-CV was also used to classify the experimental suites.

174 ## RESULTS AND DISCUSSION

175 For experimental vs theoretical classification (see Figure 6) the 46 rotamers can be classified by means of
176 backbone $^{13}C'$ CS with a maximal F_1 score of 0.34 (see Supplementary Table S5). When the 46 rotamers
177 are grouped in families based on their torsional angles distributions, the highest scores correspond to
178 the use of $\delta_{(i-1)}$ and $\delta_{(i)}$ torsional angles, where all the classifiers gave maximal scores above 0.65.
179 This result is in agreement with the fact that backbone $^{13}C'$ CS are highly sensitive to ribose puckering
180 states (Giessner-Prettre and Pullman, 1987), since the δ torsional angle keeps a direct relation with
181 the ribose puckering (Gelbin et al., 1996). The $\delta_{i-1}\delta_i\gamma$, $\delta_{i-1}\delta_i\alpha$, $\delta_{i-1}\delta_i\alpha\gamma$ and $\alpha\gamma$ families also show
182 improved scores over the classification of the 46 rotamers. The A*_noA* and A_noA families show low
183 classification scores relative to their random choice classification scores, which means that backbone
184 $^{13}C'$ CS cannot distinguish between A–form helix and no A–form helix rotamers. In general the use
185 of more complex classifier models such as Neural Networks, Support Vector Machine, Decision Tree
186 and Random Forest does not assure a better performance for the current task, thus the simpler Nearest
187 Neighbor model can be chosen for classification into RNA rotamers. In both the theoretical vs theoretical
188 and the experimental vs experimental classifications (see Figure 7 and 8, respectively), the performances
189 increase for every group of families, compared to the experimental vs theoretical classification. In the
190 theoretical vs theoretical classification the performance values are very close to 1.0 for $\delta_{i-1}\delta_i$ families
191 and A–form helix vs no A–form helix rotamers (A_noA). In the theoretical vs theoretical classification,
192 the performance value ranges are particularly narrow, except for MLP and SVM classifiers.

193 The high scores obtained for the theoretical vs theoretical classification indicates that $^{13}C'$ CS are in
194 fact very sensitive to changes of the torsional angles, the only variable we changed for the construction of
195 the theoretical dataset. At the same time the lower performance obtained in the experimental vs theoretical
196 classification, is signalling that the atomistic model used for the DFT computations is not good enough to
197 reproduce the experimental observations.

198 One reason the theoretical vs theoretical classification gives better results compared to both the
199 experimental vs experimental and the experimental vs theoretical classifications, could be that the

experimental database is very sparse and the theoretical dataset is instead dense, or in other words the coverage of the theoretical dataset is much better than the experimental one. To explore if this is in fact a reasonable explanation, we removed elements from the theoretical dataset to mimic the sparsity of the experimental dataset (see Supplementary Figure S13). We found that while the weighted accuracy decreased (on average 0.09 points) this is not enough to explain the lower performance of the experimental vs theoretical (on average 0.31 points lower) or experimental vs experimental (on average 0.16 points lower) classifications. In another experiment, noise on the order of the expected error (1.47 ppm) between experimental and theoretical ^{13}C′ CS for those rotamers correctly classified, was added to the theoretical ^{13}C′ CS and then a theoretical vs theoretical + noise classification was performed (see Supplementary Figure S14). Both tests reinforce the idea discussed in the previous paragraph, i.e we need a better model for the theoretical DFT computations. These experiments also provide indirect evidence indicating that the accuracy of the experimental vs experimental classification will be improved as more RNA conformations are deposited in databases giving another incentive to determine and deposit RNA structures and ^{13}C′ CS data.

CONCLUSION

In this work, we explored the use of RNA backbone ^{13}C′ CS to classify backbone conformations into rotamers and families of rotamers. In general, our study led us to the following conclusions: (1) the classification of the rotamer families defined by the δ torsional angles (see Table 2), which are directly related to the ribose puckering states, gives the best performances, in line with the results previously described by other authors; (2) classification of A-form helix and no A-form helix rotamers using ^{13}C′ CS is not better than a random classification; (3) the performance achieved using the simple Nearest-Neighbor method is on a par with more complex classifiers such as Neural Networks, Support Vector Machine, Decision Tree and Random Forest; (4) ^{13}C′ CS values are able to sense changes in torsional angles, but they are also affected by other factors, thus future DFT computations of RNA ^{13}C′ CS should use more complex models than the one used in this work; (5) experimental ^{13}C′ CS can be useful to identify RNA rotamers, if the rotamers are re-grouped in smaller families as the 46 rotamers seems to be a too fine description for accurate discrimination in terms of ^{13}C′ CS; (6) the usefulness of ^{13}C′ CS for rotamers identification should improve as more RNA structures and experimental ^{13}C′ CS become available.

ACKNOWLEDGEMENTS

We greatly appreciate Myriam Villegas for valuable discussions, comments and suggestions.

Competing Interests

The authors declare there are no competing interests.

Funding

This research was supported by grants from: Consejo Nacional de Investigaciones Científicas y Técnicas (Argentina) [PIP-0087 to Jorge Alberto Vila]; Agencia Nacional de Promoción Científica y Tecnológica (Argentina) [PICT-0556 to Jorge Alberto Vila, PICT-0767 to Jorge Alberto Vila, PICT-0218 to Osvaldo Antonio Martin].

REFERENCES

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The Protein Data Bank. *Nucleic acids research*, 28(1):235–42.

Brown, J. D., Summers, M. F., and Johnson, B. A. (2015). Prediction of hydrogen and carbon chemical shifts from RNA using database mining and support vector regression. *Journal of Biomolecular NMR*, 63(1):39–52.

Chesnut, D. B. and Moore, K. D. (1989). Locally dense basis sets for chemical shift calculations. *Journal of Computational Chemistry*, 10(5):648–659.

Eddy, S. R. (2001). Non–coding RNA genes and the modern RNA world. *Nature Reviews Genetics*, 2(12):919–929.

248 Frank, A. T., Law, S. M., Ahlstrom, L. S., and Brooks, C. L. (2015). Predicting protein backbone chemical
249 shifts from Cα coordinates: Extracting high resolution experimental observables from low resolution
250 models. *Journal of Chemical Theory and Computation*, 11(1):325–331.

251 Frank, A. T., Law, S. M., and Brooks, C. L. (2014). A Simple and Fast Approach for Predicting H-1
252 and C-13 Chemical Shifts: Toward Chemical Shift-Guided Simulations of RNA. *Journal of Physical*
253 *Chemistry B*, 118(42):12168–12175.

254 Frank, A. T., Stelzer, A. C., and Bae, S.-h. (2013). Prediction of RNA 1H and ^{13}C Chemical Shifts: A
255 Structure Based Approach. *J Phys Chem B*, 117(43):13497–13506.

256 Frisch, M. J., Trucks, G. W., Schlegel, H. B., Scuseria, G. E., Robb, M. A., Cheeseman, J. R., Montgomery
257 Jr., J. A., Vreven, T., Kudin, K. N., Burant, J. C., Millam, J. M., Iyengar, S. S., Tomasi, J., Barone, V.,
258 Mennucci, B., Cossi, M., Scalmani, G., Rega, N., Petersson, G. A., Nakatsuji, H., Hada, M., Ehara, M.,
259 Toyota, K., Fukuda, R., Hasegawa, J., Ishida, M., Nakajima, T., Honda, Y., Kitao, O., Nakai, H., Klene,
260 M., Li, X., Knox, J. E., Hratchian, H. P., Cross, J. B., Bakken, V., Adamo, C., Jaramillo, J., Gomperts,
261 R., Stratmann, R. E., Yazyev, O., Austin, A. J., Cammi, R., Pomelli, C., Ochterski, J. W., Ayala, P. Y.,
262 Morokuma, K., Voth, G. A., Salvador, P., Dannenberg, J. J., Zakrzewski, V. G., Dapprich, S., Daniels,
263 A. D., Strain, M. C., Farkas, O., Malick, D. K., Rabuck, A. D., Raghavachari, K., Foresman, J. B.,
264 Ortiz, J. V., Cui, Q., Baboul, A. G., Clifford, S., Cioslowski, J., Stefanov, B. B., Liu, G., Liashenko,
265 A., Piskorz, P., Komaromi, I., Martin, R. L., Fox, D. J., Keith, T., Al-Laham, M. A., Peng, C. Y.,
266 Nanayakkara, A., Challacombe, M., Gill, P. M. W., Johnson, B., Chen, W., Wong, M. W., Gonzalez, C.,
267 and Pople, J. A. (2004). Gaussian 03,Revision C.02.

268 Garay, P. G., Martin, O. A., Scheraga, H. A., and Vila, J. A. (2014). Factors affecting the computation of
269 the ^{13}C shielding in disaccharides. *Journal of Computational Chemistry*, 35(25):1854–1864.

270 Garay, P. G. and Vila, J. A.and Martin, O. A. (2018). CheSweet: An application to predict glycan's
271 chemicals shifts. *The Journal of Open Source Software*, 3(21):488.

272 Gelbin, A., Schneider, B., Clowney, L., Hsieh, S.-h., Olson, W. K., and Berman, H. M. (1996). Geometric
273 Parameters in Nucleic Acids: Sugar and Phosphate Constituents. *Journal of the American Chemical*
274 *Society*, 118(3):519–529.

275 Giessner-Prettre, C. and Pullman, B. (1987). Quantum mechanical calculations of NMR chemical shifts
276 in nucleic acids. *Quarterly Reviews of Biophysics*, 20(3-4):113.

277 Henikoff, S. and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings*
278 *of the National Academy of Sciences*, 89(22):10915–10919.

279 Icazatti, A. A., Martin, O. A., Villegas, M., Szleifer, I., and Vila, J. A. (2018). 13Check_RNA: a tool to
280 evaluate 13C chemical shift assignments of RNA. *Bioinformatics*, 34(23):4124–4126.

281 Lehninger A.L., N. D. L. and Cox, M. M. (2000). *Lehninger principles of biochemistry*. Worth Pub.

282 Martin, O. A., Arnautova, Y. A., Icazatti, A. A., Scheraga, H. A., and Vila, J. A. (2013). Physics-based
283 method to validate and repair flaws in protein structures. *Proceedings of the National Academy of*
284 *Sciences*, 110(42):16826–16831.

285 Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer,
286 P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and
287 Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning*
288 *Research*, 12:2825–2830.

289 Richardson, J. S., Schneider, B., Murray, L. W., Kapral, G. J., Immormino, R. M., Headd, J. J., Richardson,
290 D. C., Ham, D., Hershkovits, E., Williams, L. D., Keating, K. S., Pyle, A. M., Micallef, D., Westbrook,
291 J., and Berman, H. M. (2008). RNA backbone: Consensus all-angle conformers and modular string
292 nomenclature (an RNA Ontology Consortium contribution). *RNA*, 14(3):465–481.

293 Shen, Y. and Bax, A. (2010). SPARTA+: A modest improvement in empirical NMR chemical shift
294 prediction by means of an artificial neural network. *Journal of Biomolecular NMR*, 48(1):13–22.

295 Sripakdeevong, P., Cevec, M., Chang, A. T., Erat, M. C., Ziegeler, M., Zhao, Q., Fox, G. E., Gao, X.,
296 Kennedy, S. D., Kierzek, R., Nikonowicz, E. P., Schwalbe, H., Sigel, R. K. O., Turner, D. H., and Das,
297 R. (2014). Structure determination of noncanonical RNA motifs guided by 1H NMR chemical shifts.
298 *Nature methods*, 11(4):413–6.

299 Ulrich, E. L., Akutsu, H., Doreleijers, J. F., Harano, Y., Ioannidis, Y. E., Lin, J., Livny, M., Mading, S.,
300 Maziuk, D., Miller, Z., Nakatani, E., Schulte, C. F., Tolmie, D. E., Kent Wenger, R., Yao, H., and
301 Markley, J. L. (2008). BioMagResBank. *Nucleic Acids Research*, 36(SUPPL. 1):402–408.

302 van Rijsbergen, C. (1979). *Information Retrieval*. Butterworth-Heinemann.

303 Vila, J. A. and Scheraga, H. A. (2008). Factors affecting the use of ^{13}Cα chemical shifts to determine,
304   refine, and validate protein structures. *Proteins: Structure, Function, and Bioinformatics*, 71(2):641–
305   654.
306 Vila, J. A. and Scheraga, H. A. (2009). Assessing the accuracy of protein structures by quantum
307   mechanical computations of ^{13}Cα chemical shifts. *Accounts of chemical research*, 42(10):1545–53.
308 Wan, Y., Kertesz, M., Spitale, R. C., Segal, E., and Chang, H. Y. (2011). Understanding the transcriptome
309   through RNA structure. *Nature Reviews Genetics*, 12(9):641–655.

Table 1. Families of rotamers

46 rotamers	22 families $\delta_{i-1}\delta_i\alpha\gamma$	10 families $\delta_{i-1}\delta_i\alpha$	10 families $\delta_{i-1}\delta_i\gamma$	7 families $\alpha\gamma$	4 families $\delta_{i-1}\delta_i$	2 families A_noA[i]	2 families A*_noA*[ii]
&a	e	a	a	e	a	b	b
#a	q	c	c	e	c	b	b
0a	q	c	c	e	c	b	a
0b	t	d	d	e	d	b	b
0i	o	g	g	b	c	b	b
1[	l	b	b	e	b	b	b
1a	e	a	a	e	a	a	a
1b	l	b	b	e	b	b	b
1c	d	e	e	d	a	b	b
1e	f	e	e	f	a	b	b
1f	d	e	e	d	a	b	b
1g	c	a	a	c	a	b	b
1L	e	a	a	e	a	b	b
1m	e	a	a	e	a	b	b
1o	m	i	i	g	b	b	b
1t	k	f	f	d	b	b	b
1z	j	b	b	c	b	b	b
2[	t	d	d	e	d	b	b
2a	q	c	c	e	c	b	b
2h	r	g	g	f	c	b	b
2o	v	j	j	g	d	b	b
3a	e	a	a	e	a	b	b
3b	l	b	b	e	b	b	a
3d	a	a	a	a	a	b	a
4a	q	c	c	e	c	b	b
4b	t	d	d	e	d	b	a
4d	n	c	c	a	c	b	b
4g	p	c	c	c	c	b	b
4n	o	g	g	b	c	b	b
4p	s	d	d	a	d	b	b
4s	u	h	h	f	d	b	b
5d	a	a	a	a	a	b	a
5j	b	e	e	b	a	b	b
5q	h	f	f	b	b	b	b
5z	j	b	b	c	b	b	b
6d	n	c	c	a	c	b	a
6g	p	c	c	c	c	b	b
6j	o	g	g	b	c	b	b
6n	o	g	g	b	c	b	b
6p	s	d	d	a	d	b	b
7a	e	a	a	e	a	b	b
7d	a	a	a	a	a	b	b
7p	g	b	b	a	b	b	b
7r	i	i	i	c	b	b	b
8d	n	c	c	a	c	b	b
9a	e	a	a	e	a	b	b

The 46 RNA backbone rotamers were arranged in 22, 10, 10, 7 and 4 families of rotamers based on the observed distributions of $\delta_{i-1}\delta_i\alpha\gamma$, $\delta_{i-1}\delta_i\alpha$, $\delta_{i-1}\delta_i\gamma$, $\alpha\gamma$ and $\delta_{i-1}\delta_i$ torsional angles values, respectively. Additionally, the 46 rotamers were separated in RNA A–form helix vs no A–form helix rotamers in two ways: (i) RNA A-form helix rotamer 1a vs the remaining no A–form helix rotamers (A_noA families) and (ii) rotamers related to A–form helix (i.e. 1a, 3d, 3b, 5d, 0a, 6b, 4b) vs the remaining rotamers (A*_noA* families).

$\delta_{i-1}\delta_i$ families	46 rotamers	$\delta_{(i-1)}$	$\varepsilon_{(i-1)}$	$\zeta_{(i-1)}$	$\alpha_{(i)}$	$\beta_{(i)}$	$\gamma_{(i)}$	$\delta_{(i)}$
a	1a	81	212	289	295	174	54	81
b	1b	84	215	289	300	177	58	145
c	2a	145	260	289	288	193	53	84
d	2[	146	259	291	292	210	54	148

Table 2. Mean torsional angles values of the representative (i.e. most frequent) rotamers from the four $\delta_{i-1}\delta_i$ families. Values were extracted from the rotamer table of (Richardson et al., 2008).
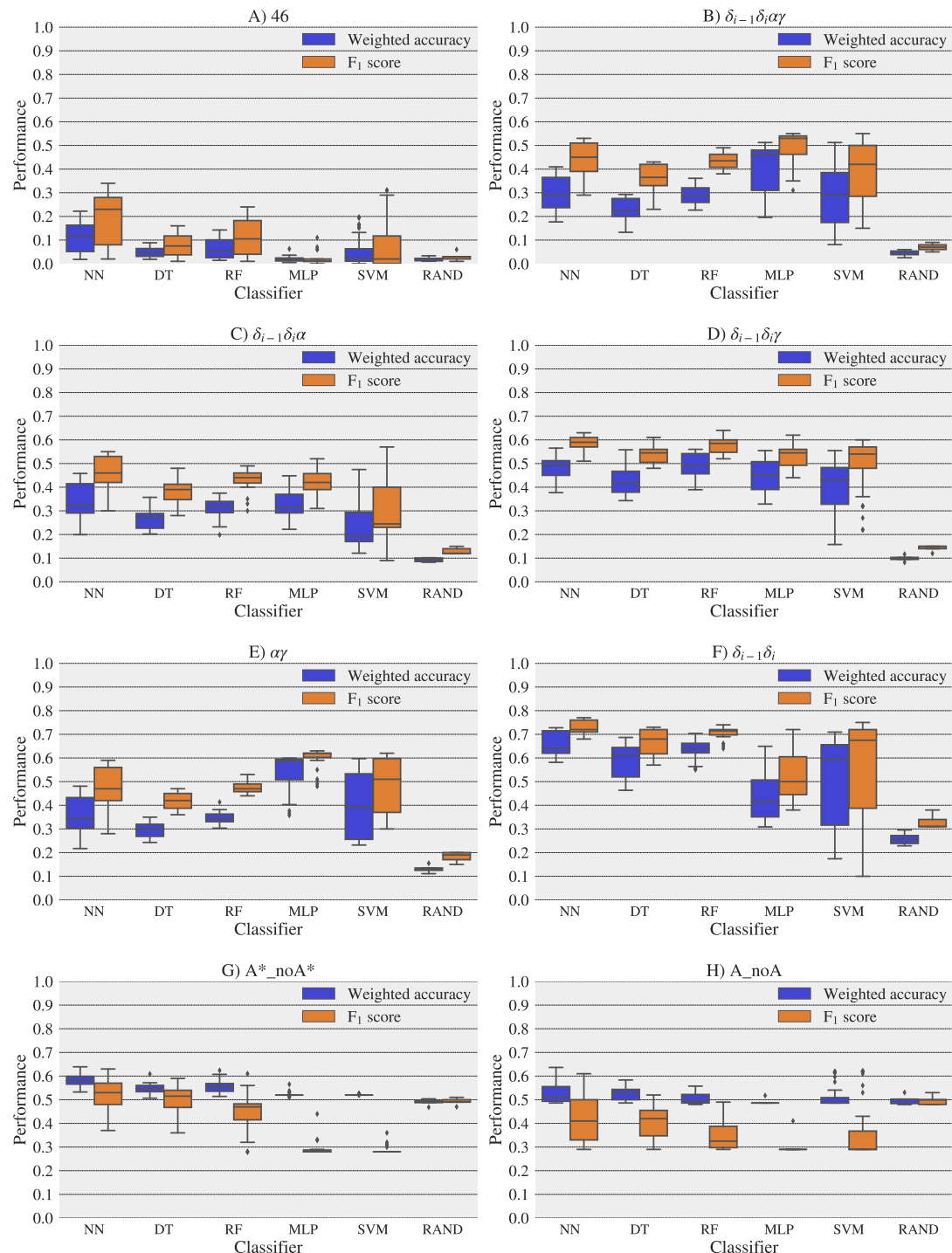
Figure 6. Box-plots with the weighted accuracy and F_1 score for the experimental vs theoretical classification of rotamers and families of rotamers, using Nearest Neighbor (NN), Decision Tree (DT), Random Forest (RF), Multi-Layer Perceptron (MLP) and Support Vector Machine (SVM) classifiers. A random-choice (RAND) algorithm was used as a baseline reference. Classification results for the 46, $\delta_{i-1}\delta_i\alpha\gamma$, $\delta_{i-1}\delta_i\alpha$, $\delta_{i-1}\delta_i\gamma$, $\alpha\gamma$ and $\delta_{i-1}\delta_i$, A_noA families and A*_noA* rotamer families are shown in A), B), C), D), E), F), G) and H) respectively. The highest values of weighted accuracy and F_1 score, for the experimental vs theoretical classification along with parameters of the classifiers are provided in Supplementary Tables S4 and S5. Precision and recall are shown in Supplementary Figure S12.
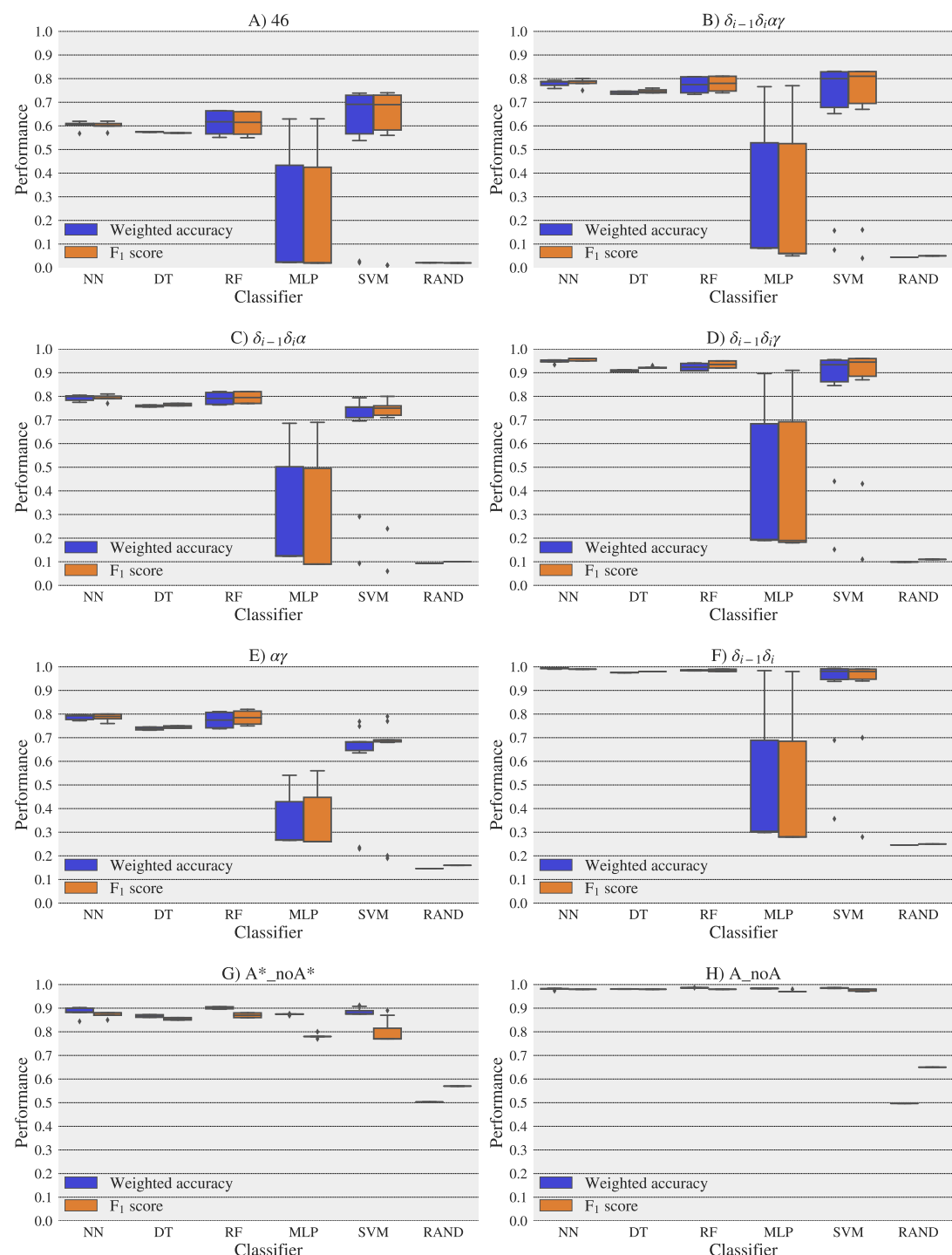
Figure 7. Box-plots with the weighted accuracy and F_1 score for the theoretical vs theoretical classification of rotamers and families of rotamers, using Nearest Neighbor (NN), Decision Tree (DT), Random Forest (RF), Multi-Layer Perceptron (MLP) and Support Vector Machine (SVM) classifiers. A random-choice (RAND) algorithm was used as a baseline reference. Classification results for the 46, $\delta_{i-1}\delta_i\alpha\gamma$, $\delta_{i-1}\delta_i\alpha$, $\delta_{i-1}\delta_i\gamma$, $\alpha\gamma$ and $\delta_{i-1}\delta_i$, A_noA families and A*_noA* rotamer families are shown in A), B), C), D), E), F), G) and H) respectively.
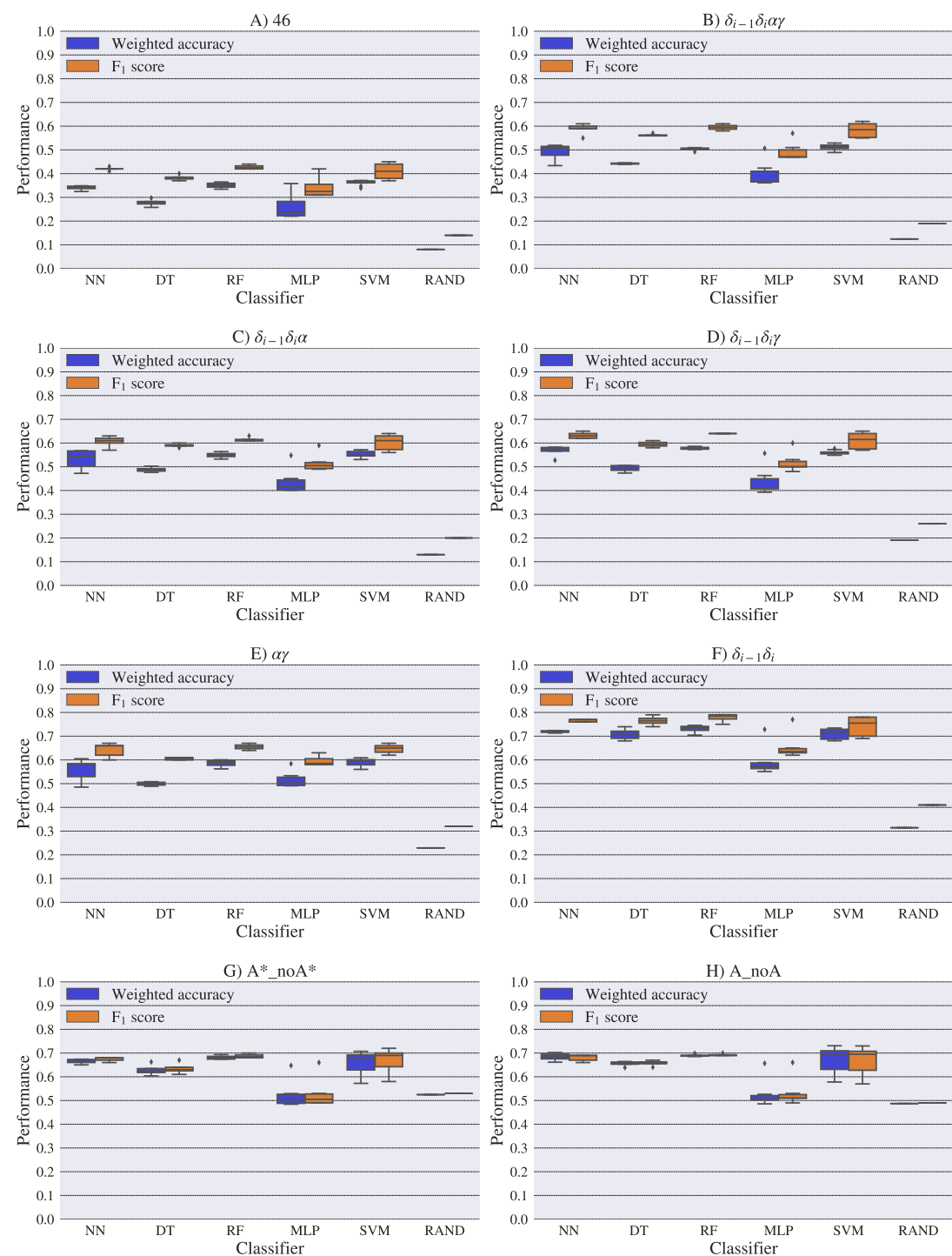
Figure 8. Box-plots with the weighted accuracy and F_1 score for the experimental vs experimental classification of rotamers and families of rotamers, using Nearest Neighbor (NN), Decision Tree (DT), Random Forest (RF), Multi-Layer Perceptron (MLP) and Support Vector Machine (SVM) classifiers. A random-choice (RAND) algorithm was used as a baseline reference. Classification results for the 46, $\delta_{i-1}\delta_i\alpha\gamma$, $\delta_{i-1}\delta_i\alpha$, $\delta_{i-1}\delta_i\gamma$, $\alpha\gamma$ and $\delta_{i-1}\delta_i$, A_noA families and A*_noA* rotamer families are shown in A), B), C), D), E), F), G) and H) respectively.