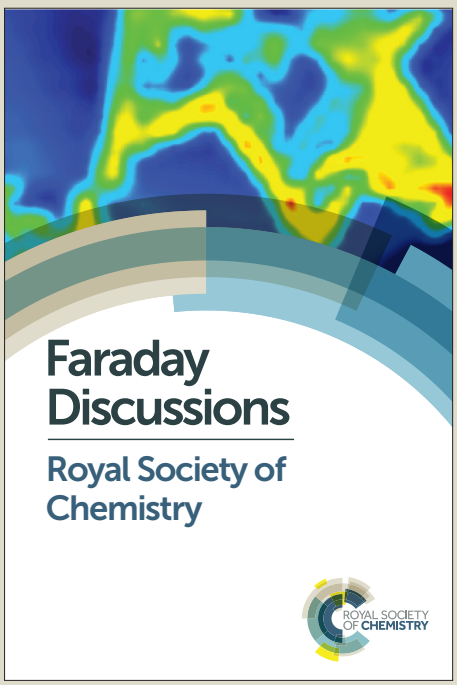
Faraday Discussions

Accepted Manuscript

This manuscript will be presented and discussed at a forthcoming Faraday Discussion meeting. All delegates can contribute to the discussion which will be included in the final volume.

Register now to attend! Full details of all upcoming meetings: http://rsc.li/fd-upcoming-meetings



This is an *Accepted Manuscript*, which has been through the Royal Society of Chemistry peer review process and has been accepted for publication.

Accepted Manuscripts are published online shortly after acceptance, before technical editing, formatting and proof reading. Using this free service, authors can make their results available to the community, in citable form, before we publish the edited article. We will replace this *Accepted Manuscript* with the edited and formatted *Advance Article* as soon as it is available.

You can find more information about *Accepted Manuscripts* in the Information for Authors.

Please note that technical editing may introduce minor changes to the text and/or graphics, which may alter content. The journal's standard Terms & Conditions and the Ethical guidelines still apply. In no event shall the Royal Society of Chemistry be held responsible for any errors or omissions in this *Accepted Manuscript* or any consequences arising from the use of any information it contains.

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/faraday_d

Vibronic structure in the far-UV electronic circular dichroism spectra of proteins

Zhuo Li, David Robinson and Jonathan D. Hirst*
DOI: 10.1039/b000000x [DO NOT ALTER/DELETE THIS TEXT]

The Franck-Condon effect is considered and the vibrational structure of the $\pi_{nb}\pi^*$ transition of the peptide backbone is incorporated into matrix method calculations of the electronic circular dichroism (CD) spectra of proteins in the far-ultraviolet. We employ the state-averaged CASPT2 method to calculate the ground and $\pi_{nb}\pi^*$ excited state geometries and frequencies of N-methylacetamide (NMA), which represents the peptide chromophore. The results of these calculations are used to incorporate vibronic levels of the excited states into the matrix method calculation. The CD spectra of a set of 49 proteins, comprising a range of structural types, are calculated to assess the influence of the vibrational structure. The calculated spectra of α-helical proteins are better resolved using the vibronic parameters and correlation between the experimental and the calculated intensity of less regular β structure proteins improves over most wavelengths in the far-UV. No obvious improvement is observed in the calculated spectra of β-sheet proteins. Our high-level *ab initio* calculations of the vibronic structure of the $\pi_{nb}\pi^*$ transition in NMA have provided some further insight into the physical origins of the nature of protein CD spectra in the far-UV.

1 Introduction

Electronic circular dichroism (CD) spectroscopy in the far-UV is a widely used technique for studying chiral molecules. Since almost all biological systems and many non-living systems are chiral, CD spectroscopy has become an important tool for structure determination in several research areas. Applications include establishing that a chiral molecule has indeed been synthesized[1], determining the absolute configuration of a molecule[2], studying molecular interactions by comparing the calculated spectra of different models with the experimental spectrum[3], and probing the structures of biological macromolecules, such as proteins[4] or DNA[5], as well as their folding[6, 7] and binding[8] processes. These studies can benefit from an accurate calculation from first principles of the CD spectrum.

Much effort has been made to improve the accuracy of calculations of CD spectra, in part to understand better the relationship between molecular structure and the associated spectra. For small molecules such calculations have become fairly routine and can be performed with density functional theory (DFT)[2, 9], Hartree-Fock[10] or coupled cluster theory[11]. High-level calculations have been performed on simple amides and dipeptides[12-14]. However, calculation of the CD spectra of proteins remains a challenge, due to their size and flexibility. There are several approximate

School of Chemistry, University of Nottingham, Nottingham NG7 2RD, UK. Fax: 44 115 9513562; Tel: 44 115 9513478; E-mail: jonathan.hirst@nottingham.ac.uk

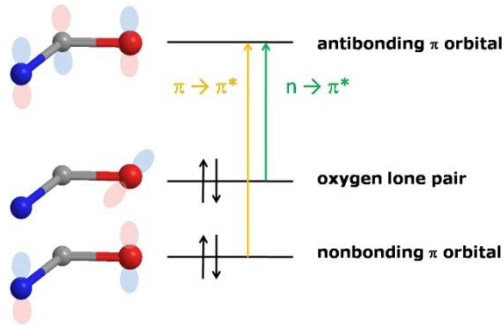


Fig.1 Molecular orbitals relevant to electronic transitions in the far-UV region. The ball and stick model represents N-methylacetamide (blue-nitrogen, grey-carbon, red-oxygen).

methods for polymeric systems like proteins, such as the dipole interaction model[15-19] and the matrix method[20-23]. The calculations in this study are based on the matrix method, an exciton approach[24], which has proven to be quite successful[25, 26]. The calculations are almost, but not fully, quantitative for proteins with a high amount of α-helical structure; the spectra for proteins with other secondary types are less well reproduced. Additional aspects may need to be considered to obtain more accurate calculated spectra.

The parameterization in the matrix method is from *ab initio* calculation of *trans* N-methylacetamide (NMA), which represents the chromophoric group of the protein backbone[27]. Fig. 1 shows the two electronic transitions of NMA located in the far-UV region, which are from the amide non-bonding π orbital (π_{nb}) to the anti-bonding π orbital (π*) and from the oxygen lone pair orbital n to the π* orbital. The parameters used in the CD calculation describe the charge distributions of the different electronic states. Deviation from planarity of the peptide bond will have some influence on the electronic structure[28], but this is not the focus of the current study. Hydrogen bonding may also be important, but previous work on the explicit consideration of this did not find any significant improvement for protein calculations[29, 30].

Vibrational fine structure is common in high-resolution gas phase spectra. These fine features may be not obvious in solution, but can contribute to the broadening of bands. The application of a (uniform) bandwidth to calculated line spectra to model the broadening is an approximation. To explore this aspect further, we have examined explicit incorporation of vibronic effects of the $\pi_{nb}\pi^*$ transition of the peptide chromophore.

The Franck-Condon effect has been considered in previous calculations of the CD spectra of various molecules. The CD spectrum of dimethyloxirane in the gas phase was calculated using time-dependent DFT (TD-DFT) including the envelopes of several states[31]. Some narrower bands arise from excitations to the 1B, 2A and 3B states, while other broader bands were assigned to transitions to the 4A and 5A states. The CD spectrum was reproduced well once the vibrational structure for the various transitions was considered. The disappearance of a negative band in the experimental spectrum was rationalised by the calculation: the band was cancelled by other positive bands because of a strong vibrational progression. Vibronic

contributions were taken into account in calculations of the CD spectra of naphthalenediimide (NDI) dimers[32] and oligomers[33]. The CD spectra of NDI helical nanotubes were calculated with the matrix method considering several vibronic transitions associated with two electronic transitions. Inclusion of the Franck-Condon effect was essential in reproducing the key features of the CD spectra.

An early study considering vibronic coupling in proteins investigated absorption in the far-UV of an idealized α-helix[34] and poly-L-lysine as *β*-sheet[35] and calculated the spectrum, focusing on the C-N stretch mode in the $\pi_{nb}\pi^*$ band. The predominant role of the C-N stretch in this transition has been confirmed by resonance Raman experiments[36]. No combination transition along the amide II' overtones was observed in D_2O. Although the role of protic solvent is not clear yet, the band of the carbonyl stretch, which is the primary mode in gas phase and in acetonitrile, was not apparent in aqueous solution.

In this study, we calculate the vibronic structure of the $\pi_{nb}\pi^*$ transition of the backbone chromophore and incorporate this into the calculations of protein CD spectra. There are various approaches to exciton-vibrational coupling[37]. Yang[38] described a reduced density–matrix theory for the case where a single vibration mode is predominant in the coupling, with a spectral density used to represent a bath of nuclear degrees of freedom. We take advantage of this model in constructing the system Hamiltonian, which will be described in detail in the next section.

2 Methods

We have considered the *trans* form of NMA, which is much more common in proteins than the *cis* form[39] (0.34% peptides are *cis* in our dataset). The electronic ground and $\pi_{nb}\pi^*$ excited state geometries and vibrational frequencies of NMA were computed using multireference perturbation theory[40] (CASPT2) based on a complete active space self-consistent field (CASSCF) reference and the 6-31G* basis set with the MOLPRO package (Version 2010.1)[41]. A multi-configurational method is necessary for a proper description of the electronic structure of amides, because of a significant interaction between the ground and $\pi_{nb}\pi^*$ states. The active space we used is four electrons in five active orbitals (two occupied and three virtual), which, although minimal, suffices for our purposes. The excited state optimization used state-averaged CASPT2, averaged over the ground, $n\pi^*$ and $\pi_{nb}\pi^*$ states, which should give molecular orbitals that are not biased toward any particular electronic state. Although the ground state of NMA is planar[42], we calculated both the ground and excited states with no symmetry imposed.

The Franck-Condon overlap integral, f_v, characterizes the vibronic transitions, which govern the intensity distribution along the progression in the absorption spectrum. The intensity is proportional to the Franck-Condon factor, which is the square of the overlap integral of the vibrational wave function between the ground and the excited states. ezSpectrum[43] was used to calculate the Franck-Condon overlap integrals within the double-harmonic approximation. The calculations require the equilibrium geometries of the ground and the excited states as well as the frequencies and vibrational mode vectors, which are obtained from the *ab initio* calculations. The Duschinsky rotation is not included in the calculation because there is no severe distortion of the C-N stretch mode between the ground and excited vibrational states. It means the structure of the excited state can align well with the ground state geometry after rotation and translation. The temperature is set to 0 K so

that the system populates the lowest vibrational level of the ground electronic state. No combinations of normal modes were considered, since the C-N stretch is the predominant vibration in the absorption spectrum of NMA in aqueous solution.

In the matrix method, the protein is considered as being composed of M chromophores. The wave function ψ^k is expressed as a linear combination of the wave function Φ_{ia}, which is a product of M chromophoric wave functions ϕ. Each chromophoric group i is considered separately and, in the simplest form of the approach, can only be excited to higher electronic states, ϕ_{ia}, of itself, while others stay in the ground state, ϕ_{j0}.

$$\boldsymbol{\Phi}_{ia} = \phi_{10} \cdots \phi_{ia} \cdots \phi_{j0} \cdots \phi_{M0} \tag{1}$$

$$\psi^k = \sum_i^M \sum_a^{n_i} c_{ia}^k \boldsymbol{\Phi}_{ia} \tag{2}$$

where c_{ia}^k are expansion coefficients, which reflect the interactions of the states, to be determined.

The ground state of the protein is

$$\psi^0 = \phi_{10} \cdots \phi_{i0} \cdots \phi_{j0} \cdots \phi_{M0} \tag{3}$$

With the above approximation, the wave function ψ^k of the electronic excited state k of the protein associated with the energy can be calculated by solving the Schrödinger equation

$$\hat{H}\psi^k = E^k\psi^k \tag{4}$$

The Hamiltonian of the system $\hat{H}$ is constructed as the sum of the local Hamiltonian $\hat{H}_i$ of the independent group i and the interactions between two groups which is denoted as intergroup potential $\hat{V}_{ij}$.

$$\hat{H} = \sum_{i=1}^M \hat{H}_i + \sum_{i=1}^{M-1} \sum_{j=i+1}^M \hat{V}_{ij} \tag{5}$$

The interaction energy between transitions a←0 on group i and transition b←0 on group j is expressed as:

$$V_{i0a;j0b} = \int\int \phi_{i0}\phi_{ia}\widehat{V_{ij}}\,\phi_{j0}\phi_{jb}\,d\tau_i\,d\tau_j \tag{6}$$

The interactions between chromophores are assumed to be purely electrostatic in nature, involving charge densities ρ of separation r.

$$V_{i0a;j0b} = \int\int \frac{\rho_{i0a}(r_i)\rho_{j0b}(r_j)}{4\pi\varepsilon_0 r_{ij}} d\tau_i\,d\tau_j \tag{7}$$

where $\rho_{i0a}(r_i)$ and $\rho_{j0b}(r_j)$ are the transition electron densities of chromophores i and j, ε_0 is the vacuum permittivity and r_{ij} is the separation of the chromophores.

The monopole-monopole approximation is used to make the calculation of $V_{i0a;j0b}$ tractable, whereby monopole charges are fitted to the atoms of the group to reproduce the permanent and transition densities. The Coulomb interactions between these monopoles account for the interaction between groups.

$$V_{i0a;j0b} = \sum_{s=1}^{N_s} \sum_{t=1}^{N_t} \frac{q_s q_t}{r_{st}} \tag{8}$$

where q_s and q_t are the monopole charges on chromophores, and N_s and N_t are the number of charges fitted on the chromophore. The inter-chromophore interaction depends only on the magnitudes and locations of these point charges and their distance and orientation with respect to each other will reflect the secondary and tertiary structure of the protein.

The incorporation of vibrational structure is within the harmonic approximation. The corresponding illustrative Hamiltonian matrix for a dipeptide in equation 9 considers the nπ* transition and two vibrational states, *v*1 and *v*2, in the $\pi_{nb}\pi^*$ transition. In our protein calculations, we consider five vibrational states. The energy spacing between these vibrational states is derived from the vibrational frequencies of the relevant mode obtained from the *ab initio* calculation.

$$\hat{H} = \begin{pmatrix} E_{1n\pi^*} & V_{1n\pi^*;1\pi\pi^*\_v1} & V_{1n\pi^*;1\pi\pi^*\_v2} & V_{1n\pi^*;2n\pi^*} & V_{1n\pi^*;2\pi\pi^*\_v1} & V_{1n\pi^*;2\pi\pi^*\_v2} \\ V_{1n\pi^*;1\pi\pi^*\_v1} & E_{1\pi\pi^*\_v1} & 0 & V_{1\pi\pi^*\_v1;2n\pi^*} & V_{1\pi\pi^*\_v1;2\pi\pi^*\_v1} & V_{1\pi\pi^*\_v1;2\pi\pi^*\_v2} \\ V_{1n\pi^*;1\pi\pi^*\_v2} & 0 & E_{1\pi\pi^*\_v2} & V_{1\pi\pi^*\_v1;2n\pi^*} & V_{1\pi\pi^*\_v2;2\pi\pi^*\_v1} & V_{1\pi\pi^*\_v2;2\pi\pi^*\_v2} \\ V_{1n\pi^*;2n\pi^*} & V_{1\pi\pi^*\_v1;2n\pi^*} & V_{1\pi\pi^*\_v1;2n\pi^*} & E_{2n\pi^*} & V_{2n\pi^*;2\pi\pi^*\_v1} & V_{2n\pi^*;2\pi\pi^*\_v2} \\ V_{1n\pi^*;2\pi\pi^*\_v1} & V_{1\pi\pi^*\_v1;2\pi\pi^*\_v1} & V_{1\pi\pi^*\_v2;2\pi\pi^*\_v1} & V_{2n\pi^*;2\pi\pi^*\_v1} & E_{2\pi\pi^*\_v1} & 0 \\ V_{1n\pi^*;2\pi\pi^*\_v2} & V_{1\pi\pi^*\_v1;2\pi\pi^*\_v2} & V_{1\pi\pi^*\_v2;2\pi\pi^*\_v2} & V_{2n\pi^*;2\pi\pi^*\_v2} & 0 & E_{2\pi\pi^*\_v2} \end{pmatrix} \quad (9)$$

where the nπ* transition energy is $E_{n\pi^*}$; the $\pi_{nb}\pi^*$ transition has two associated elements, $E_{\pi\pi^*\_v1}$ and $E_{\pi\pi^*\_v2}$, representing the different vibrational levels in the excited state.

The interactions between transitions are modified by multiplying the monopole charges by the Franck-Condon overlap integral, f_v. It is assumed that there is no interaction between the vibrational levels of the same chromophore. The Franck-Condon overlap integral of the nπ* transition was taken as unity.

$$V_{i0a;j0b} = \sum_{s=1}^{N_s} \sum_{t=1}^{N_t} \frac{q_s q_t f_a f_b}{r_{st}} \quad (10)$$

Diagonalization of the Hamiltonian matrix by a unitary matrix *U* can be used to calculate the energies and the coefficients c_{ia}^k. The eigenvalues are the transition energies of the system because the ground state energy is set to be zero and the eigenvectors form the unitary matrix.

$$U^{-1} \cdot \hat{H} \cdot U = H_{diag} \quad (11)$$

The electric and magnetic transition dipole moments from ground state to k^{th} excited state of the interacting system are calculated from the transition dipole moments of individual chromophore using the unitary matrix:

$$\overrightarrow{\mu_i} = \sum_a U_{ai}\, \vec{\mu}_a^0 \quad (12)$$

$$\overrightarrow{m_i} = \sum_a U_{ai}\, \vec{m}_a^0 \quad (13)$$

With the electric and magnetic transition dipole moments and wave functions of the excited state of the protein, the CD spectrum (rotational strengths in the interacting system) is readily calculated using Rosenfeld equation. The rotational strength (R^{0k}) of an excitation from the ground state *0* to the excited state *k* is the product of the electric transition dipole moment, $\vec{\mu}$, and the magnetic transition dipole moment, $\vec{m}$, given by Rosenfeld equation:

Table 1 Proteins considered in this study.

class	protein (PDB code)
α-helix	lysozyme (193l), myoglobin whale (1a6m), leptin (1ax8), c-phycocyanin (1ha7)*[a], hemoglobin (1hda), cytochrome C (1hrc), calmodulin (1lin), insulin (1trz), phospholipase A2 (1une), myoglobin horse (1ymb)
mixed-α*β*	aldolase (1ado)*, haloalkane dehydrogenase (1bn6), carbonic anhydrase I (1hcb), PNMT (1hnn)*, NmrA (1k6j)*, monellin (1mol), papain (1ppn), dehydroquinase type 1 (1qfe), rhodanese (1rhs), ubiquitin (1ubi), dehydroquinase type 2 (2dhq), deoxyribonuclease-1 (3dni), ribonuclease A (3rn3), carboxypeptidase A (5cpa), triosephosphate isomerase (7tim)
β-I	pectate lyase C (1air), β-bactoglobulin (1b8e), γ-D-crystallin (1elp), γ-S-crystallin (C-term) (1ha4), γ-D-crystallin (human) (1hk0), Jacalin (1ku8)*, lentil lectin (1les), γ-E-crystallin (1m8u), concanavalin A (1nls), pea lectin (1ofs), avidin (1rav), thaumatin (1thw), β-B2-crystallin (2bb2), pepsinogen (2psg), γ-B crystallin (4gcr)
β-II	Cu,Zn Superoxide dismutase (1cbj), α-bungarotoxin (1hc9), rubredoxin (1r0i), streptavidin (1stp), α-chymotrypsinogen (2cga), ferredoxin (2fdn), elastase (3est), α-chymotrypsin (5cha), aprotinin (5pti)

[a] For computational tractability, only the unique chains of multi-meric proteins (denoted by asterisk) containing more than 500 residues were considered. The validity of this approximation has been established elsewhere[44].

$$R^{0k} = \mathrm{Im}(\langle\psi^0|\vec{\mu}|\psi^k\rangle\langle\psi^k|\vec{m}|\psi^0\rangle) \quad (14)$$

where ψ^0 and ψ^k represent the wave functions of different states. *Im* represents the imaginary part of the product.

Calculations using the matrix method generate line spectra, in which the lines represent the calculated rotational strength at the specific wavelength. A Gaussian function is applied to the line spectra, broadening bands, to mimic the spectra in solution. This permits a direct comparison between the calculations and the experimental spectra. The calculated CD spectra are sensitive to the bandwidth[45], which is effectively an empirical parameter. A single bandwidth is used for all peaks in all proteins in this study. A bandwidth of 10 nm was used for spectra considering vibronic coupling, while 12.5 nm was used for the calculations without vibrational structure. Our study, thus, in part, investigates one of the contributing factors to line broadening.

CD spectra were calculated for a set of 49 proteins (Table 1) with the above approach. The crystal structures of these proteins are in the Protein Data Bank (PDB). These proteins are taken from the database SP175 of the Protein Circular Dichroism Data Bank[46] (PCDDB) in which the experimental spectra are available. To assess our calculations, the CD spectra of these proteins were calculated and compared with the experimental data in the far-UV region.

3 Results and Discussion

The active space employed in the CASPT2 calculation comprises two occupied and three virtual orbitals. The first transition is the nπ^* transition, followed by the $\pi_{nb}\pi^*$ transition. The lowest energy conformers of ground and $\pi_{nb}\pi^*$ excited states of NMA are shown in Fig. 2 and the optimized geometric parameters are summarized in Table 2. The structure of NMA has been previously studied experimentally[47-50]

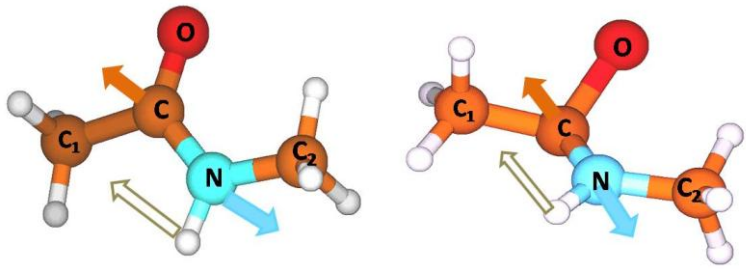


Fig.2 Calculated geometries and C-N stretching mode of NMA in the ground state (left) and the $\pi_{nb}\pi^*$ excited state (right).

Table 2 Ground and excited state geometries of NMA, optimized at the CASPT2 level, along with the frequencies of C-N stretch in both states. Experimental results[51, 52] are listed as a reference.

	Exp.	CASPT2	
	g.s.[51]	g.s.	$\pi_{nb}\pi^*$ e.s.
C-N / Å	1.386	1.368	1.326
C=O / Å	1.224	1.231	1.587
N-H / Å	1.106	1.010	1.021
C-C_1 / Å	1.520	1.517	1.485
N-C_2 / Å	1.469	1.450	1.457
O=C-N / °	121.8	123.2	108.4
C-N-H / °	110.0	118.9	120.2
N-C-C_1 / °	114.1	115.2	117.1
C-N-C_2 / °	119.6	122.2	121.3
H-N-C=O / °		-179.9	138.6
C-N stretch[52] / cm^{-1}	1581	1610	1696

and theoretically[53-55]. Our calculation suggests a planar symmetry for the ground state which is consistent with X-ray crystallography[51] and gas electron diffraction experiments[56]. The $\pi_{nb}\pi^*$ excited state has been less well studied, due to the challenges of characterising the geometries of electronically excited states. The geometry distortion and the displacements involved in the excited states are known mainly from Raman experiments[36, 55]. Chen *et al*.[55] concluded that C-N elongation is the largest displacement of NMA in the aqueous phase, which agrees with Mayne and Hudson's analysis[36] of the resonance Raman spectrum that the C-N stretch is the predominant mode in water. They estimated the vibrational frequency of the C-N stretch to be 1400 cm^{-1}, with which they obtained a good fit to the absorption spectrum of NMA. Li *et al*.[53] calculated, using the configuration interaction between all singly excited configurations (CIS) approach, that the excited state of NMA possesses a pyramidal carbonyl and a planar amine conformation. Our CASPT2 calculation suggests a similar amide geometry. However, our calculation indicated a shorter C-N bond in the excited state structure compared to the ground state geometry, while this bond appeared to be elongated in the Raman experiments. This discrepancy may have arisen from the neglect of solvent effects in our calculation. Nevertheless, the absorption spectrum of NMA in water (Fig. 3) can be reproduced quite well by the Franck-Condon factors generated with these *ab initio* data. We incorporate the vibronic coupling into the CD calculations with the Franck-Condon overlap integrals calculated from the geometries and frequencies of the current results. The calculated vibrational frequencies of the C-N stretch in the ground and

Table 3 Franck-Condon overlap integrals of the ππ* transition in NMA, considering only the C-N stretching mode.

Transition	0←0	1←0	2←0	3←0	4←0
FC integral	0.62	0.60	0.42	0.25	0.13

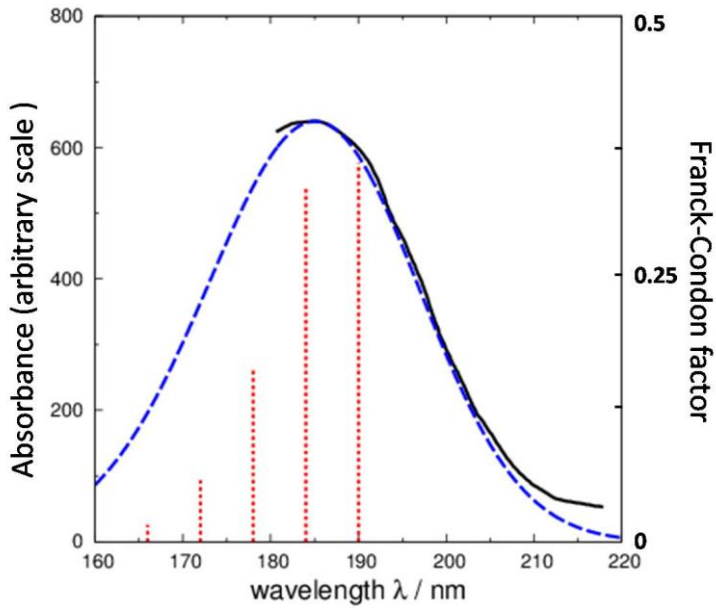


Fig.3 Absorption spectra. The black solid line is experimental spectrum, the dashed line is the calculated spectrum and the dotted vertical lines are proportional to the Franck-Condon factors.

excited states are 1610 cm^{-1} and 1696 cm^{-1}, respectively. Compared to the experimental frequency of the ground state, which is 1581 cm^{-1}, the calculated ground state frequency is within the known systematic errors[57].

The parallel approximation in ezSpectrum was used to calculate the Franck-Condon overlap integral of NMA. Only the C-N displacement was considered in this study and the Franck-Condon overlap integrals of C-N stretch mode are summarized in Table 3. The transitions are denoted as [excited vibronic state ← ground state]. Fig. 3 compares the experimental absorption spectrum[58] with the simulated absorption intensity from our *ab initio* calculation. The origin of the transition (0←0) was placed at 190 nm (52632 cm^{-1}) and the energy gap between different vibrational levels is set to 1696 cm^{-1}, taken from the excited state frequency calculation. The intensities, which are represented by the vertical lines, are proportional to the Franck-Condon factors. A Gaussian function with a bandwidth of 15 nm was applied to simulate other broadening effects.

In order to assess the CD calculations, we used the Spearman rank correlation coefficient and mean absolute error (MAE). The Spearman rank correlation coefficient measures how the calculated intensities correlated with the experimental ranks at a specific wavelength (1 corresponds to a perfect correlation, zero to no correlation, and -1 to a perfect inverse correlation). The CD calculations are

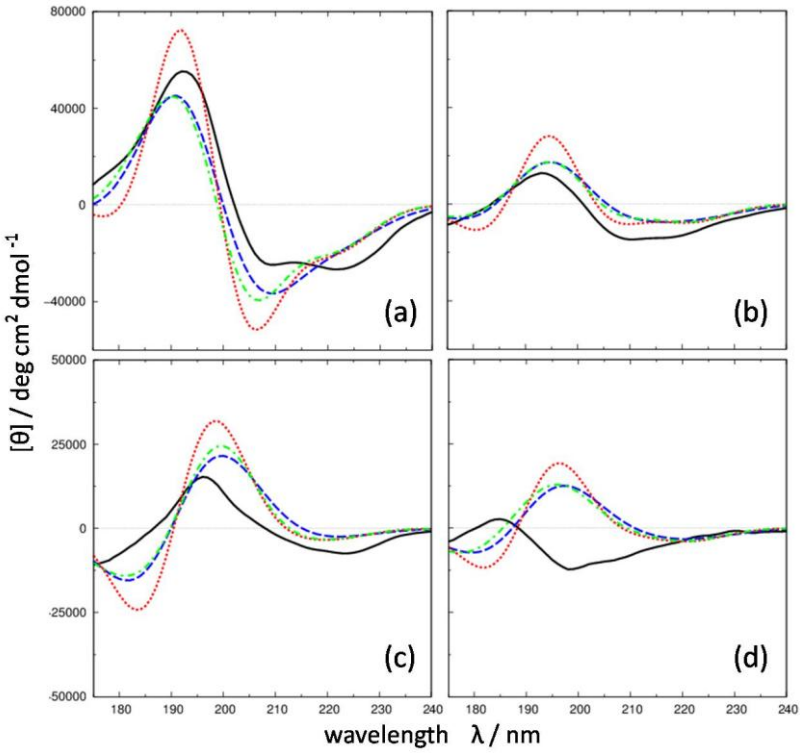


Fig.4 Experimental and calculated CD spectra for (a) myoglobin, α-helical protein, (b) deoxyribonuclease-1, mixed-α,*β* protein, (c) concanavalin A, *β*-I protein, (d) elastase, *β*-II protein: solid line, experiment; dashed line, non-vibronic calculation with a bandwidth, Δ = 12.5 nm; dotted line, non-vibronic calculation with Δ = 10 nm; dash-dotted line, vibronic calculation with Δ = 10 nm.

discussed below based on various secondary structure types. The proteins are grouped (Table 1) based on assignments by DSSP[59]. Proteins containing more than 40% α-helical or *β* structures are classified as the α-helix and *β*-sheet groups, respectively. The *β* type is further divided into *β*-I and *β*-II subtype based on the nature of the *β* structure. The *β*-I type proteins have regular sheets while the *β*-II type contains more *β*-bulges and irregular strands. The rest of the proteins in the dataset are categorized as mixed-α*β*.

Fig. 4(a) compares the experimental and calculated spectra of myoglobin (PDB code: 1YMB), which with 73% helix is a representative of the α-helix subset. The main features of the CD spectra of α-helical proteins are an intense positive band at around 190 nm and two negative bands at 208 nm and 222 nm. The bands at 190 nm and 208 nm arise from the $\pi_{nb}\pi^*$ transition, while the band at 222 nm is due to the $n\pi^*$ transition. For myoglobin, the intensity of peaks at 190 nm and 208 nm are reproduced equally well by the calculations with and without the vibronic structure considered with bandwidths of 10 nm and 12.5 nm, respectively. The double negative bands are better reproduced when the vibronic structure is included. CD spectra calculated with a 10 nm bandwidth for the non-vibronic-parameter are also

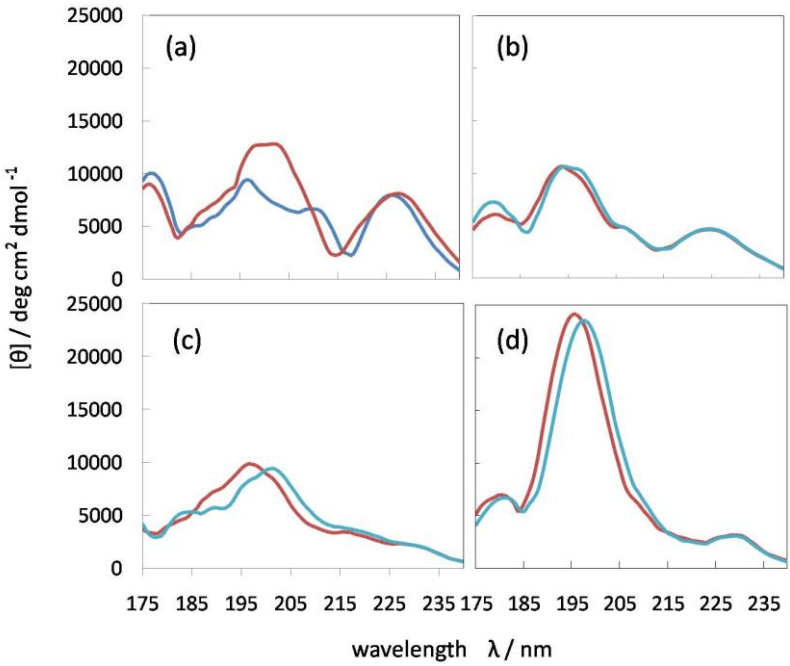


Fig.5 Mean absolute errors of calculations without (blue line) and with vibronic structure (red line). (a) α-helix; (b) mixed-α*β*; (c) *β*-I; (d) *β*-II.

shown in Fig. 4 to examine the influence of the bandwidth in the calculations. The 208 and 222 nm peaks are distinct with a narrower bandwidth without considering vibrational structure. However, both peaks arising from the $\pi_{nb}\pi^*$ transition are overestimated. This is because a narrower bandwidth is balanced by the vibronic coupling, which broadens the peak. Calculating a spectrum simply using a narrower bandwidth without incorporating any other broadening effect in the absorption process is, thus, less satisfactory. However, the MAE between 210 nm and 240 nm is a little higher for the vibronic calculation and the correlation is a little lower between the calculated and experimental intensity (Fig. 5(a)) around 215 nm and 235 nm.

Although the spectra are reproduced well for the intense positive peak, there is poor correlation around 200 nm, which is the descending edge of the most intense peak in the spectrum of α-helical proteins. This is the point that the intensity changes from positive to negative with a large gradient, which means a small error in the wavelength calculation will lead to a large difference in intensity. In the α-helical proteins subset, six out of ten calculated spectra are blue-shifted for both non-vibronic and vibronic calculations. This zero-crossing issue may also explain the difference in the MAEs (Fig. 5(a)). Compared to the non-vibronic calculations, those with vibronic structure are slightly shifted to higher energy due to the position of the origin transition. This deviation causes a larger difference when the CD intensity changes sign.

The features of the mixed-α*β* type (Fig. 4(b)) are quite similar to the spectra of

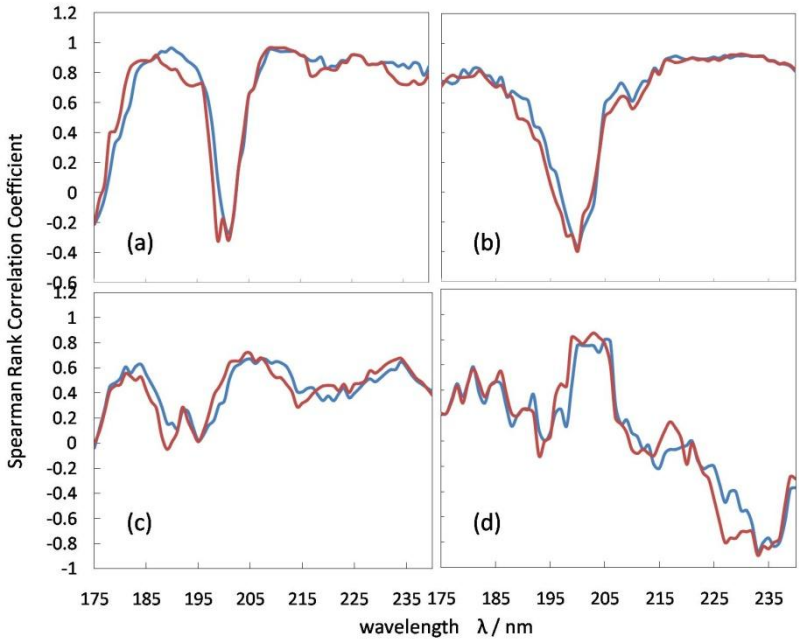


Fig.6 The Spearman rank correlation coefficient between the experimental and calculated intensities for the non-vibronic (blue line) and vibronic parameters (red line) of different structural types of proteins: (a) α-helix; (b) mix-α,*β*; (c) *β*-I; (d) *β*-II.

the α-helical proteins in most cases. However, the positive peak positions and the intensities, the zero crossing points and the ratio of the double negative bands at 208 and 220 nm all fluctuate, due to the variation in secondary structure content. The Spearman rank correlation coefficient for the mixed-α*β* proteins is similar to that of the α-helical group between 175 and 240 nm. The MAE also seems comparable. However, for 10 out of 15 mixed-α*β* proteins the peaks at 190 nm are overestimated, while these peaks in the α-helical proteins are either reproduced well or underestimated. The differences between calculated and experimental spectra for mixed-α*β* proteins with greater *β* content are much larger than the other proteins in this subset. There are also double negative bands in this subtype between 205 nm and 225 nm. The vibronic calculation tends to capture this feature in the mixed-α*β* group, although the ratio of these negative peaks is not well-reproduced, as is the case with the α-helical subset.

Fig. 4(c) and 4(d) compare the spectra for *β*-I and *β*-II type proteins, respectively. The general features of *β*-I spectra are a negative band at about 216 nm and a positive band with comparable magnitude near 195 nm. When there are irregular *β* structures, the main features are a minimum at about 200 nm and a positive band in the range of 185 to 190 nm. However, the band positions vary dramatically for different *β*-sheet proteins. For concanavalin A (Fig. 4(c), 1NLS), the intensity of the negative peak was reproduced better with the vibronic structure consider, though both calculations (with and without vibronic structure) show a red shift. The vibronic parameters tend to give more intense bands around 195 nm for *β*-I subtype

Table 4 Spearman rank correlation coefficients between the calculated and experimental intensity at the peaks close to 187 and 201 nm, for β-II proteins calculated by with and without vibronic structure.

parameter	correlation coefficient	
	187±2 nm	201±2 nm
non-vibronic	0.30	0.69
vibronic	0.36	0.81

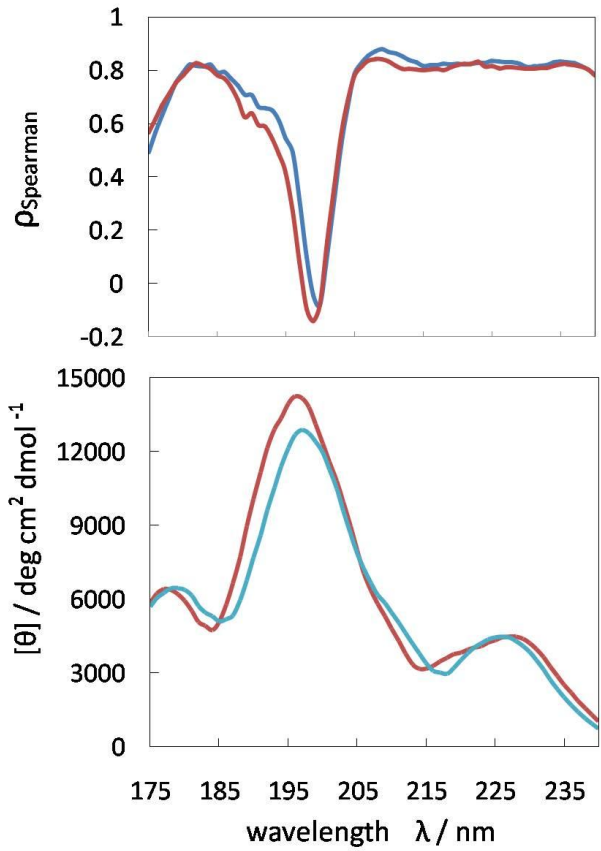


Fig.7 The Spearman rank correlation coefficient (upper) and MAE (lower) of all 49 proteins in the dataset. The blue lines correspond to the non-vibronic-parameters and the red lines to the vibronic-parameters.

proteins. This causes a larger error (Fig. 5(c)) between 185 nm and 200 nm; both calculations overestimate the positive bands at 195 nm in most cases in this protein group. For 10 out of 15 β-I proteins, the calculated intensity of the negative peak at 216 nm is too low.

The β-II type proteins have less regular structure and possibly greater conformational dynamics[60] than β-I proteins and their CD spectra are usually poorly reproduced. The MAE is up to 15,000 deg cm^2 dmol^{-1} greater in the region 185 to 210 nm, compared to other secondary structure types (Fig. 5). There is an improvement in the Spearman rank correlation (Fig. 6(d)) at most wavelengths for

the β-II proteins using the vibronic parameters; the correlations at the two peak positions, 187 and 201 nm, are summarized in Table 4. The coefficients are averaged over a range of ±2 nm to cover the appropriate peak locations. Both peaks listed in Table 4 arise from the $\pi_{nb}\pi^*$ transitions. Calculations considering vibronic coupling of the $\pi_{nb}\pi^*$ transition have a positive effect for β-II proteins. Although the correlation coefficient is over 0.75 between 200 and 205 nm, six out of nine proteins in this type, have calculated spectra with a band of the opposite sign to experiment in this region. Both parameter sets overestimate the positive peaks and underestimate the negative ones.

A comparison of the Spearman rank correlation coefficient and MAEs of the entire protein data set for the two sets of calculations (Fig. 7) reveals no obvious improvement for most regions in the calculated CD spectra with vibronic parameters. However, there is a smaller error around 208 nm for α-helix, β-I and β-II groups. Including the vibronic structure of the $\pi_{nb}\pi^*$ transition may enhance the coupling between the $\pi_{nb}\pi^*$ transition and the $n\pi^*$ transition. The MAE difference between 185 and 200 nm is mainly from the wavelength where the origin was set. The large error between 190 and 210 nm comes predominantly from the β-II group.

4 Conclusions

To incorporate the vibronic coupling in the matrix method, Franck-Condon overlap integrals have been calculated using the ground and excited state geometries and frequencies of NMA. CASPT2 calculations gave a planar geometry of the ground state; in the $\pi_{nb}\pi^*$ excited state the carbonyl is pyramidal and the amine is planar. The conformations are consistent with the previous experimental and theoretical results. Five vibronic transitions in the $\pi_{nb}\pi^*$ excited state were taken into account in the matrix method calculation.

A set of 49 proteins, whose experimental spectra is available, allows us to assess the influence of incorporation of vibrational structure in the CD calculations. The negative peaks between 205 and 225 nm are reproduced with better resolution in α-helical proteins using vibronic parameters. No obvious improvement is observed for mixed-$\alpha\beta$ and β-I proteins. For β-II proteins, whose structures are less regular and whose spectra are difficult to reproduce, the Spearman correlation coefficients between the experimental and the calculated intensities show improvement over most wavelengths in the far UV region when taking the vibrational structure into account. It is possible that the calculation may be further improved by considering the conformational dynamics and flexibility of β-II proteins. Incorporation of the vibrational structure of NMA in the matrix method calculations provides us with a template to consider Franck-Condon effects in the near-UV[23], and will be the subject of future work.

Acknowledgments

We are grateful for access to the University of Nottingham High Performance Computing facility. ZL is partially supported by a University of Nottingham Vice-Chancellor's scholarship.

References

1. N. Berova, L. Di Bari and G. Pescitelli, *Chem. Soc. Rev.*, 2007, **36**, 914.
2. D. M. McCann and P. J. Stephens, *J. Org. Chem.*, 2006, **71**, 6074.
3. G. Pescitelli, S. Gabriel, Y. K. Wang, J. Fleischhauer, R. W. Woody and N. Berova, *J. Am. Chem. Soc.*, 2003, **125**, 7613.
4. K. Matsuo, H. Hiramatsu, K. Gekko, H. Namatame, M. Taniguchi and R. W. Woody, *J. Phys. Chem. B*, 2014, **118**, 2785.
5. A. M. O'Mahony, M. F. Cronin, A. McMahon, J. C. Evans, K. Daly, R. Darcy and C. M. O'Driscoll, *J. Pharm. Sci.*, 2014, **103**, 1346.
6. J. L. S. Lopes, D. Orcia, A. P. U. Araujo, R. DeMarco and B. A. Wallace, *Biophys. J.*, 2013, **104**, 2512.
7. M. Vanhove, A. Lejeune and R. H. Pain, *Cell Mol. Life Sci.*, 1998, **54**, 372.
8. A. J. Miles, N. U. Fedosova, S. V. Hoffmann, B. A. Wallace and M. Esmann, *Biochem. Bioph. Res. Co*, 2013, **435**, 300.
9. T. Mori, Y. Inoue and S. Grimme, *J. Phys. Chem. A*, 2007, **111**, 7995.
10. K. L. Bak, A. E. Hansen, K. Ruud, T. Helgaker, J. Olsen and P. Jorgensen, *Theoretica Chimica Acta*, 1995, **90**, 441.
11. T. B. Pedersen, H. Koch and K. Ruud, *J. Chem. Phys.*, 1999, **110**, 2883.
12. J. D. Hirst and B. J. Persson, *J. Phys. Chem. A*, 1998, **102**, 7519.
13. N. A. Besley, M. J. Brienne and J. D. Hirst, *J. Phys. Chem. B*, 2000, **104**, 12371.
14. M.-P. Gaigeot, N. A. Besley and J. D. Hirst, *J. Phys. Chem. B*, 2011, **115**, 5526.
15. H. Devoe, *J. Chem. Phys.*, 1964, **41**, 393.
16. H. Devoe, *J. Chem. Phys.*, 1965, **43**, 3199.
17. J. Applequist, K. R. Sundberg, M. L. Olson and L. C. Weiss, *J. Chem. Phys.*, 1979, **70**, 1240.
18. J. Applequist, *J. Chem. Phys.*, 1979, **71**, 4332.
19. K. A. Bode and J. Applequist, *J. Phys. Chem.*, 1996, **100**, 17825.
20. P. M. Bayley, E. B. Nielsen and J. A. Schellma, *J. Phys. Chem.*, 1969, **73**, 228.
21. R. W. Woody and I. Tinoco, *J. Chem. Phys.*, 1967, **46**, 4927.
22. R. W. Woody, *J. Chem. Phys.*, 1968, **49**, 4797.
23. R. W. Woody and N. Sreerama, *J. Chem. Phys.*, 1999, **111**, 2844.
24. A. S. Davydov, *Theory of molecular excitations*, Plenum Press, New York, 1971.
25. N. A. Besley and J. D. Hirst, *J. Am. Chem. Soc.*, 1999, **121**, 9636.
26. D. M. Rogers and J. D. Hirst, *Biochemistry*, 2004, **43**, 11092.
27. J. D. Hirst, D. M. Hirst and C. L. Brooks, *J. Phys. Chem. A*, 1997, **101**, 4821.
28. N. A. Besley, M. T. Oakley, A. J. Cowan and J. D. Hirst, *J. Am. Chem. Soc.*, 2004, **126**, 13502.
29. N. A. Besley and J. D. Hirst, *J. Mol. Struct. (THEOCHEM)*, 2000, **506**, 161.
30. Z. J. Dang and J. D. Hirst, *Angew. Chem. Int. Ed.*, 2001, **40**, 3619.
31. *J. Neugebauer, E. J. Baerends, M. Nooijen and J. Autschbach, J. Chem. Phys.*, 2005, **122**, 234305.
32. T. Seidler, M. Andrzejak and M. T. Pawlikowski, *Chem. Phys. Lett.*, 2010, **496**, 74.
33. B. M. Bulheller, G. D. Pantos, J. K. M. Sanders and J. D. Hirst, *Phys. Chem. Chem. Phys.*, 2009, **11**, 6060.
34. T. Sanematu and Y. Mizuno, *J. Phys. Soc. Jpn.*, 1976, **40**, 1733.
35. T. Sanematu, *J. Phys. Soc. Jpn.*, 1977, **43**, 600.
36. L. C. Mayne and B. Hudson, *J. Phys. Chem.*, 1991, **95**, 2962.

37. M. Volkhard and K. Oliver, *Charge and Energy Transfer Dynamics in Molecular Systems*, Wiley-VCH, Weinheim, 2011.
38. M. Yang, *J. Mol. Spectrosc.*, 2006, **239**, 108.
39. M. S. Weiss, H. J. Metzner and R. Hilgenfeld, *FEBS Lett.*, 1998, **423**, 291.
40. B. O. Roos and K. Andersson, *Chem. Phys. Lett., 1995, **245**, 215.*
41. H.-J. Werner, P. J. Knowles, G. Knizia, F. R. Manby and M. Schütz, *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 2012, **2**, 242.
42. N. G. Mirkin and S. Krimm, *J. Am. Chem. Soc.*, 1991, **113**, 9742.
43. V. A. Mozhayskiy and A. I. Krylov, *ezSpectrum*, http://iopenshell.usc.edu/downloads.
44. B. M. Bulheller, A. J. Miles, B. A. Wallace and J. D. Hirst, *J. Phys. Chem. B*, 2008, **112**, 1866.
45. J. D. Hirst, K. Colella and A. T. B. Gilbert, *J. Phys. Chem. B*, 2003, **107**, 11813.
46. L. Whitmore, B. Woollett, A. J. Miles, D. P. Klose, R. W. Janes and B. A. Wallace, *Nucleic. Acids. Res.*, 2011, **39**, D480.
47. M. B. Robin, ed., *Higher Excited States of Polyatomic Molecules*, New York, 1985.
48. A. T. Hagler, L. Leiserowitz and M. Tuval, *J. Am. Chem. Soc.*, 1976, **98**, 4600.
49. K. Itoh and T. Shimanou, *Biopolymers*, 1967, **5**, 921.
50. F. Fillaux and C. Deloze, *Chem. Phys. Lett.*, 1976, **39**, 547.
51. J. L. Katz and B. Post, *Acta Crystallogr.*, 1960, **13**, 624.
52. L. C. Mayne, L. D. Ziegler and B. Hudson, *J. Phys. Chem.*, 1985, **89**, 3395.
53. Y. Li, R. L. Garrell and K. N. Houk, *J. Am. Chem. Soc., 1991, **113**, 5895.*
54. X. G. Chen, R. Schweitzerstenner, S. Krimm, N. G. Mirkin and S. A. Asher, *J. Am. Chem. Soc.*, 1994, **116**, 11141.
55. X. G. Chen, S. A. Asher, R. Schweitzerstenner, N. G. Mirkin and S. Krimm, *J. Am. Chem. Soc.*, 1995, **117**, 2884.
56. M. Kitano, T. Fukuyama and K. Kuchitsu, *B. Chem. Soc. Jpn.*, 1973, **46**, 384.
57. J. A. Pople, A. P. Scott, M. W. Wong and L. Radom, *Israel J. Chem.*, 1993, **33**, 345.
58. K. Kaya and S. Nagakura, *Theor. Chim. Acta.*, 1967, **7**, 124.
59. W. Kabsch and C. Sander, *Biopolymers*, 1983, **22**, 2577.
60. J. D. Hirst, S. Bhattacharjee and A. V. Onufriev, *Faraday. Discuss.*, 2003, **122**, 253.