




Cite this: *Anal. Methods*, 2021, 13, 359

## Bayesian inference assessment of protein secondary structure analysis using circular dichroism data – how much structural information is contained in protein circular dichroism spectra?†

Simon E. F. Spencer<sup>\*a</sup> and Alison Rodger <sup>b</sup>

Circular dichroism spectroscopy is an important tool for determining the structural characteristics of biomolecules, particularly the secondary structure of proteins. In this paper we propose a Bayesian model that estimates the covariance structure within a measured spectrum and quantifies the uncertainty associated with the inferred secondary structures and characteristic spectra associated with each secondary structure type. Furthermore, we used tools from Bayesian model selection to determine the best secondary structure classification scheme and illustrate a technique for comparing whether or not two or more measured protein spectra share the same secondary structure. Our findings suggest that it is not possible to identify more than 3 distinct secondary structure classes from CD spectra above 175 nm. The inclusion of data from wavelengths between 175 and 200 nm did not substantially affect the ability to determine secondary structure fractions.

Received 31st August 2020  
 Accepted 5th November 2020

DOI: 10.1039/d0ay01645d

[rsc.li/methods](https://rsc.li/methods)

## Introduction

Proteins are the main focus of a wide range of areas of research, from biochemistry to cellular biology to drug discovery. Since a protein's structure determines its functionality, many spectroscopic techniques have been developed, each one designed to explore an aspect of these biomolecules.<sup>1</sup> Far ultra-violet (<260 nm) circular dichroism (CD) spectroscopy is an important and successful spectroscopic technique that gives meaningful information about the secondary structure of proteins,<sup>2–6</sup> *i.e.* its local shape. CD is particularly useful when only samples in solution are available and techniques such as X-ray crystallography cannot be used. Thanks to the fast and cheap nature of the experiments, CD is an ideal tool for testing controls in many protein screening assays related to the drug discovery process.<sup>7</sup>

In recent years large datasets of CD spectra have been produced<sup>8–10</sup> which enable the relationships between secondary structure and CD data to be explored through mathematical modeling and statistical analysis. In CD spectroscopy the main approach to find the secondary structure has been to deconvolute a spectrum into a weighted sum of so-called characteristic spectra by a variety of different algorithms.<sup>11–14</sup> For a given protein the relative weight of each characteristic spectrum

enables calculation of the abundance of the respective structure element from known data about the structure of the proteins making up the reference set. From a statistical perspective these approaches share the same model structure: a linear model, and use standard regression techniques to fit the model. There are a few exceptions, such as neural network model approaches,<sup>15–17</sup> though the self-organising map neural network approaches also involve finding a best match combination of spectra and assigning secondary structures from them. The implicit assumption behind a linear model is that the errors at each wavelength are uncorrelated and have equal variance, even though evidence for a more complicated error structure has been recently pointed out,<sup>18</sup> with variance depending on the wavelength. There are also significant amounts of correlation within some parts of the spectrum. In this work we used a Bayesian approach in which we keep the linear structure but let the covariance matrix of the CD spectra be as general as possible in order to identify crucial dependencies within the CD spectra and weight the information in the most coherent way. Firstly, our approach can be used as a secondary structure estimation method or to enhance existing algorithms. Secondly, we can use tools from Bayesian model selection to investigate the secondary structure classifications schemes that can be determined successfully from CD data. In particular, we consider which secondary structures can be assigned from CD data between 175–260 nm. We also explore possible uses of the posterior uncertainty.

<sup>a</sup>Department of Statistics, University of Warwick, Coventry, UK. E-mail: [s.e.f.spencer@warwick.ac.uk](mailto:s.e.f.spencer@warwick.ac.uk)

<sup>b</sup>Department of Molecular Sciences, Macquarie University, NSW, 2109, Australia. E-mail: [alison.rodger@mq.edu.au](mailto:alison.rodger@mq.edu.au)

† Electronic supplementary information (ESI) available. See DOI: 10.1039/d0ay01645d



## Methods

### Notation

Let  $c_x$  be the CD spectrum measured from an individual protein  $x$ , formally a row with  $n_\lambda$  entries along the wavelength range, in our case usually for data measured between 175 nm and 260 nm with steps of 1 nm. The CD spectrum units are the per residue molar absorption units  $\Delta\epsilon$  measured in  $\text{mol}^{-1} \text{dm}^{-3} \text{cm}^{-1}$ . Let  $f_x$  represent the secondary structure fractions, a row vector of  $n_s$  secondary structure elements that must sum to one. Finally let  $B$  be a matrix of dimension  $n_s \times n_\lambda$ , whose rows hold the characteristic CD spectra for each secondary structure class. The common hypothesis in the linear model is:

$$c_x = f_x B + w_x \quad (1)$$

where the row vector  $w_x$  represents the error between the predicted spectrum and the observed data. Even though eqn (1) seems a typical regression problem it has challenging features to be faced. First, eqn (1) is actually a special kind of inverse problem<sup>19</sup> where both the parameters  $f_x$  and the design matrix are unknown. To overcome this lack of knowledge about the design matrix it is common practice to use a reference set, a dataset of proteins with known secondary structure and CD spectra, to estimate  $B$ . The secondary structure fractions  $f_x$  for proteins in a test set can then be determined. Many existing algorithms do not use the full information from the reference set but often have a variable selection step<sup>13</sup> from which to identify a subset of proteins closely related to the test protein  $x$  to estimate the matrix  $B$ . Second, the elements of  $f_x$  are not independent but are constrained to sum to one and usually constrained to be non-negative numbers, as they represent the proportion of the molecule that belongs to each secondary structure class:

$$0 \leq f_i \leq 1, \text{ for } i = 1, \dots, n_s \quad (2)$$

$$\sum_{i=1}^{n_s} f_i = 1 \quad (3)$$

Each existing approach has a different method for satisfying such constraints, sometimes based on *ad hoc* criteria and leading to an approximate solution, *e.g.* the sum can be in the region  $1 \pm 0.05$ .<sup>11</sup>

A third challenge relates to the common assumptions of considering normal and independent variables for the error  $w_x$ . If  $\Sigma$  is the covariance matrix within a spectrum, then it is usually chosen to be a diagonal matrix, not taking into account possible correlations that a spectrum is known to display.

In the following sections we discuss a Bayesian approach to inference where the fraction vector  $f_x$  and characteristic matrix  $B$  are estimated jointly, making full use of the information in the reference set, *i.e.* without variable selection. We introduce a Dirichlet prior for the fractions  $f_x$  to capture the constraints (2) and (3) and most importantly we estimate a general covariance matrix  $\Sigma$  for the errors, allowing the model to learn the correct covariance structure from the reference set.

### Model and likelihood

In order to use all the data in a reference set to estimate secondary structures of unknown proteins, we proceed as follows. Let  $C$  be the matrix denoting all the spectroscopic data, whose first  $n_r$  rows are the CD spectra of the reference set, and the remaining  $n_t$  are the spectra of the proteins in the test set to be analyzed, each row having length  $n_\lambda$ . In the same way let  $F$  be the  $(n_r + n_t) \times n_s$  matrix of secondary structure fractions for all the proteins, where  $n_s$  is the number of secondary structure classes. We treat all of the proteins the same, whether they are in the reference set or not, and any unknown secondary structure fractions are treated as parameters and inferred. The matrix formulation of the model for the CD spectra is:

$$C = FB + W \quad (4)$$

where the  $n_s$  rows of  $B$  can be thought of as the characteristic CD spectra for each secondary structure class and  $W$  is the random matrix representing experimental variability and any other disagreement between the observed data and predicted spectra, such as lack-of-fit.

We suppose that the CD spectra are normally distributed about  $BF$  with the same general covariance matrix  $\Sigma$ , which will be inferred from the data. Furthermore, we assume no dependence between the spectra of different proteins. This leads to the error  $W$  taking the form of a matrix normal distribution:

$$W \sim N_{n_r+n_t, n_\lambda}(O_{n_r+n_t, n_\lambda}, I_{n_r+n_t}, \Sigma) \quad (5)$$

where  $O_{n_r+n_t, n_\lambda}$  is the  $(n_r + n_t) \times n_\lambda$  zero matrix,  $I_{n_r+n_t}$  is the  $(n_r + n_t) \times (n_r + n_t)$  identity matrix. The matrix  $\Sigma$  captures the covariance structure along the rows of  $W$  (among the wavelengths in a spectrum), whilst the columns of  $W$  (representing the proteins) are assumed to be independent. Thus, the likelihood function is the matrix normal density:

$$L(C|F_t, B, \Sigma) = ((2\pi)^{n_\lambda} |\Sigma|)^{-(n_r+n_t)/2} \exp \left[ -\frac{1}{2} \text{Tr}[(C - FB)^T \Sigma^{-1} (C - FB)] \right] \quad (6)$$

where  $F_t$  is the submatrix of unknown secondary structure fractions related to the  $n_t$  proteins in the test-set.

### Prior distributions

The model parameters are the set  $\{F_t, B, \Sigma\}$ , and prior knowledge is factorized as follows:

$$\pi(F_t, B, \Sigma) = \pi(F_t) \pi(\Sigma) \pi(B | \Sigma, F_t)$$

The conditional dependence structure of the model is shown in the ESI (Fig. S1†). For every protein in the test set we chose independent Dirichlet-distributed priors:

$$f_x \sim \text{Dir}(\alpha), \text{ for } x = n_r + 1, \dots, n_r + n_t$$

and in applications we used the Jeffreys prior



$$\alpha = [1/2, \dots, 1/2]_{ms}$$

The Dirichlet distribution is a natural choice because the parameter space is a  $n_s$ -dimensional simplex defined by the constraints (2) and (3).

For  $\mathbf{B}$  and  $\mathbf{\Sigma}$  we follow the common choice for Bayesian linear models and choose conjugate priors.<sup>20</sup> The prior for  $\mathbf{B}$  is the matrix normal distribution with mean  $\mathbf{M}$  and covariance matrices  $\mathbf{U}$  and  $\mathbf{\Sigma}$ . In applications the matrix  $\mathbf{M}$  was chosen to be the zero matrix, which is symmetrical with respect to positive and negative signals – reflecting left and right-handed chirality – and has a shrinkage effect. The matrix  $\mathbf{U}$  represents the covariance between the secondary structure classes and a  $g$ -prior<sup>21</sup> is chosen to account for possible relationships within the secondary structures:

$$\mathbf{U} = g(\mathbf{F}^T \mathbf{F})^{-1}.$$

Following George and Foster,<sup>22</sup> we set the hyper-parameter  $g = n_r$ , the dimension of the reference set, this choice is referred to as the unit information prior.

The prior for the  $n_\lambda \times n_\lambda$  covariance matrix  $\mathbf{\Sigma}$ , representing the covariance structure within a CD spectrum, is the inverse-Wishart distribution. The inverse-Wishart  $W_n^{-1}(\delta, \mathbf{S})$  is the generalization of the inverse-Gamma distribution in  $n$ -dimensions, having density:

$$\pi(\mathbf{\Sigma}) = c(n, \delta) |\mathbf{S}|^{(n+\delta-1)/2} |\mathbf{\Sigma}|^{-(2n+\delta)/2} \exp \left[ -\frac{1}{2} \text{Tr}(\mathbf{\Sigma}^{-1} \mathbf{S}) \right]$$

for positive definite  $\mathbf{\Sigma}$  where

$$c(n, \delta) = \frac{2^{-n(n+\delta-1)/2}}{\Gamma[(n+\delta-1)/2]}$$

and  $\Gamma[\cdot]$  is the Euler Gamma function.

In applications we chose  $\delta = n_\lambda$ , representing the degrees of freedom, and  $\mathbf{S} = \mathbf{I}_{n_\lambda}$ , a covariance matrix with no correlation and unit variances. In summary, we can write for the priors:

$$\mathbf{B} | (\mathbf{\Sigma}, \mathbf{F}_t) \sim N_{n_s, n_\lambda}(\mathbf{M}, n_r (\mathbf{F}^T \mathbf{F})^{-1}, \mathbf{\Sigma}) \quad (7)$$

$$\mathbf{\Sigma} \sim W_{n_\lambda}^{-1}(n_\lambda, \mathbf{S}) \quad (8)$$

### McMC algorithm

The computation of the posterior distribution  $\pi(\mathbf{F}_t, \mathbf{B}, \mathbf{\Sigma} | \mathbf{C})$  is done using Markov chain Monte Carlo (McMC). Due to the conjugate prior specified in eqn (7), the full conditional distribution for  $\mathbf{B}$  follows a matrix normal distribution,<sup>20,23</sup>

$$\mathbf{B} | (\mathbf{C}, \mathbf{\Sigma}, \mathbf{F}_t) \sim N_{n_s, n_\lambda}(\mathbf{M}^*, \mathbf{U}^*, \mathbf{\Sigma}) \quad (9)$$

with updated parameters:

$$\begin{aligned} \mathbf{M}^* &= \mathbf{U}^* (\mathbf{F}^T \mathbf{C} + \mathbf{U}^{-1} \mathbf{M}) \\ \mathbf{U}^* &= (\mathbf{F}^T \mathbf{F} + \mathbf{U}^{-1})^{-1} \end{aligned}$$

The full conditional distribution for the covariance matrix  $\mathbf{\Sigma}$  is given by,

$$\mathbf{\Sigma} | (\mathbf{C}, \mathbf{F}_t) = W^{-1}(\delta^*, \mathbf{S}^*) \quad (10)$$

with

$$\begin{aligned} \delta^* &= \delta + n_r \\ \mathbf{S}^* &= \mathbf{S} + \mathbf{M}^T \mathbf{U}^{-1} \mathbf{M} + \mathbf{C}^T \mathbf{C} - (\mathbf{M}^*)^T (\mathbf{U}^*)^{-1} \mathbf{M}^* \end{aligned}$$

The conjugate priors for  $\mathbf{B}$  and  $\mathbf{\Sigma}$  allow us to integrate out these parameters and obtain a closed-form expression for the likelihood:

$$\pi(\mathbf{C} | \mathbf{F}_t) = \frac{1}{\pi^{n_r n_\lambda}} \left( \frac{|\mathbf{U}^*|}{|\mathbf{U}|} \right)^{n_r} \frac{\Gamma_{n_\lambda}(\delta^*/2) |\mathbf{S}|^{\delta^*/2}}{\Gamma_{n_\lambda}(\delta/2) |\mathbf{S}^*|^{\delta^*/2}} \quad (11)$$

where  $\Gamma_p(a) = \pi^{p(p-1)/4} \prod_{j=1}^p \Gamma(a + (1-j)/2)$  is the multivariate gamma function. Integrating out these parameters greatly improves the mixing of the resulting MCMC algorithm.

The prior for  $\mathbf{F}_t$  is not conjugate and so these parameters are updated using a Metropolis–Hastings step, for each protein individually. Proposals are generated using an adaptive Dirichlet random walk algorithm. If  $\mathbf{f}_x$  is the row within  $\mathbf{F}_t$  for protein  $x$ , then the proposal is

$$\mathbf{f}'_x \sim \text{Dirichlet}(\alpha + \beta_x \mathbf{f}_x) \quad (12)$$

where the scaling factor  $\beta_x$  is increased by  $0.234 \times 10/i$  on iteration  $i$  if the proposal is rejected and decreased by  $0.776 \times 10/i$  if the proposal is accepted. This adapts each proposal to target an acceptance rate of 0.234. If samples for  $\mathbf{\Sigma}$  and  $\mathbf{B}$  are required then these can be drawn from eqn (9) and (10) respectively given the samples from  $\mathbf{F}_t$ .

### Protein data and reference sets

Open access CD datasets are available in the Protein Circular Dichroism Data Base (PCDDb)<sup>9</sup> with high quality data obtained using Synchrotron Radiation Circular Dichroism. From the PCDDb we took the SP175 dataset<sup>8</sup> containing spectra for 71 globular proteins. We considered three secondary structure classification schemes. The first comes from the DSSP algorithm<sup>24,25</sup> which includes 8 classes:  $\alpha$ -helix, 3-10-helix,  $\beta$ -strand, turn, bend,  $\pi$ -helix,  $\beta$ -bridge, irregular. In the SP175 database there is almost no contribution from  $\pi$ -helix, so in all analyses we combined this class with irregular, leaving 7 classes in total. The second scheme we considered included just 3 classes taken from DSSP:  $\alpha$ -helix,  $\beta$ -strand and other (referred to as DSSP<sub>red</sub>), comprising the sum of the remaining categories. The third scheme was defined through the CD scheme<sup>26</sup> and following.<sup>27</sup> We refer to this as the SELCON scheme whose six secondary structure classes are: regular helix (the middle of any helix), distorted helix (the two residues on each end of a helix), regular  $\beta$ -strand, distorted  $\beta$ -strand (1 residue on the end of each strand), turn and other. Finally, we also consider the BeStSel classification scheme from ref. 27. Their eight classes are regular  $\alpha$ -helix, distorted  $\alpha$ -helix, left-twisted  $\beta$ -strand, relaxed  $\beta$ -strand, right-twisted  $\beta$ -strand, parallel-strand, turn and other.



For details of the interrelationship between these classification schemes, see ref. 27.

### Performance indices and cross validation

Algorithms are usually tested with leave-one-out cross validation. In this methodology one protein at a time is removed from the reference set and is specified to be the test set. Repeating this for each protein in the original reference set gives  $n_r$  estimated vectors of secondary structure, which we will refer to as the cross-validation set.

The performance of an algorithm, *i.e.* comparing a secondary structure estimate with the X-ray experiment value, taken as “truth”, is performed with two measures commonly accepted in the CD literature: the root-mean-square deviation (RMSD)  $\delta$  and the Pearson correlation coefficient  $r$ . They are defined for a protein  $x$  with true value  $\mathbf{f}_x^*$  as

$$\delta_x = \left( \frac{1}{n_s} \sum_{i=1}^{n_s} (f_{x,i} - f_{x,i}^*)^2 \right)^{1/2}$$

$$r_x = \frac{\sum_{i=1}^{n_s} (f_{x,i} - \bar{f}_x)(f_{x,i}^* - \bar{f}_x^*)}{\sigma_{f_x} \sigma_{f_x^*}} \quad (13)$$

where  $\bar{f}$  denotes the mean of the entries in the vector  $\mathbf{f}$  and  $\sigma_f$  denotes the standard deviation. These two quantities are related to single protein estimation only. In order to measure the performance of the algorithm, averages over the cross-validation set are taken. In particular the average RMSD,  $\delta = \frac{1}{n_r} \sum_{x=1}^{n_r} \delta_x$  has been used by other authors, see for example,<sup>11</sup> as prediction error of any (future) protein structure estimates.

As well as measuring global performance, it is interesting to know how the algorithm behaves for each secondary structure class. The RMSD  $\delta_i$  and correlation coefficient  $r_i$ , for class  $i$ , are calculated similarly to (12) and (13), but performing the sum over the proteins  $x$  in the cross-validation set. The measures are known in the literature as performance indices of the analysis and used for comparisons between methods. A variety of more sophisticated performance measures also exist, which attempt to normalise by the amount of variation inherent within each class. One such measure is  $\zeta_i = \sigma_i / \delta_i$ ,<sup>28</sup> where  $\sigma_i$  is the standard deviation of the secondary structure fractions for class  $i$ . If  $\zeta_i$  is greater than 1 then the average error is less than a random choice from the reference set.

Our Bayesian approach produces a posterior distribution over secondary structures rather than a single estimated secondary structure vector. Unless otherwise stated we have used the posterior mean secondary structure as the estimated structure.

### Model comparison *via* marginal likelihoods

An advantage of our Bayesian approach is that it becomes possible to use Bayesian model comparison to answer questions of scientific interest, such as which secondary structure classes can be identified from the reference proteins and whether two

CD spectra share the same secondary structure. In order to choose between models  $\{M_i : i \in I\}$ , we examine the posterior probability in favour of model  $M_i$ , given by:

$$P(M_i|C) = \frac{\pi(C|M_i)P(M_i)}{\sum_{j \in I} \pi(C|M_j)P(M_j)} \quad (14)$$

where  $\pi(C|M_i)$  is the marginal likelihood for model  $i$ .

There is no closed form available for the marginal likelihoods for these models. However, due to the conjugate priors for  $\mathbf{B}$  and  $\Sigma$ , we can integrate out these two parameters to obtain an expression for  $\pi(C|F_t)$ , see eqn (11).

To obtain an estimate of the full marginal likelihood

$$\pi(C) = \int \pi(C|F_t)\pi(F_t) dF_t$$

we apply methodology that uses samples from the McMC to inform an importance sampling estimator for the marginal likelihood.<sup>29,30</sup> First samples are obtained from the marginal posterior  $\pi(F_t|C)$  using the usual McMC algorithm. Secondly a parametric distribution (with known normalising constant) is fitted to the McMC samples, usually a multivariate normal distribution. Let  $q(F_t)$  denote the density of this distribution. In this application only the first  $(n_s - 1)$  components of each secondary structure vector are used, as the final component can be recovered from constraint (3). Thirdly,  $N$  samples (labelled  $F_t^{(1)}, \dots, F_t^{(N)}$ ) are drawn from  $q(F_t)$ . Finally, we obtain the importance sampling estimator for the marginal likelihood for a specific model  $M_i$ :

$$\pi(C|M_i) = \frac{1}{N} \sum_{k=1}^N \frac{\pi(C|F_t^{(k)}, M_i)\pi(F_t^{(k)}|M_i)}{q(F_t^{(k)})} \quad (15)$$

It is desirable to make  $q(F_t)$  over-dispersed relative to the true posterior, to make the variance of the importance sampling estimator as small as possible. This can easily be achieved by replacing  $q(\cdot)$  with a multivariate  $t$ -distribution or a mixture of the multivariate normal and the prior  $\pi(F_t)$ . For full details of this methodology see Touloupou *et al.*<sup>29</sup>

## Results and discussion

### Secondary structure estimation

There is significant debate in the literature as to whether CD spectra from 260–175 nm contain enough information to give different spectral signatures for any folds more than  $\alpha$ -helix and  $\beta$ -sheet. So, to avoid trying to answer multiple questions simultaneously we chose to assess the accuracy of secondary structure estimation using our Bayesian approach by performing a leave-one-out cross validation over the reference set SP175 with the simplest classification scheme, DSSP<sub>red</sub>. In Table 1 we have compared our model predictions with some of the other algorithms, including SELMAT3,<sup>8,26</sup> Partial Least Squares (PLS), Principal Component Regression (PCR), Neural Network (NN), and Support Vector Machines (SVM) using results taken from ref. 31. Broadly speaking, our Bayesian approach is competitive with the other approaches for  $\alpha$ -helix, but does not do as well for





**Table 1** Cross-validation results for the SP175 proteins with 3 secondary structure classes from DSSP:  $\alpha$ -helix,  $\beta$ -sheet and other structure. Results for competing approaches (SELMAT3, PCR, PLS, NN and SVM, taken from ref. 31) are shown. The best performing approach for each measure is given in bold.  $\delta$  is RMSD and  $r$  is the correlation

Method	$\alpha$ -helix		$\beta$ -sheet		Other	
	$\delta$	$r$	$\delta$	$r$	$\delta$	$r$
Bayesian	0.061	0.96	0.127	0.77	0.137	0.65
SELMAT3	0.063	0.96	0.083	0.86	0.078	0.70
PCR	0.057	0.97	0.069	0.91	0.066	0.80
PLS	<b>0.053</b>	<b>0.97</b>	0.073	0.90	0.068	0.78
NN	0.055	0.97	<b>0.067</b>	<b>0.91</b>	<b>0.062</b>	<b>0.82</b>
SVM	0.057	0.97	0.069	0.91	0.066	0.79

**Table 2** Cross-validation results for the SP175 proteins with the SELCON secondary structure scheme. Results for competing approaches SELMAT3 and PCR are taken from ref. 31. The best performing approach for each measure is given in bold

Structure	Bayesian		SELMAT3		PLS	
	$\delta$	$r$	$\delta$	$r$	$\delta$	$r$
Regular helix	0.090	0.836	0.048	0.956	<b>0.040</b>	<b>0.971</b>
Distorted helix	0.129	0.043	<b>0.035</b>	<b>0.809</b>	<b>0.036</b>	0.791
Regular $\beta$ -strand	0.090	0.695	0.073	0.792	0.063	<b>0.853</b>
Distorted $\beta$ -strand	0.281	−0.081	<b>0.020</b>	<b>0.913</b>	0.023	0.889
Turn	0.201	0.098	<b>0.052</b>	0.325	<b>0.052</b>	<b>0.332</b>
Other	0.169	0.278	<b>0.050</b>	0.717	<b>0.050</b>	<b>0.720</b>

$\beta$ -sheet. Results for the normalised measure  $\zeta$  are given in the ESI (Tables S1 and S2†). Overall, there is no clear best approach.

Table 2 (which is rotated with respect to Table 1) shows the leave-one-out cross validation results using the SELCON secondary structure scheme, over the SP175 reference set. This time the Bayesian approach does not perform well, particularly for the classes turn and other. Results for the normalised measure  $\zeta$  are given in the ESI† where a value above one indicates an improvement above choosing at random from the reference set. For both SELMAT3 and PLS the value of  $\zeta$  is just 1.04 for turn. The results of Table 2 and the normalised measure  $\zeta$  in the ESI† indicate that there is not enough information, even in spectra down to 175 nm to differentiate the 6 SELCON secondary structure motifs.

We hypothesise that the underlying cause of this is that in our Bayesian approach we do not perform variable selection to identify a subset of the reference set for each cross-validation step. Variable selection has been found to avoid inconsistencies between CD spectra and secondary structure schema to improve the quality of the analysis.<sup>13,27</sup> We have chosen not to select a subset of proteins for the reference set that are similar to the test protein as we wanted to use all the information from the reference set. If an  $\alpha$ -helix, for example, has a characteristic signal then it should be consistently present in spectra for all proteins containing  $\alpha$ -helices. We do believe that other protein features, such as distortions at the ends/joins, side chains and higher order structures, can obscure or modify this signal.

These features are not necessarily represented in any of the secondary structure characterisation schemes. Bayesian analyses usually consider information from the whole data set and use the parameter uncertainty to weight the information content rather than discarding information that does not fit the pattern.

Another consideration is that early, landmark papers that found a need for variable selection<sup>32,33</sup> were using techniques such as partial least squares (PLS) to identify the basis vectors from a comparatively small number of reference proteins. Whilst PLS will always succeed in producing  $n$  basis vectors from  $n$  (linearly independent) protein measurements, it is not clear from this kind of analysis how many of the resulting vectors contain only contributions from the underlying signal. The variability inherent in the dataset will certainly dominate in the last basis vectors and (hopefully) the signal will dominate in the first few vectors, but it may be the case, especially when the number of proteins in the reference set is small, that the basis vectors in the middle are actually representing the variability in the dataset as much as clearly identified secondary structures. The immediate conclusion that could be made is that variable selection is important for making predictions using existing classification schemes. However, the apparent success of the variable selection approaches depends on having 'like' proteins in the reference set which is simply not always possible with unknown proteins.

### Identifiability of secondary structure classes

To explore what can be identified from the CD spectra, we used the model selection methodology to determine which secondary structure classes can be identified from the amount of information in a given reference set. We used the 3 classification schemes: DSSP, SELCON and BeStSel, and we also defined simpler schemes from within these by summing together components. Since DSSP, SELCON and BeStSel have 7, 6 and 8 secondary structure classes respectively, we considered 5220 schemes in total.

Each potential secondary structure scheme is represented by a different design matrix of secondary structure fractions  $F$ . Given  $F$ , we can evaluate the marginal likelihood for the model analytically from eqn (11) since the test set is empty here. These marginal likelihoods can be used to produce Bayes factors comparing any pair of models and, once prior probabilities for each model have been specified, the posterior probability in favour of each secondary structure scheme can be identified. Following Scott *et al.* and Spencer *et al.*<sup>34,35</sup> we adjusted for multiplicity caused by different numbers of structures in the different models by first assigning a prior distribution over the number of classes, and then dividing the mass equally amongst the models that have the same number of classes. In applications we chose a uniform prior over the number of classes between 3 and the maximum as summarised in Table 3.

We performed model comparison to find the most appropriate model within the three secondary structure schemes individually and also amongst all three schemes. Again, to avoid bias towards schemes with larger numbers of models, we first



**Table 3** Models with posterior probability greater than 0.001 within each classification scheme and in a comparison between all schemes combined

Classification schemes	Posterior probability	Log marginal likelihood	Secondary structure classes		
DSSP	0.510	−7296.235	$\alpha$ -helix	$3_{10}$ -helix+ $\beta$ -strand + bend	$\beta$ -bridge + turn + other
	0.448	−7296.366	$\alpha$ -helix	$\beta$ -strand + bend	$3_{10}$ -helix+ $\beta$ -bridge + turn + other
	0.028	−7299.131	$\alpha$ -helix+ $\beta$ -bridge	$\beta$ -strand + bend	$3_{10}$ -helix + turn + other
	0.013	−7299.892	$\alpha$ -helix+ $\beta$ -bridge	$3_{10}$ -helix + $\beta$ -strand + bend	Turn + other
SELCON	0.999	−7319.677	Regular $\alpha$ -helix	Regular $\beta$ -strand	Distorted $\alpha$ -helix + distorted $\beta$ -strand + turn + unordered
BeStSel	1.000	−7205.936	Regular $\alpha$ -helix	Left $\beta$ -strand + relaxed $\beta$ -strand + parallel	Distorted $\alpha$ -helix + right $\beta$ -strand + turn
DSSP, SELCON, BeStSel	1.000	−7205.936	Regular $\alpha$ -helix	Left $\beta$ -strand + relaxed $\beta$ -strand + parallel	Distorted $\alpha$ -helix + right $\beta$ -strand + turn

assigned a prior probability of one third to each scheme and then divided this prior mass amongst the models stemming from each scheme as before. Table 3 gives all the schemes with posterior probability greater than 0.001. For DSSP the posterior probability is largely split between 2 very similar models that differ only in where 3–10 helix is included. For the SELCON scheme the best model included regular helix and regular  $\beta$ -strand as separate components and combined the remaining classes together. The best model for the BeStSel scheme included 3 basis spectra: regular  $\alpha$ -helix; the sum of left- $\beta$ -strand + relaxed and  $\beta$ -strand + parallel, and the remaining components. This model had by far the largest marginal likelihood and therefore it also dominates the comparison between the 3 classification schemes. Interestingly under all three classification schemes just 3 basis spectra were needed to explain the data. These all included a single  $\alpha$ -helix class in the best model and combined class including  $\beta$ -strand as the second basis spectrum. It may be concluded that no more than three distinct structures that can be assigned from the data between 175–260 nm.

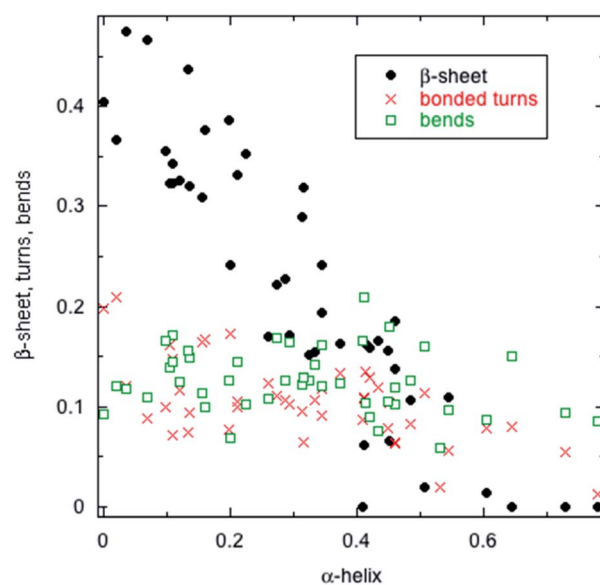
To investigate this further we performed a cross-validation with the BeStSel secondary structure scheme with the highest marginal likelihood for 3, 4 and 5 secondary structure classes. Results for the normalised measure  $\zeta$  are given in ESI.† We repeated each analysis using data down to 175, 180, 185, 190, 195 and 200 nm. The results show that for all these ranges our approach can successfully identify 3 secondary structure classes, as the  $\zeta$  values are all substantially above 1. This provides at least some evidence that our model selection approach, which favours the BeStSel secondary structure scheme, does so with good reason. The  $\zeta$  values decay slightly as the lowest 5 nm of the spectra are sequentially removed, but not substantially. However, when we try to infer more than 3 secondary structure classes then there is always at least one class that has a  $\zeta$  value below one.

These results are an interesting contrast with the accepted literature consensus *e.g.*<sup>36</sup> which is that data to 200 nm contain two independent pieces of information (which with the requirement of the sum of components adding to 1 makes three

pieces of information), data to 190 nm contain 3–4, and data to 178 nm provide 5. The origin of this consensus is in a single value decomposition approach based on 16 reference spectra performed in 1985 by Manavalan and Johnson.<sup>33</sup> Our work indicates that although CD spectra to lower wavelengths do contain more information about protein structure, it cannot be translated into increasing numbers of traditional well-defined structures.

### Spectra covariance matrix and spectral quality

As shown in Fig. 1,  $\alpha$ -helix and  $\beta$ -sheet content are in practice fairly anti-correlated, whereas turns and bends scatter about a mean value largely independent of helix and sheet content until high helix content where bonded turns decrease. So, we wished to characterise the covariance structure of a CD spectrum. We used the SP175 reference set to calculate the posterior



**Fig. 1** Plots of fraction (out of a total of 1) of secondary structure motifs versus  $\alpha$ -helical content for the proteins in reference set SP175 using the DSSP structure annotation.



distributions of the transformation matrix  $\mathbf{B}$  and covariance matrix  $\mathbf{\Sigma}$  using the DSSP<sub>red</sub> scheme ( $\alpha$ -helix,  $\beta$ -sheet and other). In this case the test set is empty and so there is no need for McMC – samples from the posterior can be drawn directly from eqn (9) and (10).

Fig. 2(a) shows the three characteristic spectra that were estimated from the SP175 reference set: the thick lines are the posterior median of the three columns of  $\mathbf{B}$  and the shaded area captures a 95% credible interval. In Fig. 2(b) we show an image representation of the posterior mode for  $\mathbf{\Sigma}$ , given by  $\mathbf{S}^*/(n_r + n_\lambda + 1)$ . Thus, we conclude that despite the tendency for  $\alpha$ -helix and  $\beta$ -sheet to act like a see-saw (Fig. 1), they have distinct typical shapes that combine to give an observed spectrum.

Fig. 2(a) and (b) show that there is more uncertainty/variability for lower wavelengths and relatively little variability above 250 nm where the spectra approach zero. The wide diagonal green band in Fig. 2(b) indicates a strong correlation between errors at similar wavelengths. Conversely the red patches indicate a negative correlation between very low wavelengths and the middle of the spectrum, indicating that if the observed spectrum is lower than predicted at around 190 nm, for example, then it will be higher than expected in the 210–230 nm region.

A key feature of our modelling approach is that we estimate covariance structure of a CD spectrum directly from the data, allowing the measurement uncertainty to change with wavelength and errors at similar wavelengths to be correlated. Most existing algorithms implicitly assume that the errors are uncorrelated so that  $\mathbf{\Sigma}$  is forced to be a diagonal matrix.<sup>11–14,27</sup> This forces errors at similar wavelengths to be uncorrelated, when in reality we expect them to be similar. Fig. 2(b) shows the estimated error structure of a spectrum and strong correlation structure is clearly present. Furthermore, the diagonal elements, representing variances, have a strong dependence on the wavelength, with larger variances at lower wavelengths. This feature is expected due to how the CD instrument works. At shorter wavelengths the high-tension voltage of the photomultiplier tube, which transforms the light signal into an electrical signal, is increased to compensate for the lower power of the light source. Thus, an increased variability in the low UV region data ( $\lambda < 200$  nm) is a known feature of this kind of

spectrum, but gives a worrying indication that the reference spectra are not as perfect as one might hope in the low wavelength region. Above 200 nm the covariance gradually fades to zero along the diagonal indicating higher quality data in this region.

Fig. 2(b) also shows that close wavelengths are positively correlated but further wavelengths are negatively correlated. Two potential mechanisms for generating a negative correlation structure are shown in Fig. 3. Negative correlation could be due to differences in the spectra of proteins in the reference set with similar overall secondary structure content. For example, if two proteins, with close secondary structure, display two similar spectra with a small scaling factor (concentration error) or a slight translation on the wavelength axis (poor wavelength calibration),<sup>37</sup> these differences would lead to a fitted spectrum somewhere in between and the residual would be positive correlated in the short distance but negative for further wavelengths as the spectra change sign or gradient. Another source of disagreement between the observed and fitted spectra could be a lack of fit of the model, which might stem from the different between,  $n$  helical residues being in 5 small helices or 1 large one.<sup>27</sup> Nevertheless, our methodology is able to identify these features of the CD data covariance structure, allowing it to properly weight the information when performing secondary structure estimation or spectral comparison.

### Do two proteins with similar spectra have the same secondary structure?

The second model selection question we address is to determine whether two or more similar looking protein CD spectra

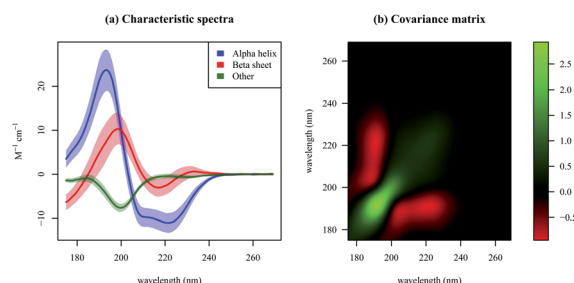


Fig. 2 (a) Plot of the estimated characteristic spectra for  $\alpha$ -helix,  $\beta$ -sheet and other structure based on an analysis of the SP175 reference set. The lines represent the posterior median and the shaded areas represent 95% credible intervals for the characteristic spectra. (b) The posterior mode for the covariance matrix within a CD spectrum  $\mathbf{\Sigma}$ , based on the same analysis.

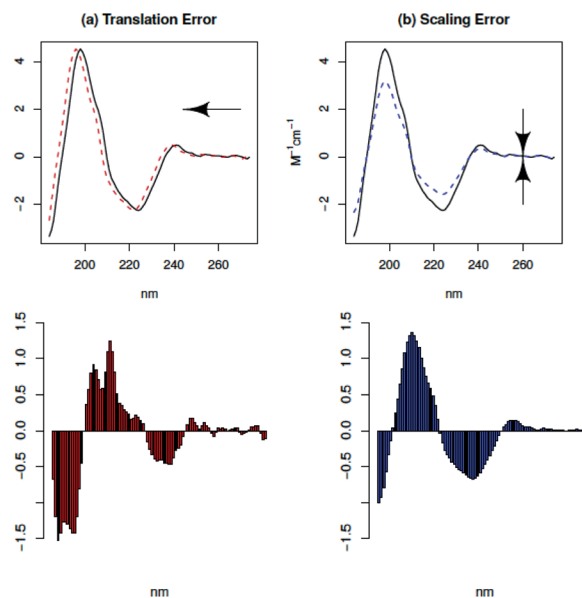


Fig. 3 Schematic representation of two kinds of error that lead to the correlation structure observed in Fig. 2(b). In (a) a translation to the left of 2 nm and (b) a rescaling of the spectrum by a factor of 0.7 (dash line) was applied to the true spectrum (solid line). The true spectrum was bovine gamma-E protein (PDB entry 1m8u). The bottom row plots show the difference between the true spectrum and the spectrum with error.



correspond to proteins with the same secondary structure or not. Let  $c_x$  and  $c_y$  be the spectra from two protein samples  $x$  and  $y$ . Let  $M_0$  be the model assuming the two proteins have the same secondary structure. Let  $M_1$  be the alternative model in which we look for two separate secondary structures for  $x$  and  $y$ . Schematically we have:

$$\begin{aligned} M_0: f_x &= f_y \\ M_1: f_x &\neq f_y \end{aligned}$$

Using eqn (14) and (15) we evaluate the marginal likelihood and the posterior model probability for the two models. As a model prior we chose  $P(M_i) = 1/2$  for  $i \in \{0,1\}$ .

Under model  $M_0$  the spectra  $c_x$  and  $c_y$  are assumed to have the same secondary structure  $f: f_x = f_y$ . Here  $F_t$  has two identical columns, both equal to  $f$ . Under model  $M_1$  we obtain posterior samples for  $f_x$  and  $f_y$ , which represent the two columns of  $F_t$ . For both models the marginal likelihood can be estimated with the importance sampling estimator (15). However, for model  $M_0$  the importance proposal must be fitted to posterior samples from just one column of  $F_t$ , and the second column is set equal to the first; whilst in  $M_1$  the two columns are sampled independently.

We tested the comparison method for two CD spectra with two simulated examples. For our first spectrum  $c_x$ , we take the spectrum of sucrose porin protein (scrY, PDBID: 1a0s) from the MP180 dataset.<sup>38</sup> To obtain our second spectrum  $c_y$ , we added noise to  $c_x$ . First, we added white noise (Fig. 4 left column), with standard deviation equal to  $0.12\Delta\epsilon$ . Second, Fig. 4 right column, we added multivariate Gaussian noise with zero mean and covariance matrix given by the posterior mode for  $\Sigma$ , representing the usual experimental variability. For both comparisons we fitted the competing models using 1000 iterations of the MCMC and then used 100 000 importance samples drawn from a  $t$ -distribution with 3 degrees of freedom. The model

comparison with white noise suggests that the secondary structures are different (posterior probability of a difference 0.987). For the experimental noise, the model comparison favours the simpler model, that the two secondary structures are the same (posterior probability of a difference 0.033). Although the magnitude of the white noise is much smaller than the experimental variability (see Fig. 4 lower plots), the model comparison exercise has correctly identified that the white noise does not conform to experimental variability.

## Conclusions

In this paper we first validated our Bayesian approach for CD structure fitting, then we used the model selection methodology to compare secondary structure classification schemes. We found that the BeStSel scheme was better at explaining the SP175 reference set than the competing schemes. We also found that the preferred model included just 3 basis spectra, which suggests that attempting to predict more than 3 types of structure will lead to much greater uncertainty in the estimation. By looking at the normalised measure  $\zeta$  we showed that CD data between 175–260 nm contain only enough information to assign 3 secondary structure motifs (some version of  $\alpha$ -helix and  $\beta$ -strand, and the rest). In contrast to the general consensus, our work indicates that although CD spectra to lower wavelengths do contain more information about protein structure, it cannot be translated into increasing numbers of traditional well-defined structures that can be determined. We would advise using data down to at least 195 nm, with lower cut-offs slightly improving the structure fitting. In practice more structures can be assigned if the reference set is reduced to include only spectra similar to the unknown protein.

We have showed the importance of capturing the correct covariance structure within a spectrum. Our Bayesian approach accounts carefully for the uncertainty in measurement as well as the unknown basis spectra. Three basis spectra and their uncertainty envelopes were generated. The experimental uncertainty could be removed by replacing the smoothed averaged reference and test spectra in our calculations with multiple individual repeats, preferably from different experiments and instruments. The result would be the spectral/structural variation for a revised Fig. 1(a). With replicate spectra the model could be further developed by adding a multivariate random effect associated with each protein, so that it would become possible to quantify the variation due to measurement error and poor model fit separately.

Whether our Bayesian method identifies two spectra as different depends strongly on the kind of noise in the two spectra and we were able to distinguish between a typical CD error, inferred from the reference set, and another kind of noise, *i.e.* a Gaussian random error. This comparison method has potential wide applications, as a suitable tool to monitor structural changes in protein screening assays, production processes or drug discovery processes.

A second direction for future development would be to combine data from different techniques such as linear dichroism, infrared absorbance spectroscopy, Raman, NMR

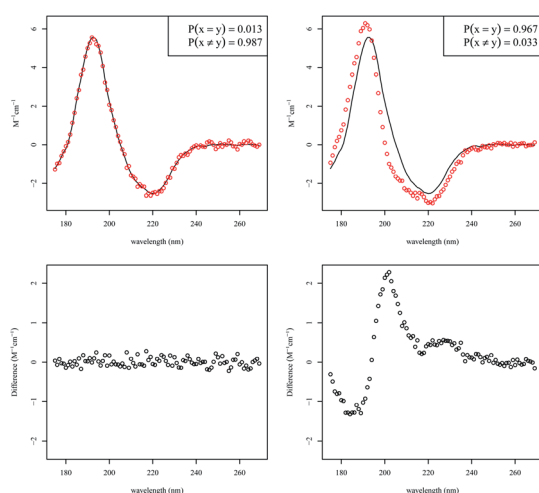


Fig. 4 Top row: plots of the scrY spectrum (black) and the scrY spectrum with added noise (red). Bottom row: plots of the difference between the two spectra. Left column: Gaussian white noise, right column: multivariate Gaussian noise with covariance structure representing usual experimental variability (see text for details).





*etc.*<sup>3,39–46</sup> that might provide orthogonal information about protein structure. By fully characterising the measurement uncertainty with each technique, our Bayesian approach provides a natural way to combine and to correctly weight information across techniques, unlike existing approaches<sup>28,47</sup> in which the influence of each technique depends only on the relative numbers of points observed in each spectrum.

## Conflicts of interest

There are no conflicts to declare.

## Acknowledgements

Jacopo Franco's contributions to the early stages of this project are gratefully acknowledged, however he is unaware that this paper is being submitted for publication as we have no contact details for him. Therefore, he does not take any responsibility for the contents. JF is grateful for funding from the Centre for Analytical Science Marie Curie Innovative Doctoral Programme, funded by the EU, whilst at the University of Warwick.

## References

- 1 J. T. Pelton and L. R. McLean, *Anal. Biochem.*, 2000, **277**, 167–176.
- 2 S. M. Kelly, T. J. Jess and N. C. Price, *Biochim. Biophys. Acta*, 2005, **1751**, 119–139.
- 3 B. Nordén, A. Rodger and T. R. Dafforn, *Linear dichroism and circular dichroism: a textbook on polarized spectroscopy*, Royal Society of Chemistry, Cambridge, 2010.
- 4 R. W. Woody, in *Circular dichroism principles and applications*, ed. K. Nakanishi, N. Berova and R. W. Woody, VCH, New York, 1994.
- 5 N. Sreerama and R. W. Woody, *Anal. Biochem.*, 1993, **209**, 32–44.
- 6 W. J. Johnson, *Annu. Rev. Biophys. Biophys. Chem.*, 1988, **17**, 145–166.
- 7 C. Bertucci, M. Pistolozzi and A. De Simone, *Anal. Bioanal. Chem.*, 2010, **398**, 155–166.
- 8 J. G. Lees, A. J. Miles, F. Wien and B. A. Wallace, *Bioinformatics*, 2006, **22**, 1955–1962.
- 9 L. Whitmore, B. Woollett, A. J. Miles, D. Klose, R. W. Janes and B. A. Wallace, *Nucleic Acids Res.*, 2011, **39**, D480–D486.
- 10 K. A. Oberg, J.-M. Ruysschaert and E. Goormaghtigh, *Protein Sci.*, 2003, **12**, 2015–2031.
- 11 N. Sreerama and R. W. Woody, *Anal. Biochem.*, 2000, **287**, 252–260.
- 12 L. A. Compton and W. C. Johnson, *Anal. Biochem.*, 1986, **155**, 155–167.
- 13 P. Manavalan and C. W. Johnson, *Anal. Biochem.*, 1987, **167**, 76–85.
- 14 I. H. van Stokkum, H. J. Spoelder, M. Bloemendal, R. van Grondelle and F. C. Groen, *Anal. Biochem.*, 1990, **191**, 110–118.
- 15 M. A. Andrade, P. Chacon, J. J. Merelo and F. Moran, *Protein Eng.*, 1993, **6**, 383–390.
- 16 V. Hall, A. Nash, E. Hines and A. Rodger, *J. Comput. Chem.*, 2013, **34**, 2774–2786.
- 17 V. Hall, M. Sklepari and A. Rodger, *Chirality*, 2014, **26**, 471–482.
- 18 N. P. Chmel, P. Scott and A. Rodger, *Chirality*, 2012, **24**, 699–705.
- 19 R. C. Aster, B. Borchers and C. H. Thurber, *Parameter estimation and inverse problems*, Academic Press, 2011.
- 20 P. J. Brown, M. Vannucci and T. Fearn, *J. Roy. Stat. Soc. B Stat. Methodol.*, 1998, **60**, 627–641.
- 21 A. Zellner, *Bayesian inference and decision techniques: Essays in honor of Bruno De Finetti*, 1986, vol. 6, pp. 233–243.
- 22 E. George and D. P. Foster, *Biometrika*, 2000, **87**, 731–747.
- 23 D. G. T. Denison, *Bayesian methods for nonlinear classification and regression*, Wiley, Chichester, 2002.
- 24 R. P. Joosten, T. A. Te Beek, E. Krieger, M. L. Hekkelman, R. W. Hooft, R. Schneider, C. Sander and G. Vriend, *Nucleic Acids Res.*, 2011, **39**, D411–D419.
- 25 W. Kabsch and C. Sander, *Biopolymers*, 1983, **22**, 2577–2637.
- 26 N. Sreerama, S. Y. Venyaminov and R. W. Woody, *Protein Sci.*, 1999, **8**, 370–380.
- 27 A. Micsonai, F. Wien, L. Kernya, Y.-H. Lee, Y. Goto, M. Refregiers and J. Kardos, *Proc. Natl. Acad. Sci. U. S. A.*, 2015, **112**, E3095–E3103.
- 28 K. A. Oberg, J.-M. Ruysschaert and E. Goormaghtigh, *Eur. J. Biochem.*, 2004, **271**, 2937–2948.
- 29 P. Touloupou, N. Alzahrani, P. Neal, S. E. Spencer and T. J. McKinley, *Bayesian Analysis*, 2019, 1–32.
- 30 M. Clyde, J. Berger, F. Bullard, E. Ford, W. Jfferys, R. Luo, R. Paulo and T. Lored, *Statistical Challenges in Modern Astronomy IV*, 2007, vol. 371, p. 224.
- 31 J. G. Lees, A. J. Miles, R. W. Janes and B. A. Wallace, *BMC Bioinf.*, 2006, **7**, 507.
- 32 J. P. Hennessey and W. C. Johnson, *Biochemistry*, 1981, 1085–1094.
- 33 P. Manavalan and W. C. Johnson, *Proc. Int. Symp. Biomol. Struct. Interactions, Suppl. J. Biosci.*, 1985, **8**, 141–149.
- 34 J. G. Scott and J. O. Berger, *Ann. Stat.*, 2010, **38**, 2587–2619.
- 35 S. E. Spencer, S. M. Hill and S. Mukherjee, *Ann. Appl. Stat.*, 2015, **9**, 507–524.
- 36 B. A. Wallace and R. W. Janes, *Curr. Opin. Chem. Biol.*, 2001, **5**, 567–571.
- 37 M. G. Cox, J. Ravi, P. D. Rakowska and A. E. Knight, *Metrologia*, 2014, **51**, 67.
- 38 A. Abdul-Gader, A. J. Miles and B. A. Wallace, *Bioinformatics*, 2011, **27**, 1630–1636.
- 39 A. Rodger, M. J. Steel, S. C. Goodchild, N. P. Chmel and A. Reason, *Q. Rev. Biophys.*, 2020, **1**, e8.
- 40 M. Sklepari, A. Rodger, A. Reason, S. Jamshidi, I. Prokes and C. A. Blindauer, *Anal. Methods*, 2016, **8**, 7460–7471.
- 41 M. Kinalwa, E. W. Blanch and A. J. Doig, *Protein Sci.*, 2011, **20**, 1668–1674.
- 42 M. Pinto-Corujo, M. Sklepari, D. Ang, M. Millichip, A. Reason, S. Goodchild, P. Wormell, D. P. Amarasinghe, R. Dukor, V. Lindo, N. P. Chmel and A. Rodger, *Chirality*, 2018, **30**, 957–965.



- 43 P. I. Haris, in *Encyclopedia of Biophysics*, ed. G. K. Roberts, European Biophysical Societies' Association, 2013.
- 44 K. K. Chittur, *Biomaterials*, 1998, **19**, 357–369.
- 45 F. Dousseau, M. Therrien and M. Pezolet, *Appl. Spectrosc.*, 1986, **43**, 538–542.
- 46 J. Rajendra, A. Damianoglou, M. Hicks, P. Booth, P. Rodger and A. Rodger, *Chem. Phys.*, 2006, **326**, 210–220.
- 47 B. M. Bulheller, A. Rodger and J. D. Hirst, *Phys. Chem. Chem. Phys.*, 2007, **9**, 2020–2035.

