SESCA: Predicting the Circular Dichroism Spectra of Proteins from Molecular Structure

Gabor Nagy[1], Søren V. Hoffmann[2], Nykola C. Jones[2], Helmut Grubmüller[1*]

[1]: Department of Theoretical and Computational Biophysics, Max Planck Institute for Biophysical Chemistry, Am Fassberg 11, D-37077 Göttingen, Germany

[2]: ISA, Department of Physics & Astronomy, Aarhus University, Ny Munkegade 120, DK 8000 Aarhus C, Denmark

,

[*]Corresponding author
Email: hgrubmu@gwdg.de

Keywords: protein structure, CD spectrum prediction, semi-empirical, secondary structure

Abstract

Circular dichroism spectroscopy is a highly sensitive, but low-resolution technique to study the structure of proteins. Combed with molecular modelling and other complementary techniques, CD spectroscopy can also provide essential information at higher resolution. To this aim, we introduce a new computational method to calculate the electronic circular dichroism spectra of proteins from a three dimensional-model structure or structural ensemble. The method determines the CD spectrum from the average secondary structure composition of the protein using a pre-calculated set of basis spectra. We derived several basis spectrum sets obtained from the experimental CD spectra and secondary structure information of 71 reference proteins and tested the prediction accuracy of these basis spectrum sets through cross-validation. Furthermore, we investigated how prediction accuracy is affected by contributions from amino acid side chain groups and protein flexibility, potential experimental errors of the reference protein spectra, as well as the choice of the secondary structure classification algorithm and the number of basis spectra. We compared the predictive power of our method to previous spectrum prediction algorithms – such as DichroCalc and PDB2CD – and found that SESCA predicts the CD spectra with up to 50% smaller deviation. Our results indicate that SESCA basis sets are robust to experimental error in the reference spectra, and the choice of the secondary structure classification algorithm. For over 80% of the globular reference proteins, SESCA basis sets could accurately predict the experimental spectrum solely from their secondary structure composition. To improve SESCA predictions for the remaining proteins, we applied corrections to account for intensity normalization, contributions from the amino side chains, and conformational flexibility. For globular proteins only intensity scaling improved the prediction accuracy significantly, but our models indicate that side chain contributions and structural flexibility are pivotal for the prediction of shorter peptides and intrinsically disordered proteins.

Author summary

Proteins are biomolecules that perform almost all of active task in living organisms, and how they perform these task is defined by their structure. By understanding the structure of proteins, we can alter and regulate their biological functions, which may lead to many medical, scientific, and technological advancements. Here we present SESCA, a new method that allows the assessment, and refinement of protein model structures. SESCA predicts the expected circular dichroism spectrum of a proposed protein model and compares it to an experimentally determined CD spectrum, to determine the model quality. CD spectroscopy is an experimental technique that is very sensitive to the secondary structure of the protein, and widely used as a quality control in protein chemistry.

We demonstrate that our method can accurately and robustly predict the spectrum of globular proteins from their secondary structure, which is necessary for a rigorous model assessment. The SESCA scheme can also address protein flexibility and contributions from amino acid side chains, which further enhance the accuracy of the method. In addition, this allows SESCA predictions to target disordered proteins. For these proteins, flexibility is part of their function, but it also renders their structural characterization much more challenging.

Introduction

Electronic circular dichroism (CD) spectroscopy is a widely applied optical method to study the structure and structural changes of biomolecules such as proteins, nucleic acids, and carbohydrates [1]. CD spectroscopy is a very sensitive tool, often used as a quality control of recombinant proteins or to monitor changes of the protein structure during folding, aggregation, and binding events. Because of this sensitivity, CD spectroscopy does not require large amounts of protein or special labelling and can be readily used in aqueous solutions. These qualities also render CD spectroscopy a good tool for verifying proposed structural and mechanistic models for proteins, provided that a direct, quantitative comparison is possible between the models and the observed spectra.

The CD spectra of proteins in the far ultraviolet (UV) range (180-250 nm) depend strongly on the backbone conformation, and therefore, on their secondary structure [2–5]. The main contributor to a protein's CD spectrum is the electronic excitation of the partially delocalized peptide bonds, which form the backbone of the polypeptide chain. Isolated amino acids, except glycine, also show a CD signal in this wavelength range [6–8].-Therefore, amino acid side chains contribute to the protein CD spectrum as well, although this contribution is typically smaller than that of the protein backbone. Since the 1980‘s, several methods have been proposed to quantitatively connect the secondary structure composition of a protein and its CD spectrum. CD spectra were collected and compiled into data banks and reference data sets [9,10] to improve and assess the quality of predictions. Two major categories of methods - spectrum deconvolution and spectrum prediction - were established to provide quantitative predictions related to CD spectra. Spectrum deconvolution methods aim at predicting the secondary structure of a protein from its CD spectrum. Spectrum prediction methods, *vice versa*, determine the CD spectrum from the structure, often by quantum mechanics (QM) calculations, or QM-derived parameters (*ab initio* methods).

Deconvolution of CD spectra is a very convenient method of gaining structural information on proteins as it requires no special labelling or crystallization, and several different approaches (e.g. CCA, K2D3, BestSel) have been developed and implemented for it [11–13]. The measured CD spectrum is decomposed into a linear combination of basis spectrum components (basis spectra). The basis spectra usually reflect the CD signal of secondary structure elements, and are derived either from the CD spectra of model peptides or from a larger set of reference proteins with known CD spectra and secondary structure composition. Once derived, they are used to estimate the secondary structure composition of proteins with unknown structure by fitting a linear combination of basis spectra to the measured CD spectrum. The main drawback of this approach is the fitting procedure which is sensitive to experimental error of measured the CD spectrum. In the absence of additional information, different secondary structure estimates may provide fits of similar quality, which renders the comparison to model structures difficult.

Ab initio spectrum prediction methods typically require advanced time-dependent QM or density functional methods [14–16]. The large computational effort limits such calculations to rather small peptides, especially because the CD signal is sensitive to the conformation of the molecule as well as the structure and fluctuations of several solvent shells. A simplified algorithm based on *ab initio* calculations, called the matrix method [17], was implemented in the program DichroCalc [18]. DichroCalc can determine the most important features of the CD spectrum of a protein based on its conformation, albeit with limited accuracy. Recently, a new empirical spectrum prediction algorithm named PDB2CD [19] was proposed which combines secondary and tertiary structure information obtained from a three-dimensional structure of the protein to predict its CD spectrum. PDB2CD is based on a representative set of globular proteins, where the predicted CD spectrum is calculated as the weighted average of spectra from structurally similar proteins. By combining structural and spectral information, this web-based empirical implementation achieved significantly improved accuracy.

Generalizing this approach here, we developed and cross-validated a semi-empirical method to predict the CD spectra of proteins from their three dimensional structures using empirically derived basis spectra. Our approach combines the structural and spectral information of a reference protein set to systematically derive structure-related basis spectra.

The basis spectra are then used to predict the CD spectra of proteins based on their three dimensional structure, or to determine how well proposed structural models agree with the measured spectrum. This Semi-Empirical Spectrum Calculation Approach (SESCA) is computationally efficient and allows accurate prediction of protein CD spectra both from a single protein structure as well as from a set or an ensemble of structures to account for structural flexibility. We compare the main steps of the SESCA scheme, spectrum deconvolution, and ab initio spectrum prediction methods in Fig. 1.

In this study, our approach will be evaluated and optimized using multiple, freely available structure classification algorithms. In addition, we will address the effects of structural flexibility as well as the contribution of amino acid side chains in the far UV region. SESCA eliminates the uncertainty of deconvolution based reconstructions, predicts the experimental CD spectra of globular proteins more accurately than DichroCalc, and matches the accuracy of PDB2CD. In addition, the increased calculation efficiency gained from using pre-calculated basis spectra renders SESCA more suitable for calculating the CD spectra from structural ensembles. This advantage is particularly important for the ensemble refinement of disordered proteins where model verification by comparison to experimental observables is crucial.

Theoretical background

2.1 Semi-empirical spectrum calculations

Here, we describe our semi-empirical CD prediction method (Fig. 2), and summarize our optimization and cross-validation procedure (Fig. 3). We will initially assume that the CD spectra are mainly determined by the local conformation of the peptide bonds, and subsequently also consider the effects of the amino acid side chain groups. In each case, the local backbone conformation will first be grouped into secondary structure elements with established methods (Fig. 2A) and secondly, these secondary structure elements will be combined into broader classes (Fig. 2B) for which basis spectra are determined (Fig. 2C). The CD spectra of proteins will be calculated from weighted averages of the basis spectra (Fig. 2D), each reflecting the CD signal of one of the secondary structure classes averaged over all other conformational degrees of freedom, such as solvent shell arrangements, side-chain conformers, and local conformational variations of the protein backbone.

We will derive and assess several basis spectrum sets – henceforth referred to as "basis sets" – according to the scheme shown in Fig. 3. The secondary structure elements from five different available secondary structure classification methods will be combined into classes in two different ways ("hard" and "soft" optimization). The optimal basis spectra $B_i(\lambda)$ will be derived for each secondary structure class *i*, such that the reference CD spectra $S_j(\lambda)$ measured for *N* globular proteins of a reference set are approximated by a weighted sum of *F* basis spectra

$$S_j(\lambda) = \sum_{i=1}^{F} C_{ji}\, B_i(\lambda) \qquad (1)$$

as accurately, as possible measured by the "fitting" accuracy. The fitting accuracy is quantified by the average root-mean-squared deviation (RMSD) between the calculated and experimental reference spectra. For each obtained optimal basis set, cross-validation against measured CD spectra that have not been used for the optimization will be carried out to determine its prediction accuracy.

To calculate the coefficients for the basis spectra C_{ji} we utilize W_{jk}, the fraction of residues classified as secondary structure element *k* in a structural model of protein *j*. Grouping secondary structure elements into secondary structure classes *i* is achieved via an assignment matrix **A**=$\{\alpha_{ki}\}$, combining the *K* secondary structure elements into *F* structural classes, such that

$$C_{ji} = \sum_{k=1}^{K} W_{jk}\, \alpha_{ki} \ . \qquad (2)$$

This assignment is also subject to optimization, and the constraints on the assignment matrix separate the hard and soft optimization approaches. In the more conventional hard approach, each secondary structure element is assigned to exactly one structural class (and, therefore basis spectrum), indicated by entries "0" and "1" in the assignment matrix (e.g. Fig. 2C). In the more general soft approach, the secondary structure elements are assigned to multiple structural classes and the assignment factors α_{ki} can be any real number.

Combining the above two equations relates the CD spectrum of a protein to its secondary structure composition

$$S_j(\lambda) = \sum_{k=1}^{K} \sum_{i=1}^{F} W_{jk}\, \alpha_{ki}\, B_i(\lambda), \qquad (3)$$

such that for N reference proteins j with known CD spectra $S_j^{\mathrm{exp}}(\lambda)$, secondary structure composition W_{jk}, and a given assignment α_{ki} , the optimal basis spectra $B_i(\lambda)$ are readily calculated from minimizing $RMSD_{\mathrm{set}}$, the root-mean-squared deviation between the measured spectra and those calculated from the secondary structure $S_j^{\mathrm{calc}}(\lambda)$,

$$RMSD_{\mathrm{set}} = \frac{1}{N} \sum_{j=1}^{N} \sqrt{\int_{\lambda\mathrm{min}}^{\lambda\mathrm{max}} \left[S_j^{\mathrm{calc}}(\lambda) - S_j^{\mathrm{exp}}(\lambda)\right]^2 d\lambda}. \qquad (4)$$

We note that in spectrum deconvolution methods [11,12,20] basis spectra are derived via the same notion, albeit applied in reverse direction. Whereas in deconvolution methods, the basis spectrum coefficients are treated as fit parameters which yield the secondary structure content (as shown in Fig.1A), in our approach the secondary structure fractions are extracted from the known structure and combined into the basis spectrum coefficients. By calculating the spectrum from the structure, our method avoids the (numerically often unstable) fitting procedure, and rather proceeds by direct comparison to the CD spectrum as the primary experimental observable (as depicted in Fig. 1B). In this respect it resembles *ab initio* methods (shown in Fig. 1C).

We also note that the level of coarse graining of secondary structure information is given by the assignment matrix α_{ik}. Extreme cases are (a) combining all secondary structure elements provided by the particular secondary structure classification method in use into F=1 class, and (b) into F=K classes. In case (a), only very little (likely too little) information is retained – typically the α-helical content – whereas in the "naive" case (b), the full secondary structure information is provided with the possible risk of over-fitting. Therefore, subsequent cross validation is crucial for determining the optimal level of coarse graining.

Finally, we note that the hard combination of secondary structure elements is a special case of the more general soft combination approach and therefore, one might expect the latter to yield more accurate calculated spectra for the reference proteins from the same amount of structural information. Because in the soft optimization approach the assignment factors α_{ki} can adopt any real number without further constraints, eq. 2 yields linear combinations of the secondary structure fractions W_{kj}. Hence, each basis spectrum $B_i(\lambda)$ can be understood as a ”collective" secondary structure class, such as "0.3 α-helical + 0.7 β-sheet". Of course, the collective secondary structure classes introduce another layer of complexity to the optimization problem, and therefore increase the chances of over-fitting the basis spectra.

2.2 Basis spectrum optimization: "Hard approach"

For the hard basis set optimization approach (Fig. 3A), our aim was to find basis spectrum sets that provide the most accurate prediction of protein CD spectra. To trade-off the fitting accuracy for reduced over-fitting, we applied a Monte Carlo (MC) approach with a cross-validation, during the search for assignments and the number of basis spectra. To this aim, the reference protein reference set was divided into two sub-sets. The larger sub-set (training set) was used to derive the basis spectra, and the basis set accuracy was evaluated by the average RMSD of the calculated CD spectra of the smaller sub-set (evaluation set) according to eq. 4. During each optimization cycle, random changes were applied to the assignment matrix, the corresponding basis spectra for the given assignment were calculated (described in Section 2.3), and the new assignment was accepted or rejected the change based on its effect on the obtained basis set accuracy of the evaluation set (RMSD$_{\mathrm{eval}}$). At the end of the optimization, the five assignments with the lowest RMSD$_{\mathrm{eval}}$ and the complete reference set were used to fit basis spectra and obtain the final optimized basis sets. These basis sets were subsequently assessed by cross-validation (Fig. 3C) on a protein set not used in the optimization procedure (cross-validation set) to estimate their prediction accuracy (RMSD$_{\mathrm{cross}}$), and by calculating their fitting accuracy (RMSD$_{\mathrm{ref}}$) on the reference set (Fig. 3D).

We imposed two constraints on the assignment factors of the hard basis sets: 1) $\sum_{k=1}^{K} \alpha_{ki} = 1$, and 2) $\alpha_{ki} \in \{1,0\}$. These constraints ensured that the resulting basis spectra are normalized, and that there are no overlaps between the structural classes the basis spectra represent, significantly reducing the search space of the MC algorithm. Initially, the hard optimization procedures were started from a naïve assignment (F=K) for each classification method, in which case **A** is the identity matrix (α_{ki} is 1 if i=j and 0 otherwise). However, the basis sets resulting from the first optimization were used as initial guesses for subsequent optimization rounds until convergence was reached both for the number of basis spectra and RMSD$_{\mathrm{eval}}$.

2.3 Calculation of basis spectra

For a given assignment matrix **A**, coefficients of the basis spectra C_{ji} are readily calculated via eq. 2 from the fraction of secondary structure elements W_{jk}. The basis spectra $B_i(\lambda)$ are derived using eq. 1 independently for each available wavelength λ from a sufficiently large training set of protein structures and their CD-spectra $S_j(\lambda)$. Because typically the number of basis spectra F is smaller than the number of available training spectra N (here, F=1…20 and N=64*)*, eq. 1 represents an over-determined linear equation system. The basis spectra that minimize the average RMSD between calculated and experimental CD spectra according to eq. 4, where $S_j^{\text{calc}}(\lambda) = \sum_{i=1}^{F} C_{ji}\, B_i(\lambda)$, are obtained via

$$\boldsymbol{b}(\lambda) = (\boldsymbol{C^T}\,\boldsymbol{C})^{-1}\,\boldsymbol{C}^T\,\boldsymbol{s}(\lambda). \qquad (5)$$

We have used matrix notation for the coefficients **C = {**C_{ij}**}** and the vector notation for the basis spectra **b**(λ)={B_i(λ)}, and CD spectra **s**(λ)={S_j(λ)}, respectively. Figures 2 and S1-S14 show basis spectrum sets that were derived by determining the basis set coefficients for different assignment and applying eq. 5 on the far UV (175-269 nm) wavelength range sampled in 1 nm steps, for all 64 proteins in the TR64 set (see section 3.1).

2.4 Assignment optimization details

In this section, we describe how the changes in the secondary structure element assignment were evaluated during the MC search. During each hard optimization step, a random change was introduced to the assignment matrix **A**, by reassigning one of the secondary structure elements to another structural class. Then, the basis spectra $B_i(\lambda)$ were recalculated and the average deviation (RMSD$_{\text{eval}}$) from the experimental CD spectra was computed for the evaluation set both before and after the change was applied. If $e^{-\beta*(\Delta RMSD_{eval})}$ was larger than a randomly generated number between 0 and 1, the new assignment was accepted, otherwise rejected. In the next optimization step, a new random change was applied to the last accepted assignment. The acceptance ratio in this notation was controlled by β, the strictness parameter determining how often changes with an unfavourable $\Delta RMSD_{eval}$ are accepted. By default, β = 8.0 was applied to optimizations, which was lowered (down to 1.0) if the acceptance rate in an optimization dropped below 20%. Accepted assignments with the lowest five $RMSD_{eval}$ during the MC search were saved and used to calculate the basis spectra of optimized basis sets.

The search space for the hard optimization contains F^K possible **A** matrices, where *F* is the number of structural classes/basis spectra and *K* is the number of the secondary structure elements. For example, assigning five structural elements to three classes defines a search space of 3^5 = 243 assignments, whilst 19 structural elements assigned to 10 classes result in a search space of 10^{19}. When optimizing small basis sets with 5-8 secondary structure elements, a single optimization process with 500 accepted moves was sufficient to completely explore the search space, often visiting the global optimum of the assignment space multiple times. In the case of more than 10 structural elements, several 10000-step optimizations were started from multiple initial assignments described in Section 3.3. In these cases, assignments resulting from the initial optimization procedure were used to start new parallel processes to more effectively explore the search space. To further increase the efficiency of the hard optimization, important secondary structure elements – such as the α-helix and at least one of β-strand elements – were assigned to different classes and then excluded from being reassigned (effectively decreasing K). In addition, if the move resulted in a more favourable $RMSD_{eval}$, both structural classes with no assigned secondary structure elements and the secondary structure elements themselves could be temporarily eliminated from the basis set. Eliminated classes and secondary structure elements could be reintroduced to the basis set through random changes during the same optimization process, and missing secondary structure elements were reintroduced between subsequent optimization processes to conserve the normalization of basis spectra. We have performed several optimization processes for each secondary structure classification method, until the number of basis spectra in the best optimized basis sets stabilized, and RMSD$_{\mathrm{ref}}$ values similar to the soft basis sets of the same basis set size were reached (described below).

2.5 Basis set determination: The "soft approach"

The hard optimization scheme introduced in Sections 2.2-2.4 is limited to a restricted assignment factor space ($\alpha_{ki} \in \{0,1\}$) and, therefore, it should be possible to further improve the accuracy of reconstructing the CD spectra from the secondary structure information by removing this limitation. Accordingly, in our more general soft optimization approach, the assignment factors can be any real number ($\alpha_{ki} \in R$). During the soft optimization, we simultaneously derived the basis spectra and assignment factors that most accurately reproduced the CD spectra of the reference protein data set (best fitting accuracy). Consequently, besides the spectral and structural information of the reference data set, only the desired number of basis spectra is specified for the soft optimization, and no "internal" cross-validation is required to trade-off the accuracy of the fit for an improved general predictive power. To obtain the optimal basis sets, the non-linear equation system defined by eqs. 3 and 4 has to be solved simultaneously for all wavelengths of each protein spectrum in the reference data set. In matrix notation, this optimization problem reads as

$$\left|\left|\boldsymbol{W}\,\boldsymbol{A}\,\boldsymbol{B}-\boldsymbol{S}\right|\right|^{2} \overset{!}{=} \min, \qquad (6)$$

where $\mathbf{S}$=(S_{jl}) and $\mathbf{W}$(=W_{jk}) are the matrices containing the spectral and structural information of the reference set, respectively, and the matrix $\mathbf{B}$ ={B_{il}} describes the basis spectra. The matrix elements S_{jl} and B_{il} are obtained by discretizing the experimental CD spectra $S_j(\lambda)$ and basis spectra $B_i(\lambda)$ at L wavelengths. This optimization problem is solved simultaneously for the matrices **A** and **B** by setting their element-wise matrix derivatives to zero:

$$\frac{\partial}{\partial \boldsymbol{A}}\mathrm{tr}[(\boldsymbol{W}\,\boldsymbol{A}\,\boldsymbol{B}-\boldsymbol{S})^{T}\ (\boldsymbol{W}\,\boldsymbol{A}\,\boldsymbol{B}-\boldsymbol{S})] =$$

$$2\,\boldsymbol{B}\,\boldsymbol{B}^{\mathbf{T}}\,\boldsymbol{A}^{\mathbf{T}}\,\boldsymbol{W}^{\mathbf{T}}\,\boldsymbol{W}-2\,\boldsymbol{B}\,\boldsymbol{S}^{\mathbf{T}}\,\boldsymbol{W} \overset{!}{=} 0 \qquad (7)$$

$$\frac{\partial}{\partial \boldsymbol{B}}\mathrm{tr}[(\boldsymbol{W}\,\boldsymbol{A}\,\boldsymbol{B}-\boldsymbol{S})^{T}\ (\boldsymbol{W}\,\boldsymbol{A}\,\boldsymbol{B}-\boldsymbol{S})] =$$

$$2\,\boldsymbol{B}^{\mathbf{T}}\,\boldsymbol{A}^{\mathbf{T}}\,\boldsymbol{W}^{\mathbf{T}}\,\boldsymbol{W}\,\boldsymbol{A}-2\,\boldsymbol{S}^{\mathbf{T}}\,\boldsymbol{W}\,\boldsymbol{A} \overset{!}{=} 0 \qquad (8)$$

which, yields two coupled non-linear matrix equations

$$\boldsymbol{A} = (\boldsymbol{W}^{\mathbf{T}}\,\boldsymbol{W})^{-\mathbf{1}}\,\boldsymbol{W}^{\mathbf{T}}\,\boldsymbol{S}\,\boldsymbol{B}^{\mathbf{T}}\,(\boldsymbol{B}^{\mathbf{T}}\,\boldsymbol{B})^{-\mathbf{1}} \qquad (9)$$

and

$$\boldsymbol{B} = (\boldsymbol{A}^{\mathbf{T}}\,\boldsymbol{W}^{\mathbf{T}}\,\boldsymbol{W}\,\boldsymbol{A})^{-\mathbf{1}}\,\boldsymbol{A}^{\mathbf{T}}\,\boldsymbol{W}^{\mathbf{T}}\,\boldsymbol{S} \qquad (10)$$

Equations 9 and 10 are solved iteratively, starting from a random generated matrix **A** ($0.0 \leq \alpha_{ki} \leq 1.0$) to obtain an initial **B** via eq. 10, which is inserted into eq. 9 to obtain an improved **A**, repeated until convergence. A summary of the soft optimization scheme is shown in Fig. 3B

This soft optimization procedure was systematically repeated for each secondary structure classification method K times to obtain optimized basis sets with 1-K basis spectra (K being the number of secondary structure elements in the classification method). These series of basis sets determine the best fitting accuracy as the function basis set size and secondary structure classification. For each optimization procedure, the convergence criterion was to reach less than Y = 0.0001 x 10^3 deg cm^2/dmol change between iterations in the average RMSD of the CD spectra calculated for the reference set (ΔRMSD$_{\mathrm{ref}}$).

2.6 Spectral component analysis

The overall accuracy of our method is limited by two factors, first, the information content of the secondary structure composition, and second, the applicability of linear combinations of basis spectra in approximating the experimental CD spectra. The first factor was addressed by our soft optimization approach (section 2.5). The second factor determines an upper limit for the fitting accuracy (lowest $\mathrm{RMSD}_{\mathrm{ref}}$) given a set of reference CD spectra and the number of used basis spectra. To this aim, we carried out a principal component analysis (PCA) on CD spectra of the SP175 reference set (see Section 3.1). PCA is a mathematical method to describe a (multidimensional) data set of N members by a basis set of N orthogonal principal component (PC) vectors. How much the data points differ from the average of the set (the variance of the data set) along a PC vector is quantified by its eigenvalue. It is possible to describe a data set with just a few (F) PC vectors of the highest eigenvalues (dimensionality reduction) [21], which – by construction – retains the maximum possible variance of the data set, and consequently, provides the reconstruction with the smallest possible deviation. Here, we used PCA to describe the reference CD spectra (a set of L dimensional data points) by basis sets constructed from 1-10 PC vectors of the highest eigenvalues. The basis spectrum coefficients (C_{ij}) of the protein j for these basis sets were defined as the projection of the CD spectrum along the particular PC vector i (described in Section 3.5). Figure 4 shows the first ten principal components with highest eigenvalues, the fitting accuracy (R_i) of reconstructions for selected CD spectra, as well as the SP175 protein set on average. Note that this analysis is based solely on CD spectra of the reference data set, and does not account for any possible source of inaccuracy related to structure, secondary structure calculations, or scaling errors within the reference set.

Materials and Methods

3.1 Structures and CD spectra used for calibration

To derive and assess the required basis sets for our CD spectrum calculation method, several protein data sets were compiled of which both the CD spectra and the structure of the proteins were experimentally determined. We used seven protein data sets throughout this study, for which comprehensive lists are provided in supplementary materials (Tables S1-S3).

The protein data set SP175 (Table S1) was the standard reference set to determine basis sets derived only from secondary structure information. It also represented globular proteins, e.g. during the principal component analysis of protein CD spectra, as was used to determine the fitting accuracy of all SESCA basis sets. This data set is comprised of 71 globular protein structures and their corresponding CD spectra, assembled by Lees *et al.* [10] such that its secondary structure distribution reflects that of the full collection of proteins in the protein databank [22] (PDB). In addition, the proteins for SP175 were selected according to the following criteria: 1) high resolution PDB structure available (average resolution 1.9 Å), 2) high quality CD spectrum available (wavelength range 175-269 nm), 3) the set represents the major protein folds as defined by the CATH [23] database, 4) the set covers proteins with diverse secondary structure compositions.

The SP175 data set was divided into two sub-sets for the hard optimization approach, a larger training set for calculating the basis spectra, and smaller evaluation set for testing the predictive power of basis set. The second protein set termed TR64 is comprised of 64 proteins, was the standard training set for the hard basis spectrum optimization approach. The third data set is labelled EV9 (Table S2), and was used as the standard evaluation set for the hard basis spectrum optimization. The EV9 set consists of nine proteins, seven of which were part of SP175, and two additional proteins with a β-sheet architecture. The evaluation set was selected such, that it contains three α-helical proteins, three β-sheet containing proteins, and three proteins with an α/β fold. In addition, the proteins of the evaluation set did not contain gaps in the structure, and had to be small enough for visual inspections and quick evaluation during basis set optimizations.

The fourth protein set was used for cross-validation to assess the prediction accuracy of both the hard and soft basis sets (Fig 3). The cross-validation set (Table S3) – labelled TS8 for test set – contains eight globular proteins, which were not part of the previously mentioned data sets. The proteins of the TS8 set were selected from a set of 22 proteins, previously used to determine basis spectrum sets for CD spectrum deconvolution [24]. The CD spectra were obtained from an example spectrum set provided for the deconvolution algorithm CCA by Hollósi *et. al*. [12], whilst their crystallographic structures (crystal structures) were retrieved from the PDB [22]. The globular proteins of the TS8 set had slightly truncated spectra (178-260 nm) compared to the SP175 proteins. The crystal structures did not contain any gaps or missing residues, and had an average resolution of 1.7 Å.

The fifth data set – labelled as GXG20 – consists of the CD spectra and structural ensembles of 20 short peptides with the consensus sequence of Ac-GXG-NH_2 (X stands for any amino acid). This reference set was used to estimate the contribution of amino acid side chains in a protein environment. The CD spectra of these peptides were recorded on the AU-CD beam line at the ASTRID2 synchrotron radiation source in Aarhus Denmark, under similar conditions (298 K, in 50 mM NaF solution with Na_2HPO_4 buffer, pH = 7.1) within the wavelength range of 178-300 nm. Peptide concentrations (0.5-2.0 mg/ml) were determined based on the light absorption at 214 nm [25] and, when possible, at 280 nm (for GYG and GWG). The structural ensembles for each peptide were generated using a 10 μs long molecular dynamics simulation (recorded at every 2 ns) using the GROMACS simulation package [26] (version 5.06) and the Charmm 36M [27] parameter set with explicit TIP3P water modified for the force field. The simulations were performed under periodic boundary conditions on 298 K, with Na^+ and Cl^- ions appropriate for a 50 mM ionic strength and protonation states dominant at pH = 7. The size of the simulation box was chosen such to keep ~2 nm distance between any solute atom and the box boundaries, resulting in a simulation box of ~5500 atoms.

There were two more data sets that were used to derive mixed basis sets which include both backbone (secondary structure) and side chain related basis spectra. The sixth protein set is a sub-set of the SP175 reference set, containing 59 globular proteins that provide a wide variation secondary structure contents, designated as GP59 (globular protein set). The 12 proteins excluded from the SP175 set to form the GP59 set were hard to predict by several spectrum prediction algorithms (see section 5.1) and may have hindered the determination of side chain basis spectra. The seventh data set contained all 20 peptides of the GXG20 data set and the 59 proteins of the GP59 data set, resulting in a mixed polypeptide set with 79 entries (designated as MP79). The MP79 set was used as a reference set to derive the average contribution of side chain groups, as well as our mixed basis sets.

In addition to the protein data sets to derive and cross-validate basis sets, we prepared a system to probe the effects of conformational dynamics has on the quality of predicted CD spectra described in Section 5.2. The chosen system was the complex of CBP-NCBD and P53-AD2, two disordered protein domains which form an ordered crystallisable complex. These protein domains were produced by the company Karebay using solid state peptide synthesis, and the CD spectrum of their 1:1 molar ration complex was measured under the same conditions as described for the peptides of the GXG20 data set. Three structural models were prepared for the P53/CBP complex based on an NMR solution structure obtained from the protein data bank (PDB code 2L14). The three models included the original NMR bundle with 20 conformations, the first extracted conformation of the bundle, and a structural ensemble obtained from a molecular dynamics simulation. The details of the simulation were similar to those described for the peptides of GXG20 reference set, except that the Charmm 22* parameter set [28] was used instead of the Charmm 36M, and the simulation box contained ~82 000 atoms. The simulation was started using the first conformation of the NMR bundle, and protein conformations were recorded after every 10 ns throughout a 10 us long simulation trajectory, resulting in an ensemble of 1000 conformations.

CD spectra in all data sets were converted to Mean Residue Ellipticity (MRE). The CD spectra themselves as well as the deviation between the experimental and calculated spectra in this work are shown in the units of 10^3 degree*cm^2/dmol, abbreviated as kMRE. Prior to the analysis, crystallographic water, non-standard residues, and cofactors were removed from the crystal structures of the data sets. Residue numbers and chain codes were relabelled to ensure compatibility with the analysis software. For all entries of the reference protein sets, the amino acid composition and secondary structure contents were determined (section 3.2). Additionally, CD spectra of globular proteins of the reference sets were also calculated by Dichrocalc and PDB2CD software. A principal component analysis was performed on CD spectra of the SP175 data set to determine the number of necessary spectral components and to probe correlations between the principal components, secondary structure elements and amino acid composition (see sections 3.5 and 3.6).

3.2 Secondary structure determination

The secondary structure of proteins comprising the data sets described in section 3.1 was determined from the protein structure using the algorithms DSSP (Dictionary of Secondary Structure for Proteins) [29] as well as DISICL (DIhedral based Segment Identification and CLassification) [30] and an in-house algorithm HbSS (Hydrogen-bond based Secondary Structure). DSSP is an algorithm based on identifying secondary structure elements based on their distinctive backbone hydrogen-bonding patterns. DSSP classifies each amino acid in the protein as one of the eight secondary structure elements shown in Table S4. The DISICL algorithm classifies tetra-peptide segments of the protein based on two (ϕ,ψ) backbone dihedral angle pairs. The detailed DISICL (DS_det) library contains nineteen secondary structure elements, which are grouped into eight broader secondary structure classes in the simplified DISICL library (DS_sim). Table S5 lists the detailed and simplified DISICL secondary structure elements. The HbSS algorithm was used to distinguish between parallel and antiparallel β-strands (Fig. S15), determined based on backbone hydrogen bonding patterns. In addition, HbSS determined helical and turn-based secondary structure elements (listed in Table S6) similarly to DSSP. Furthermore, the HbSS classification was also extended (HBSS_ext) based on the β-strand twist to determine the amount of left-handed, relaxed (non-twisted) and right-handed β-strands as described in Ref [31] with boundaries of 0° and 23°, respectively, for both parallel and anti-parallel strand arrangements. This extended structural classification is directly comparable to the estimates of the deconvolution algorithm BestSel (Table S7). For comparison, the secondary structure content of each protein was estimated from their CD-spectrum using the deconvolution algorithms SELCON [20] and BestSel [11]. These estimates were also included in the spectral component analysis (section 2.6).

3.3 Initial basis sets

Three deconvolution basis sets (Figs. S16-S18) were used to assess the applicability of our method without extensive optimization. The first basis set (Set_Perczel-1) was derived by Hollósi and Perczel [12] and contains five basis spectra (α-helix. β-strand, Turn type I/III, unordered, and other contributions). The second basis set, determined by Shreerama and Woody (Set_Sreer-1) [32], contains six basis spectra (regular helix, irregular helix, regular strand, irregular strand, poly-proline helix, and disordered). Finally, the third basis set (Set_BestSel-1) was derived for the BESTSEL program by Micsonai and Kardos [11], with eight basis spectra (regular helix, irregular helix, left-handed anti-parallel, relaxed anti-parallel, and right-handed anti parallel β-strands, parallel β-strand, turn structures, and others). For each of these basis spectra, secondary structure elements from the structure classification algorithms (DSSP and DISICL for the first two and DISICL and HbSS for the third) were assigned based on the description of the basis set in their original publications. Once the assignment was complete, the CD spectra for the proteins of the TS8, EV9, TR64 and SP175 sets were calculated using the secondary structure content of their crystal structure and were compared to the experimental spectra.

Furthermore, we derived naive basis sets for the classification algorithms (Figs. S1-S5) DSSP (Set_DSSP-F), simplified and detailed DISICL (Set_DS-simF and Set_DS-detF, respectively), normal and extended HbSS (Set_HBSS-F and Set_HBSS-E) and the deconvolution algorithm BESTSEL (Set_Bestsel-der). These basis sets contained one basis spectrum for each of the algorithm's secondary structure elements, and the SP175 data set was used as a reference set to calculate their basis spectra. These basis sets were used as initial guesses for the hard and soft optimization procedures.

3.4 Spectrum prediction quality

We determined the basis set quality based on the average accuracy of the calculated spectra ($\mathrm{RMSD_{set}}$) for the proteins of the TS8 cross-validation set ($\mathrm{RMSD_{cross}}$) and the SP175 reference set ($\mathrm{RMSD_{ref}}$). However, it was necessary to assess the quality of the calculated spectra for individual proteins as well. The RMSD of a single calculated spectrum of protein *j* (R_j) was determined as the root-mean-square deviation between a spectrum calculated from the structure (S_{jl}^{calc}) and the experimental CD spectrum (S_{jl})

$$R_j = \sqrt{\frac{1}{L} * \sum_{l=1}^{L} \left(S_{jl}^{calc} - S_{jl}\right)^2}. \qquad (11)$$

The indices *j* (1…N) denote the protein, whist *l* (1…L) denote the wavelength of the discretized spectra. By comparing R_j of a protein to the $\mathrm{RMSD}_{\mathrm{set}}$, it was possible to identify the proteins whose the CD spectra are hard to predict using a given methodology. In addition, the standard error of the mean RMSD (SE_{RMSD}) was determined as $SE_{RMSD} = \frac{\sigma}{\sqrt{N}}$, where σ is the standard deviation of R_j within the data set.

3.5 Principal Component Analysis of CD spectra

We performed a PCA on the CD spectra of the SP175 protein reference set, treating each spectrum as an L dimensional vector (where L is the number of wavelengths). The resulting PC vectors were described by the matrix $\mathbf{V}=\{V_{pl}\}$, where the indices *p* (1….P) and *l* (1….L) stand for the principal component (in order of their eigenvalue) and wavelength, respectively. In our case, each $\mathbf{v}_{\mathrm{rp}}$ row vector of the matrix **V** is one of the discretized PC vectors. The spectra of a reference protein data set were reconstructed using the first P={1-10} principal components

$$S_{jl} = S_l^{ave} + \sum_{p=1}^{P} C_{jp} V_{pl}, \qquad (13)$$

where S_{jl} is the circular dichroism of the j^{th} reconstructed protein spectrum at the wavelength *l*, C_{jp} is the projection of that spectrum along the PC vector *p*, V_{pl} and S_l^{ave} are the value of the PC vector and the average CD signal of the data set at wavelength *l*, respectively. The projection of spectrum *j* along the principal component *p* can be calculated by taking the scalar product of the normalized spectrum and the PC vector

$$C_{jp} = (\boldsymbol{s}_{rj} - \boldsymbol{s}^{ave})^T \, \boldsymbol{v}_{rp}. \qquad (14)$$

The vector $\boldsymbol{s}^{\mathrm{ave}} = \{S_l^{\mathrm{ave}}\}$ is the averaged CD spectrum of the data set.

The projections along the PC vectors are analogous to the basis spectrum coefficients. Therefore, Pearson correlation (R$_{\text{pearson}}$) between the secondary structure composition, amino acid composition, and the projections were calculated for the proteins in the SP175 reference set to estimate the importance of these structural descriptors in calculating the CD spectra. The pearson correlation between these descriptors were calculated according to

$$R_{\text{pearson}} = \frac{\sum(X_j - \bar{X})\cdot(Y_j - \bar{Y})}{\sqrt{\sum(X_j - \bar{X})^2} \cdot \sqrt{\sum\left(Y_j - \bar{Y}\right)^2}}, \qquad (15)$$

where X_j and Y_j are either the fraction of an amino acid, the fraction of amino acids classified as a secondary structure element, or the projection of the CD spectrum along a principal component for the protein j, whilst $\bar{X}$ and $\bar{Y}$are the calculated averages for the whole reference set.

3.6 Side chain contributions

To assess the contribution of amino acid side chains, we assumed that the two main contributors to the CD spectra of proteins are the secondary structure and the chromophores of the amino-acid side chains, with no coupling between the side chains and the rest of the protein. This assumption allows the calculation of a backbone independent side-chain correction baseline. The side chain baseline of a protein was determined by the weighted average of the individual side-chain CD signals, where the weighing factor was the corresponding amino acid content for the protein (similarly to eq. 1).

The individual side-chain contributions were estimated from the CD spectra of the MP79 reference set. First, the secondary structure contributions were calculated using an initial basis set (either DS5-4, DS-dT, DSSP-1 or DSSP-T, see the Sup. Mat. for further details on these basis sets) and subtracted from the experimental spectra. Then, the “secondary-structure-free” CD spectra and the amino acid composition of the proteins and peptides were used to derive one basis spectrum for each amino acid side chain. We also derived basis sets with more simplified representations of the side chain contributions. These mixed basis sets were derived from the MP79 reference set in three steps. First, the secondary structure contributions were calculated and subtracted from the CD spectra. Second, basis spectra for the side chains were derived and optimized using the amino acid composition and the secondary-structure free CD spectra of the reference proteins. Third, the side chain contributions were calculated and subtracted from the experimental CD spectra, and these “side-chain free” CD spectra were used to re-optimize the basis spectra for backbone contributions (secondary structure).

The optimization of the side chain and backbone basis spectra was performed by the hard optimization scheme separately (as described in section 2.4) with the following modifications. Before the optimization, the MP79 reference set was separated into six sub-sets (each containing 13 or 14 proteins). In each optimization step, after the secondary structure elements / amino acids were grouped and assigned to basis spectra, one of the MP79 sub-sets was designated as the evaluation set, whilst the rest of the reference proteins were used to derive the basis spectra (as a training set). The derived basis spectra were used to calculate the CD spectra of the evaluation set. This process was repeated six times such that each of the sub-sets was predicted once from the rest of the MP79 reference set. After calculating each of the evaluation sub-sets, their RMSD was averaged and used as $\mathrm{RMSD}_{\mathrm{eval}}$ to determine if the assignment is accepted or rejected. The optimization process was continued until 250 - 5000 accepted moves were reached (depending on the basis set size), with the five best assignments recorded for further use. The recorded assignments were recalculated from the full MP79 reference set. These finalized basis spectra were used to predict the “secondary-structure free” or “side chain free” CD spectra of the TS8 protein set as cross validation. The combination of side chain and backbone basis spectra that predicted the TS8 protein set with lowest RMSD$_{cross}$ were combined into mixed basis sets. These mixed basis sets were used to calculate the CD spectra of the SP175, GXG20, GP59, and TS8 data sets, so that they can be compared with initial the basis sets, PDB2CD, DichroCalc, and BestSel algorithms.

Results and Discussion

We present our results in two sections. Section 4 is focused on the optimization and assessment of our semi-empirical spectrum calculation approach, SESCA. In Section 5, we compare the impact of different contributions on the CD spectra of our reference proteins, in order to identify the largest sources of discrepancies, which might support further improvements.

4. Secondary structure based CD calculations

We derive the optimal basis spectra required for our semi-empirical spectrum calculations, using the SP175 reference set including the CD spectra and secondary structure classification of 71 proteins. To assess the average accuracy of SESCA predictions, we proceeded in three steps. First, we applied a principal component analysis (PCA) to determine the best achievable accuracy at which the CD spectra can be described using basis sets of a given size. Second, we used our soft optimization approach to derive basis sets to optimally reproduce the CD spectra of reference proteins from their secondary structure information. Third, we derived basis sets optimized for prediction accuracy using the hard optimization approach and assessed the predictive power of the obtained basis sets through cross validation using the TS8 data set. In addition, we compared SESCA with other published CD prediction methods, and assessed the sensitivity of our basis sets with respect to the secondary structure composition.

4.1 Estimate of best possible accuracy

As the main determinants of the accuracy, we considered the number of used basis spectra, the experimental error, both on the structure and the CD spectrum level, as well as the secondary structure classification method applied for spectrum calculation. We quantified the best possible accuracy of our basis sets by the fitting accuracy ($\mathrm{RMSD}_{\mathrm{ref}}$), the $\mathrm{RMSD}_{\mathrm{set}}$ calculated for the reference set used to derive the basis set. For a new protein with a crystal structure of similar quality, the RMSD of the predicted CD spectrum is expected to be larger than the fitting accuracy.

We first determined the best achievable accuracy for a given number of basis spectra (Fig. 4). To this end, basis spectra were calculated as eigenvectors of a PCA of the SP175 reference CD spectra, which by construction minimize the RMSD to the reference spectra as described in Section 3.5. In Fig. 4A the first ten obtained PCA basis spectra are illustrated. In line with previous results [13,16,33], the first two PCA basis spectra are similar to the CD spectrum of purely α-helical and β-sheet proteins, and represent already about 94% of the variance within the spectra of the reference data set. As the sorted eigenvalues (Fig. 4B) suggest, only a few basis spectra should be required to achieve good to very high accuracy. Indeed, almost 99% of the variance of the SP175 CD spectra are represented by only the first five basis spectra, and the first ten basis spectra essentially describe the full data set. This expectation is confirmed by the reconstruction of the α-amylase precursor spectrum (#3 of the SP175) shown in Fig. 4C, which corresponds to using one to ten PCA basis spectra. For this spectrum already the first three basis spectra allow a good reconstruction with an average RMSD of 2.105 kMRE units (10^3 deg*cm^2/dmol), and using more than six or seven basis spectra essentially recovers the reference spectrum. For comparison, the average spectrum (brown curve) is shown, corresponding to using no basis spectra at all, which serves as a lower limit of how well the spectra can be 'predicted' without any information. The table in

Fig. 4D quantifies the changes in fitting accuracy for three sample spectra, taken from representative proteins of the three main structure classes (α-helical, β-sheet, and mixed α/β) and also provides the average RMSD for all 71 spectrum reconstructions (RMSD_ref). For RMSD_ref a rapid decrease from an initial 6.395 to 1.335 kMRE units is observed for using the first three components, followed by a more gradual decrease from 0.955 to 0.182 for using up to ten components.

Depending on the desired accuracy, these results suggest that three to eight basis spectra should be used to construct highly accurate basis sets. Further in this study, we will use the deviations 0.237 kMRE and 6.395 kMRE obtained for eight and zero basis spectra, respectively, as an estimate for the 'best' and 'worst' achievable accuracy using all structural information but a limited set of up to eight basis spectra. The actual achievable accuracy is reduced by the fact that only limited structural information is contained in the secondary structure and by potential experimental error.

4.2 Accuracy limits of the secondary structure based CD spectrum prediction

After determining the best possible accuracy by PCA, we probed the accuracy CD spectrum calculations based on the limited structural information given by the secondary structure composition. To this end, we determined the secondary structure composition from the reference structures obtained by X-ray crystallography using five secondary structure classification methods (DSSP, DS_det, DS_sim, HbSS and HbSS_ext) described in Section 3.2. For each of the secondary structure classification methods, various basis sets were derived and their fitting accuracy was tested.

The fitting accuracy ($\mathrm{RMSD}_{\mathrm{ref}}$) of our basis sets is shown as the function of used basis spectra (basis set size) in Fig. 5A. We compared the optimized soft (solid lines) and hard (crosses) basis sets – coloured according to the underlying structure classification method – to the best possible fitting accuracy from the PCA basis sets (depicted as a dotted line). The more general soft basis sets were optimized for the lowest possible $\mathrm{RMSD}_{\mathrm{ref}}$ and represent the best fitting accuracy achievable with the limited structural information provided by the secondary structure classification algorithms.

For all five classification algorithms, the fitting accuracy of soft basis sets improves monotonously with the increasing basis set size. However, the gain in accuracy above six to eight basis spectra becomes increasingly smaller, and converges to values between 3.7 (for HbSS) and 2.8 (DS_det) kMRE units depending on the classification method. Notably, the best fitting accuracy of 2.8 kMRE is achieved for basis sets based on the DS_det classification method (blue), underscoring the trend that better fits are achieved with more fine grained secondary structure classification schemes. In comparison, the best possible fitting accuracy outlined by the PCA basis set converges to 0.17 kMRE. These trends indicate that predicting the CD spectra exclusively from the secondary structure of the protein crystal structure is possible, but imposes a significant limitation on the accuracy of the calculated spectra (~3.2 kMRE). This limitation is further influenced (± 0.5 kMRE) by the secondary structure classification scheme.

In addition, Fig. 5A shows that the hard basis sets with three to eight basis spectra converged to fitting accuracies of 3.2 - 3.8 kMRE, which are comparable to the limits set by the soft basis sets of the same size and classification method (2.8 - 3.7 kMRE). As expected, the two optimization methods yield basis sets of the same fitting accuracy if the number of secondary structure elements in the classification is equal to the number of basis spectra (F=K). These results indicate that the basis sets obtained by the hard optimization method accurately reconstruct the reference CD spectra, despite the additional restraints used during the optimization to improve the prediction accuracy.

4.3 Cross-validation of the prediction accuracy

We assessed the prediction accuracy of the optimized basis sets by cross validation. To this end we used each of these basis sets to calculate the CD spectra for the TS8 cross-validation set, comprising eight selected proteins with high quality CD spectra (between 178 - 260 nm), and high resolution crystal structures (< 2.5 Å). The prediction accuracy of each basis set was determined by computing the average RMSD between the calculated and measured CD spectra of the cross validation set ($\mathrm{RMSD}_{\mathrm{cross}}$).

Figure 5B shows the obtained $\mathrm{RMSD}_{\mathrm{cross}}$ for our basis sets: hard basis sets are depicted as crosses and soft basis set series as solid lines, coloured according to the underlying classification algorithm. The resulting prediction accuracies show different trends compared to the fitting accuracies calculated for the SP175 reference spectra (Fig. 5A), and they allow us to determine whether or not the results were influenced by over-fitting to the experimental error of the reference data set.

The TS8 CD spectra calculated from our soft basis sets (solid lines on Fig. 5B) show the best prediction accuracy between 2-6 basis spectra (depending on the classification algorithm). Including additional basis spectra into our basis sets results in larger deviations from the experimental CD spectra, although the decrease in accuracy for more than eight basis spectra is small. Additionally, the trend that classification methods with more secondary structure elements yield smaller RMSDs, as depicted in Fig. 5A, is not observed in Fig. 5B. Instead, classification algorithms with eight or less secondary structure elements (DSSP (8), DS_sim (8), and HbSS (7)) are the most suitable for predicting the CD spectra with soft basis sets. In contrast, the prediction accuracy of soft basis sets based on more fine-grained classification methods (namely DS_det (19, extended turn definitions) and HbSS_ext (11, extended β-sheet classification)) were markedly worse than their respective fitting accuracy, as seen from the 1.2 and 0.9 kMRE larger average RMSD of the cross validation (compared to the SP175 results). Unexpectedly, for some basis sets – particularly those based on DSSP

– their prediction accuracy was better than their fitting accuracy, which we attribute to the higher average quality of crystal structures in the cross-validation data set.

The restraints and ‘internal cross-validation’ during the evaluation step applied during the hard optimization scheme significantly reduced over-fitting in most of our hard basis sets (crosses in Fig. 5B), and produced basis with prediction accuracies of 3.034, 3.124, 3.042, and 3.288 kMRE units for the DSSP (DSSP-1), DS_sim (DS3-1), DS_det (DS6-1) and HbSS_ext (HBSS-3) classification algorithms, respectively. These basis sets – regardless of the underlying classification algorithm – consist of three to eight basis spectra (again, in line with the PCA results), and predict the CD spectra of the SP175 reference set with a comparable accuracy. These common features suggest that our hard basis sets indeed minimized the over-fitting to reference proteins, and reached the best prediction accuracies possible based on the experimental information of the reference data set.

4.4 Performance comparison

Above, we derived SESCA basis sets and reported the estimated fitting and prediction accuracy of our semi-empirical CD calculation scheme. We use these accuracy values to compare SESCA with other available CD calculation methods, DichroCalc, and PDB2CD. For this comparison, we also calculated the CD spectra of the SP175 and TS8 proteins from their crystallographic structures using both DichroCalc and PDB2CD. We emphasize that these algorithms represent different approaches of quantitative predictions based on CD spectroscopy. Note that PD2CD was also developed based on the SP175 reference protein set, thus our proteins sets provide an even ground for a comparison to SESCA, while DichroCalc – being an *ab initio* spectrum calculation method – was not parametrized to reproduce any particular protein reference set.

Dichrocalc is a heuristic *ab initio* CD spectrum calculation algorithm that predicts a spectrum from the protein conformation using QM derived parameters. The average RMSD-s of CD spectra predicted by DichroCalc were 6.095 and 6.124 kMRE units for the SP175 and TS8 data sets (indicated by the red dashed lines in Figs 5A and 5B), respectively. Note that as expected, the average accuracy of DichroCalc was similar for both datasets (no over-fitting), however, this accuracy was close to the PCA determined RMSD limit of a predictive method (6.4 kMRE). This indicates that DichroCalc can only determine the most prominent spectral features and likely sacrificed some of the accuracy of typical *ab initio* methods to be applicable for proteins.

PDB2CD ($RMSD_{set}$ values shown as brown dashed lines in Fig. 5) is a purely empirical method, which calculates the CD spectrum of a target protein by selecting structurally similar reference proteins based on secondary and tertiary structure information, and taking the weighted average of their spectrum. For the SP175 reference set PDB2CD was markedly more accurate ($RMSD_{ref}$ 2.395 kMRE) than any of the SESCA basis sets, or DichroCalc. However, in contrast to DichroCalc and most of hard SESCA basis sets, the prediction accuracy of PDB2CD ($RMSD_{cross}$ 4.725 kMRE) was significantly worse than its fitting accuracy. These results suggest that PDB2CD has similar or less predictive power compared to our SESCA basis sets ($RMSD_{cross}$ 3.0 - 3.9 kMRE), and may suffer from over-fitting to the SP175 reference set. This outcome was in contrast with the results of the cross-validation performed by Mavridis *et al.* [19] which showed very similar fitting and prediction accuracies for PDB2CD. Therefore, we performed a second cross validation using the same 14 protein structures, on which both the SESCA basis sets and PDB2CD achieved and $RMSD_{set}$ of ~3.8 kMRE units, whilst Dichrocalc performed somewhat worse (5.6 kMRE). We also found that four of the best eight cases where PDB2CD predicted a very accurate spectrum were β-crystallin proteins with a very similar fold, all of which were part of the SP175 reference set as well, although with a different crystal structure.

In Fig. 6 we present a comparison between the CD spectra calculated by three SESCA hard basis sets, DichroCalc, and PDB2CD for selected proteins: one α-helical, one β-sheet, and one α/β protein, in Figs. 6D - 6F, respectively. Although the number and shape of the basis spectra can differ significantly (Figs. 6A - 6C) depending on the assignment and classification method, the figure illustrates that the best performing SESCA basis sets often yield very similar calculated spectra. The calculated CD spectra from different spectrum prediction methods often have a comparable average RMSD for the same protein, and all correctly reproduce the overall shape of the experimental CD spectrum.

As an additional technical remark, we would like to highlight the speed advantage of the SESCA approach over PDB2CD and DichroCalc. We tested the speed of the algorithms by providing a single conformation for a protein of average size (490 amino acids) in PDB format, and measuring the time to receive the CD spectrum. While it took PDB2CD and DichroCalc servers nineteen and eight minutes respectively – queuing time not included – to predict a CD spectrum, SESCA predicted the spectrum in 0.3 seconds using the DSSP classification, and determined the average CD spectrum of an ensemble of 1000 conformations of the same protein just under five minutes. This three orders of magnitude difference in the calculation speed is due to the relatively simple geometric terms required for determining the secondary structure composition and the pre-calculation of basis sets in the SESCA scheme. The speed advantage in CD predictions may be particularly important for the iterative refinement of structural ensembles, an approach often used in the modelling of intrinsically disordered proteins.

4.5 Sensitivity to changes in secondary structure

We quantified the prediction accuracy of SESCA basis sets, PDB2CD, and Dichrocalc, based on the average deviation (RMSD) from experimental CD spectra. In the following, we estimate the sensitivity of this metric with respect to changes in the secondary structure composition. For this purpose, we selected a very simple basis set (DS-dT) with only three basis spectra (α-helix, β-strand, and coil) and three reference proteins which were predicted accurately by this basis set (alkaline phosphatase RMSD: 0.61 kMRE, met-myoglobin RMSD: 1.77 kMRE, and prealbumin RMD: 2.38 kMRE). We systematically altered the secondary structure information of these reference proteins to see how the RMSD of the resulting calculated spectrum is affected. Our results in Fig. 7A show an almost prefect linear dependence between the RMSD of the calculated spectrum and the deviation from the ideal secondary structure composition, with slightly different slopes (m) for α-helix to coil (A->C), α-helix to β-strand (A->B) and β-strand to coil (B->C) deviations. The ideal secondary structure composition in this context is the composition with the lowest RMSD from the experimental spectrum, which was identical to the secondary structure composition of the crystal structure in the case of alkaline phosphatase. For met-myoglobin and prealbumin, the ideal structure composition was a slightly altered secondary structure composition (A->C -4 %, and B->C +8 %, respectively).

The Table in Fig. 7 shows the expected error in the secondary structure composition of our model structure at a given RMSD between the calculated and experimental spectra. For example, if we obtained a calculated spectrum which differs from the experimental CD spectrum by 0.6 kMRE units, the secondary structure composition of our model should be within 2.5% of the true secondary structure composition. If the protein does not contain β-strands, however, the real composition should be within 2%, since the RMSD is more sensitive to A->C deviations. Applying the same calculations to the prediction accuracy of our best basis sets (RMSD ~3.1 kMRE), we can claim that the secondary structure composition of crystal structures of the cross-validation set is within 10 – 15 % from the secondary structure that best describes the CD spectrum (depending on the particular protein).

Using the same principles enabled us to assess the quality of the crystal structures of the SP175 proteins as models to predict the CD spectrum.. The RMSD distribution of CD spectra predicted by the DS-dT basis set for all proteins in the reference set is shown in Fig. 7B. We found two reference proteins with an RMSD less than 1.2 kMRE, which would mean an excellent agreement with the CD spectrum, and less than 5 % deviation in the secondary structure composition (ΔSS). There were 14 proteins in the SP175 set with a good agreement between the CD spectrum and crystal structure (RMSD: 1.2 - 2.4 kMRE, ΔSS less than 10 %), 27 proteins with average agreement (RMSD: 2.4 - 3.6 kMRE, ΔSS less than 15 %), 11 proteins with poor agreement (RMSD: 3.6 - 4.8 kMRE, ΔSS less than 20 %), and 17 proteins with very poor agreement (RMSD: larger than 4.8 kMRE and ΔSS likely more than 20 %).

The presence of 17 proteins with quite large RMSDs suggests that either the secondary structure composition of these proteins change significantly upon crystallization, or that additional factors affect the CD spectra of the reference proteins. In the next sections, we investigate several potential sources of such deviations, in order to identify potential routes for improving the accuracy of CD spectrum calculations.

4.6 Estimating the accuracy from solution structures

The analysis presented in Section 4.5 shows that even for proteins whose CD spectrum was predicted very accurately from their crystal structure, the secondary structure composition obtained from the structure did not necessarily provide an optimal description of the CD spectrum. So far in our study, we assumed that the crystal structure accurately reflects the protein structure under CD measurement conditions. This is of course not necessarily true, as the crystal structure typically reflects the minimum-energy conformation of the protein at low temperatures (~70 K), while the CD spectrum is usually measured near room temperature (~300 K) in aqueous solution, where larger fluctuations and structural heterogeneity are expected. This difference in structure and dynamics will likely result in differences of the average secondary structure composition and contribute to the RMSD between the measured and predicted CD spectra in our protein sets. In this section, we will estimate the difference between the average crystal and solution structures of our reference proteins, as well as its impact on the average accuracy of CD spectrum predictions.

A straightforward way to address the above mentioned problem would be to determine the solution structure of proteins using an independent method (such as NMR), and compare their secondary structure composition to those obtained from crystal structures. However, NMR solution structures are not available for most of the reference proteins used in this study. Therefore, we estimated the secondary structure compositions of the average solution structure from the CD spectrum of the reference proteins by using the well-established spectrum deconvolution method BestSel [11]. This algorithm was also trained on the SP175 protein set and provides detailed secondary structure predictions with eight structural elements (details in Section 3.3) with a particular focus on the structure of β-sheets. The secondary structure composition of the crystal structures were obtained by HbSS_ext classification method (described in Section 3.2), because it shares the detailed β-sheet classification with BestSel, based on the parity and local twist of the β-strands.

We obtained the secondary structure composition from both methods for the proteins of the SP175 reference set, as well as the TS8 cross-validation set, then computed and compared the average compositions to quantify the differences. Compared to the crystal structures, the estimated secondary structure composition of the solution structures showed lower average α-helix content (-4.9% for SP175 and -7.7 % for TS8) and a higher β-strand content (+7.7 % for SP175 and 7.9 % for TS8) for both data sets. These average differences in the secondary structure composition would translate to an average RMSD of up to 2.0 kMRE units according to sensitivity of SESCA predictions shown in Fig. 7. This is more than half of the 3.6 kMRE average deviation of the predicted CD spectra based on the optimized

SESCA basis sets, suggesting that the difference between solution and crystal structures is one of the major sources of error for SESCA predictions.

To provide a more direct comparison to the spectrum prediction methods discussed in this study, we used eq. 5 to derive a specialized SESCA basis set (BestSel_der) that reconstructed the CD spectra from the BestSel secondary structure compositions. This basis set indeed yielded good fits ($\mathrm{RMSD_{ref}}$ 2.931 kMRE) to the SP175 spectra, and even better ones to the TS8 spectra ($\mathrm{RMSD_{cross}}$ 1.828 KMRE). Next, we compared the average RMSD of the CD spectra predicted by the BestSel _der basis set with the accuracy of hard SESCA basis sets listed in Table S8. The HBSS-3 basis set was the most accurate from those based on the HbSS_ext algorithm ($\mathrm{RMSD_{ref}}$ 3.754 kMRE and $\mathrm{RMSD_{cross}}$ 3.288 kMRE), it's fitting and prediction accuracies are 0.8 and 1.5 KMRE units worse than what BestSel_der achieved on the same proteins. The difference between the average accuracy of the BestSel_der and HBSS-3 is smaller than expected for the proteins of SP175 reference set. However, BestSel_der reconstructed most of the SP175 spectra more accurately, except for seven proteins with exceptionally large RMSDs between their measured and calculated CD spectra. These proteins were also poorly predicted by the HBSS-3 algorithm, but their presence reduced the average difference between the $\mathrm{RMSD_{ref}}$ of the two basis sets. The improved accuracy for the rest of reference proteins agrees well with the estimated difference in the average secondary structure composition between the solution and crystal structures of the data sets, and thus confirms its impact on the accuracy of SESCA predictions.

Interestingly, the basis spectrum of the right-handed anti-parallel β-strand secondary structure element (Anti3 in Fig 6A) in BestSel_der showed a distinctive negative peak around 195 nm, as is typical for random coil proteins. This secondary structure element was also the most populated one (10 %) among the β-strand elements in the SP175 reference set, whereas HbSS_ext classified only 5 % of the residues as such. The 5 % overestimation of this particular secondary structure element indicates that the difference between the solution and crystal structures is most likely due to the higher occurrence of unfolded/disordered residues in solution, rather than due to the larger fraction of β-strands.

From the above results we conclude that the secondary structure composition of a globular protein in aqueous solution may differ by 5 - 10 % from its composition in crystal structures, and that this difference contributes up to 2.0 kMRE to the RMSD of the CD spectra predicted from the crystal structures of the proteins in our study. Furthermore, for several proteins of the SP175 reference set, the CD spectra were predicted with relatively poor accuracy even from the ideal secondary structure composition. This points to either problems related to the measured CD spectra of these proteins, or to strong contributions to the spectrum that cannot be predicted through the secondary structure composition. We will investigate these possibilities in the following sections.

5. Improving the CD prediction accuracy

In section 4, we derived several SESCA basis sets to predict the CD spectra of globular proteins and determined that their best achieved prediction accuracy is 3.0 ± 0.6 kMRE. In this section, we focus on whether the prediction accuracy of our basis sets can be improved by changing the reference protein set. First, we consider how the "hard-to-predict" CD spectra in our reference set influence the robustness of SESCA predictions. Then, we determine if replacing crystal structures with structural ensembles can improve the accuracy of the predicted spectra. Finally, we expand the reference set with a series of short peptides and include the amino acid composition into the basis set determination process.

5.1 Potential measurement errors of the reference set

The RMSD distribution shown in Fig. 7B suggests that the CD spectra of certain proteins in the SP175 data set are hard to predict based on their respective crystal structure. In this section we will identify these proteins and assess their effect on the SESCA basis sets. To this end, we calculated a method-independent mean RMSD (R_j^{mean}) for each protein as the average accuracy of six different prediction methods: four SESCA basis sets (DSSP-1, HBSS-3, DS5-4, DS-dT) as well as PDB2CD and the BestSel reconstruction basis set (BestSel_der). This method-independent R_j^{mean} value and the standard deviation (σ_j or scatter) of the individual RMSDs of the predicted spectra were calculated for the SP175 and TS8 proteins, and were averaged over the data sets to obtain mean fitting and prediction accuracies. The method-independent mean RMSD ($RMSD_{\mathrm{set}}^{\mathrm{mean}}$) and scatter ($\sigma_{\mathrm{set}}^{\mathrm{mean}}$) were similar for the SP175 ($RMSD_{\mathrm{fit}}^{\mathrm{mean}}$ = 3.3 kMRE, $\sigma_{\mathrm{fit}}^{\mathrm{mean}}$ = 0.9 kMRE) and TS8 data sets ($RMSD_{\mathrm{cross}}^{\mathrm{mean}}$ = 3.2 kMRE, $\sigma_{\mathrm{cross}}^{\mathrm{mean}}$ =1.2 kMRE). We considered proteins difficult to predict, if their R_j^{mean} value were larger than the mean RMSD and scatter of the TS8 cross-validation set combined ($RMSD_{\mathrm{cross}}^{\mathrm{mean}} + \sigma_{\mathrm{cross}}^{\mathrm{mean}}$ = 4.4 kMRE).

Figure 8A shows R_j^{mean} of the calculated spectra for each of the 71 proteins of the SP175 data set. As can be seen, 12 proteins (annotated in grey) show marked deviations from the mean prediction accuracy and, hence, were classified as difficult to predict based on their secondary structure. Closer inspection of these 12 proteins (average R_j^{mean} ~6.0 kMRE) shows that in many cases the peak positions and relative peak heights were similar, but the absolute intensity of the experimental spectra differed significantly from that of the calculated spectra.

Therefore, we applied scaling factors to the experimental spectra of all 12 proteins which minimize the deviation from the calculated spectra. Indeed, as can be seen from Fig. 8B, for eight proteins (marked, magenta) scaling factors between 0.3 and 1.5 improved the agreement with the calculated spectrum on average to 3.1 kMRE units. The largest improvement (more than 12 kMRE) was observed for Subtilisin Carlsberg (SP175/67) shown in Fig. 8C. For the other five hard-to-predict proteins, such as Jacalin (SP175/41) shown in Fig. 8D, the shape of experimental and calculated spectra differed significantly and a simple scaling factor did not yield a good agreement between the two. In addition, when we applied the same procedure to the TS8 data set, we found that Hemerythrin (TS8/1) was also difficult to predict ($R_j^{\mathrm{mean}} = 6.4$ kMRE with $\sigma_j = 1.7$ kMRE), but a scaling factor of 1.3 greatly improved the RMSD of its predicted spectra (to R_j^{mean} = 3.3 kMRE with σ_j =0.6 kMRE).

To assess how much these outlier proteins affect the accuracy of our CD spectrum calculations, we removed them from the SP175 data set and recalculated the SESCA basis sets with the remaining 59 proteins. As shown by the black and dark blue lines in Fig. 9A, the resulting mean RMSD of the modified reference set improved from 3.3 to 2.7 kMRE units, whereas the mean prediction accuracy of the basis sets shown in Fig. 9B was reduced slightly (by 0.03 kMRE) due to changes in the basis spectra of rarely occurring secondary structure classes. These results demonstrate that the prediction accuracy of our basis sets is robust with respect to the presence of the hard-to-predict proteins, although the shape of some basis spectra is sensitive to the changes in the reference set, especially if the average occurrence of its structural elements is below 1%.

Because the above results suggest that inaccurate normalization of the experimental spectra may generally limit the accuracy of our CD spectrum calculations, we also applied scaling factors to the experimental spectra of all proteins in the SP175 and TS8 data sets. As expected and shown in Fig. 9 (light blue lines), the mean RMSDs improved markedly for both data sets, from 3.3 to 2.2 and from 3.4 to 2.5 kMRE units, respectively.

These observations suggest that the main source of the normalization problems is the inaccurately determined soluble protein concentration during the CD measurements. Protein precipitation and aggregation may both affect the soluble protein concentrations in the measurement cell, which are difficult to account for experimentally. If the applied scaling factors indeed indicate errors of the assumed soluble protein concentrations, it would usually translate to errors up to ±30 % between the assumed and actual protein concentrations, with a few exceptions as large as 60 % within the SP175 data set.

5.2 The impact of conformational flexibility on model quality

As discussed in Section 4.6, the crystal structure of a protein may differ from its solution structure both in terms of average structure as well as structure fluctuations and heterogeneity. We also proposed that these effects may alter the average secondary structure composition of proteins, and that therefore, the neglect of these fluctuations in our models reduced the accuracy of our CD spectrum predictions. In this section we test this possibility by analysing how conformational flexibility affects the average secondary structure of a model protein and the accuracy of predicted macroscopic observables such as CD spectra and NMR chemical shifts.

To this aim, we chose a highly flexible protein complex formed by the two disordered protein domains P53-AD2 and CBP-NCBD. These domains form an ordered complex for which we obtained three structural models that all describe average structure, but differ in the level of the conformational flexibility. The models are based on the P53/CBP complex structure determined by NMR spectroscopy and deposited in the protein databank by Lee *et al.* (PDB code 2L14). This model contained a bundle of 20 protein conformations, which fulfil the NMR distance restraints in an aqueous solution. For all these structure models, we calculated average secondary structure, CD spectra, and NMR chemical shifts, and compared them to the respective experimental values.

The three structural models of the P53/CBP complex to probe the effect of the conformational fluctuations are depicted in Fig. 10A. In an ascending order of conformational flexibility, the first model was the first conformation of the NMR bundle, with no explicit information on conformational fluctuations. This model mimicked the minimum-energy conformation of a crystallographic structure (‘Cryst’). The second model was the full NMR bundle with 20 conformations, which described conformational fluctuation near the minimum-energy structure. The third model was a structural ensemble of 1000 conformations, obtained from a molecular dynamics (MD) simulation described in Section 3.1. The MD ensemble explored the conformational dynamics and fluctuations of the system further away from the average, to describe the average protein structure in an aqueous solution at room temperature.

First, we analysed the differences in the secondary structure composition of the three models. A summary over secondary structure composition of each structural model is shown below their cartoon representation in Fig. 10A. As the figure shows, the model Cryst was the most structured of the NMR conformations and 49 % of its residues were α-helical. In the case of the NMR model the termini of domains were more flexible, which lead to a slightly lower average helix content of 47 %. Although no β-sheets appeared in these models, a low percentage amino acids adopted a local conformation typical for an extended β-strand at the termini of the two protein domains.

The P53/CBP complex was very dynamic during the MD simulations. The two domains remained strongly bound during the simulation, but the conformational fluctuations resulted in a 38 % average helix content. In addition, while total β-strand content decreased slightly in the MD model compared to the NMR bundle, 2.8 % of the residues in the MD model was in a regular β-strand conformation, and established the hydrogen bonds to form two short β-sheets which appeared with ~15 % probability in the MD ensemble. These short β sheets connected the N-terminus of CBP-NCBD with residues 25-27 of P53-AD2, and the two termini P53-AD2.

In line with our expectations, the added conformational flexbility of the MD ensemble indeed changed the average secondary structure composition of the P53/CBP complex by up to 15 % compared to the Crys model it was started from. To show that these changes improved the quality of the structure model, we predicted the CD spectrum from all three models using several optimized SESCA basis sets (DSSP-1, DS5-4, and DS-dT), and compared them with a high-quality synchrotron radition CD spectrum of the P53/CBP complex.

Figure 10B shows a comparison between the measured CD spectrum of the P5/CBP complex, and the CD spectra which were predicted from the three structural models by the DSSP-1 basis set. The lower average helix content in the MD ensemble was also reflected in the predicted CD spectra of this model (red line in Fig. 10B), as it shows a less pronounced positive peak at 192 nm, typical for α-helical proteins. Comparison of the spectra shows that this decreased helix content of the MD ensemble agrees better with the recorded CD spectrum (RMSD: 3.1 kMRE), than either the original NMR bundle (RMSD: 5.4 kMRE) or the single-conformation model (RMSD: 6.0 kMRE). The RMSD values clearly show that the Cryst and NMR models are rather poor representations of the secondary structure, whilst the MD ensemble reflects the average structure composition much better. However, the RMSD of its predicted spectrum is still not better than that of the average globular protein model with no conformational flexibility (3.0 ± 0.6 kMRE). We speculate that this relatively large RMSD of MD model is due to the missing slower conformational dynamics of the protein. These conformation fluctuations may decrease the average helix content further, but are not captured during a 10 μs long simulation trajectory. This speculation is also in line with the ideal secondary structure composition estimated by BestSel based on the measured CD spectrum, which predicted a 29 % average helix content.

To avoid possible biases from inaccurate normalization, we also applied scaling factors to fit the intensity of the experimental spectrum to each of the predicted spectra. The scaling factors (1.519, 1.463, and 1.244 for Cryst, NMR and MD, respectively) highlight the differences between the shapes of the predicted spectra, but did not change their RMSD order. The MD ensemble reproduced the scaled experimental spectrum most accurately (RMSD: 2.4 kMRE), followed by NMR bundle (RMSD: 3.9 kMRE), and the single-conformation model (RMSD: 4.2 kMRE). Similar trends were obtained, when the CD spectra were predicted using other optimized SESCA basis sets - such as DS5-4 and DS-dT - as well, underlining the conclusion that the most flexible MD ensemble is best in line with the CD spectrum.

From this trend we conclude that the use of structural ensembles to include protein conformational flexibility improves the accuracy of our CD spectrum calculations for the P53/CBP complex substantially (by ~3.0 kMRE). This protein complex was chosen because dynamics was expected to be important for its average structure, and consequently the impact of conformational flexibility on typically less flexible globular proteins is likely to be smaller (between 1.0 and 2.0 kMRE), but still significant.

To assess whether or not inclusion of conformational flexibility generally improves not only the accuracy of the calculated CD spectra, but also the quality of the structure model, we compared our structural models to the experimental chemical shifts from the original NMR measurements (obtained from biological magnetic resonance databank, entry no. 17073). We computed the backbone chemical shifts (including those for the N, C_α, C_β, C, H_N, and H_α atoms) for the three models using the chemical shift predictor Sparta+ [34]. Figure 10C shows the comparison between the experimental and calculated C_α secondary chemical shifts. Secondary chemical shift values are corrected for the average random coil chemical shift of the amino acid, and therefore indicative of the local protein (secondary) structure. A sequence of large positive secondary C_α shifts indicates a high propensity for α-helix in that region, whilst a sequence of large negative values shows a preference towards β-strands. The overall agreement between the measured and predicted chemical shifts was quantified the through average RMSD of their secondary chemical shift profiles.

The comparison in Fig. 10C also revealed that the RMSD of the MD ensemble chemical shift (1.057 ppm) was lower than that of the NMR bundle (1.385 ppm) or the single-conformation model (1.419 ppm). This trend is expected, and is also in line with RMSD of the predicted CD spectra. The same trends were observed for the average RMSD of all backbone chemical shifts as well, which again suggests that our conclusions about the effects of conformational flexibility are robust.

The chemical shifts also provide information on where the secondary structure elements are located along the protein sequence. The C_α chemical shifts predicted from our models agree well with the experimental chemical shifts on the position of the helical regions, but significantly overestimate the helix propensities, especially for the C-terminal helix of CBP-NCBD, and the helical regions in P53-AD2. These regions are also the ones where the average secondary structure composition is considerably less helical in the MD ensemble than the other two models. Additionally, the residues of the short β-sheets observed only in the MD model possess some of the largest negative C_α secondary chemical shifts of the experimental profile, suggesting that presence of these β-sheets also contribute to the lower average RMSD of the MD model.

In summary, both the predicted CD spectra and chemical shifts suggested a clear trend: the MD ensemble model which includes conformation dynamics in aqueous solutions most accurately reproduced all considered experimental observables. In contrast, the crystal model, which ignores structure fluctuations, is the least accurate. The example of the P53/CBP complex presented above strongly supports our previous conclusions, that including conformational flexibility improves our structural models, which in turn allow more accurate predictions of CD spectra as well as other experimental observables (such as NMR chemical shifts).

5.3 Side chain CD spectrum calculations

Comparison of the best achievable prediction accuracy (Section 4.1) with the much lower accuracy achievable based solely on the secondary structure composition (Section 4.2) suggests that including additional information should improve the CD spectrum calculations. Amino acid side chain groups are the second most common type of chromophores in proteins. Side chain contributions are also considered as optional corrections in DichroCalc, and some deconvolution basis sets have side chain related basis spectra [5]. Here, we will therefore attempt to determine the contribution of side chain groups to the protein CD spectra in the far-UV range, and include those contributions into the SESCA scheme to improve the prediction accuracy of our method.

To determine how much the side chains contribute to the CD spectra of the SP175 reference set, we analysed the correlations between the principal components describing the shape of the CD spectra (see Section 2.6) and the occurrence of amino acids and secondary structure elements in the reference proteins. To this aim, we calculated the Pearson correlation coefficients between the projections of the first ten PC vectors (details in Section 3.5), the amino acid composition of the proteins, as well as the secondary structure compositions determined by the BestSel, DISICL, DSSP and HBSS algorithms.

Table 1 shows those structural properties which correlate most strongly with the principal components (PCs) of the CD spectra. As can be seen, the first three principal components involve mainly secondary structure elements: PC 1 – which accounts for over 80 % of the spectral variance of the reference set – was very strongly correlated ($R_{pearson}$ ~0.9) to the presence of α-helices in the protein structure, whilst PC 2 and 3 are moderately correlated to β-strand and turn structures. However, PCs 4, 6, 9, and 10 correlate more strongly with the presence of amino acids than secondary structure elements. Since these principal components describe ~3 % of the spectral variance, one would expect a somewhat smaller but still notable contribution from side chain groups. In addition, the most commonly considered correction to CD spectra are associated with the aromatic side chains of tryptophan, phenyl-alanine, and tyrosine because these amino acids have the strongest CD signals in isolation. Our analysis also suggests that amino acid side chains with weaker CD activity, particularly arginine, histidine, cysteine and serine, may also contribute significantly to the CD spectra.

To also include the amino acid side chains into our SESCA predictions, we assumed that their average contribution is not strongly affected by couplings to the local structure of the protein backbone, or by the adjacent side chains. This assumption allowed us to assign one SESCA basis spectrum to each side chain, and to determine the average contribution of side chains from the amino acid composition of the protein sequence.

Our first attempt was to use measured CD spectra of isolated natural amino acids to estimate the contribution of amino acid side chains. The amino acid CD spectra (except for glycine) were measured by Nisihno *et al* [35]. at neutral, acidic and basic *pH*. We used the CD spectra at neutral *pH* (7.0) shown in Fig. 11A as a basis set to calculate side chain dependent baseline corrections similarly to eq. 1, with weighing coefficients for the basis spectra proportional to the fraction of amino acids in the protein sequence. The calculated baselines were then subtracted from the CD spectra of proteins in the SP175 and TS8 data sets, and the side-chain corrected data sets were used to derive and cross-validate basis sets based on the “pure” secondary structure contributions. This procedure, however, resulted in basis sets with lower prediction accuracies in all cases, when they were compared to non-corrected basis sets with the same assignment. This observation suggests that the average contribution of side chain groups may differ significantly from the CD signal of isolated amino acid when they are attached to a polypeptide chain in a protein.

To test this hypothesis, and to obtain improved side chain signals more representative for a polypeptide environment, we prepared a new reference set of twenty short tri-peptides (designated as the GXG20 set), each consisting of the same capped backbone, and one of twenty side chain groups ('X') of the natural amino acids.

As shown in Fig. S19, the CD spectra of the GXG20 peptide set differ substantially from one another, despite the fact that the peptides were too short to form the hydrogen bonds required for stable α-helices and β-sheets, and therefore mostly adopted a random coil structure. We therefore assumed that the spectra of these peptides are largely defined by their side chain group, and although the spectra differed considerably from the CD spectra shown in Fig. 11A, the influence of the phenyl-alanine tyrosine, tryptophan, and histidine side chains is indeed remarkably strong in both cases. The GXG20 spectra indicate that aromatic side groups – and particularly phenyl-alanine and tyrosine – have strong positive contributions to the CD spectra, which differs from the signals of other side chains. The CD spectrum of the GAG peptide, on the other hand, shows the largest negative peak at ~195 nm, similar to CD signal that is associated with a random coil protein, whereas the CD signal of the GGG peptide – in the absence of a chirality centre – is very weak.

We derived the average contribution of side chain groups to the CD signal of proteins as described in Section 3.6 from a new mixed reference set (MP79), which included 59 globular proteins of the SP175 reference set and the 20 tri-peptides of the GXG20 set. The resulting "pure" side chain basis spectra shown in Fig. 11B are very similar for the same amino acid regardless which secondary structure basis set was used to derive them. The pure basis spectra are significantly larger than the CD spectra of the independent amino acids (Fig 11A), and confirm the large contributions of the phenyl-alanine and tyrosine side chains. In addition, the basis spectra show moderate contributions from the amino acid side groups of asparagine, aspartate, glutamate, histidine, leucine, serine, and tryptophan, while the side chains of other amino acids such as glycine, valine, isoleucine and threonine had weaker CD signals.

Finally, we quantified the effects of the derived side chain contributions on the prediction accuracy of SESCA basis sets. Using the derived side chain contributions as our basis set, the side chain dependent baselines were calculated once again and subtracted from CD spectra of the SP175 and TS8 data sets. Then, the basis spectra of our optimized basis sets were recalculated and the accuracy of the basis sets were cross-validated using the side-chain corrected CD spectra. Including the side chain contributions of the twenty amino acids now resulted in small improvement in the prediction accuracy (RMSD$_{cross}$) on the order of ~0.05 kMRE units, compared to the secondary-structure-only basis sets. This improvement is almost an order of magnitude smaller than expected, based on our correlation analysis. This result is particularly surprising in the light of the large contributions of the individual amino acid side chains to the protein CD spectra. In the following section we will therefore ask if and how the contributions of side chains to the CD spectra can be described even more accurately.

5.4 Combining side chain and backbone contribution

To that aim we hypothesized that one of the reasons for the limited success might be over-fitting. Indeed, we used twenty independent basis spectra to describe the contribution of side chain groups to the protein CD spectra, whilst the PCA analysis (Section 5.3) showed that already four basis spectra represent these 20 contributions quite accurately. To avoid such over-fitting, we applied optimization schemes to obtain basis spectra for both the secondary structure of the protein backbone and side-chain contributions, and then combined them in an optimal “mixed” basis set.

To this aim, we used the hard optimization scheme in a three-stage process (described in Section 3.6) to reduce the number of required basis spectra and – hopefully – to improve the prediction accuracy. In this protocol, the side chain basis spectra were optimized first, followed by an independent optimization of secondary structure-based backbone basis spectra (including the secondary structure assignments). The resulting optimized basis sets (examples shown in Figs. S20-S23) typically included 3 - 6 backbone basis spectra and 4 - 7 side chain basis spectra, with one or two basis spectra representing the positive CD signals of the aromatic residues.

Figure 12A compares the average RMSDs achieved by optimized basis sets with and without side chain contributions. The comparison shows small improvements (>0.2 kMRE) in the quality of the calculated spectra for both the cross-validation (TS8) and the globular reference (SP175) proteins. This improvement persisted when both side-chain corrections and scaling (described in section 5.1) were applied, further reducing $\mathrm{RMSD}_{\mathrm{set}}$ for cross-validation proteins from 2.6 kMRE to 2.4 kMRE units. The relatively small influence of the side groups is now more in line with the PCA analysis of the SP175 spectra (Fig. 4 and Table 1), which suggests that over 95% of the spectral variance is mainly associated with the backbone secondary structure. On the other hand, the $\mathrm{RMSD}_{\mathrm{set}}$ calculated for the GXG20 peptides shows significant improvements from side chain corrections (from > 5.5 kMRE to < 3.5 kMRE), because their CD spectrum is largely defined by the side chain signals.

Figures 12B and 12C show the backbone and side chain basis spectra of an optimized basis set (DSSP-dT1SC), respectively. Clearly, the strength of the CD signals is comparable between the basis spectra of side chain groups and secondary structure elements. This observation is again unexpected, as the influence of backbone basis spectra on the accuracy of CD spectrum predictions is twentyfold larger. To explain the smaller impact of the side chain basis spectra on globular proteins, we calculated the total contribution of the side chain basis spectra to the calculated CD spectra for each of the SP175 proteins (Fig. 12D). These contributions typically vary between -5 and +5 kMRE units, depending on the protein and the wavelength, thus amounted to approximately one tenth of the total contribution from the protein backbone.

Closer analysis revealed mainly three reasons that combine to produce this unexpected outcome. First, the side chain basis spectra have opposite signs and therefore partially cancel out in the total side-chain contributions. Second, the amino acid compositions of the globular proteins in our reference sets are rather similar, which further decrease the variance of the already small total contributions. Third, the secondary structure contents correlate with the amino acid composition (in our reference set, Pearson correlations coefficients between 0.2 and 0.6 were calculated) such that part of the side chain information is already encoded within the secondary structure information.

One possible reason for the cancellation of side chain basis spectra may be that the side chain contributions strongly depend on their environment, and an averaged side-chain signal cannot accurately represent the actual contribution of buried and solvent accessible side chains or side chains in different protonation states. Accordingly, one would expect more accurate CD spectrum predictions, if the different relevant side chain signals were identified and separated from each other. This possibility, however, will not be further explored in this study.

As a side note, the correlation between the amino acid composition and the backbone secondary structure can be exploited to predict the CD spectrum even in the absence of a structural model. Relying on the strong amino acid preferences of the secondary structure elements, we used the hard optimization scheme to derive “amino-acid only” basis sets, which predict the CD spectra of proteins using only the amino acid composition of their sequence. These basis sets (marked by the type “Seq” in Table S8) achieved fitting accuracies between 3.9 - 4.7 kMRE units on the SP175 reference proteins and their prediction accuracies on the TS8 proteins amounted to 5.1 - 6.2 kMRE depending on the amino acid grouping.

Although the accuracy of structure-based spectrum predictions is better as expected, the $\text{RMSD}_{\text{crosss}}$ of sequence-based basis sets shows they retain some predictive power.

The above mentioned three factors combined such that the predictive power of our mixed basis sets improved only moderately beyond the accuracy achieved by using secondary-structure exclusive basis sets. Of course, the limited impact of side chain contributions to CD spectra of globular proteins also underlines the robustness of the secondary-structure based SESCA predictions. Including the side chain corrections will certainly be helpful in certain cases, but in our view not essential for the accurate prediction of most globular protein CD spectra.

In contrast, the example of the GXG20 peptides also suggests that for small or disordered peptides, mixed basis sets – including the side chain contributions – can be pivotal for the accurate prediction of their CD spectra. This may be particularly true for proteins with unusual amino acid compositions such as the low complexity regions and sequence repeats often found in intrinsically disordered proteins. Because disordered proteins rarely form stable α-helices or β-strands, the backbone contributions to their CD spectra are less pronounced than for globular proteins. Moreover, most of the amino acid side chains in IDPs are solvent accessible and, therefore, their average CD signals may more closely resemble those of the GXG20 peptides.

Conclusions

In this study we presented a new semi-empirical spectrum calculation approach (SESCA) to predict the electronic circular dichroism (CD) spectra of globular proteins from their model structures. We derived basis spectrum sets which can be used to predict the CD spectrum of a chosen protein from the secondary structure composition determined by various structure classification algorithms (including DSSP, DISICL, and HbSS), to render the method more versatile and broadly applicable.

The basis spectra were derived and optimized using a reference set consisting of 71 globular proteins; then the prediction accuracy of the basis sets was determined by cross-validation on a second, non-overlapping set of eight selected proteins, covering a broad range of secondary structure contents. The experimental CD spectra of these proteins were predicted with an average root-mean-squared deviation (RMSD) as small as of 3.0 ± 0.6 x 10^3 degree·cm^2/dmol in mean residue ellipticity units or 0.9 ± 0.2 M^{-1}cm^{-1} in Δε units. This deviation is on average 50 % smaller than what is achieved by the best currently available algorithm (PDB2CD average deviation ~4.7 x 10^3 degree·cm^2/dmol).

Our analysis of the optimized basis sets have shown that the accuracy of the CD predictions does not depend strongly on the underlying secondary structure classification method. In contrast, is strongly dependent on the number basis spectra in the basis set. Our results suggest that 3 - 8 basis spectra which describe the backbone structure of the protein provide the optimal trade-off between model complexity and possible over-fitting to our reference data, and thus allow the most accurate prediction of the protein CD spectrum.

We attempted to further improve the accuracy of SESCA predictions by including basis spectra into our basis sets which reflect the average contribution amino acid side chain groups. Unexpectedly, for globular proteins the inclusion of side chain information did not markedly improve the accuracy of the predicted CD spectra. This finding is particularly surprising because the side chain CD signals, in the context of the proteins and peptides investigated, were significantly larger than the CD spectra of the isolated amino acids. Apparently, prediction methods based purely on the secondary structure are rather robust against the variation of side chain contributions, due to the cancellation of side chain signals, similarity of the amino acid composition, and correlations between the presence of amino acids and the structure of the protein backbone. In summary, although side chain contributions can be neglected for the CD calculation of the typical globular protein, we expect markedly improve the spectrum prediction accuracy for short peptides, and possibly disordered proteins. For these molecules the inclusion of 4 - 7 side chain basis spectra may provide the optimum of spectrum prediction accuracy.

Analysis of deviations between calculated and experimental spectra of the reference proteins showed that ~15 % of the predicted globular protein CD spectra agree rather poorly with the measured spectra. The main source of these deviations seems to be the uncertainty in the intensity of the experimental CD signal, most likely due to the often challenging concentration-dependent normalization of the CD spectra. By scaling the experimental CD spectra, the average RMSD of both the TS8 cross-validation set and the SP175 reference protein sets were reduced to below 2.6 x 10^3 degree·cm^2/dmol. Although this scaling had a large impact on the RMSD of individual ”hard-to-predict” proteins, SESCA basis sets turned out to be robust to the presence of these proteins in the reference set.

Due to the simple secondary structure calculations and the pre-calculation of basis sets, SESCA can be efficiently applied to rather large structural ensembles. This allows us to account for the conformational flexibility of a protein when calculating its CD spectrum. Indeed, for the test case studied here, including conformational flexibility of the protein, as obtained from an extended molecular dynamics trajectory, considerably improved the accuracy of the calculated CD spectrum. Whether this encouraging result is true in general is an interesting question which will be addressed in a separate study.

By exploiting the high sensitivity of CD spectra to the average secondary structure of proteins, SESCA basis sets can be used for evaluating and improving protein structural models in biology and biophysics. As our example of the P53/CBP complex demonstrated, the accuracy of CD predictions, the inclusion of conformational flexibility, and the robustness of the secondary structure based CD predictions enables SESCA basis sets to target not only the average structures of globular proteins, but also their structural flexibility and heterogeneity.

Furthermore, by accounting for both flexibility and side chain contributions, SESCA basis sets may be particularly helpful in modelling intrinsically disordered protein (IDP) ensembles, as they can provide information about the transient secondary structure patterns of these molecules. These biologically highly relevant molecules are notoriously hard to characterize, and also the modelling of IDP ensembles based on experimental input is particularly challenging.

A python implementation of our semi-empirical CD calculation method SESCA, as well as basis sets and tools compatible with the secondary structure classification algorithms DISICL and DSSP are publicly available online: http://www.mpibpc.mpg.de/sesca.

Acknowledgements:

The authors would like to thank J. Kardos and Cs. Micsonai for prodiving CD spectra for the SP175 protein data set and the BestSel algorithm, to J. Hritz, S. Becker, and C. Griesinger for providing protein samples for the CD measurements, B. Gyurcsik for the aid in the experimental design, and P. Kellers for editing the manuscript. We would like to thank the storage ring facilities of Aarhus University (ISA) for awarding beam time on the AU-CD beam line, ASTRID2, for the measurements of the CD spectra of the short peptides.

References

1. Fasman GD, editor. Circular Dichroism and the Conformational Analysis of Biomolecules [Internet]. Boston, MA: Springer US; 1996. Available: http://link.springer.com/10.1007/978-1-4757-2508-7

2. Brahms S, Brahms J. Determination of Protein Secondary Structure in Solution by Vacuum Ultraviolet Circular Dichroism. J Mol Biol. 1980;138: 147–178.

3. Kelly SM, Jess TJ, Price NC. How to study proteins by circular dichroism. Biochim Biophys Acta BBA - Proteins Proteomics. 2005;1751: 119–139. doi:10.1016/j.bbapap.2005.06.005

4. Johnson Jr. WC. Protein Secondary Structure and Circular Dichroism: A Practical Guide. PROTEINS Struct Funct Genet. 1990;7: 205–214.

5. Hennessey Jr JP, Johnson Jr WC. Information content in the circular dichroism of proteins. Biochemistry (Mosc). 1981;20: 1085–1094.

6. Goodman M, Toniolo C. Conformational Studies of Proteins with Aromatic Side-Chain Effects. Biopolymers. 1968;6: 1673–1689.

7. Strickland EH, Beychok S. Aromatic Contributions To Circular Dichroism Spectra Of Protein. Crit Rev Biochem. 1974;2: 113–175. doi:10.3109/10409237409105445

8. Chakrabartty A, Kortemme T, Padmanabhan S, Baldwin RL. Aromatic Side-Chain Contribution to Far-Ultraviolet Circular Dichroism of Helical Peptides and Its Effect on Measurement of Helix Propensities. Biochemistry (Mosc). 1993;32: 5560–5565.

9. Whitmore L, Wallace BA. Protein secondary structure analyses from circular dichroism spectroscopy: Methods and reference databases. Biopolymers. 2008;89: 392–400. doi:10.1002/bip.20853

10. Lees JG, Miles AJ, Wien F, Wallace BA. A reference database for circular dichroism spectroscopy covering fold and secondary structure space. Bioinformatics. 2006;22: 1955–1962. doi:10.1093/bioinformatics/btl327

11. Micsonai A, Wien F, Kernya L, Lee Y-H, Goto Y, Réfrégiers M, et al. Accurate secondary structure prediction and fold recognition for circular dichroism spectroscopy. Proc Natl Acad Sci. 2015;112: E3095–E3103. doi:10.1073/pnas.1500851112

12. Hollósi M, Fasman GD, others. Convex constraint analysis: a natural deconvolution of circular dichroism curves of proteins. Protein Eng. 1991;4: 669–679.

13. Louis-Jeune C, Andrade-Navarro MA, Perez-Iratxeta C. Prediction of protein secondary structure from circular dichroism using theoretically derived spectra. Proteins Struct Funct Bioinforma. 2012;80: 374–381. doi:10.1002/prot.23188

14. Štěpánek P, Bouř P. Multi-scale modeling of electronic spectra of three aromatic amino acids: importance of conformational averaging and explicit solute–solvent interactions. Phys Chem Chem Phys. 2014;16: 20639–20649. doi:10.1039/C4CP02668C

15. Fukuyama T, Matsuo K, Gekko K. Vacuum-Ultraviolet Electronic Circular Dichroism of L-Alanine in Aqueous Solution Investigated by Time-Dependent Density Functional Theory. J Phys Chem A. 2005;109: 6928–6933. doi:10.1021/jp051763h

16. Oakley MT, Bulheller BM, Hirst JD. First-Principles Calculations of Protein Circular Dichroism in the Far-Ultraviolet and Beyond. Chirality. 2006;18: 340–347. doi:10.1002/chir.20264

17. Bayley PM, Nielsen EB, Schellman JA. The Rotatory properties of molecules containing two peptide groups: theory. J Phys Chem. 1969;73: 228–243.

18. Bulheller BM, Hirst JD. DichroCalc--circular and linear dichroism online. Bioinformatics. 2009;25: 539–540. doi:10.1093/bioinformatics/btp016

19. Mavridis L, Janes RW. PDB2CD: a web-based application for the generation of circular dichroism spectra from protein atomic coordinates. Bioinformatics. 2017;33: 56–63. doi:10.1093/bioinformatics/btw554

20. Sreerama N, Venyaminov SY, Woody RW. Estimation of the number of α-helical and β-strand segments in proteins using circular dichroism spectroscopy. Protein Sci. 1999;8: 370–380.

21. Leskovec J, Rajaraman A, Ullman JD. Mining of Massive Datasets. 2nd ed. Cameridge: Camebridge University Press; 2014.

22. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res. 2000;28: 235–242. doi:10.1093/nar/28.1.235

23. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH–a hierarchic classification of protein domain structures. Structure. 1997;5: 1093–1109.

24. Johnson WC. Analyzing protein circular dichroism spectra for accurate secondary structures. Proteins Struct Funct Bioinforma. 1999;35: 307–312.

25. Kuipers BJH, Gruppen H. Prediction of Molar Extinction Coefficients of Proteins and Peptides Using UV Absorption of the Constituent Amino Acids at 214 nm To Enable Quantitative Reverse Phase High-Performance Liquid Chromatography-Mass Spectrometry Analysis. J Agric Food Chem. 2007;55: 5445–5451. doi:10.1021/jf070337l

26. Abraham MJ, Murtola T, Schulz R, Páll S, Smith JC, Hess B, et al. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX. 2015;1–2: 19–25. doi:10.1016/j.softx.2015.06.001

27. Huang J, Rauscher S, Nawrocki G, Ran T, Feig M, de Groot BL, et al. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. Nat Methods. 2016;14: 71–73. doi:10.1038/nmeth.4067

28. MacKerell Jr AD, Bashford D, Bellott M, Dunbrack Jr RL, Evanseck JD, Field MJ, et al. All-atom empirical potential for molecular modeling and dynamics studies of proteins. J Phys Chem B. 1998;102: 3586–3616.

29. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983;22: 2577–2637.

30. Nagy G, Oostenbrink C. Dihedral-Based Segment Identification and Classification of Biopolymers I: Proteins. J Chem Inf Model. 2014;54: 266–277. doi:10.1021/ci400541d

31. Ho BK, Curmi PMG. Twist and shear in β-sheets and β-ribbons. J Mol Biol. 2002;317: 291–308. doi:10.1006/jmbi.2001.5385

32. Sreerama N, Woody RW. Poly (Pro) II helixes in globular proteins: Identification and circular dichroic analysis. Biochemistry (Mosc). 1994;33: 10022–10025.

33. Reed J, Reed TA. A Set of Constructed Type Spectra for the Practical Estimation of Peptide Secondary Structure from Circular Dichroism. Anal Biochem. 1997;254: 36–40.

34. Shen Y, Bax A. SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network. J Biomol NMR. 2010;48: 13–22. doi:10.1007/s10858-010-9433-9

35. Nishino H, Kosaka A, Hembury GA, Matsushima K, Inoue Y. The pH dependence of the anisotropy factors of essential amino acids. J Chem Soc Perkin Trans 2. 2002; 582–590. doi:10.1039/b108575c

Figures:

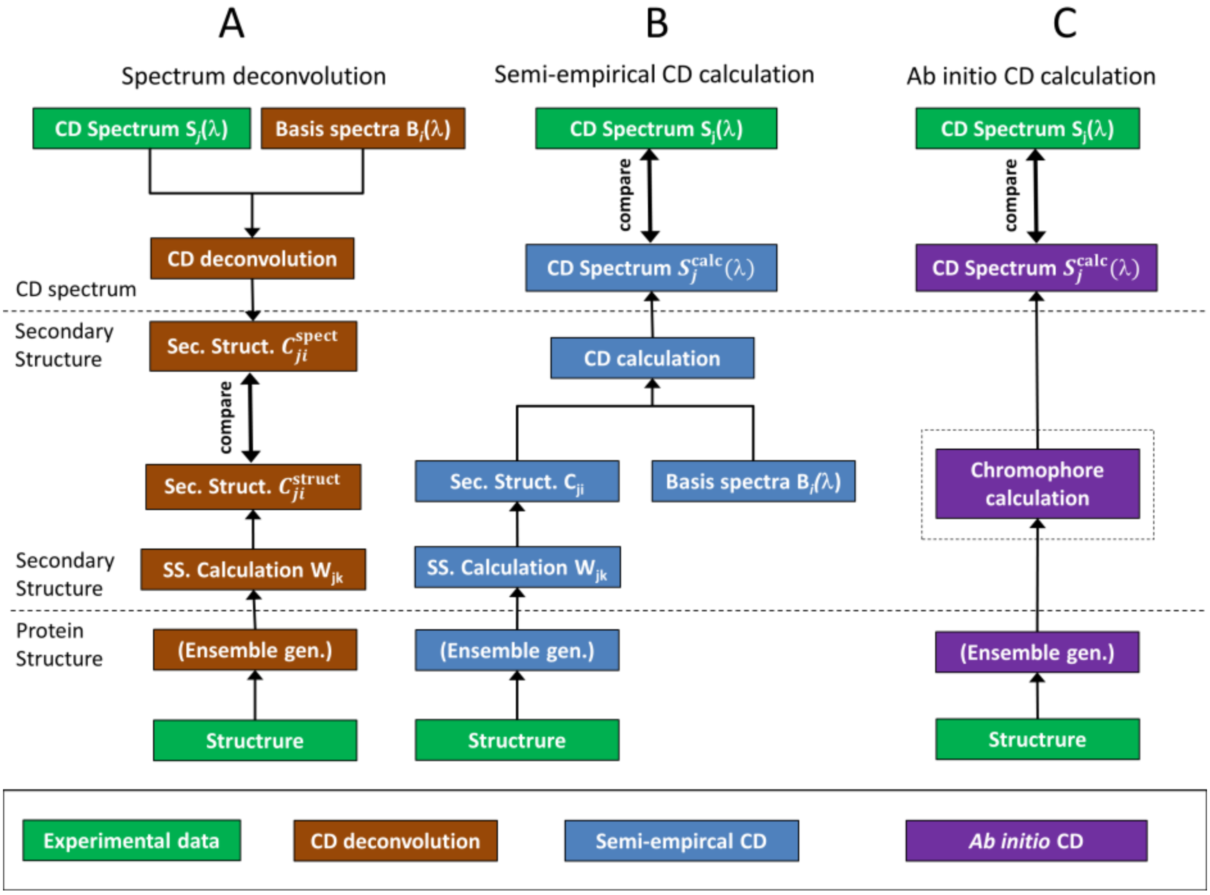
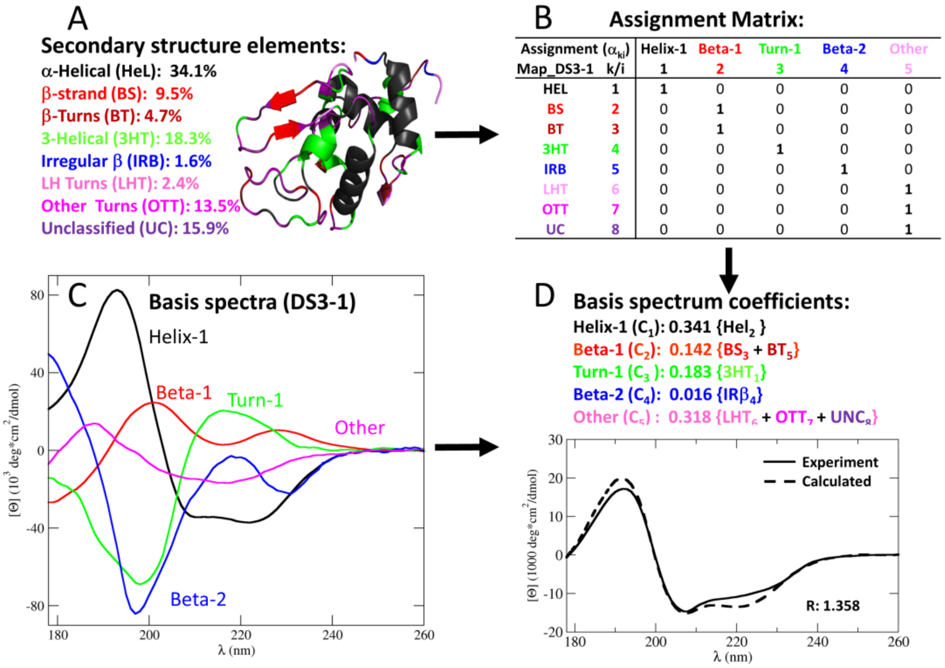


Figure 1: Schemes to compare a protein structure with its circular dichroism spectrum. Green rectangles represent experimental data, brown, blue, and purple fields are related to spectrum deconvolution, semi-empirical- and *ab initio* spectrum calculation, respectively. During spectrum deconvolution (panel A), the secondary structure is estimated from the CD spectrum and calculated from the structure independently, then compared on the secondary structure level. In contrast, during the semi empirical (Panel B) and *ab initio* (Panel C) prediction methods a CD spectrum is computed from the structure and compared directly to the experimentally observable spectrum.



Assignment (α_{ki})		Helix-1	Beta-1	Turn-1	Beta-2	Other
Map_DS3-1	k/i	1	2	3	4	5
HEL	1	1	0	0	0	0
BS	2	0	1	0	0	0
BT	3	0	1	0	0	0
3HT	4	0	0	1	0	0
IRB	5	0	0	0	1	0
LHT	6	0	0	0	0	1
OTT	7	0	0	0	0	1
UC	8	0	0	0	0	1

Figure 2: Semi-empirical CD spectrum calculation scheme. Panel A shows the cartoon representation and secondary structure composition of Lysozyme (pdb code: 4lzt), coloured according to the structural elements of the simplified DISICL library. The Secondary structure information is translated into a theoretical CD spectrum by a basis set (Map_DS3-1), consisting of an assignment matrix (panel B) and a set of basis spectra (panel C). Panel D shows the CD spectrum (dashed line) calculated as the weighted average of basis spectra. The secondary structure composition and assignment matrix determine the basis spectrum coefficients (*C_i*, on panel D) for weighing the basis spectra. The deviation between the experimental (solid line in panel D) and calculated (dashed line) CD spectrum (R:) is shown in mean residue ellipticity units (10^3 degree*cm^2/dmol). The table displays the ID (*k*) and abbreviation of the secondary structure element, the name and ID (*i*) of the basis spectra, and the assignments matrix of structure coefficients (α_{ki}) connecting them. The basis spectra are shown as coloured lines in Panel C, and the same colour coding is used in Panel D to display their coefficients.

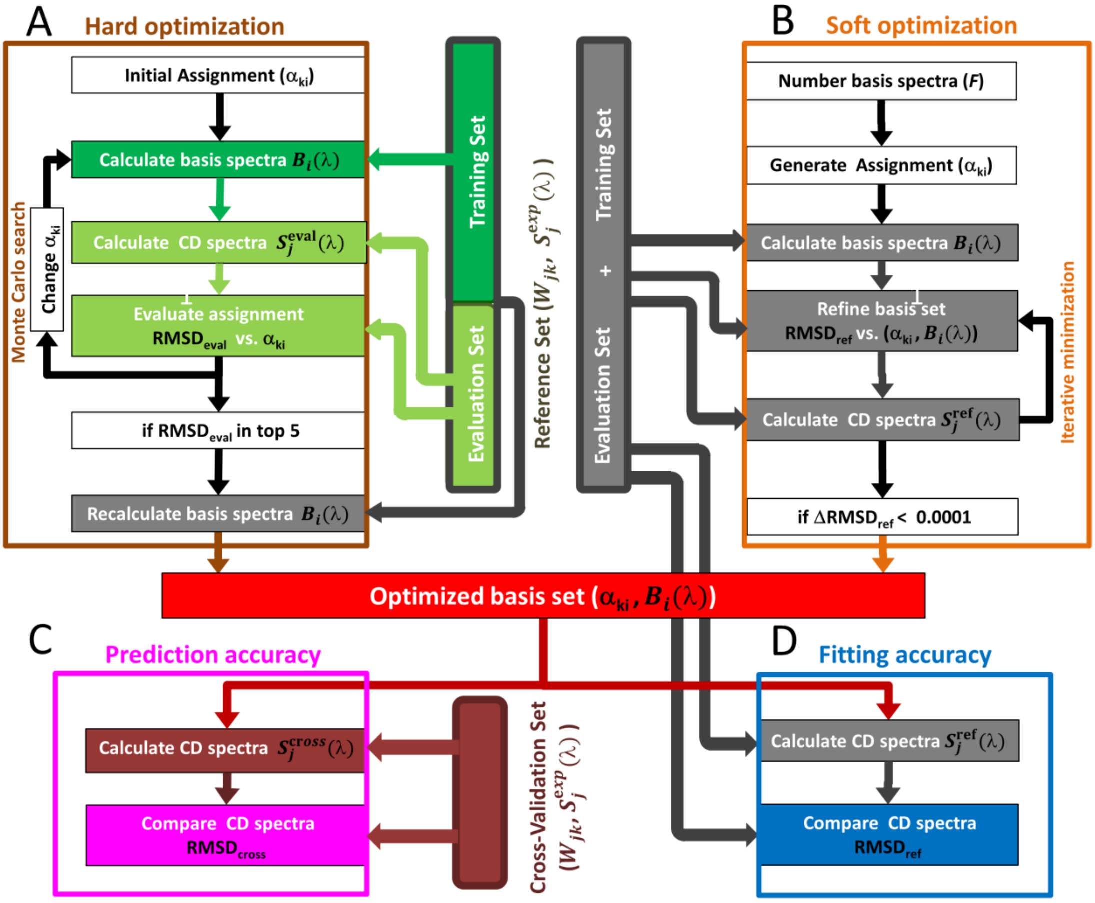
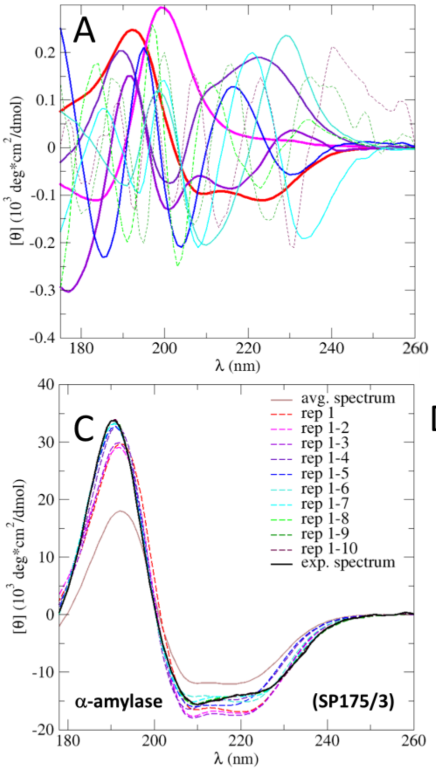

A
Hard optimization
Initial Assignment (α_{ki})
Calculate basis spectra $B_i(\lambda)$
Calculate CD spectra $S_j^{eval}(\lambda)$
Evaluate assignment RMSD$_{eval}$ vs. α_{ki}
Monte Carlo search
Change α_{ki}
if RMSD$_{eval}$ in top 5
Recalculate basis spectra $B_i(\lambda)$
Training Set
Evaluation Set
Reference Set (W_{jk}, $S_j^{exp}(\lambda)$)
Training Set + Evaluation Set
B
Soft optimization
Number basis spectra (F)
Generate Assignment (α_{ki})
Calculate basis spectra $B_i(\lambda)$
Refine basis set RMSD$_{ref}$ vs. (α_{ki}, $B_i(\lambda)$)
Calculate CD spectra $S_j^{ref}(\lambda)$
Iterative minimization
if ΔRMSD$_{ref}$ < 0.0001
Optimized basis set (α_{ki}, $B_i(\lambda)$)
C
Prediction accuracy
Calculate CD spectra $S_j^{cross}(\lambda)$
Compare CD spectra RMSD$_{cross}$
Cross-Validation Set (W_{jk}, $S_j^{exp}(\lambda)$)
D
Fitting accuracy
Calculate CD spectra $S_j^{ref}(\lambda)$
Compare CD spectra RMSD$_{ref}$

Figure 3: Basis set optimization and assessment schemes. The basis sets (shown in red) are derived and optimized either though the hard or the soft optimization approach, using the same reference set of proteins, including the secondary structure information (W_{jk}) and CD spectra ($S_j^{\text{exp}}(\lambda)$) of each protein. During the hard optimization (panel A) the reference set was divided into a training set (dark green) and an evaluation set (light green) to perform an "internal" cross validation during the search for optimal assignments. The undivided reference set (shown as grey boxes and arrows) was used during the soft optimization (panel B) as well as at the end of the hard optimization to calculate basis spectra for the best assignments. The same undivided reference set was used to assess the fitting accuracy (panel D) of the optimized basis set (regardless of the optimization method), where CD spectra calculated from the structural information were compared with the experimental CD spectra of the reference proteins. In contrast, during the assessment of the prediction accuracy (panel C), a different set of proteins (shown in dark red) were used for cross-validating the predictive power of the optimized basis sets.



B

	PC #	Eigenvalue	Var %	Cumulative
━━	1	4417	81.0	81.0
━━	2	677.6	12.4	93.4
━━	3	138.8	2.55	96.0
━━	4	103.3	1.89	97.9
━━	5	47.7	0.87	98.8
━━	6	42.8	0.79	99.6
━━	7	11.4	0.21	99.8
- - -	8	8.4	0.15	99.9
- - -	9	1.6	0.029	99.9
- - -	10	0.8	0.015	100.0

D

Basis set	Aamy	Cah1	Cona	SP175 av.
#	R_3	R_{14}	R_{20}	$RMSD_{ref}$
0	5.455	12.534	6.131	6.359
1	2.459	3.170	4.236	2.820
2	2.180	2.259	1.083	1.680
3	2.105	2.245	1.044	1.335
4	1.359	1.179	1.023	0.955
5	1.121	0.705	1.018	0.720
6	0.682	0.221	0.458	0.481
7	0.466	0.212	0.350	0.350
8	0.207	0.207	0.243	0.237
9	0.201	0.191	0.194	0.203
10	0.157	0.156	0.176	0.182

Figure 4: Principal component analysis of the SP175 protein CD spectra. A) graphical representation of the first 10 principal component vectors sorted by their contribution to the spectral variance. B) Eigenvalue, contribution to variance, and cumulative contribution to the spectral variance for the same PC vectors. C) Reconstruction of the CD spectrum of α-amylase (Aamy) by its projection on the first 0-10 PC vectors. The original spectrum is shown in black, the average spectrum of SP 175 data set is shown in brown. The reconstructed spectra are shown as coloured dashed lines. D) RMSD between the reconstruction of three selected proteins – α-amylase, carbonic anhydrase I (Cah1), and Concanavalin A (Cona) – and their original CD spectrum as function of PC vectors used. The column SP175 av. shows average RMSD for all 71 proteins in the data set.

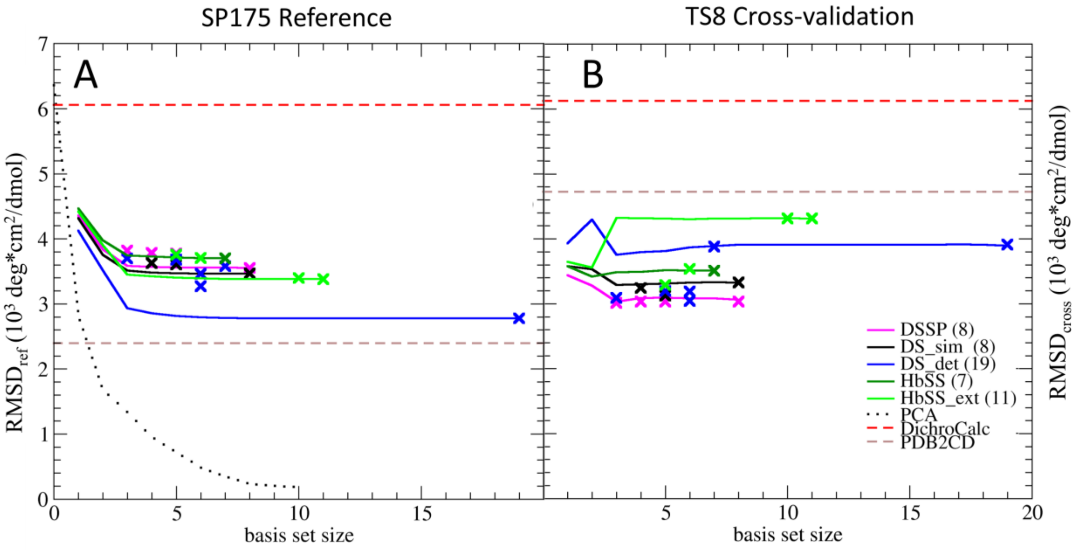


Figure 5: Basis set performance on globular proteins. The panels show the basis set accuracy for A) the reference set for globular proteins (SP175), and B) a small independent set of globular proteins used for cross-validation (TS8). The average deviation between the CD spectra calculated by a basis set and experimental CD spectra (RMSD) is shown as the function of the number of basis spectra in the respective basis set. Series of basis sets derived using the soft basis set optimization approach are shown as solid lines coloured according to the underlying secondary structure classification method. Basis sets derived using the hard optimization approach are shown as crosses also coloured according to the underlying secondary structure classification. The average deviation of published CD prediction algorithms DichroCalc and P2CD are shown as red and brown horizontal dashed lines, respectively. The highest limit of fitting accuracy defined by PCA basis sets is shown as a black dotted line in panel A. The numbers in brackets behind the secondary structutre classification methods (DSSP, DS_sim, DS_det, HbSS, HbSS_ext) denote the number structural elements of the classification.

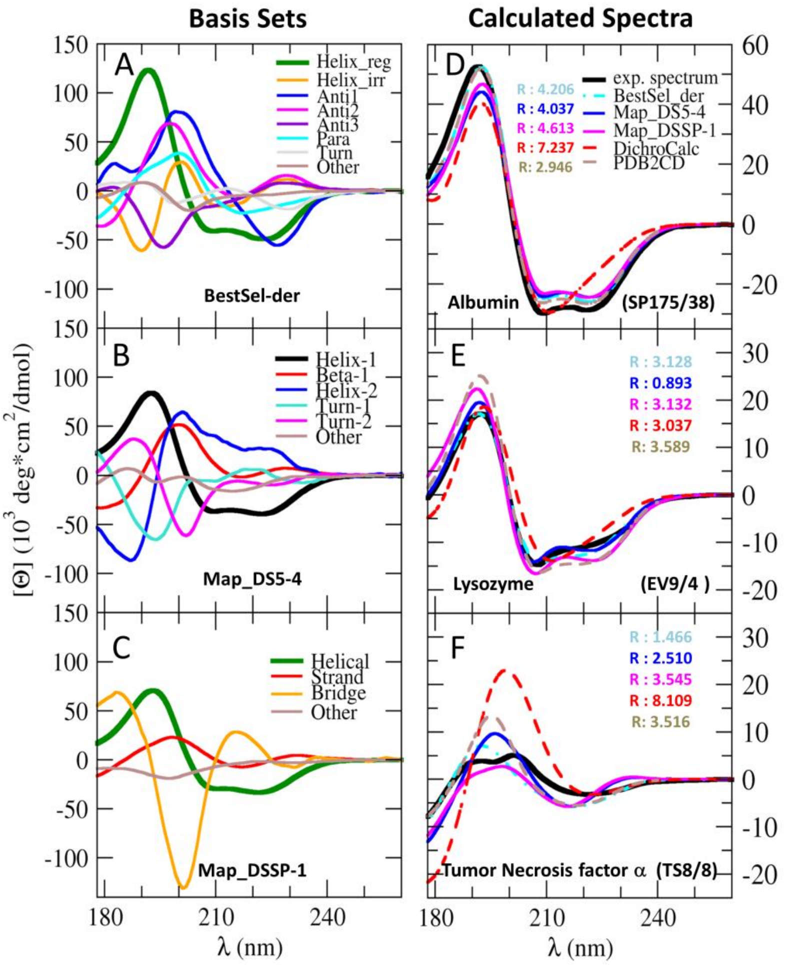
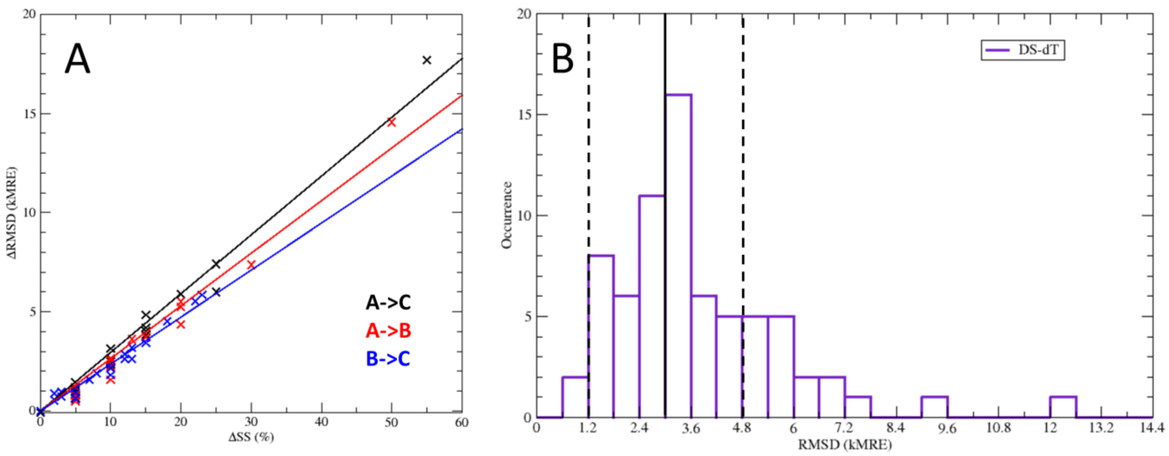
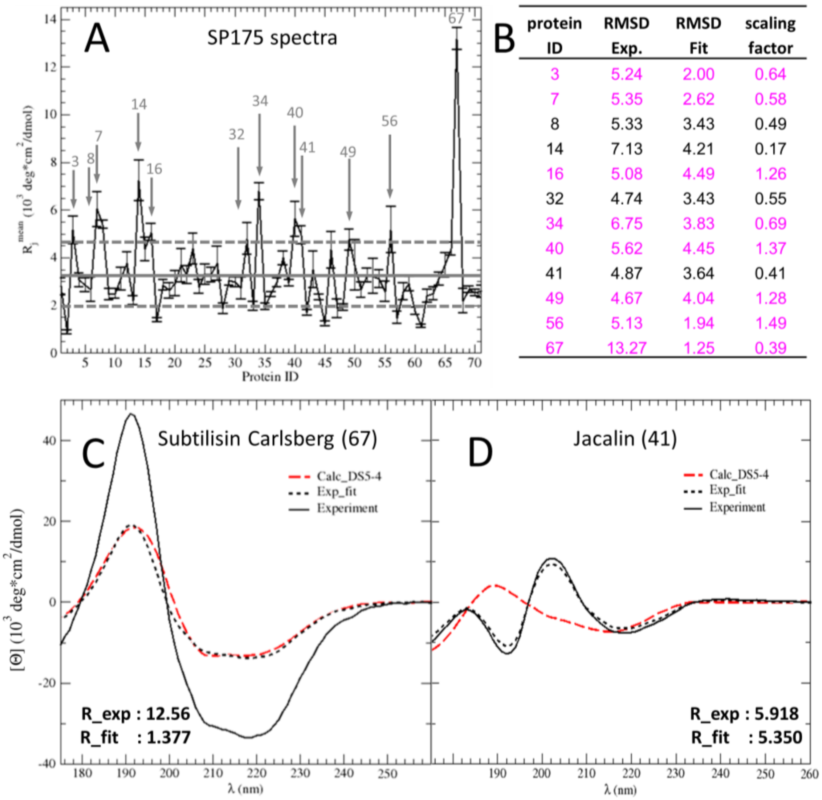


Figure 6: Basis spectrum sets, experimental and calculated CD spectra of selected proteins. The basis spectra of three high-accuracy basis sets with nine, six, and four components is shown in panels A - C, respectively. Panels D - F show the experimental (solid black line) and calculated CD spectra of human serum albumin, lysozyme, and tumor necrosis factor α, respectively. The accuracy of the CD spectra calculated from these basis sets was compared with spectra from two competing algorithms Dichrocalc and PDB2CD. The average RMSD (R:) from the experimental spectrum is displayed in the corresponding colour. All RMSD values are in 10^3 deg*cm^2/dmol (kMRE) units.



Change	m	R^2	RMSD (kMRE)	0.6	1.2	2.4	3.6	4.8	6.0	7.2
A->C	0.296	0.985	ΔSS (%)	2.0	4.1	8.1	12.2	16.2	20.3	24.3
A->B	0.265	0.984	ΔSS (%)	2.3	4.5	9.0	13.6	18.1	22.6	27.1
B->C	0.237	0.989	ΔSS (%)	2.5	5.1	10.1	15.2	20.3	25.3	30.4

Figure 7: Linear correlations between the deviation of the calculated and expeirmental CD spectra and the deviation from the ideal secondary structure composition. The table displays the slope (m) and the square of the Pearson correlation coeffcient (R^2) of the fitted linear functions that connect the deviation from the experimental CD spectra (RMSD) to the deviation in secondary structure (ΔSS) for α-helix to coil (A->C), α-helix to β-strand (A->B) and β-strand to coil (B->C) type deviations. A) The linear fitting functions obtained from systematically altering the secondary structure composition of three selected proteins. B) The RMSD distribution of predicted spectra of the SP175 reference proteins, calculated with the SESCA basis set DS-dT. The vertical lines on the plot indicate the average RMSD (solid) and the standard deviation (dashed) of the predicted spectra for the TS8 cross validation set. The right side of the Table was used to estimate the maximal deviation in the secondary structure composition between the crystal structure and the ideal solution structure of the protein, based on the RMSD of its predicted spectrum.



protein ID	RMSD Exp.	RMSD Fit	scaling factor
3	5.24	2.00	0.64
7	5.35	2.62	0.58
8	5.33	3.43	0.49
14	7.13	4.21	0.17
16	5.08	4.49	1.26
32	4.74	3.43	0.55
34	6.75	3.83	0.69
40	5.62	4.45	1.37
41	4.87	3.64	0.41
49	4.67	4.04	1.28
56	5.13	1.94	1.49
67	13.27	1.25	0.39

Figure 8: Analysis of the spectrum prediction quality for the proteins of the SP175 data set. A) Mean deviation (RMSD) between the experimental CD spectra and spectra calculated by six different CD prediction methods (described Section 5.3)). The grey line in the Figure represents the average RMSD of the TS8 cross-validation set, and the dashed lines show standard deviation from that mean of the six RMSDs. Twelve hard-to-predict proteins with unusually large mean RMSD are highlighted by grey arrows. B) Mean RMSD of twelve hard-to predict proteins before ($\mathrm{RMSD}_{\mathrm{exp}}$) and after ($\mathrm{RMSD}_{\mathrm{fit}}$) the experimental spectra were rescaled, as well as the scaling factors yielding the lowest RMSD. Proteins for which scaling could yield a significantly better agreement with the calculated spectra are marked with magenta. C) Example protein 1: significant RMSD improvement by scaling the experimental CD spectrum and D) Example protein 2: where scaling could not improve the RMSD significantly. For panels C and D the experimental CD spectrum is shown as a solid black line, the rescaled experimental spectrum is shown as a dotted black line, and the spectrum calculated by the basis set DS5-4 is shown as a red dashed line. The name and index number of the protein is shown on the top of the panel, while the unscaled (R_exp) and scaled (R_fit) RMSD of the DS5-4 spectrum in kMRE units is shown on the bottom.

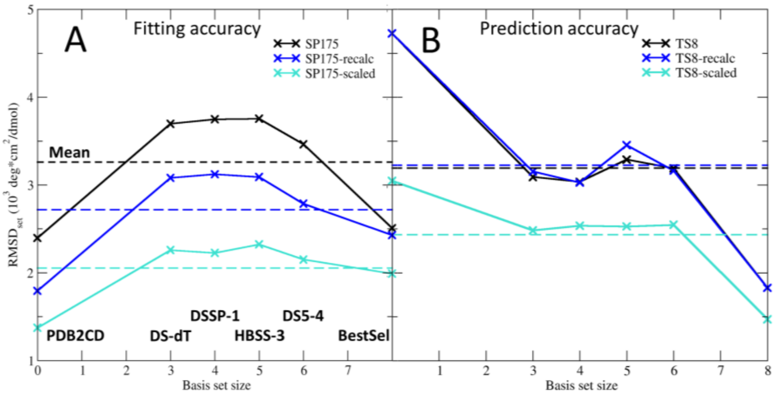


Figure 9: Changes in the mean fitting accuracy (Panel A) and prediction accuracy (Panel B). The method independent mean RMSDs (shown as dashed lines) for the SP175 and TS8 data sets were calculated as the average RMSD$_{\text{set}}$ of six spectrum prediction methods (crosses) including PDB2CD, four optimized SESCA basis sets of different sizes and underlying classification schemes (DS-dT, DSSP-1, HBSS-3 and DS5-4), and the BestSel reconstruction basis set. The accuracy calculated for the original unmodified data sets are shown in black, whilst the accuracies calculated after the removal of hard-to-predict proteins from the SP175 reference set and recalculation of the SESCA basis spectra are shown in dark blue. The cyan accuracies were obtained by applying scaling factors to the experimental spectra of both data sets to account for normalization problems.

*

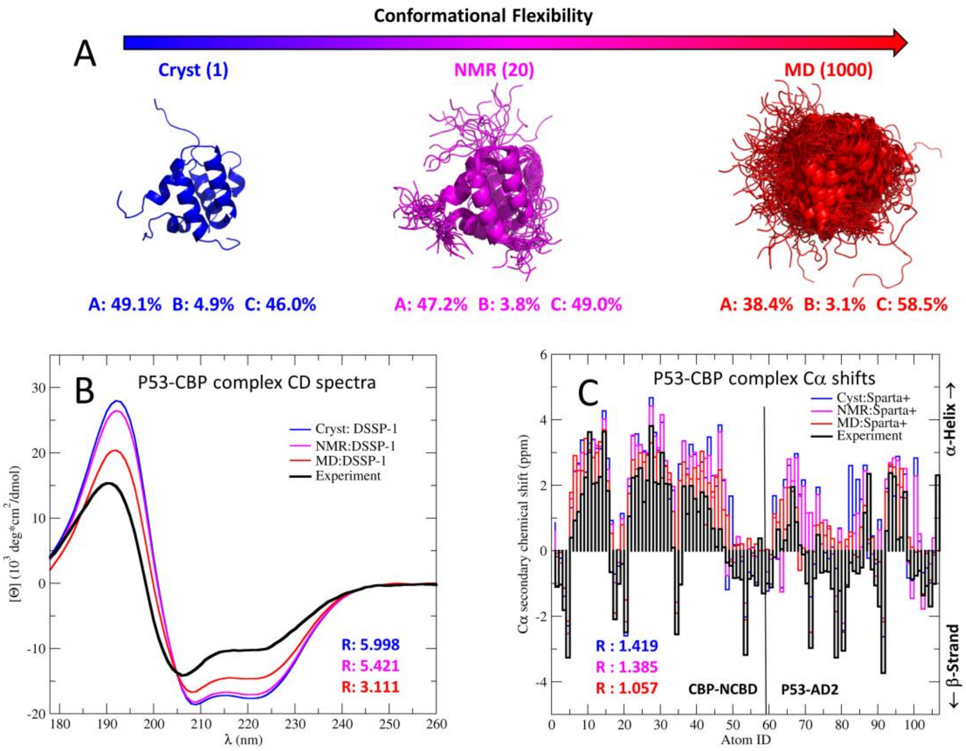


Figure 10: The impact of conformational flexibility: Comparison between measured experimental observables and the same observables calculated from three structural models including different levels of protein dynamics. Panel A shows the three structural models: one model with no conformational flexibility, consisting of a single structure (Cryst), one model with limited flexibility, consisting of a bundle 20 structures from NMR (NMR, PDB code 2L14), and one highly flexible model with 1000 structures obtained from an MD simulation (MD, 100 are shown). The line at bottom of panel A shows the average secondary structure composition of the models where A, B, and C abbreviates fractions of α-helices, β-strands, Coil structures, respectively. Panels B and C depict the comparison for the calculated CD spectra and C_α secondary chemical shifts of the P53-CBP complex, respectively. The measured experimental observables on panels B and C are shown as black solid lines, calculated observables are shown in different colours according to the underlying model. The RMSD (R:) from the experimental observable is also shown in the corresponding colour.

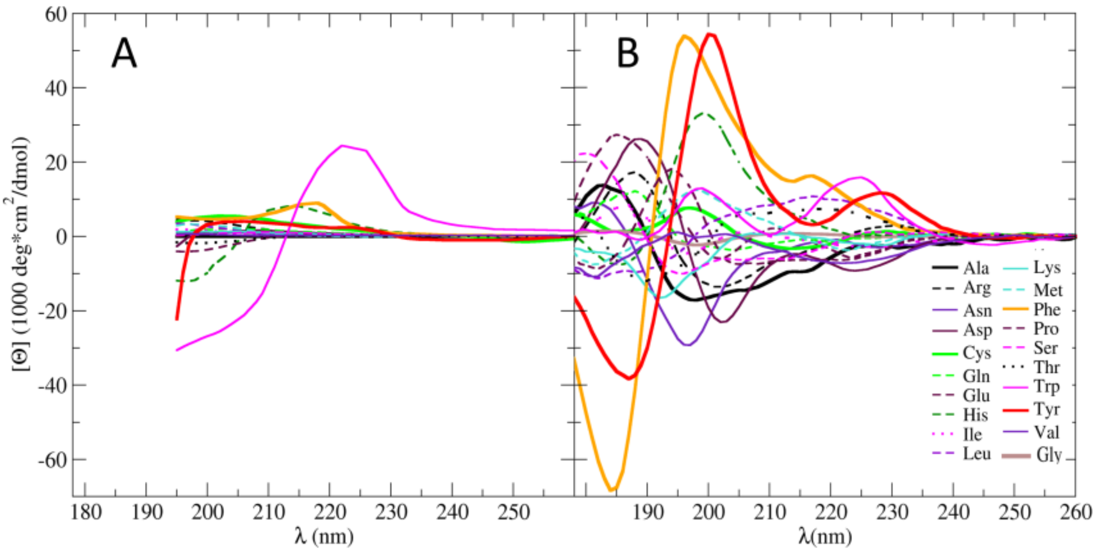


Figure 11: Circular dichroism contribution of amino acid side chains. A) Experimentally measured CD spectra for natural amino acids at pH = 7.0 adapted from Nishino *et. al.* [35]. B) Calculated side chain contributions for each amino acid side chain, derived from the CD spectra of 59 globular proteins and the 20 Ac-GXG-NH_2 peptides. The (basis) spectra are colour coded according to the amino acid side chain groups they represent.

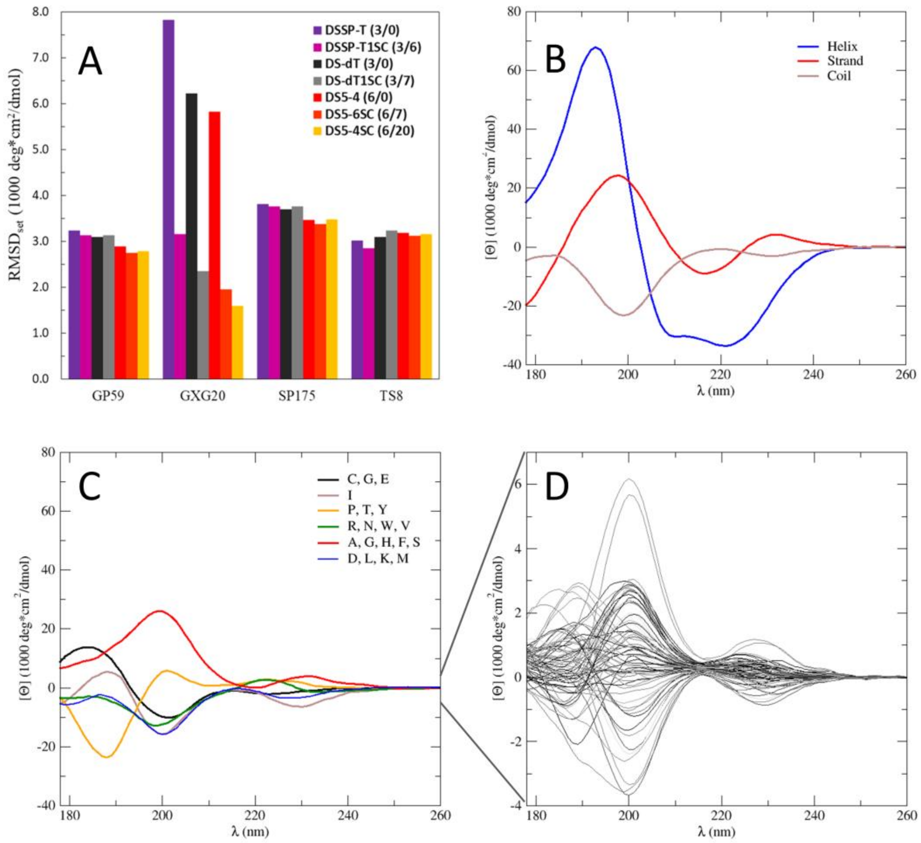


Figure 12: Comparison of backbone and side chain contributions. A) Comparison between selected basis sets with and without side chain corrections. The legends denote the name of the basis set followed by the number of backbone and side chain basis spectra in brackets. The accuracy (RMSD$_{set}$) of the basis sets achieved on the globular protein (GP59) and short peptide (GXG20) sub-sets of their training set, as well as the accuracy for the full SP175 reference set and the TS8 cross-validation set. B) Backbone and C) side chain basis spectra of the basis set DSSP-dT1SC. The amino acids assigned to the side chain basis spectra are abbriviated with on-letter codes. D) Combined side chain contributions of the basis set DSSP-dT1SC for the SP175 reference set. The scale of side chain contributions was changed for better visibility.

Tables:

Table 1: Correlation analysis of the spectral components. The six best correlated structural properties are listed for each of the first principal components of the SP175 CD spectra. The table displays the abbreviated code of the structural property (Prop), the Pearson correlation score (Corr.) between the projections of the PC vector, and the coefficients of the structural property (the fraction of secondary structure element or amino acid in a protein), the type and a short description of the structural property. The type (in parenthesis) defines the source algorithm for secondary structure elements (DSSP, HbSS, DISICL or BestSel algorithms) and (AA) for amino acids. The short description shows if the secondary structure element is either associated with α-helix, irregular helix (Helix), β-strand or turn structures

PC1	Corr.	Prop	Desc.	PC6	Corr.	Prop	Desc.
1	0.921	Hel1 (Best)	α-helix	1	0.201	SER (AA)	Amino A.
2	0.906	Hel1 (SEL)	α-helix	2	0.163	CYS (AA)	Amino A.
3	0.9	ALH (DISICL)	α-helix	3	0.138	RHA (HbSS)	Strand
4	0.898	Hel (DISICL)	α-helix	4	0.157	Hel2 (Best)	Helix
5	0.892	4H (DSSP)	α-helix	5	0.126	Hel1 (SEL)	α-helix
6	0.891	4H (HbSS)	α-helix	6	0.116	ALH (DISICL)	α-helix
PC2	**Corr.**	**Prop**	**Desc.**	**PC7**	**Corr.**	**Prop**	**Desc.**
1	0.532	EBS (DISICL)	β-strand	1	0.285	RHP (HbSS)	β-strand
2	0.513	Anti1 (BEST)	β-strand	2	0.274	BSP (HbSS)	β-strand
3	0.444	NBA (HbSS)	β-strand	3	0.25	Para (Best)	β-strand
4	0.418	Anti2 (Best)	β-strand	4	0.23	Turn (Sel)	Turn
5	0.395	BS (HbSS)	β-strand	5	0.205	Bend (DSSP)	Turn
6	0.352	HIS (AA)	Amino A.	6	0.169	GXT (DISICL)	Turn
PC3	**Corr.**	**Prop**	**Desc.**	**PC8**	**Corr.**	**Prop**	**Desc.**
1	0.31	BS (HbSS)	β-strand	1	0.386	3H (DSSP)	Helix
2	0.299	SCH (DISICL)	Turn	2	0.344	3H (HbSS)	Helix
3	0.254	NBS (DISICL)	β-strand	3	0.3	5H (HbSS)	Helix
4	0.23	Bend (DSSP)	Turn	4	0.273	HC (DISICL)	Turn
5	0.216	NBA (HbSS)	β-strand	5	0.253	MET (AA)	Amino A.
6	0.205	THR (AA)	Amino A.	6	0.139	Other (Best)	Turn
PC4	**Corr.**	**Prop**	**Desc.**	**PC9**	**Corr.**	**Prop**	**Desc.**
1	0.471	ARG (AA)	Amino A.	1	0.223	ASP (AA)	Amino A.
2	0.397	LHH (DISICL)	Turn	2	0.202	3H(HbSS)	Helix.
3	0.306	Anti2 (Best)	β-strand	3	0.192	GLU (AA)	Amino A
4	0.293	NBA (HbSS)	β-strand	4	0.152	ILE (AA)	Amino A.
5	0.299	SCH (DISICL)	Turn	5	0.152	3H (DSSP)	Helix
6	0.272	LHT (DISICL)	Turn	6	0.126	PIH (DISICL)	Helix
PC5	**Corr.**	**Prop**	**Desc.**	**PC10**	**Corr.**	**Prop**	**Desc.**
1	0.394	3HT(DISICL)	Helix	1	0.214	PHE (AA)	Amino A.
2	0.376	3H (DISICL)	Helix	2	0.15	TRP (AA)	Amino A.
3	0.33	3H (DSSP)	Helix	3	0.14	SER (AA)	Amino A.
4	0.321	3H (HbSS)	Helix	4	0.133	RHA (HbSS)	β-strand
5	0.296	Cys (AA)	Amino A.	5	0.116	Bend (DSSP)	Turn
6	0.294	Hel2 (SEL)	Helix	6	0.102	LHT (DISICL)	Turn