Received: 1 December 2020 | Revised: 24 February 2021 | Accepted: 4 April 2021

DOI: 10.1002/jcc.26556

FULL PAPER

Journal of COMPUTATIONAL CHEMISTRY WILEY

Exploring the quality of protein structural models from a Bayesian perspective

Agustina Arroyuelo | Jorge A. Vila | Osvaldo A. Martin

Instituto de Matemática Aplicada San Luis, CONICET-UNSL, San Luis, Argentina

Correspondence
Osvaldo A. Martin, Instituto de Matemática Aplicada San Luis, CONICET-UNSL, San Luis, Argentina.
Email: omarti@unsl.edu.ar

Funding information
Agencia Nacional de Promoción Científica y Tecnológica, Grant/Award Numbers: PICT-02212, PICT-0767; Consejo Nacional de Investigaciones Científicas y Técnicas, Grant/Award Number: PIP-0087

Abstract

We explore how ideas and practices common in Bayesian modeling can be applied to help assess the quality of 3D protein structural models. The basic premise of our approach is that the evaluation of a Bayesian statistical model's fit may reveal aspects of the quality of a structure when the fitted data is related to protein structural properties. Therefore, we fit a Bayesian hierarchical linear regression model to experimental and theoretical $^{13}C^{\alpha}$ chemical shifts. Then, we propose two complementary approaches for the evaluation of such fitting: (a) in terms of the *expected differences* between experimental and posterior predicted values; (b) in terms of the *leave-one-out cross-validation point-wise predictive accuracy*. Finally, we present visualizations that can help interpret these evaluations. The analyses presented in this article are aimed to aid in detecting problematic residues in protein structures. The code developed for this work is available on: https://github.com/BIOS-IMASL/Hierarchical-Bayes-NMR-Validation.

KEYWORDS
$^{13}C^{\alpha}$ chemical shifts, Bayesian hierarchical models, NMR protein structure validation

1 | INTRODUCTION

Bayesian statistics offers very suitable theoretical advantages for developing models involving bio-molecular structural data and has been applied in numerous tools and methods in this context.[1–6] Furthermore, Bayesian methods are capable of accounting for errors and noise of variable source and nature, which is suitable for working with bio-molecular experimental data.[7]

In statistics, partially pooling data means to separate observations into groups, while allowing the groups to remain somehow linked in order to *influence each other*. In a Bayesian setting, such sharing is achieved naturally through hierarchical modeling. In hierarchical models (also called multilevel models), parameters of the *prior* distributions are shared among groups, inducing dependencies and allowing them to effectively *share information*.[8–10] Advantages of hierarchical Bayesian modeling include obtaining model parameter estimates for each group as well as for the total population. In addition, using shared *prior* distributions helps prevent the models from over-fitting.[11]

As the word *model* is used in both Bayesian statistics and protein science, throughout this article we deliberately use the word *model* to discuss statistical models and *structure* to discuss protein 3D models, thus avoiding potential confusion.

In this study, we fit Bayesian models to experimental $^{13}C^{\alpha}$ chemical shifts. We focus specifically on $^{13}C^{\alpha}$ chemical shifts because they have proven to be informative on protein structure at the residue level and have been used in determination and validation in previous work by us and others.[12–15] Another convenient aspect of working with $^{13}C^{\alpha}$ chemical shifts is that their theoretical values can be computed with high accuracy using quantum-chemical methods.[13]

Bayesian model comparison and evaluation is standard in Bayesian applications as it constitutes an essential part of the Bayesian workflow.[16] A variety of methods have been proposed for this task, helping Bayesian practitioners evaluate, critique, and ultimately understand their models. In the present work, we propose two complementary approaches for the evaluation of protein structures. Both are related to different ways of analyzing the results of a Bayesian hierarchical linear model linking experimental and theoretical $^{13}C^{\alpha}$ chemical shifts. For the first approach, we compare structures in terms of their residuals (i.e., the difference between the observed and predicted values). To ease the comparison, we put the residuals in the context of

reference densities which are pre-computed from a data set of high-quality protein structures (see Methods and Software section for details). In the second approach, we evaluate the statistical model's fit in terms of its out-of-sample predictive accuracy, that is, the predicted accuracy computed from data not used to train the model. The out-of-sample predictive accuracy can be estimated using leave-one-out cross-validation, which requires to re-fit a model *n* times, with *n* being the size of the data set (i.e., the number of $^{13}C^{\alpha}$ chemical shifts). As this can be too costly and cumbersome, in this work we use an alternative; the Pareto smoothed importance sampling leave-one-out cross-validation (LOO for short).[17,18] LOO offers an accurate, reliable, and fast estimate of the out-of-sample prediction accuracy from a single model fit. Additionally, the predictive accuracy is computed per observation, this is equivalent to computing the predictive accuracy per residue, as the observations are $^{13}C^{\alpha}$ chemical shifts. This allows us to make statements of the quality of the structure at both global and per residue level.

We expect the methods and visualizations presented here to introduce Bayesian model checking tools to protein scientists. In addition, we hope these methods are adopted by protein scientists to help them evaluate the quality of a given structure. This may include the structure determination process before structure deposition at the Protein Data Bank (PDB), ideally as part of the PDB's validation pipeline.[19] Or even after deposition, such as evaluating a structures quality before further research like, for example, performing docking or template-based modeling. For this, the code developed for this analysis (see Abstract) can take as input **.pdb* or BMRB files before deposition in these databases, as long as they have a correct format in the case of **.pdb* files. For BMRB files, NMR-Star or column-separated format files can be taken as input.

2 | METHODS AND SOFTWARE

2.1 | Reference data set

A reference data set of 111 high-quality protein structures was obtained from the Protein Data Bank. Each structure in this set has a resolution $\leqslant$2.0 Å and R-factors $\leqslant$0.25. The structures were solved in the absence of DNA, RNA, or glycan molecules. Additionally, every structure in our set has a corresponding entry at the Biological Magnetic Resonance Bank (BMRB) from which experimental $^{13}C^{\alpha}$ chemical shift data were obtained.[20] Theoretical $^{13}C^{\alpha}$ chemical shift data were computed from the Cartesian coordinates of each structure in this set, using *Che*Shift-2.[15] The residues with the largest absolute differences between theoretical and experimental $^{13}C^{\alpha}$ chemical shifts, representing 1% of the total data set size, were removed from the analysis. These differences were larger than 3.77 ppm.

2.2 | Target structures

Theoretical $^{13}C^{\alpha}$ chemical shift data were obtained for two structures of protein Ubiquitin under PDB ids: 1UBQ and 1D3Z.[21,22] Code id.: 1UBQ corresponds to an X-ray crystallography determined structure of Ubiquitin, while 1D3Z corresponds to an NMR determined structure. The latter contains 10 different conformations, corresponding to the structures with the lowest energy from 54 computed conformations.[21,22] We computed and averaged the theoretical $^{13}C^{\alpha}$ chemical shifts for those 10 conformations. Additionally, the experimental $^{13}C^{\alpha}$ chemical shift set used in this analysis was taken from BMRB entry No: 6457, and is the same experimental data set used in the NMR determination of 1D3Z. In this study, the same experimental $^{13}C^{\alpha}$ chemical shift set was used for both structures, but the theoretical $^{13}C^{\alpha}$ chemical shift set is different for each structure, given that it was obtained from the Cartesian coordinates of the PDB entries 1UBQ and 1D3Z. This particular data set construction for the target structures allows us to compare 1D3Z and 1UBQ based solely on the differences between the 3D coordinates of their structures.

In order to further strengthen the demonstrations presented in this article, theoretical and experimental $^{13}C^{\alpha}$ data were obtained for an obsolete target structure under PDB id.: 1WDB, associated with BMRB No 5745.

2.3 | Hierarchical linear model

A Bayesian hierarchical linear regression model was fitted to the experimental (CSe) and theoretical (CSt) $^{13}C^{\alpha}$ chemical shifts contained in the reference data set. The full model is described by Expression 1 using standard statistical notation and it is also represented in Figure 1 in Kruschke's diagrams.[23]

The model groups the data into *r* groups, in total there are 19 groups, one for each amino acid with Cysteine being excluded given that *Che*Shift-2 does not offer reliable calculations for this amino acid.[15] For each group we fit a linear regression by finding the parameters α_r and β_r. These parameters are partially pooled, meaning that they are not free to vary but instead they are restricted by the parameters α_σ and β_σ respectively. We assume that the experimental $^{13}C^{\alpha}$ chemical shifts are conditionally Gaussian with the mean being a linear function of the theoretical $^{13}C^{\alpha}$ chemical shifts, and unknown *SD*, which we also assume to take a different, but partially pooled value σ_r.

The data was normalized before fitting by subtracting the empirical mean and dividing by the empirical *SD*.

$$\begin{aligned}
\alpha_\sigma &\sim \mathcal{HN}(1)\\
\beta_\sigma &\sim \mathcal{HN}(1)\\
\sigma_\sigma &\sim \mathcal{HN}(1)\\
\alpha_r &\sim \mathcal{N}(0,\alpha_\sigma)\\
\beta_r &\sim \mathcal{HN}(\beta_\sigma)\\
\sigma_r &\sim \mathcal{HN}(\sigma_\sigma)\\
\mu_r &= \alpha_r + \beta_r \mathrm{CSt_r}\\
\mathrm{CSe_r} &\sim \mathcal{N}(\mu_r,\sigma_r)
\end{aligned} \tag{1}$$

where $\mathcal{HN}(1)$ stands for half-normal distribution with *SD* 1. $\mathcal{N}(0,X_\sigma)$ is a normal distribution with mean 0 and *SD* X_σ. $\mathrm{CSt_r}$ represents

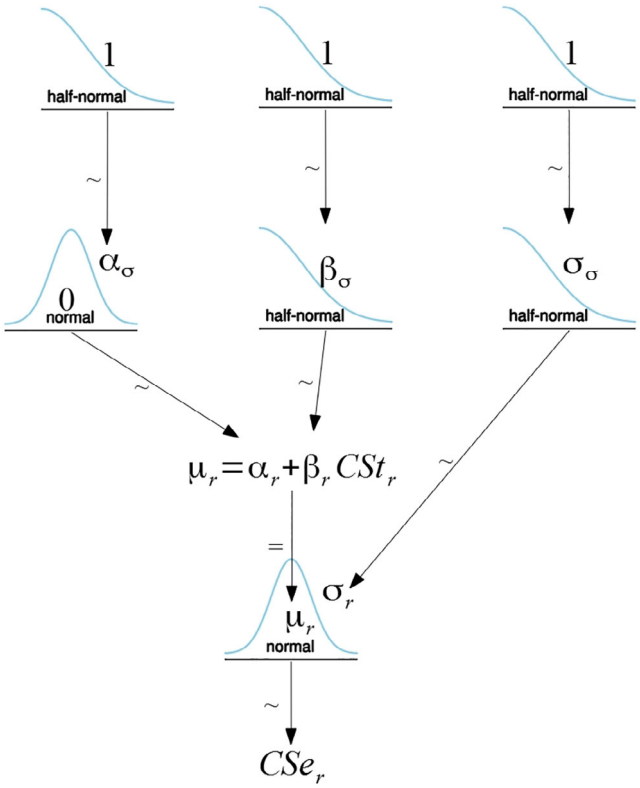

FIGURE 1 Krushke diagram representing the hierarchical linear model featured in this work

theoretical $^{13}C^{\alpha}$ chemical shifts and CSe_r the experimental ones. The sub index *r* denotes each of the at most 19 groups present in a given structure.

2.3.1 | Hierarchical linear model for target structures

For each one of the target structures, that is, the structures we want to validate, their theoretical and experimental $^{13}C^{\alpha}$ data were included as part of the reference data set to then, run Model 1. This is done in order to use the data from the reference structures to help regularize the inference for the target structure. For subsequent analysis, only the observations corresponding to the target structures were considered.

2.4 | Computation of the posterior predictive distribution of $^{13}C^{\alpha}$ chemical shifts and reference densities

From the fitted hierarchical model, the posterior predictive distribution was computed, that is the distribution of $^{13}C^{\alpha}$ chemical shifts as predicted by the statistical model. We will refer to this set as corrected $^{13}C^{\alpha}$ chemical shifts. Then, the reference densities were computed as the difference between the corrected and experimental $^{13}C^{\alpha}$ chemical shifts from the reference data set for each of the 19 most common amino acids present in proteins (with Cysteine excluded as previously explained). Intuitively, the reference densities are an approximation to the expected distribution of the difference between experimental and corrected $^{13}C^{\alpha}$ chemical shifts. The difference between corrected and experimental $^{13}C^{\alpha}$ chemical shifts was also computed for the target structures, where the corrected set was defined from the posterior predictive distribution of the model fitted to the structure's $^{13}C^{\alpha}$ chemical shift data (see Section *Hierarchical linear model for Target structures*).

2.5 | Model comparison and cross-validation

When faced with more than one model for the same data it is natural to ask which model is the best at explaining the data, and more broadly, how are models different from each other and what they have in common. One way to assess a model is through its predictions. In order to do so, we can compare a model's predictions to experimental data. If we use the same experimental data used to fit the model, that is, we compute the within-sample error, we may become overconfident in our model. The simplest solution is to compute the out-of-sample error, that is, the error that a model makes when evaluated on data not used to fit it. Unfortunately, leaving a portion of the data aside just for validation is most often than not a very expensive luxury (e.g., during NMR structure determination).

The log predictive density has an important role in model comparison because of its connection to the Kullback–Leibler divergence, a measure of closeness between two probability distributions.[11] For historical reasons, measures of predictive accuracy are referred to as information criteria and they are a collection of diverse methods that allow to estimate the out-of-sample error without requiring external data. In a Bayesian context, one such measure is LOO.[17,18,24] In the next subsections, we will briefly explain some of the details related to LOO, specifically those more relevant for this study.

2.5.1 | LOO

The cross-validated leave-one-out predictive distribution $p(y_i|y_{-i})$ (or most commonly its logarithm) can be used to assess the out-of-sample prediction accuracy. In the present work this means the probability, according to the model, of observing the *i*th $^{13}C^{\alpha}$ chemical shift when that $^{13}C^{\alpha}$ chemical shift is not included in the fitting.

Computing $p(y_i|y_{-i})$ can become costly as it requires to fit a model *n* times (where *n* is the data set's size). Fortunately, the leave-one-out predictive distribution can be approximated by using importance weights. The variance of these importance weights can be large or even infinite, LOO applies a smoothing procedure that involves replacing the largest importance weights with values from an estimated Pareto distribution. For details on how this is done and why it

works see Vehtari et al. (2017).[17] What is most important for our current discussion is that the $\hat{\kappa}$ parameter of such a Pareto distribution can be used to detect highly influential observations, that is, observations that have a large effect on the predictive distribution when they are left out. In general, higher values of $\hat{\kappa}$ can indicate problems with the data or model, especially when $\hat{\kappa} > 0.7$.[18,25]

2.5.2 | LOO-PIT

PIT (Probability Integral Transform), known as the universality of the uniform distribution, states that given a random variable with an arbitrary continuous distribution, it is possible to create a uniform distribution in the interval [0, 1]. Specifically, given a continuous random variable X for which the cumulative distribution function is F_X, then $F_X(X) \tilde{\mathcal{U}}(0,1)$.

LOO-PIT is obtained by comparing the observed data y to posterior predicted data $\tilde{y}$. The comparison is done point-wise. Given:

$$p_i = P(\tilde{y}_i \leq y_i \mid y_{-i})$$

LOO-PIT is computing the point-wise probability that the posterior predicted data $\tilde{y}_i$ has a lower value than the observed data y_i. For a well-calibrated model the expected distribution of p is the uniform distribution over the [0, 1] interval, that is, a standard uniform distribution (see Figure 2). Deviations from uniformity indicate different mismatches between the data and the predictions made by the model, see Figure 2 for a few idealized examples. LOO-PIT density. An important advantage of using the leave-one-out predictive distribution instead of just the predictive distribution is that with the former, we are not using the data twice (once to fit the model and once to validate it).[11,25]

2.5.3 | Expected log predictive density

Finally, we can compare models point-wise using their expected log predictive density (ELPD), where the expectation is taken over the whole posterior. In other words, the predictions take into account the parameter's uncertainty, as expressed by the statistical model and data. Notice that the value of the ELPD is not useful by itself as it cannot be interpreted in absolute terms, but can be used to compare the relative fit of residues within a same structure and/or to compare the relative fit of residues from two or more structures, as long as the models are fitted to the same data. While the values of $\hat{\kappa}$ indicate how influential an observation is, that is, how much the predictive distribution changes when they are left out, the ELPD indicates how difficult it is for the model to predict a particular observation. In this sense, large absolute ELPD differences indicate disparities in how two models fit the same observation. Thus, if two models are exactly the same, the ELPD difference will be zero. Although, there is no hard-threshold for analyzing ELPD differences, values lower than ±4 can be considered small.

2.6 | Reference data set model fit, B-factors, and structural dissimilarity

The quality of the described model fit to the reference data set was assessed in order to ensure the reliability of the analysis presented in this article for the target structures. This was done through the computation of the expected log predictive density, $\hat{\kappa}$ parameter values, and LOO-PIT.

Correlation with the C_α B-factors of the reference and target structures was also investigated for the ELPD and $\hat{\kappa}$ values.

In order to search for a correlation between the ELPD values and the structural dissimilarities between the 1D3Z and 1UBQ target

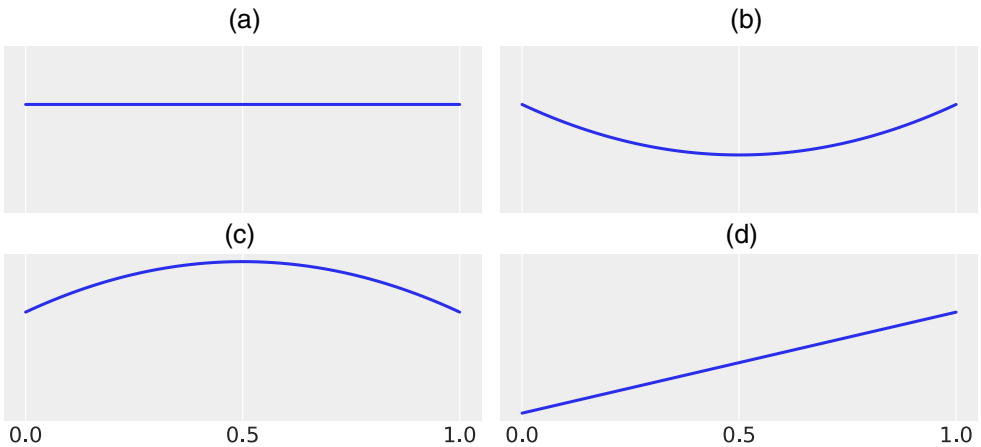

FIGURE 2 Schematic representation of a few possible LOO-PIT curves. On panel (A), the uniform distribution indicates that the model is well-calibrated. On panel (B) there are more observations for low and high values (and less in the middle) compared with what the model is predicting. In other words, the predictive distribution is narrower than the observed one. On panel (C) there are more observations around intermediate values (and less for low and high values) compared with what the model is predicting, in other words the predicted distribution is broader than the observed one. Finally, on panel (D), there are more observations on the right tail than in the left tail compared with what the model is predicting, that is, the model is biased toward predicting lower values than observed

structures, the C_α RMSD (root-mean-squared deviation) was computed between the structures. For 1D3Z, the RMSD was computed against 1UBQ for each of the 10 conformations and then averaged.

2.7 | Software

All Bayesian models were solved with PyMC3.[26] ArviZ was used to compute LOO, ELPD, and related plots.[27] PyMOL was used to visualize 3D protein structures.[28]

3 | RESULTS

In this work, we explore the quality of protein structures using $^{13}C^{\alpha}$ chemical shifts through the evaluation of a Bayesian hierarchical linear model's fit. An important aspect of fitting this particular model is the estimation of the effective reference value for the $^{13}C^{\alpha}$ chemical shifts. This is important as wrong referencing can be an issue when working with chemical shifts. In our study, the estimated reference value is unique for every protein in our data set. Moreover, by using a hierarchical model, we obtained a correction specifically for every one of the 19 most common amino acids that constitute proteins (as already mentioned Cysteine is excluded). Said effective reference is accounted for in every posterior analysis made on the model's fit. Moreover, the eventual circumstance of wrong referencing of $^{13}C^{\alpha}$ chemical shifts in our reference data set is automatically fixed through the linear model, that is, if there is a systematic error in the data set, Model 1 accounts for it. Issues related to misassignments of $^{13}C^{\alpha}$ chemical shifts are still problematic. Nevertheless, they are uncommon so, on average, over the 111 reference structures, misassignments should be rare. Additionally, we removed the 1% residues with the largest differences which is a simple way to account for gross errors from different sources, including misassigments.

We present two approaches for the Bayesian model's fit evaluation. The first approach analyses differences between corrected and experimental $^{13}C^{\alpha}$ chemical shifts. This is highly appealing as the comparison is done using a familiar metric for protein scientists, especially NMR spectroscopists. The second approach instead evaluates the model's fit using the LOO predictive distribution, which is a general and widely accepted way to assess Bayesian statistical models.[11,17,18,25] Both methods complement each other, the first one focuses on how well the corrected $^{13}C^{\alpha}$ chemical shift agrees with the expected distribution while the second approach is based on how well the model predicts the data. The combined usage of both approaches can help spectroscopists and protein scientists in general to flag problematic residues that may deserve further attention.

3.1 | Hierarchical linear model fit to the reference data set

Figure S1 in SM shows that most of the ELPD values for the reference data set take values between 0 and −4 and that they exhibit no correlation with the C_α B-factors of the reference structures. Similar conclusions can be drawn from Figures S2 and S3 in SM for 1D3Z and 1UBQ.

In Figure S4 in SM we observe that the hierarchical model used in this work is well calibrated for the reference data set as the LOO-PIT density is similar to panel (A) in Figure 2. Furthermore, Table S1 in SM shows parameter estimates for the reference data set. Effective Sample Size (ESS) and $\hat{R}$ diagnostics are also shown as proof of convergence of the hierarchical Bayesian model.

3.2 | Correlation between B-factors with ELPD and $\hat{\kappa}$ values

Figure S5 in SM shows no correlation between the $\hat{\kappa}$ values and the C_α B-factors for the target structures. Figure S6 in SM shows the RMSD between 1D3Z and 1UBQ versus the absolute ELPD difference, also showing no correlation between these values.

3.3 | Difference between experimental and corrected $^{13}C^{\alpha}$ chemical shifts

Using reference densities on the differences between corrected and experimental $^{13}C^{\alpha}$ chemical shifts obtained from high-quality structures can help contextualize particular differences found in the $^{13}C^{\alpha}$ chemical shifts of any given NMR structure. This results in a straightforward way to assess how problematic is each residue in the structure.

The reference densities for the amino acid types present in Ubiquitin are plotted in light blue in Figure 3. As expected, these distributions have zero-mean but most importantly they have different variances. Even when we did not perform any formal test to evaluate if and how these densities depart from a Gaussian distribution we can see that most of the reference densities are skewed or have more than one peak (most likely reflecting sub-populations of $^{13}C^{\alpha}$ chemical shift differences corresponding to α-helix and β-strands). These distributions also reflect variations among amino acids related to natural abundance as well as chemical features. For example Glycine, which is the most abundant amino acid and the one spanning broader regions on the Ramachandran map, has the smoothest curve. Given all these particularities, comparing differences between observed and corrected $^{13}C^{\alpha}$ chemical shifts in terms of a single common variance can be misleading. Instead, we use quantiles computed per each amino acid's distribution. Specifically, we used the 0.05, 0.2, 0.8, and 0.95 quantiles. Thus, we divide the reference densities into a central 60% (between the 0.2 and 0.8 quantiles) a 30% (15% between 0.05 and 0.2 and another 15% between 0.8 and 0.95) and finally the remaining 10% for values below 0.05 and above 0.95. In Figure 3 we can also see the differences for every residue in structures 1D3Z and 1UBQ represented as circles and crosses below each reference density. We use color to help interpret such differences, green if the $^{13}C^{\alpha}$ chemical shifts difference is found in the central 60%, yellow if it is

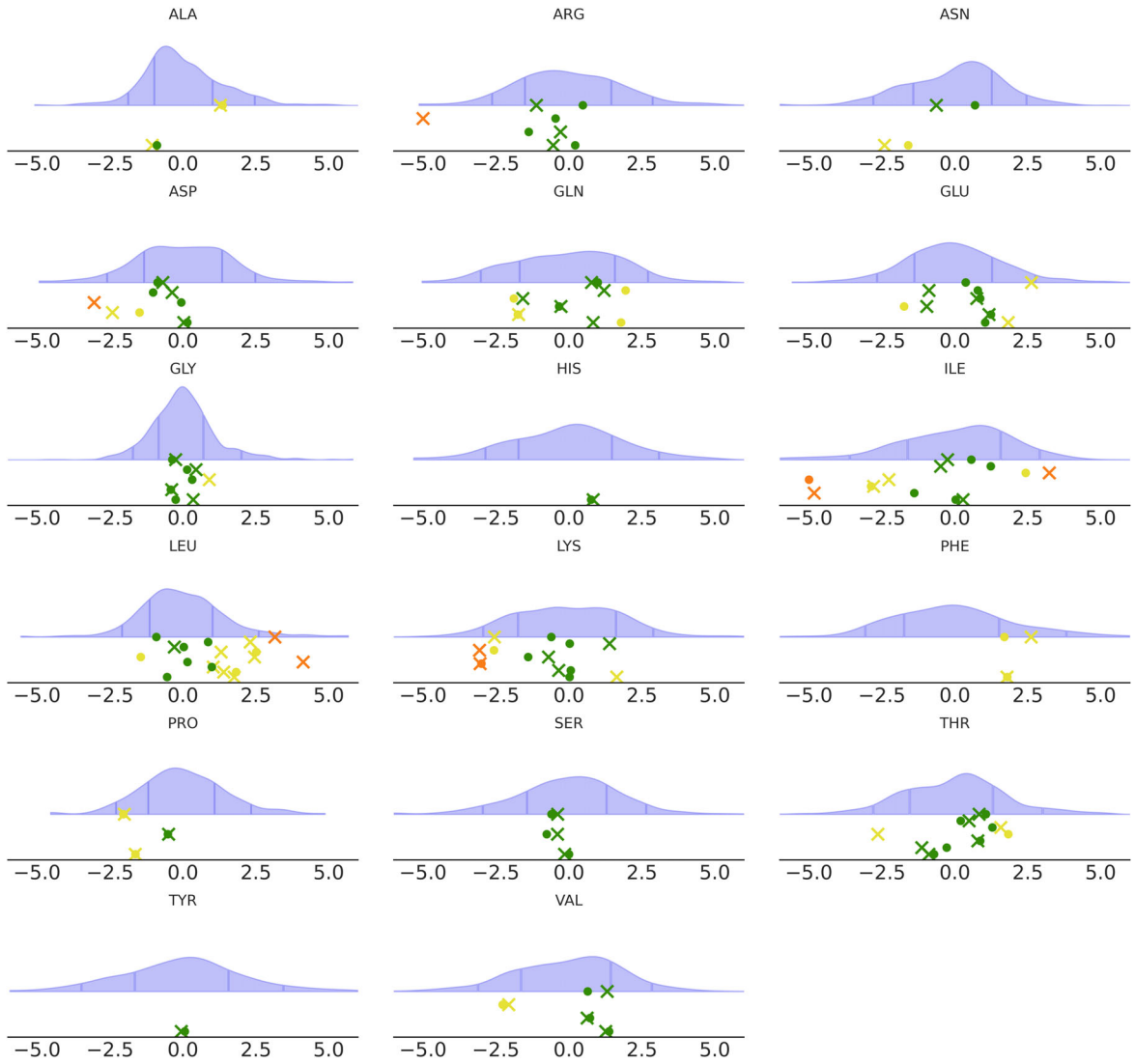

FIGURE 3 Differences between experimental $^{13}C^{\alpha}$ chemical shifts and corrected $^{13}C^{\alpha}$ chemical shifts, in ppm units. The reference densities in light blue are used as a representation of the expected differences between experimental and corrected $^{13}C^{\alpha}$ chemical shifts in high-quality NMR resolved structures. They are divided into a central 60% (between the 0.2 and 0.8 quantiles) a 30% (15% between the 0.05 and 0.2 quantiles and another 15% between 0.8 and 0.95 quantiles) and finally the remaining 10% for those values below and above 0.05 and 0.95 quantiles. The markers are displayed in color orange when found in the 10% most extreme values, yellow if they are placed in the 30% around the central values and green for the central 60%. Circles and crosses represent the $^{13}C^{\alpha}$ chemical shifts differences for structures 1D3Z and 1UBQ, respectively

found in the 30% around the central values, and orange if it is found in the 10% most extreme values. It is important to note that even when a residue is marked orange that does not automatically indicate it is a poorly determined residue, as in fact we expect that 10% of the residues from good quality structure to appear orange using the presented method. Instead, we consider them as residues that may deserve further attention.[15]

As we can see the differences in general are small between the two target structures, with a few exceptions such as Isoleucine 30 and Lysine 27 for 1D3Z and Isolecuine 13, Lysine 27, Lysine 33, Isoleucine 36, Aspartic acid 39, Leucine 50, Arginine 72, and Leucine 73 for 1UBQ. Figure 4 uses the same color-schema from Figure 3 in the context of 3D structures. The accompanying code at the repository (see Abstract) can automatically generate a file containing the colored structures as in Figure 4 and can be loaded with PyMOL or VMD.[29]

3.4 | LOO predictive distribution

As previously mentioned the LOO predictive distribution is a general way to assess Bayesian statistical models and is not related to protein structures in any direct way.[11,17,18,25] For a calibrated model, that is, a model which predictive distribution is in good agreement with the observed data, the distribution of the LOO probability integral transform (LOO-PIT) is uniform. As this only holds asymptotically, a way to

empirically assess calibration for a finite sample is to compare the density of LOO-PIT against the density of uniform samples of the same size as the data used to compute LOO-PIT. Such comparison is done in Figure 5 for structures 1D3Z and 1UBQ. We can see that both models seem to be overall well-calibrated. 1UBQ seems to be slightly worse, but that difference is within the expected margins and thus we must conclude both structures are on par according to this diagnostic.

3.5 | Analysis of the $\hat{\kappa}$ parameter

The parameter $\hat{\kappa}$ of the Pareto distribution used in the computation of LOO can help spot influential observations, that is, observations that have a large effect on predictions if left out from the analysis. The higher this value, the more influential the observation is, with values above 0.7 being of particular interest (see subsection LOO in Methods and Software section). Figure 6 shows $\hat{\kappa}$ values for 1D3Z and 1UBQ. In both structures, no residue exceeds the value of 0.7.

3.6 | Expected log predictive density

Figure 7 shows the differences of ELPD between structures 1D3Z and 1UBQ. Globally, these structures seem to be on par, except for Isoleucine 30 that is showing a better agreement for 1UBQ. Residues Isoleucine 13, Leucine 50, and Arginine 72 show better agreement in 1D3Z. It is worth noting that Figure 7 exhibits a correspondence with Figures 3 and 4. Residues highlighted in orange for 1UBQ in Figures 3 and 4 are located in the positive region of Figure 7. While 1D3Z's Isoleucine 30, located in the tail of the reference distribution in Figure 3, has a negative ELPD difference value in Figure 7. Interestingly, residue Lysine 27 flagged by the expected $^{13}C^{\alpha}$ chemical shifts differences analysis for both 1D3Z and 1UBQ, appears very near zero on Figure 7 indicating that both structures are equal at resolving this observation. The violin plot in Figure 7 shows that most of the ELPD differences are positive, indicating that 1D3Z seems to be a better structure than 1UBQ.

3.7 | Summary

Figures 3, 4, and 7 suggest that 1D3Z is a better structure than 1UBQ. This confirms the statements made in previous work, where the authors even indicated that structure 1UBQ can be improved by computing an ensemble of conformations.[30] The analysis of LOO-PIT in Figure 5 showed that both structures quality is on par, with 1UBQ having a slightly worse LOO-PIT. The analysis of $\hat{\kappa}$ in Figure 6 shows that neither 1D3Z nor 1UBQ have influential observations, as expected for structures of such good quality. In contrast, Figures S7 to S10 in SM, show what to expect of the outcomes of the presented methods for a target structure of poor quality as the obsolete structure 1WDB.

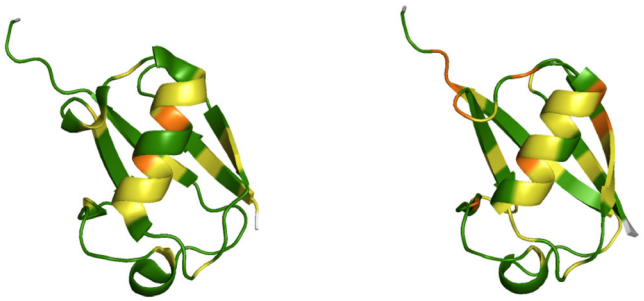

FIGURE 4 Differences between experimental $^{13}C^{\alpha}$ chemical shifts and corrected $^{13}C^{\alpha}$ chemical shifts superimposed in 3D structures. 1D3Z is shown on the left and 1UBQ is shown on the right. The amino acid residues are colored using the same criteria used in Figure 3. That is, orange when the $^{13}C^{\alpha}$ chemical shifts differences are found in the 10% most extreme values, yellow if they are placed in the 30% around the central values and green for the central 60%

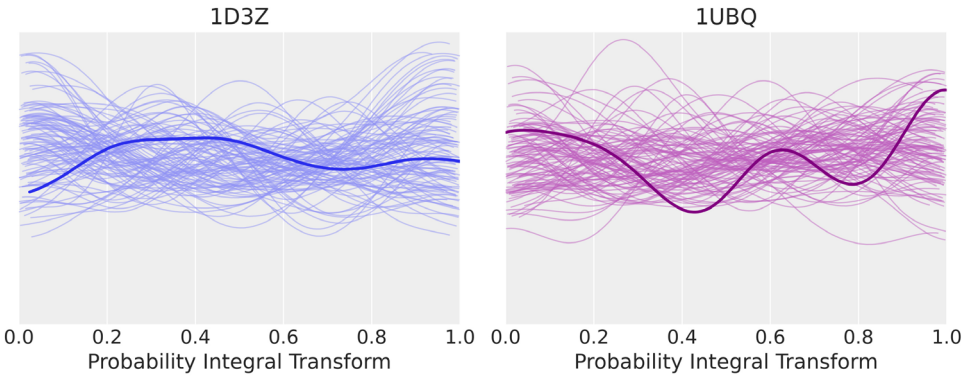

FIGURE 5 LOO-PIT for 1D3Z and 1UBQ. The thick line corresponds to the observed LOO-PIT density and the thin lines represent simulations from the standard uniform distribution in the [0, 1] interval for a data set of the same size as the one used to computed LOO-PIT. From comparison with these simulations we can define what constitutes a deviation from uniformity larger than expected. Both 1D3Z and 1UBQ are within the expected margins

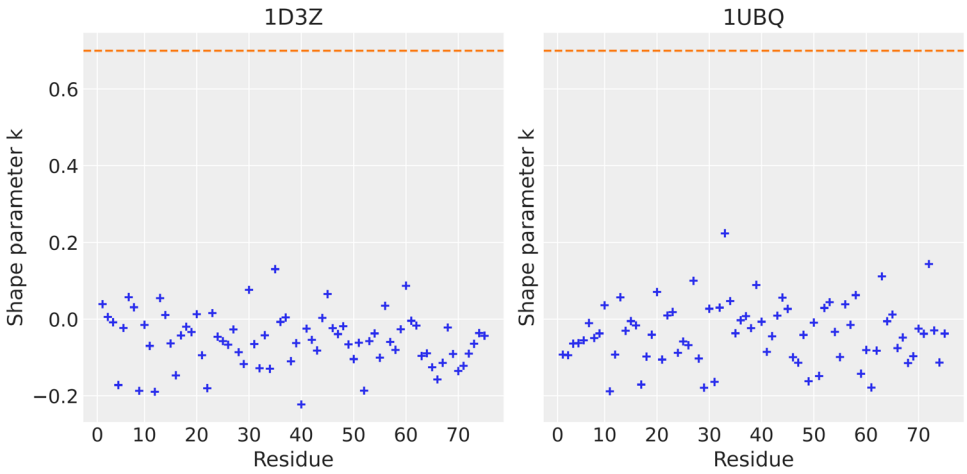

FIGURE 6 $\hat{k}$ values for every residue in 1D3Z and 1UBQ. The dashed orange line indicates the value of 0.7. Residues above this value can be considered as highly influential. As expected for good quality structures like 1UBQ and 1D3Z, highly influential observations are not present

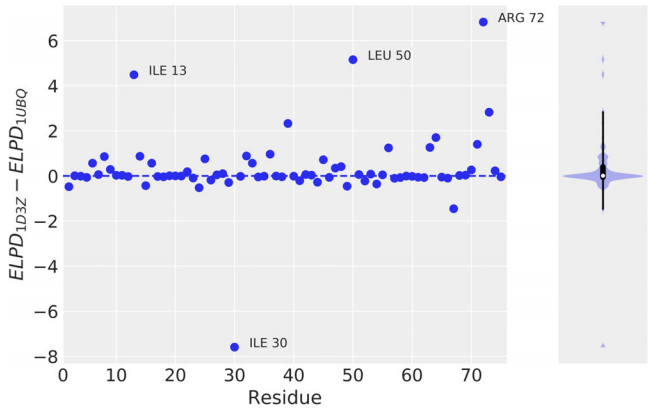

FIGURE 7 Difference for the pointwise expected log predictive density between structures 1D3Z and 1UBQ. Positive values indicate that a particular residue is better resolved by structure 1D3Z, and in turn negative values indicate that structure 1UBQ better resolves them. We have annotated the residues with the largest absolute differences. In general, differences around ±4 can be considered small

4 | CONCLUSIONS

We have presented a collection of tools and visualizations for NMR protein structure assessment. All of these tools are based on Bayesian statistical models and established validation methods from the Bayesian statistics field. We consider such visualizations as useful additions to the current toolbox for protein structure validation. We note that we are using these Bayesian model comparison tools differently from standard Bayesian model comparison routines. That is to say, the statistical model is kept fixed and the 3D structures vary. Thus, when observing a potential problem we are directly attributing it to the structure's quality, as we consider that the hierarchical linear model and the method to compute theoretical $^{13}C^{\alpha}$ chemical shifts are in general good enough for the purpose of protein structure validation.

The hierarchical model proposed in this work seems to correctly model $^{13}C^{\alpha}$ chemical shifts and can be considered an extension to the constant and linear corrections we have applied in previous works.[13,15,31] With the important addition of the partial pooling of data provided by the hierarchical structure. Nevertheless, statistical models should always be considered provisional and thus we encourage researchers to explore alternative models. In the meantime, if a researcher desires to use these tools as part of their validation process we want to emphasize that the tools and visualizations presented here are not intended to provide categorical answers about the quality of structures but instead help experts explore it and, when possible, guide them to make improvements of such structures. For example, when observing a high value of $\hat{\kappa}$, this could be indicative of either a problem with the 3D structural model, indicating that a residue needs further refinement, or a problem with the hierarchical statistical model not being able to accurately resolve a particular residue. Given the structures we have studied in this work, the latter seems unlikely to occur, and a high value of $\hat{\kappa}$ will most likely indicate a problem with the 3D structural model.

One limitation of $^{13}C^{\alpha}$ chemical shifts is that they cannot be mapped in a 1 to 1 fashion to torsional angle values. That is, they are multivalued functions of torsional angles. Nevertheless, they can be useful for validation as we and others have already shown, as they still provide useful information about torsional angles.[12,13–15,32–34] As different observables reveal different aspects of protein structures, we encourage researchers to perform similar analyses to the ones presented here using other observables than $^{13}C^{\alpha}$ chemical shifts. Furthermore, among the limitations of the presented analysis, we can mention that if residue Cystein plays a key role in a particular study posterior to validation with the presented methods, a different method for computing theoretical $^{13}C^{\alpha}$ chemical shifts must be included in the workflow.

ACKNOWLEDGMENTS

We are honored to dedicate this manuscript to the memory of Harold A. Scheraga, Professor of Chemistry, Emeritus at Cornell University. He achieved leadership in the world of science, and high respect among colleagues, as a result of his colossal experience in Experimental and Theoretical Chemistry, Physics and Mathematics, research in Protein Chemistry, and, in particular, due to his tireless efforts in the

search of possible solutions to the Protein Folding problem. Harold Scheraga passed away at the age of 98 in Ithaca, NY, on August 1, 2020. Funding was provided by National Agency of Scientific and Technological promotion—ANPCyT, Grant PICT-0767, PICT-02212. And National Scientific and Technical Research Council—CONICET, Grant PIP-0087.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available in Protein Data Bank at https://www.rcsb.org/ and Biological Magnetic Resonance Bank at https://bmrb.io/.

ORCID


Agustina Arroyuelo https://orcid.org/0000-0002-7249-755X

REFERENCES

[1] C. K. Fisher, A. Huang, C. M. Stultz, *J. Am. Chem. Soc.* **2010**, *132*, 14919.
[2] T. Hamelryck, K. Mardia, J. Ferkinghoff-Borg Eds., *Bayesian Methods in Structural Bioinformatics*, 1st ed., Springer, Berlin **2012**.
[3] S. Olsson, J. Frellsen, W. Boomsma, K. V. Mardia, T. Hamelryck, *PLoS One* **2013**, *8*, 1.
[4] M. Vendruscolo, A. Cavalli, M. Bonomi, C. Camilloni, *Sci. Adv.* **2016**, *2*, e1501177.
[5] P. G. Garay, O. A. Martin, H. A. Scheraga, J. A. Vila, *Peer J.* **2016**, *4*, e2253.
[6] A. Arroyuelo, O. A. Martin, H. A. Scheraga, J. A. Vila, *J. Phys. Chem. B* **2020**, *124*, 735.
[7] M. Bonomi, G. T. Heller, C. Camilloni, M. Vendruscolo, *Curr. Opin. Struct. Biol.* **2017**, *42*, 106.
[8] O. Martin, *Bayesian Analysis with Python: Introduction to statistical modeling and probabilistic programming using PyMC3 and ArviZ*, 2nd ed., Packt Publishing, Birmingham **2018**.
[9] R. McElreath, *Statistical Rethinking: A Bayesian Course with Examples in R and STAN*, 2nd ed., Chapman and Hall/CRC, Boca Raton **2020**.
[10] Y. W. Teh, M. I. Jordan, M. J. Beal, D. M. Blei, Sharing clusters among related groups: Hierarchical dirichlet processes. in *Advances in neural information processing systems*. NIPS, Vancouver **2005**, p. 1385.
[11] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, D. B. Rubin, *Bayesian Data Analysis*, 3rd ed., Chapman and Hall/CRC, Boca Raton **2013**.
[12] D. S. Wishart, C. G. Bigam, J. Yao, F. Abildgaard, H. J. Dyson, E. Oldfield, J. L. Markley, B. D. Sykes, *J. Biomol. NMR* **1995**, *6*, 135.
[13] J. A. Vila, Y. A. Arnautova, O. A. Martin, H. A. Scheraga, *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 16972.
[14] L. A. Bratholm, A. S. Christensen, T. Hamelryck, J. H. Jensen, *Peer J.* **2015**, *3*, e861.
[15] O. A. Martin, J. A. Vila, H. A. Scheraga, *Bioinformatics* **2012**, *28*, 1538.
[16] A. Gelman, A. Vehtari, D. Simpson, C. C. Margossian, B. Carpenter, Y. Yao, L. Kennedy, J. Gabry, P.-C. Bürkner, and M. Modrák, Bayesian workflow, *arXiv*:2011.01808, **2020**.
[17] A. Vehtari, A. Gelman, J. Gabry, *Stat. Comput.* **2017**, *27*, 1413.
[18] A. Vehtari, D. Simpson, A. Gelman, Y. Yao, and J. Gabry, Pareto Smoothed Importance Sampling, *arXiv*:1507.02646, **2019**.
[19] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne, *Nucleic Acids Res.* **2000**, *28*, 235.
[20] J. L. Markley, E. L. Ulrich, H. M. Berman, K. Henrick, H. Nakamura, H. Akutsu, *J. Biomol. NMR* **2008**, *40*, 153.
[21] S. Vijay-Kumar, C. E. Bugg, W. J. Cook, *J. Mol. Biol.* **1987**, *194*, 531.
[22] G. Cornilescu, J. L. Marquardt, M. Ottiger, A. Bax, *J. Am. Chem. Soc.* **1998**, *120*, 6836.
[23] J. Kruschke, *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*, Academic Press, New York **2014**.
[24] S. Watanabe, M. Opper, *J. Mach. Learn. Res.* **2010**, *11*, 3571.
[25] J. Gabry, D. Simpson, A. Vehtari, M. Betancourt, A. Gelman, *J. R. Stat. Soc. A*, **2019**, *182*, 389.
[26] J. Salvatier, T. V. Wiecki, C. Fonnesbeck, *Peer J. Comput. Sci.* **2016**, *2*, e55.
[27] R. Kumar, C. Carroll, A. Hartikainen, O. Martin, *J. Open Sour. Soft.* **2019**, *4*, 1143.
[28] W. L. DeLano, Pymol: An open-source molecular graphics tool. in *CCP4 Newsletter on Protein Crystallography*, Vol. 40 DeLano Scientific, San Carlos, Californina **2002**, p. 82.
[29] W. Humphrey, A. Dalke, K. Schulten, *J. Mol. Graph.* **1996**, *14*, 33.
[30] Y. A. Arnautova, J. A. Vila, O. A. Martin, H. A. Scheraga, *Acta Crystallogr. Sect. D: Biol. Crystallogr.* **2009**, *65*, 697.
[31] O. A. Martin, J. A. Vila, H. A. Scheraga, *Bioinformatics* **2012**, *28*, 1538.
[32] O. A. Martin, Y. A. Arnautova, A. A. Icazatti, H. A. Scheraga, J. A. Vila, *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, 16826.
[33] B. Han, Y. Liu, S. W. Ginzinger, D. S. Wishart, *J. Biomol. NMR* **2011**, *50*, 43.
[34] Y. Shen, A. Bax, *J. Biomol. NMR* **2010**, *48*, 13.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: A. Arroyuelo, J. A. Vila, O. A. Martin, *J Comput Chem* **2021**, *42*(21), 1466. https://doi.org/10.1002/jcc.26556