Supplementary Information for:

Tidying up the conformational ensemble of a disordered peptide by computational prediction of spectroscopic fingerprints

M. Michaelis,[†,‡] L. Cupellini,[*,¶] C. Mensch,[§] C. C. Perry,[‡] M. Delle Piane,[*,†,‖] and L. Colombi Ciacchi[†]

†*Hybrid Materials Interfaces Group, University of Bremen, Faculty of Production Engineering, Bremen Center for Computational Materials Science, Center for Environmental Research and Sustainable Technology (UFT), and MAPEX Center for Materials and Processes, Am Fallturm 1, 28359 Bremen, Germany*

‡*Biomolecular and Materials Interface Research Group, Interdisciplinary Biomedical Research Centre, School of Science and Technology, Nottingham Trent University, Clifton Lane, Nottingham NG11 8NS, United Kingdom*

¶*Dipartimento di Chimica e Chimica Industriale, University of Pisa, Via G. Moruzzi 13, I-56124 Pisa, Italy*

§*Molecular Spectroscopy Research Group, Department of Chemistry, University of Antwerp, Groenenborgerlaan 171, 2020 Antwerp, Belgium*

‖*Department of Applied Science and Technology, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy*

E-mail: lorenzo.cupellini@unipi.it; massimo.dellepiane@polito.it

Experimental and computational methods

Circular dichroism (CD) experiments

CD experiments were performed with the HSSHHQPKGTNP oligopeptide synthezised by JPT (purity $\geq$ 95 % Berlin, Germany). The as-purchased oligopeptide was dissolved in ddH_2O (18 MΩ cm^{-1}, MilliQ, Synergy, Millipore, Germany) at a stock concentration of 1 mg ml^{-1}, as quantified by absorbance at 205 and 214 nm.[1] Measured pH of the peptide solutions was 6.4$\pm$1.0. CD spectra were recorded at the Disco beamline of Synchrotron Soleil. At least three repeat scans for each sample were measured at 25°C, over the wavelength range of 175 to 250 nm using intervals of 1 nm, in Suprasil quartz cells (Hellma UK Ltd.) with a path length of 0.1 mm. The scans were averaged, the respective baseline subtracted and the resulting net spectra smoothed with a Savitsky-Golay Filter using smoothing windows of 5 to 10 data points.

The mean residue ellipticity (Θ_{MRE}) was defined as

$$\Theta_{MRE} = \frac{\Theta}{c \cdot l \cdot n} \, , \tag{S1}$$

where Θ is the raw CD ellipticity (mdeg), n is the number of amino acids in the solvated peptide (12, in our case), l is the path length of the used quartz cuvette (0.1 mm) and c the molar concentration of the peptides. To estimate the relative amount of specific secondary conformational elements in the samples, the CD spectra were analysed using the BeStSel webserver.

Vibrational spectroscopy experiments

Prior to the Raman and ROA measurement, the peptide was dissolved in H_2O (18 MΩ cm^{-1}, MilliQ, Synergy, Millipore, Germany) at a concentration of 50 mg mL^{-1}. The Raman and ROA spectrum were measured at ambient conditions using a previously described ChiralRA-

MAN scattered circular polarization (SCP) ROA instrument (BioTools Inc).[2] The Raman spectrum is displayed as the circular intensity sum (I_R + I_L) and the ROA spectrum as the circular intensity differences (I_R - I_L), with I_R and I_L denoting the scattered Raman intensities with right- and left-circular polarization, respectively. The instrument excitation wavelength was 532 nm; the laser power at the source was in the range of 200 to 800 mW; the spectral resolution 7 cm^{-1}; and the acquisition time 12 to 35 hours. The solvent spectrum was first subtracted from the Raman spectrum, and then the baseline correction procedure by Boelens et al.[3] was applied. The ROA spectrum was smoothed using a 3rd order 5-point Savitzky-Golay filter.

For IR acquisition, the peptide was first dissolved in high purity D_2O (Sigma), lyophilized after overnight incubation, and re-dissolved in D_2O. This process was repeated twice to ensure full H/D exchange of the exchangeable protons. In the peptide "HSSHHQPKGTNP", the labile protons that can readily exchange with deuterium when dissolved in D_2O are found in the O-H and N-H groups of the side chains of histidine, serine, threonine, glutamine, asparagine, and in the terminal amino and carboxyl groups (can deprotonate to carboxylate in solution), as well as in all the peptide bond amide group N-H's. A peptide concentration of 25 mg mL^{-1} in D_2O was placed in a 50 μm path-length cell and IR spectra were acquired for 8×60 minutes on a Bruker PMA 37 VCD attachment coupled to a Bruker IFS 66v/s spectrometer.

Enhanced sampling simulations

All MD simulations were performed with Gromacs 2018,[4] employing the Charmm36m force field[5] for the peptide in combination with the Charmm-modified TIP3P force field[6] for water. Bonds involving hydrogen atoms were constrained by means of the LINCS algorithm.[7] Electrostatic interactions were treated with Particle Mesh Ewald with a cutoff of 1.2 nm. The cubic simulation box (periodic in all directions) had an equilibrated cell vector $a = 6.28$ nm and included 8,253 H_2O molecules. The peptide was protonated using standard GROMACS

tools and the charge of the system was neutralized by adding Cl- ions randomly distributed in the solvent box. The total number of particles (atoms) in the system was ca. 25,000 (peptide + solvent + ions). Prior to the production runs, the system was equilibrated in a first step with an energy minimisation of the solvent molecules keeping the peptide fixed, followed by a NVT run of 100 ps at 300 K for the solvent molecules (keeping the peptide fixed) and a NPT run of the solvent for 100 ps at 1 bar. In a second step we performed an energy minimisation of the peptide (with a fixed solvent), followed by a NVT run of all systems components at 300 K for 100 ps and a NPT at 300 K and 1 bar for 100 ps. A Verlet integration time step of 2 fs was used. The trajectory coordinates were saved every 2 ps, but the total number of configurations obtained during the production stage was reduced to 5,000 for each replica in REST.

After equilibration, the conformational landscape of the peptide was sampled via Replica Exchange with Solute Tempering (REST) simulations, in an NVT ensemble with a modified Berendsen thermostat with a coupling constant of 0.1 ps.[8–10] Simulation time was 300 ns for each replica in REST (total simulation time: 2.4 μs). The peptide was defined as the 'solute' in the simulations, whose temperature was scaled in the different system replicas ('hot' system region). The water molecules remained at the base temperature $T_0 = 300$ K ('cold' system region). We used 8 replicas at temperatures T_i corresponding to 300.0, 322.7, 347.1, 373.4, 401.7, 432.1, 464.8 and 500 K, respectively, following a geometric distribution. Defining $\beta_i = 1/(k_\mathrm{B}T_i)$, where k_B is the Boltzmann constant, the Lennard-Jones parameters ϵ of the hot atoms in the i-th replica were scaled by the factor β_i/β_0, and their charges q by the factor $\sqrt{\beta_i/\beta_0}$. Of the bonded interactions, only the dihedral force constants were scaled. Exchanges between the replicas were attempted every 1.6 ps, following a Metropolis-Hastings acceptance criterion. The geometric progression of the temperatures T_i ensured a nearly uniform overlap of the potential energy distributions and thus a uniform acceptance probability across the replica ladder,[11] with average value of 32.4 ± 2.7 %. The round-trip time, which is defined as the time needed by one replica to move along the complete

temperature ladder from 300 to 500 K and back, was 1.03 ± 0.07 ns.

Structural and cluster analysis

Since only the ground temperature replica in the REST2 simulation represents the canonical distributions of interest, we utilized this part of the trajectory data for analysis (referred to as "REST trajectory" in this work). Visualization and analysis of the trajectories were performed with VMD. Cluster analysis of the conformational microstates sampled by the ground temperature REST trajectory in selected regions of the free energy landscapes was performed using the GROMOS algorithm, as implemented in Gromacs, according to the differences in the root-mean square displacement (RMSD) values of individual conformers, using an RMSD cutoff of 2 Å applied to the atoms of the backbone. The conformers were aligned by a rigid rotation and translation to minimize the RMSD variation prior to clustering. The secondary structure of the conformers was analyzed via the STRIDE algorithm. The Φ and Ψ dihedrals of the peptides backbones along the trajectories were computed via the *rama* program in Gromacs to create Ramachandran plots.

As a better visualization of the ensemble, the sketch-map algorithm, as implemented in PLUMED, was used for dimensionality reduction. Sketch-map disregards the information that corresponds to thermal fluctuations around a (meta)stable structure, by doing a multi-dimensional scaling in which projections for a set of N high-dimensionality landmark points are found by minimizing the stress function

$$\chi^2 = \sum_{i \neq j} \left[F\left(R_{ij}\right) - f\left(r_{ij}\right) \right]^2 \ , \tag{S2}$$

where $F(R_{ij})$ and $f(r_{ij})$ are two sigmoid functions.[? ?] The form of these functions ensures that points lying close together in the high-dimensional space, which most likely belong to the same conformational basin, are then projected close together. Points that are far apart, and are thus likely to be in basins that are not connected by a single transition state, are

projected far apart. For our analysis, the high-dimensionality data set consisted of the 20 backbone dihedrals of the peptide, excluding the terminal residues. The σ of both sigmoids was set at 5.5 and the a and b parameters were chosen as 12 and 7, and 1 and 2, for the high (20D) and low (2D) dimensionality functions, respectively. The resulting maps were colored with respect to: the GROMOS clustering, the radius of gyration, and the left-handed polyproline II helix (PPII) secondary-structure content. The latter was computed using the ALPHABETA collective variable implemented in PLUMED, that measures a distance between the instantaneous values of a set of torsional angles and set of reference values. These reference values were -75° and +150° for the PPII Φ and Ψ dihedrals, respectively.

For comparison, we also opted to employ a widely used dimensionality reduction method, performing the Principal Component Analysis (PCA) algorithm in the implementation from scikit–learn[12], and to project the backbone dihedrals on the principal components (PCs). However in all studied cases, the first two PCs accounted for only ca. 20% of the variance. We also colored the PCA representation by unsupervised clustering using HDBSCAN*,[13] a density-based clustering algorithms that identify clusters based on a search of high density peaks surrounded by regions where there is a lower density of points.

Calculation of theoretical CD spectra

The CD spectrum of each selected configuration was calculated employing an excitonic model in the matrix method formulation.[14–16] This approach is based on the construction of an exciton Hamiltonian, whose parameters are derived by multiscale QM/MM calculations (see below for the details). Ab initio exciton models have proven successful in describing the optical properties, in particular the CD spectra, of several classes of multi-chromophoric systems.[16–18] In such models, the exciton Hamiltonian $\hat{\mathcal{H}}_{ex}$ of N interacting chromophores is constructed on the n excitation energies $\mathcal{E}_i^a$ of each non-interacting chromophore (site energies), and on the electronic coupling V_{ij}^{ab} between two transitions of different chromophores

as

$$\hat{\mathcal{H}}_{ex} = \sum_{i}^{N}\sum_{a}^{n} \mathcal{E}_i^a |ia\rangle\langle ia| + \sum_{ij}^{N}\sum_{ab}^{n} V_{ij}^{ab} |ia\rangle\langle jb| \quad . \tag{S3}$$

In this equation, $|ia\rangle$ represents the electronic state in which chromophore i is in its a^{th} excited state, whereas all the other chromophores are in their ground state. Diagonalization of the Hamiltonian matrix yields as eigenvalues the energy levels E_K of the exciton states, and as eigenvectors the expansion coefficients C_{ia}^K of the exciton wave function on the excited states of the chromophores.

Calculation of the exciton Hamiltonian As the CD spectrum in the far-UV region arises from amide transitions,[14] we considered the amide bonds as chromophoric units for building the exciton model. Each chromophoric unit is modeled as a planar N-methylacetamide (NMA), optimized at the B3LYP/6-31+G(d) level of theory in PCM continuum solvent[?] and aligned to the amide atoms of each unit. This strategy removes the dependence of the site energies on the internal chromophore geometries sampled from MD, and has proven successful for the calculation of CD spectra of nucleic acids and proteins.[19–21]

The elements of the exciton Hamiltonian (S3), $\mathcal{E}_i^a$ and V_{ij}^{ab}, as well as the transition dipole moments $\boldsymbol{\mu}_{ia}$ and $\mathbf{m}_{ia}$, are directly obtained by multiscale quantum mechanics/molecular mechanics (QM/MM) calculations with polarizable embedding.[22] The site energies $\mathcal{E}_i^a$ and all the other site properties are obtained from TD-DFT calculations of each chromophoric unit at the ωB97X-D[23]/6-31+G(d) level of theory, including the first 6 excited states in the calculations. The effect of all the other chromophoric units, as well as of the side chains and the solvent, is accounted for by a polarizable QM/MM embedding (QM/MMPol)[22,24,25]. In the QM/MMPol approach, the MM part is described as a set of fixed point charges and isotropic polarizabilities. The polarization of the environment is described as a set of induced dipoles generated at each MM site by the QM electron density and all the other MM fixed charges and induced dipoles. In this way, the mutual polarization between the QM and MM parts is explicitly considered, and the instantaneous response of the MM part to the

electronic transition is taken into account.

The exciton couplings V_{ab}^{ij} are computed as the environment-mediated Coulomb interaction between the transition densities of the single chromophoric units.[25]

$$V_{ij}^{ab} = \int d\mathbf{r}_1 d\mathbf{r}_2 \frac{\rho_{0a}^{i}(\boldsymbol{r}_1)\rho_{0b}^{j}(\boldsymbol{r}_2)}{r_{12}} - \sum_{p} \int \mathrm{d}\mathbf{r}_1 \rho_{0a}^{i}(\mathbf{r}_1) \frac{(\mathbf{r}_1 - \mathbf{r}_l)}{|\mathbf{r}_1 - \mathbf{r}_l|^3} \cdot \boldsymbol{\mu}_l^{\mathrm{ind}} \left[\rho_{0b}^{j}\right] \ . \tag{S4}$$

In this equation, the first term represents the bare Coulomb coupling between transition densities, and the second term represents the environment-mediated contribution; ρ_{0a}^{i} (ρ_{0b}^{j}) is the $0 \to a$ ($0 \to b$) transition density of chromophore i (j), and $\boldsymbol{\mu}_l^{\mathrm{ind}}[\rho_{0b}^{j}]$ is the dipole induced on the MM atom l by the transition density ρ_{0b}^{j}.[25] All calculations were performed with a locally modified version of Gaussian 16[26] with an efficient QM/MMPol implementation[27].

Calculation of the transition intensities The rotational strength R_{0K} of a CD electronic transition between a ground state 0 and the excited state K is given by the Rosenfeld equation:[14,15]

$$R_{0K} = \Im\langle 0|\hat{\boldsymbol{\mu}}|K\rangle \cdot \langle K|\hat{\boldsymbol{m}}|0\rangle \quad , \tag{S5}$$

where $\Im$ denotes the imaginary part, and $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{m}}$ are the operators of the electric and magnetic moment vectors respectively. Their matrix elements in the exciton basis are obtained by combining the electric ($\boldsymbol{\mu}_{ia}$) and magnetic ($\mathbf{m}_{ia}$) transition dipoles of the chromophores. A common approximation is to neglect the intrinsic magnetic dipole moments of the chromophore transitions.[16] However, considering the intrinsic magnetic moments is actually fundamental to describe the $n \to \pi^*$ transitions of the peptide bonds, which have a strong magnetic character.[14] Here we employ an origin-invariant expression for the rotational strength, which takes into account the intrinsic magnetic moments, and can be obtained by using the velocity formulation of the electric transition dipoles:[28,29]

$$R_{0K} = -\frac{e\hbar^2}{2\pi m_e \nu_{K0}} \cdot \Im \left[\sum_{i,j} \sum_{a,b} C_{ia}^{K} C_{jb}^{K} \boldsymbol{\nabla}_{ia} \cdot \boldsymbol{m}_{jb} \right] \quad . \tag{S6}$$

In this equation, $\boldsymbol{\nabla}_{ia}$ is the a-th electric dipole moment of chromophore i in the velocity formulation, e and m_e are the electron charge and mass, respectively, and ν_{K0} is the wavenumber of the electronic transition E_K.

Calculation of the CD spectrum The CD spectrum is finally obtained by broadening all transitions with energy E_K and intensity R_{0K} with a Gaussian function (half width at half maximum 2000 cm^{-1}) to phenomenologically account for the line broadening. All excitonic calculations were performed using the EXAT tool.[29] For comparison with experiments, we shifted all computed spectra by 1.0 eV. This shift mainly accounts for the intrinsic error of the DFT functional compared to the experimental $\pi \rightarrow \pi^*$ energy[14].

Choice of the QM method for ECD calculations Table S1 shows the excitation energies of NMA computed at different levels of theory. Although the B3LYP/D95V+(d) combination gives a $\pi \rightarrow \pi^*$ energy slightly closer to the experimental value of 6.53 eV (190 nm)[14], but still overestimated. On the contrary, ωB97X-D/6-31+G(d) has a larger overestimation. However, due to the larger fraction of exact exchange, ωB97X-D is less prone to spurious state intrusion when employed in the disordered MMPol environment. That is, although there is a larger systematic error in the NMA site energies, we expect the ωB97X-D description of the excited states to be more balanced.

For this reason, we chose the ωB97X-D functional, noting that there is a systematic overestimation of the excitation energies by $\sim$1 eV. This is not a problem in our simulations, since all NMA chromophores are chemically identical, and thus the excitation energy shift is completely systematic. This shift can be corrected *a posteriori* when comparing simulated and experimental spectra. A similar approach has been successfully employed for the calculation of CD spectra of DNA G-quadruplexes[29].

Table S1: Excitation energies (eV) of NMA calculated in vacuo for the first 6 excited states with TD-DFT and different functional and basis set combinations. Oscillator strengths are given in parentheses. The $\pi \rightarrow \pi^*$ is the fourth excited state in all cases.

State	ωB97X-D/6-31+G(d)	B3LYP/6-31+G(d)	B3LYP/D95V+(d)
1	5.9430 (0.001)	5.8276 (0.001)	5.7576 (0.001)
2	7.0474 (0.001)	6.4035 (0.001)	6.3305 (0.001)
3	7.2068 (0.017)	6.4969 (0.021)	6.4133 (0.020)
4	7.3829 (0.328)	7.1006 (0.250)	7.0096 (0.245)
5	7.9727 (0.047)	7.1847 (0.022)	7.1293 (0.020)
6	8.2740 (0.000)	7.2843 (0.025)	7.2167 (0.021)

Calculation of theoretical IR, Raman and Raman Optical Activity spectra

The geometry optimization and spectral calculations for the IR, Raman and ROA spectra were all performed using the Gaussian16 (A.03) software.[26]

Geometry optimization The geometries of different conformations of the peptide were partially optimized in normal coordinates following the approach by Bouř and Keiderling.[30] With this method the backbone conformation of the peptide is mostly retained, while the modes of spectroscopic interest are fully relaxed. Fully optimizing the structure using DFT calculations would significantly alter the conformation. Because of the conformational sensitivity of these spectroscopic techniques, the meaning of the calculated spectra would otherwise be lost.[31] The geometry optimizations and frequency calculations were performed with the B3PW91 functional and 6-31G(d) basis set. No explicit water molecules were included in these calculations. Rather, the C-PCM implicit solvent model[?] was used to account for the presence of water molecules. To account for the H/D exchange in the computed IR spectra, the labile protons were replaced by deuterons (the O-H and N-H groups of the side chains of histidine, serine, threonine, glutamine, asparagine, and in the terminal amino, as well as in all the peptide bond amide group N-H's).

Calculation of the spectral properties The spectra were calculated using the two-step procedure (polar=ROA) with a 6-31++G(d) basis set, as diffuse functions are much more important for the calculation of the spectral properties than they are for the optimization and frequency calculation.[32] Realistic line shapes were simulated using one Lorentzian function for each normal mode with a full-width at half maximum of $10\,\mathrm{cm}^{-1}$ for the IR spectra and $20\,\mathrm{cm}^{-1}$ for the Raman and ROA spectra, respectively.[31,33]

Wavenumber scaling of the computed IR, Raman and ROA spectra Since the wavenumbers in the DFT calculated spectra are overestimated due to *inter alia* the har-

monic approximation, a global scaling factor is typically multiplied to the entire wavenumber dimension to allow a direct comparison of the experimental and the simulated spectra.[31,34] Since the amide I (and I') band (mainly amide C=O stretching) is known to be even more overestimated due to, for example, basis set dependence, and interaction with the solvent (see also below) a separate scaling factor needs to be applied to the amide I vibrations compared to the remainder of the spectrum.[31,35]

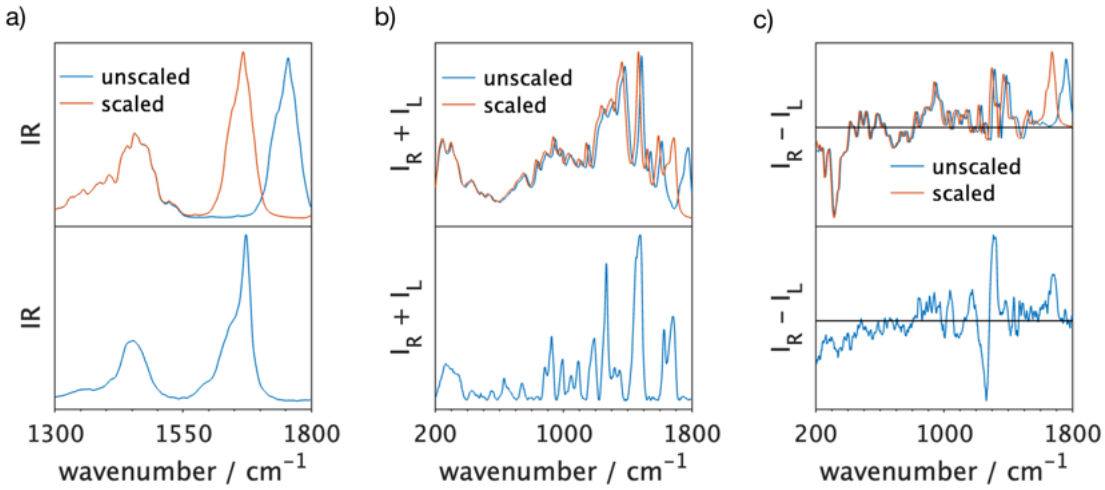

Figure S1: *Alignment of the simulated (top panels; weighted average of the 20 cluster central structures) with the experimental (bottom panel) (a) FT-IR spectrum, (b) Raman and (c) ROA spectrum. A scaling factor of 0.95 is used for the amide I and I' region and a scaling factor of 1.00 is applied for the amide II' region in the IR and 0.987 for the remainder of the Raman and ROA spectrum. The blue line in the top panels displays the unscaled weighted average of the 20 clusters and the orange shows the scaled spectrum. Realistic line shapes were simulated using a FWHM of 10 cm^{-1} for the IR and 20 cm^{-1} for the Raman and ROA.*

As shown in Figure S1 for the FT-IR spectra, a scaling factor of 0.95 for the amide I' and a global scaling factor of 1.0 (so no scaling) needs to be multiplied with the wavenumber dimension (x-axis). The top panel shows the weighted average of the 20 most populated clusters. The spectrum of each of the 20 clusters is shown in Figure S2. Since the Raman and ROA spectra have more complex spectral features than the FT-IR spectrum, the comparison of the simulated spectra with experiment is more challenging. Furthermore, the calculated spectra show some bands that are not found in experiment. Since this complicates the determination of a proper scaling factor, the scaling factor of 0.95 is again applied to

the amide I vibrations and a global scaling factor of 0.987 that is commonly reported in Raman/ROA calculations at this level of theory is applied to the remainder of the spectrum ($<$1675 cm^{-1}).[31] In Figure S1, it is shown that using these scaling factors, the amide I around 1680 cm^{-1} aligns better with experiment, while for the remainder of the spectrum, we must be careful in assigning specific bands in the experiment based on the calculated bands. Each individual spectrum after scaling is plotted in Figure S2.

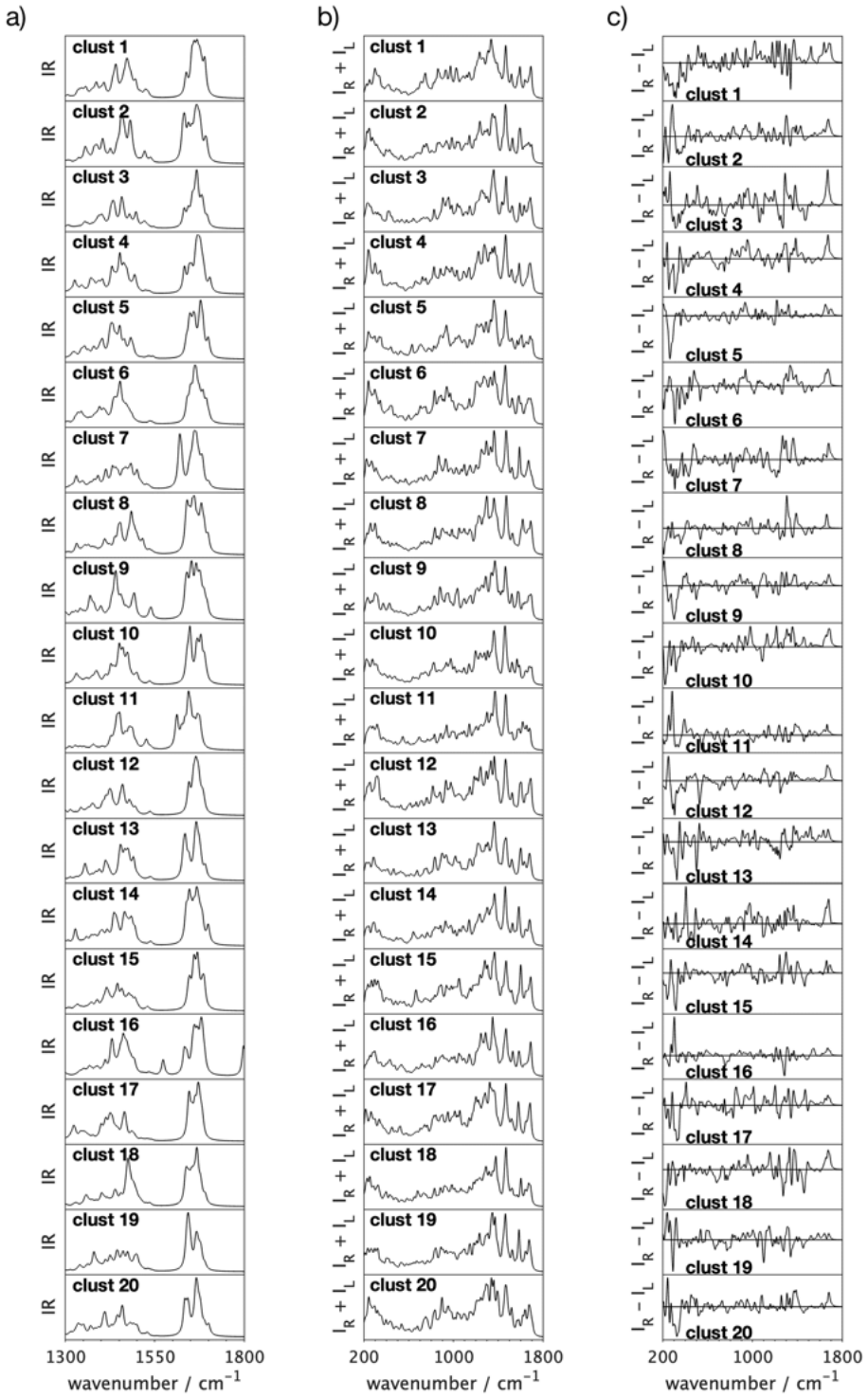

Figure S2: *Simulated (a) IR, (b) Raman and (c) ROA spectra for each of the 20 cluster center structures. A scaling factor of 0.95 is used for the amide I and I' region and a scaling factor of 1.00 is applied for the amide II' region in the IR and 0.987 for the remainder of the Raman and ROA spectrum. Realistic line shapes were simulated using a FWHM of 10 cm^{-1} for the IR and 20 cm^{-1} for the Raman and ROA.*

SUPPLEMENTARY DATA AND FIGURES

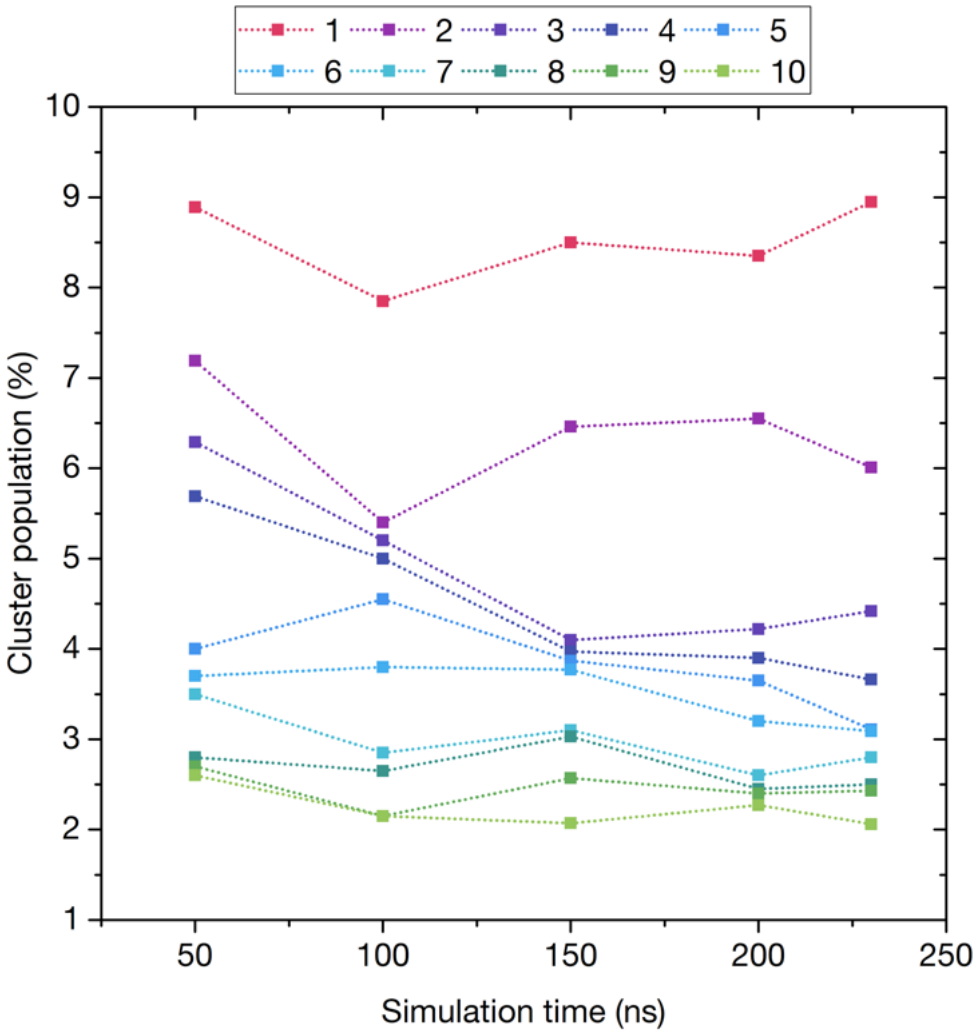

Figure S3: *Population of the 10 most populated clusters over simulation time.*

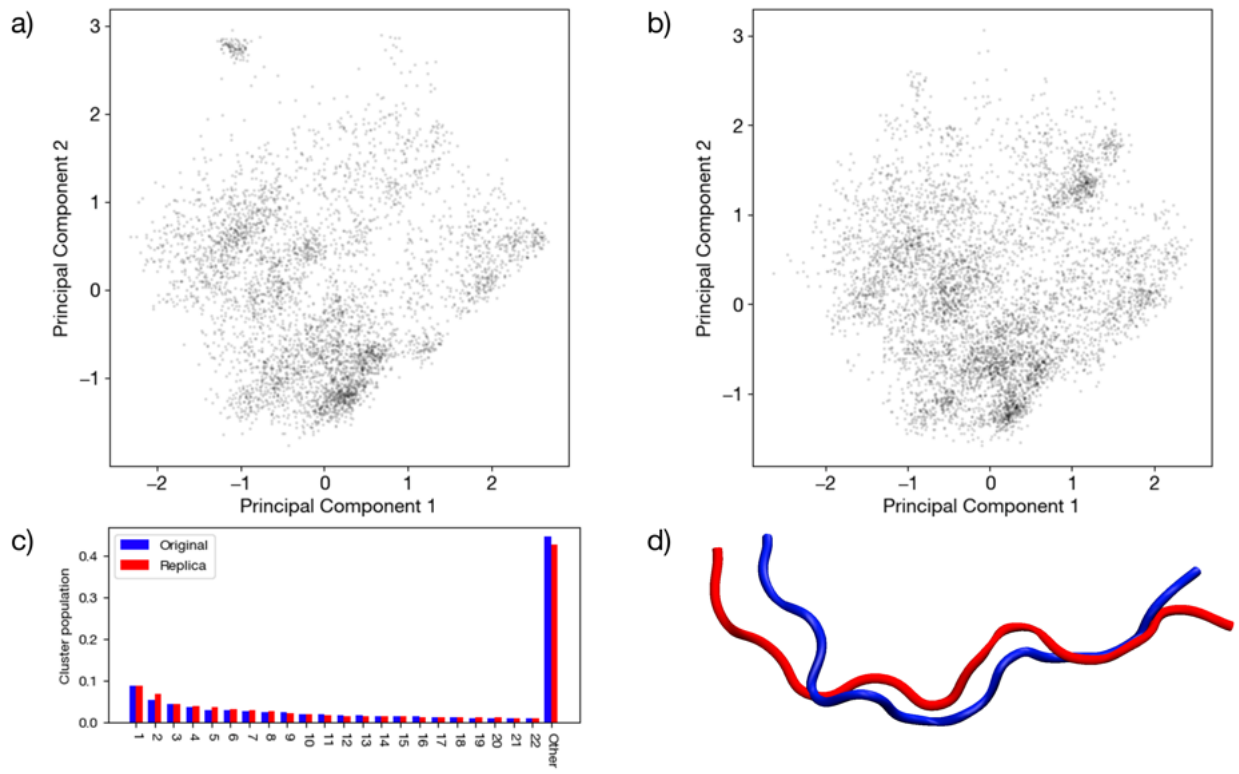

Figure S4: *a) Projection on the first two principal components of the backbone dihedral PCA of the ensemble obtained from a REST simulation started from a conformation predicted by a structure prediction server (i.e. trajectory discussed in the main manuscript, "Original"). b) Projection on the same two principal components of the dihedral ensemble obtained from a REST simulation started from a fully elongated coil conformation ("Replica). c) Comparison of the cluster occupancies after cluster analysis with the GROMOS algorithm using an RMSD cut-off of 2.0 Å. The 'other' histogram bin collects clusters with populations smaller than 1 %. d) Comparison between the backbone conformations.*

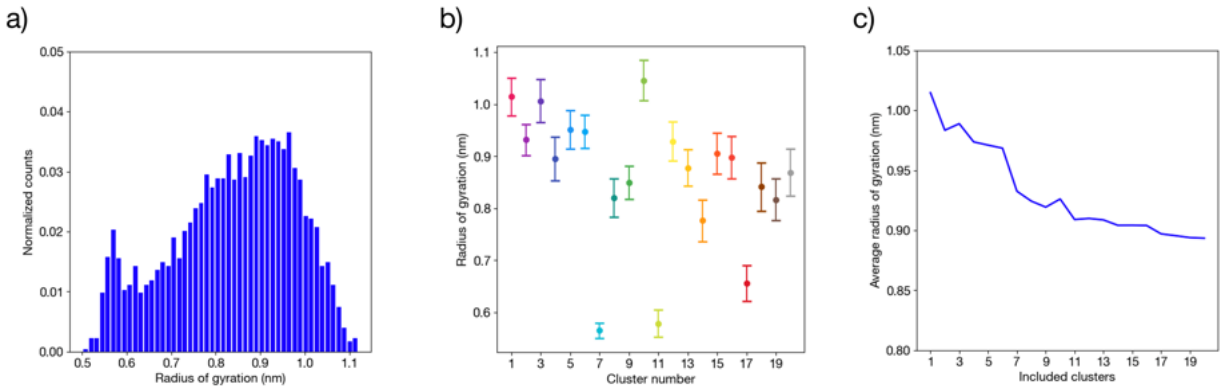

Figure S5: *Radius of gyration (Rg) in the peptide's ensemble. a) Histogram of the backbone Rg of all structures in the REST trajectory. b) Graph of the average Rg and standard deviation for each of the 20 most populated clusters in the REST trajectory. c) Evolution of the weighted-averaged Rg by adding subsequent central structures of the 20 most populated clusters of the REST trajectory.*

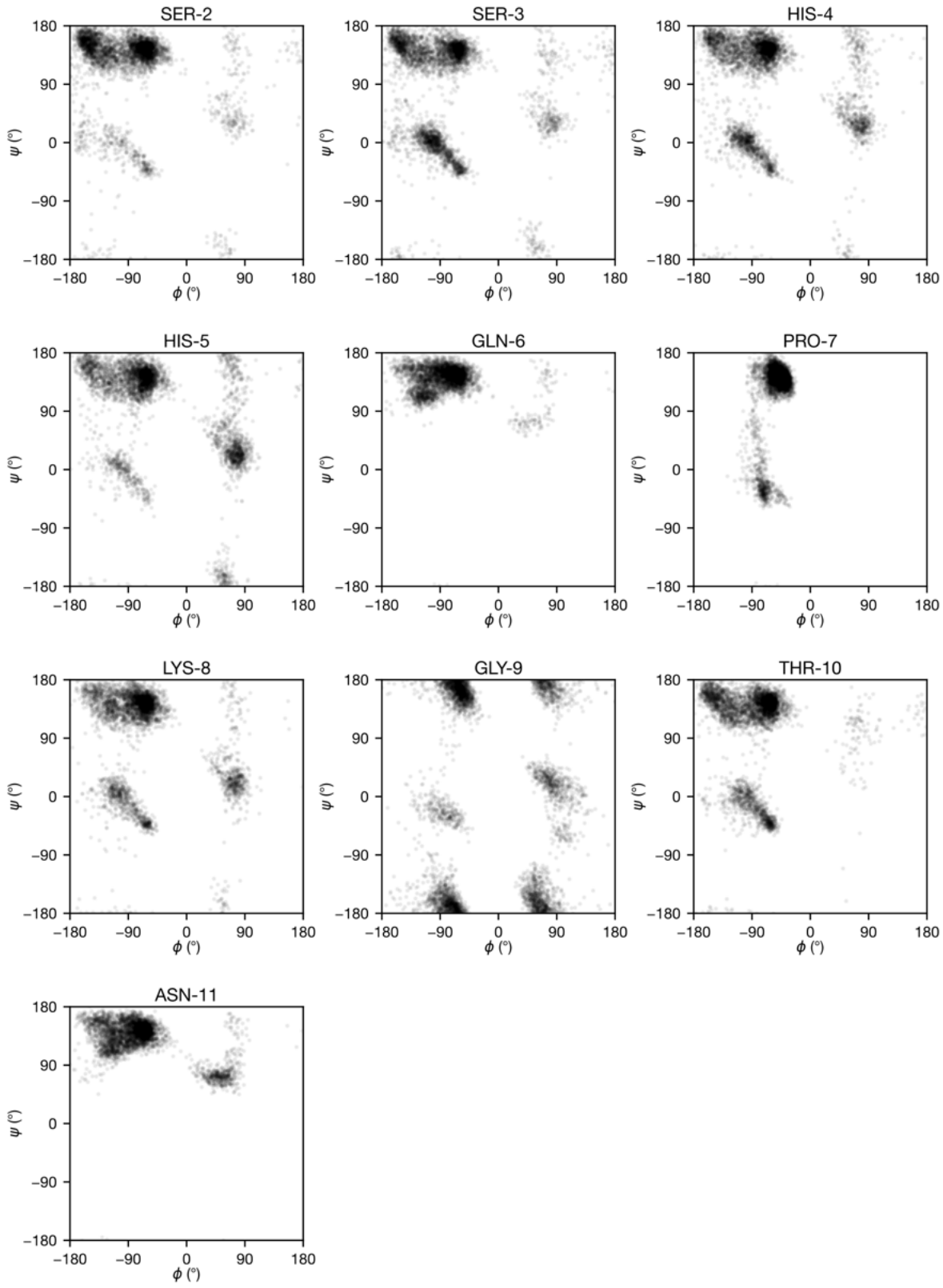

Figure S6: *Ramachandran plots for each residue, showing the dihedral distribution over the ensemble.*

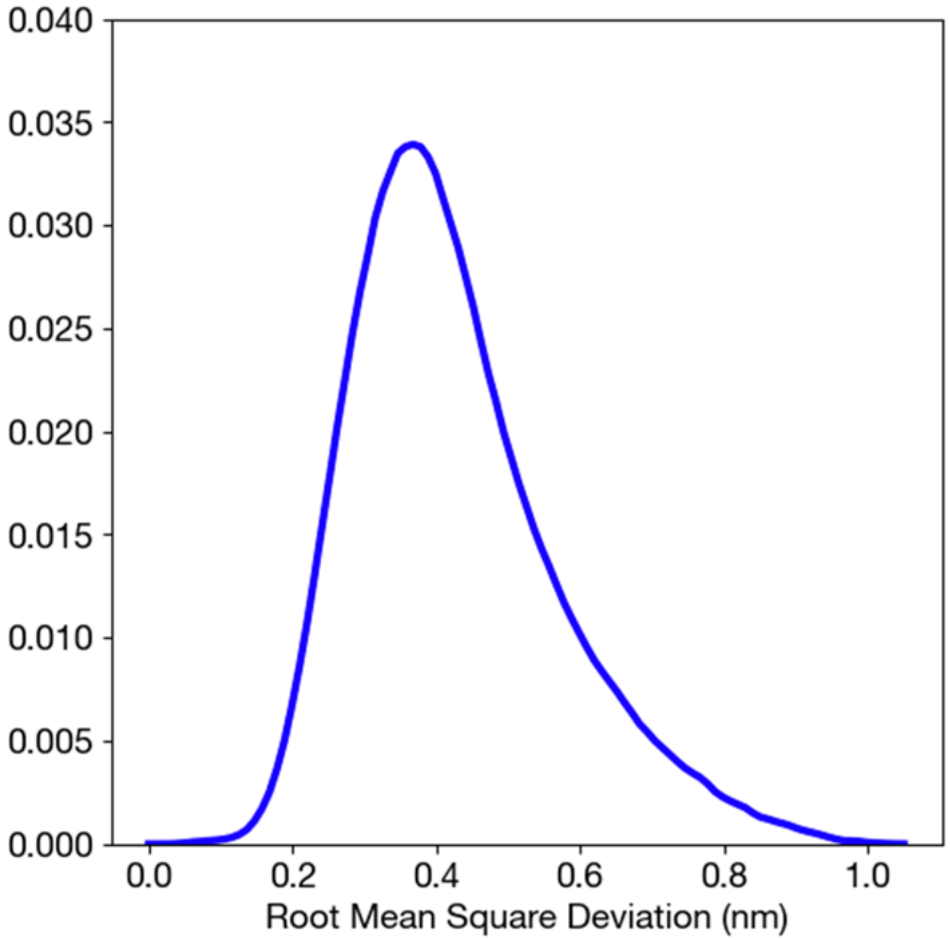

Figure S7: *Histogram of the pairwise RMSD of all structures in the REST trajectory.*

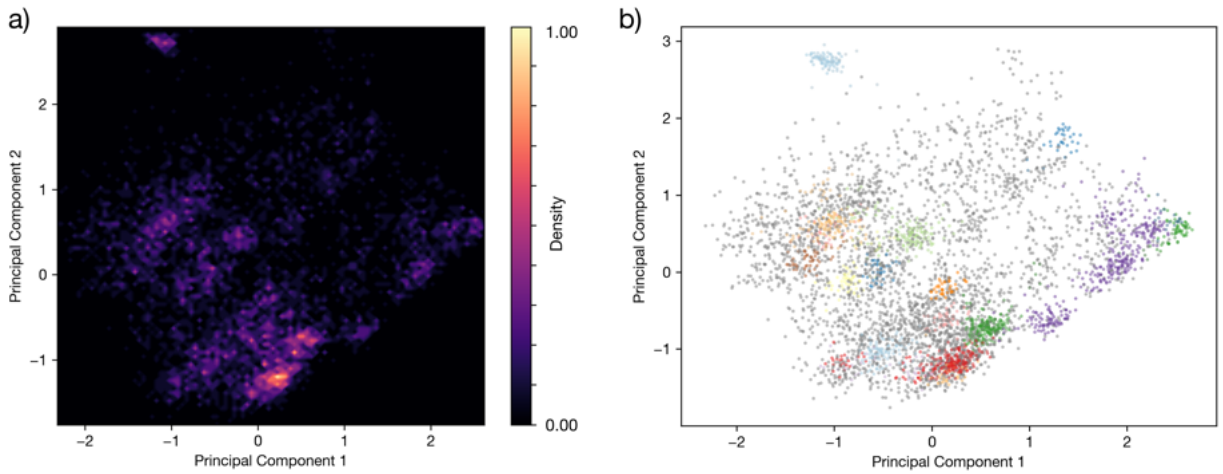

Figure S8: *a) Projection on the first two principal components of the backbone dihedral PCA of the REST ensemble, colored according to the density of points. b) Projection on the first two principal components of the backbone dihedral PCA of the REST ensemble, colored according to unsupervised clustering (HDBSCAN*).*

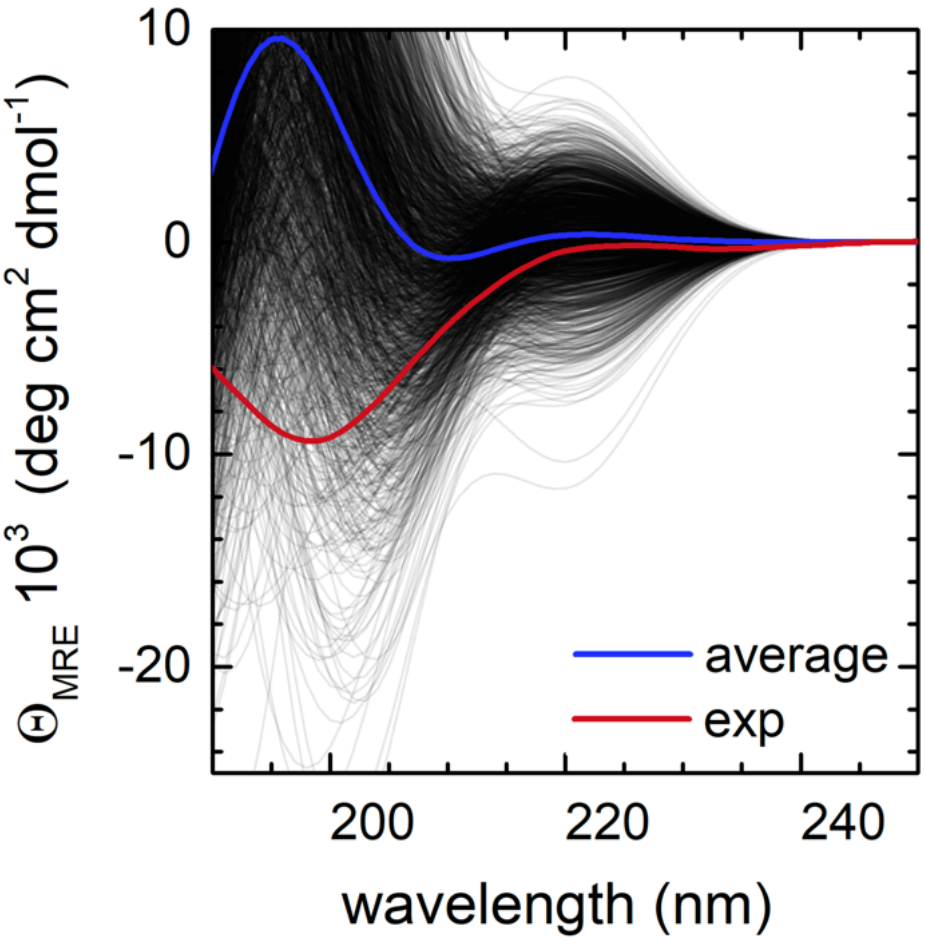

Figure S9: *Experimental CD spectrum (red curve) and simulated spectra (black lines) of 2000 conformers of the peptide in water, obtained with the DichroCalc web server. The blue curve is the calculated ensemble average.*

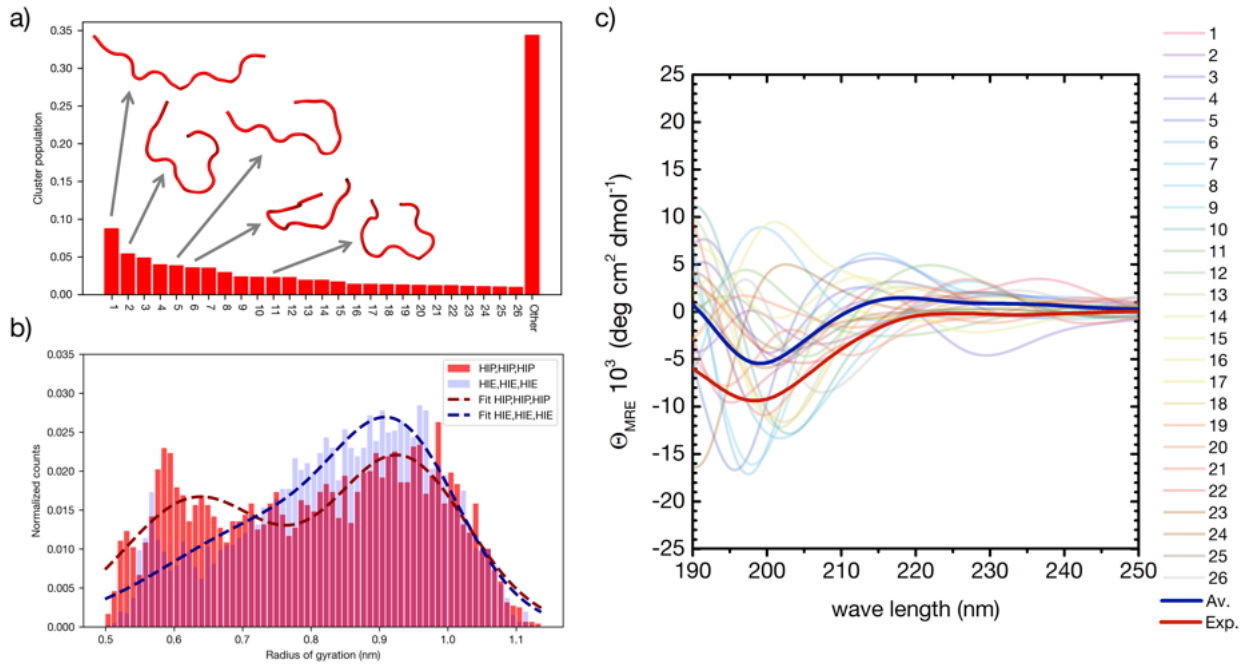

Figure S10: *Conformational ensemble of the peptide with all histidines protonated (HIP protonation state). a) Cluster occupancies after cluster analysis with the GROMOS algorithm using an RMSD cut-off of 2.0 Å. The 'other' histogram bin collects clusters with populations smaller than 1 %. Selected backbone conformations are reported. b) Histograms of the backbone Rg of all structures in the REST trajectories, both in the case of all ϵ-protonated histidines (HIE, HIE, HIE), and all protonated histidines (HIP, HIP, HIP), together with multi peak gaussian fitting. c) Computed CD spectra of the central structures of the 26 most populated clusters of the REST trajectory with all histidines protonated (HIP protonation state), with their weighted average (blue) compared to the experiment (red).*

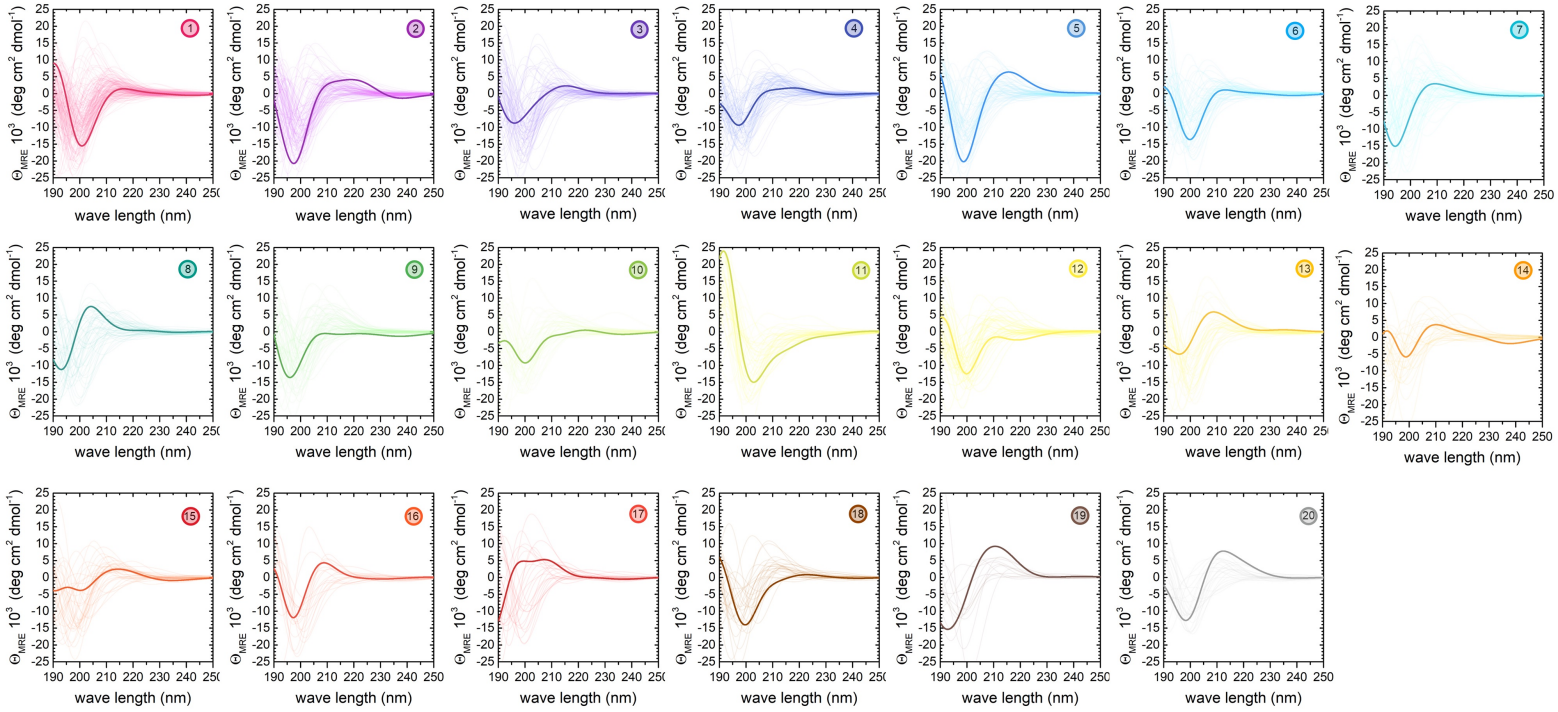

Figure S11: *Computed CD spectra for the structures belonging to the 20 most populated clusters of the REST trajectory (color code as in Fig. 3): all spectra are reported, with the spectrum of the central structure drawn in bold.*

Microsolvation

Although the implicit solvent model (C-PCM) accounts for solvent-solute interactions of the peptide with water, hydrogen bonding with the solvent is not explicitly taken into account. As the amide modes are specifically sensitive to hydrogen bonding with water,[35,36] we also recalculated the 20 IR, Raman and ROA spectra including microsolvation by taking the water molecules that are hydrogen bonded to the peptide amide groups in the respective MD snap shots explicitly into account. These weighted averages of the calculated spectra are shown in Figure S12 (orange; top panel), compared to the spectra that were calculated only with the implicit solvent model (blue; top panel). In Figure S13, the individual calculated spectra with explicit solvation are compared to the implicit calculations. All the amide band shapes and positions are influenced by the microsolvation, as described in literature.[35,36] Yet, as the water molecules are explicitly calculated with DFT (peptide and water molecules are the QM system), also the spectroscopically active modes of the explicit water molecules are included. For example, the symmetric bending mode of water overlaps with the calculated amide I', and prohibits to make further detailed comparisons of the amide regions in experiment and theory. As the spectra were simulated using the same scaling factor of 0.95 for the amide I (and I'), the microsolvation shows that the amide I (and I') modes using only C-PCM are indeed overestimated (as noted above) and need this higher scaling factor to align this spectral region with experiment.

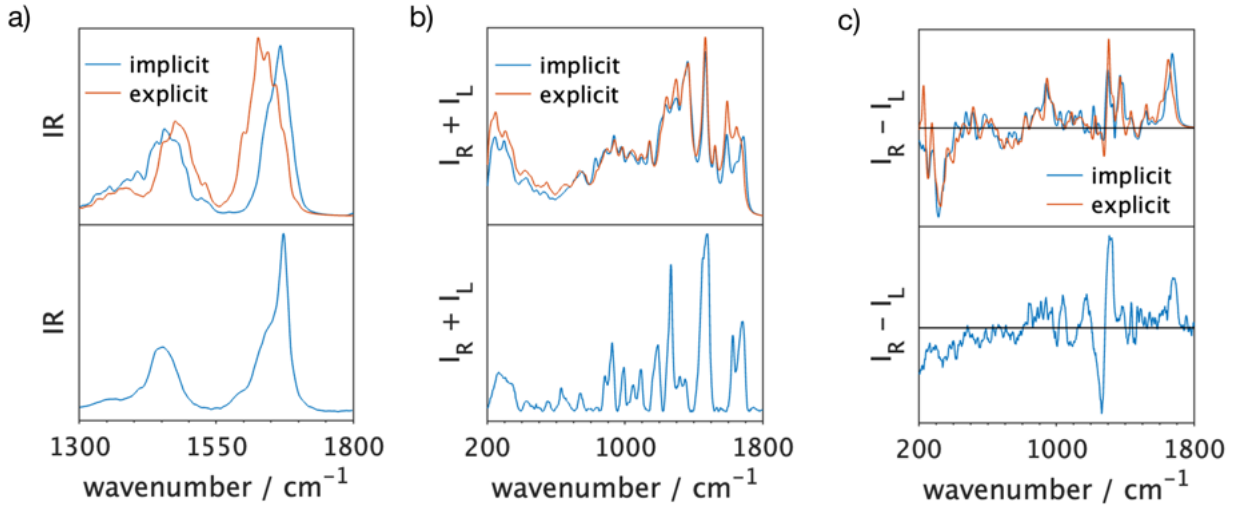

Figure S12: *Influence of solvation of the amide groups: simulated spectra where only C-PCM was used (blue) compared to when the water molecules that are hydrogen-bonded to the amide groups are explicitly included (orange) in the simulated spectra (top panels; weighted average of the 20 cluster central structures) compared to the corresponding experimental (bottom panel) (a) FT-IR spectrum, (b) Raman and (c) ROA spectrum. A scaling factor of 0.95 is used for the amide I and I' region and a scaling factor of 1.00 is applied for the amide II' region in the IR and 0.987 for the remainder of the Raman and ROA spectrum.*

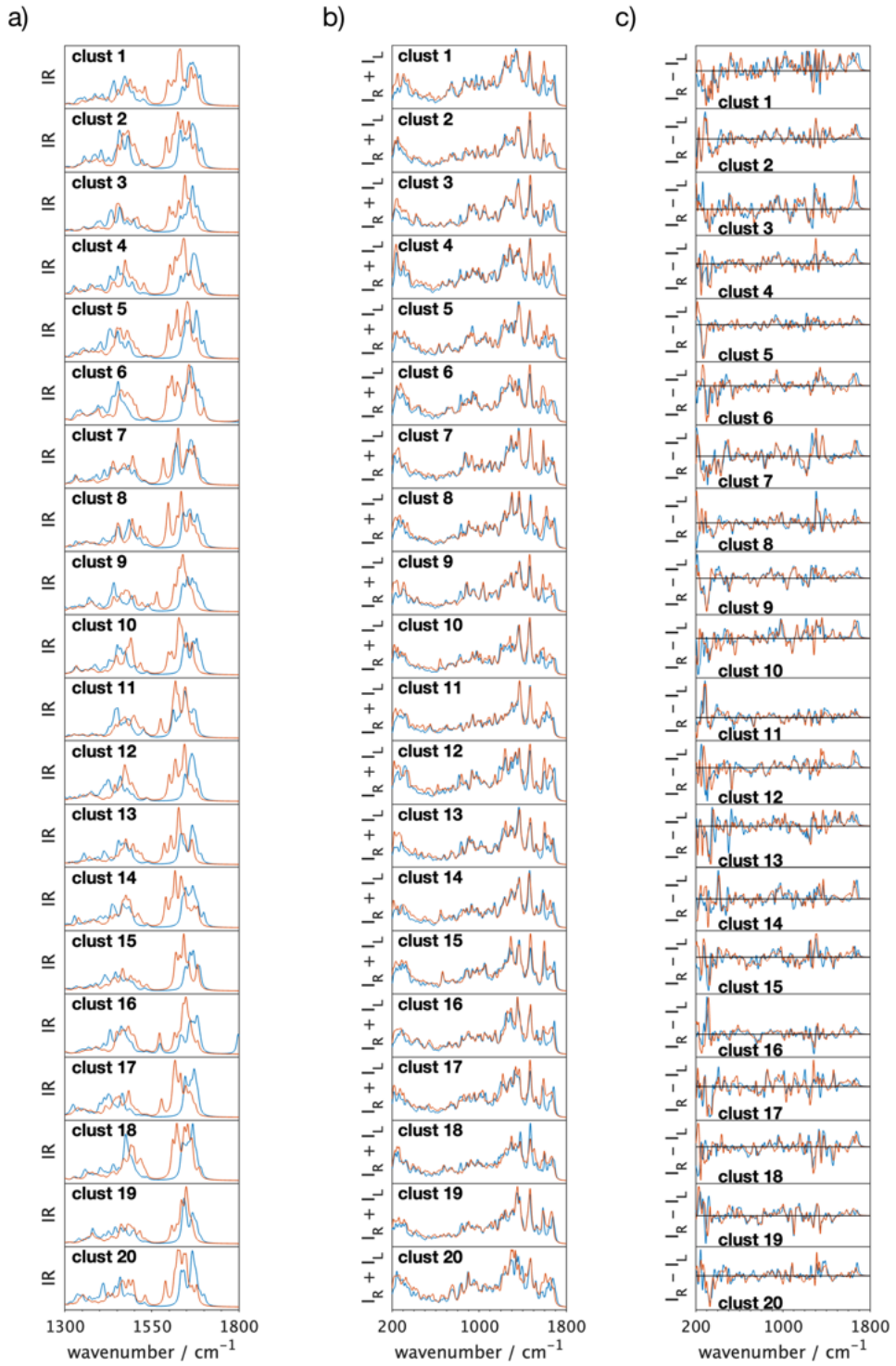

Figure S13: *DFT calculated (a) IR spectrum, (b) Raman and (c) ROA spectrum. for each cluster center, with a scaling factor of 0.95 for the amide I' region ($\geq 1550\,cm^{-1}$) and a scaling factor of 1.0 for the amide II' (1300 to $1550\,cm^{-1}$). The blue spectra correspond to the peptide calculated only using C-PCM. The blue spectra were calculated for the peptide with microsolvation and C-PCM, including the water molecules that are hydrogen bonded to the amide groups.*

Protonation state of the Histidine imidazole groups

The sensitivity of Raman to the peptide's side chains, and specifically the aromatic side chains could have a pronounced contribution to the calculation of the spectra. Since the peptide has three histidines, we demonstrate the influence of these in the calculated spectra. The protonation state of the imidazole ring of histidine in the spectral calculations was taken to be neutral. Since the experimental pH is close to the pKa of histidine, we calculated the 20 cluster structures again with one of the three histidines protonated to HIP (red H1, yellow H4, purple H5) and with all three histidines protonated to HIP (green) in Figure S14. While the influence on the IR and ROA is limited, the Raman bands that are affected are those around 1600 cm^{-1} (imidazole stretching deformations: C=C stretching), around 1510 cm^{-1} (imidazole stretching deformations: C=N stretching), 1360 cm^{-1} (complex coupled vibrations, C-H deformations and imidazole stretching deformations), 1190 cm^{-1} (skeletal deformations, also in the protonated imidazole).[37]

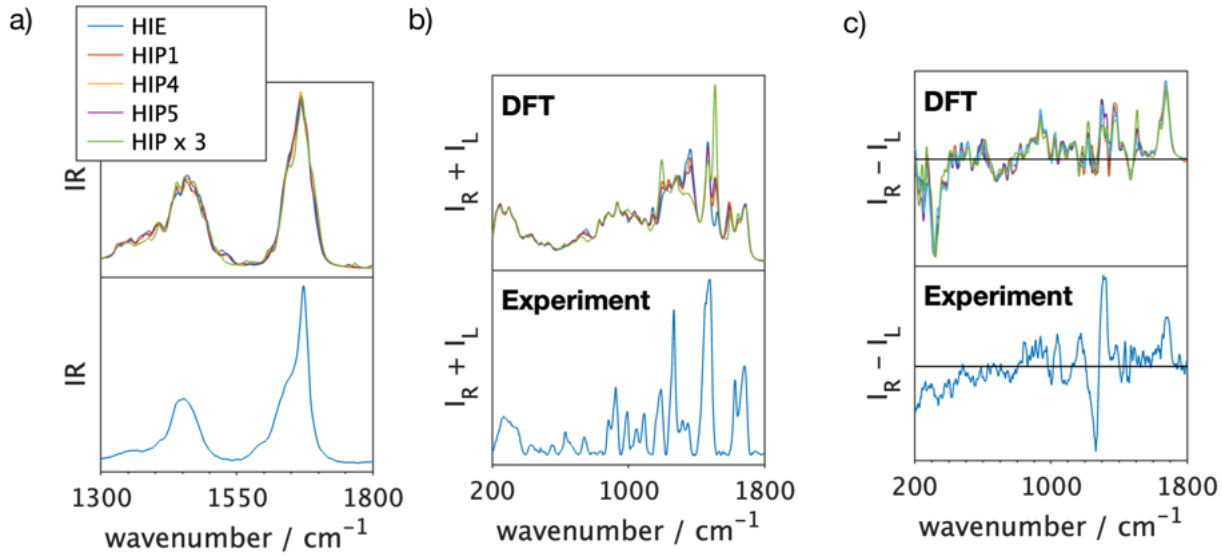

Figure S14: *Top panels: weighted averages of the DFT calculated spectra of all 20 cluster structures with (blue) neutral histidine side chains, (red) HIP 1 , (yellow) HIP 4, (purple) HIP 5, (green) all three histidine side chains protonated to HIP in the peptide: H1SSH4H5QPKGTNP.*

Proof-of-concept: the peptide in a PPII conformation

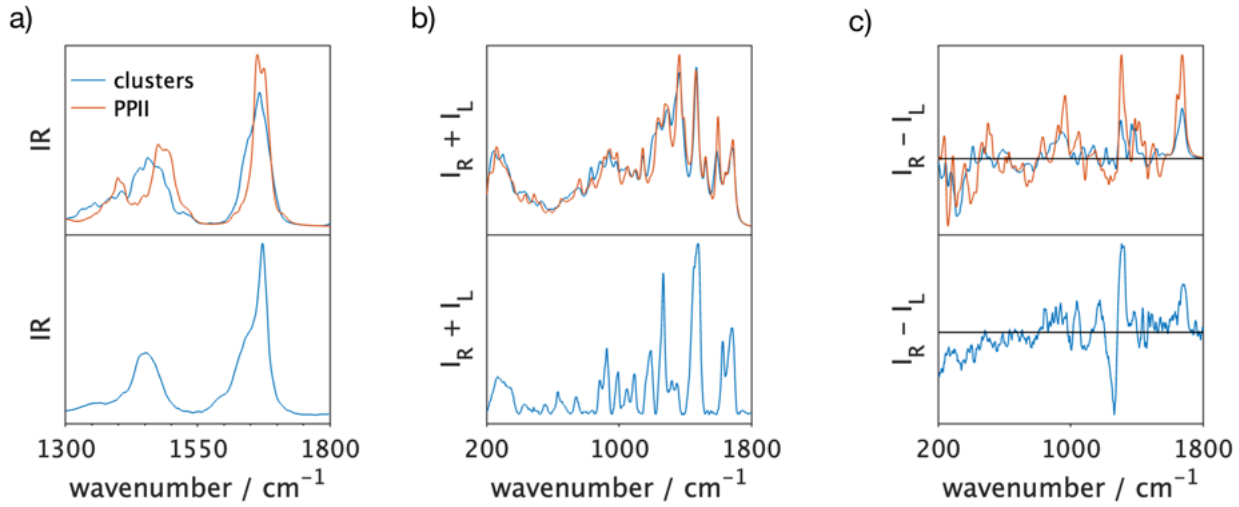

Figure S15: *Computed (a) IR, (b) Raman and (C) ROA spectra for the peptide fixed in a PPII conformation (ϕ and ψ values of $-75°$ and $+145°$, respectively) with 20 different configurations of the side chains (top panels) and their average compared to experiment (bottom panels).*

Proof-of-concept: randomized side chains

We extended the small proof-of-concept calculations with the PPII canonical conformation to the cluster center structures. The first four cluster center structures were selected, and their backbone conformation kept the same, while all torsion angles of the residues side chains were randomized over 20 different side chain configurations (different χ-angles). As shown in Figure S16, a single side chain configuration has a specific IR, Raman and ROA pattern. Averaging over different side chain conformations, leads to broader bands and a better comparison with experiment, yet for the IR and Raman, averaging over the 20 cluster center structures (Figure S1) yields a better simulation of the experiment. For the ROA, due to the bisignate nature of the spectra, a lot more spectral averaging seems necessary to simulate the experiment.

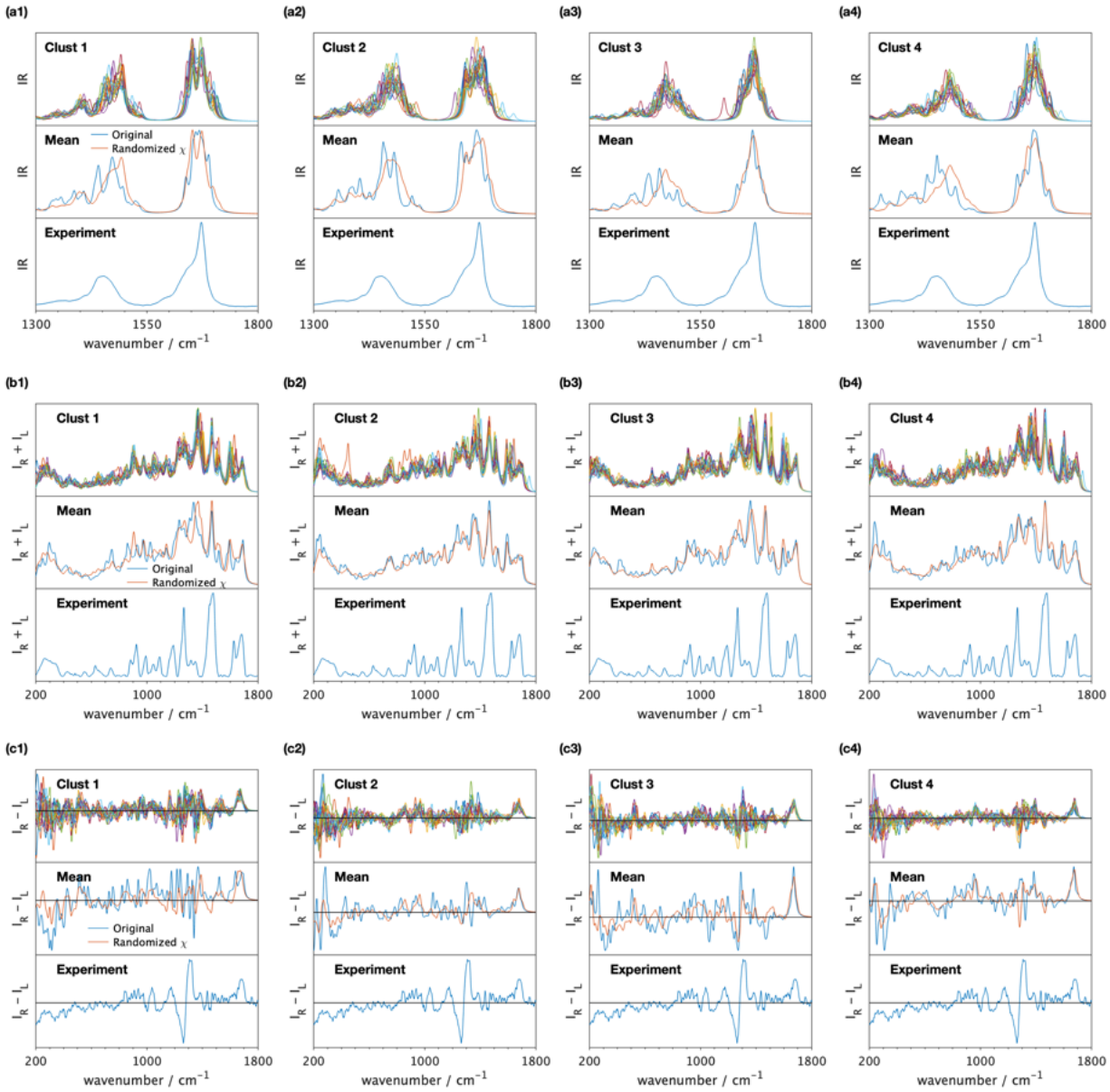

Figure S16: *Randomization of the side chain conformations for cluster 1-4 in (a1-4) IR, (b1-4) Raman and (c1-4) ROA. The top panel shows the 20 spectra calculated using the backbone conformation of the respective cluster with 20 random configurations of the side chain χ-angles. The middle panels show the mean spectra of those 20 spectra (orange) compared to the spectrum of the original cluster center structure (blue). The bottom spectra are the respective experimental spectra. The amide I and I' is scaled by 0.95, the amide II' in the IR by 1.0 and the remainder of the Raman and ROA spectra by 0.987.*

References

(1) Anthis, N. J.; Clore, G. M. Sequence-specific determination of protein and peptide concentrations by absorbance at 205 nm. *Protein Science* **2013**, *22*, 851–858, DOI: 10.1002/pro.2253.

(2) Hug, W.; Hangartner, G. A novel high-throughput Raman spectrometer for polarization difference measurements. *Journal of Raman Spectroscopy* **1999**, *30*, 841–852.

(3) Boelens, H. F.; Dijkstra, R. J.; Eilers, P. H.; Fitzpatrick, F.; Westerhuis, J. A. New background correction method for liquid chromatography with diode array detection, infrared spectroscopic detection and Raman spectroscopic detection. *Journal of Chromatography A* **2004**, *1057*, 21 – 30, DOI: https://doi.org/10.1016/j.chroma.2004.09.035.

(4) Pronk, S.; Pall, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; van der Spoel, D.; et al., GROMACS 4.5: a High-Throughput and Highly Parallel Open Source Molecular Simulation Toolkit. *Bioinformatics* **2013**, *29*, 845–854, DOI: 10.1093/bioinformatics/btt055.

(5) Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; de Groot, B. L.; Grubmüller, H.; MacKerell Jr, A. D. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nature methods* **2017**, *14*, 71, DOI: 10.1038/nmeth.4067.

(6) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics* **1983**, *79*, 926–935.

(7) Hess, B. P-LINCS: A Parallel Linear Constraint Solver for Molecular Simulation. *Journal of Chemical Theory and Computation* **2008**, *4*, 116–122, DOI: 10.1021/ct700200b.

(8) Affentranger, R.; Tavernelli, I.; Di Iorio, E. E. A Novel Hamiltonian Replica Exchange MD Protocol to Enhance Protein Conformational Space Sampling. *J. Chem. Theory Comput.* **2006**, *2*, 217–228, DOI: 10.1021/ct050250b.

(9) Wang, L.; Friesner, R. A.; Berne, B. J. Replica Exchange with Solute Scaling: a More Efficient Version of Replica Exchange with Solute Tempering (REST2). *J. Phys. Chem. B* **2011**, *115*, 9431–9438, DOI: 10.1021/jp204407d.

(10) Bussi, G. Hamiltonian Replica Exchange in GROMACS: a Flexible Implementation. *Mol. Phys* **2013**, *112*, 379–384, DOI: 10.1080/00268976.2013.824126.

(11) Rathore, N.; Chopra, M.; de Pablo, J. J. Optimal allocation of replicas in parallel tempering simulations. *J. Chem. Phys.* **2005**, *122*, 024111, DOI: 10.1063/1.1831273.

(12) Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

(13) Campello, R. J.; Moulavi, D.; Sander, J. Density-based clustering based on hierarchical density estimates. *Lect. Notes Comput. Sci.* **2013**, *7819 LNAI*, 160–172, DOI: 10.1007/978-3-642-37456-2_14.

(14) Sreerama, N.; Woody, R. W. Computation and analysis of protein circular dichroism spectra. *Methods Enzymol.* **2004**, *383*, 318–351, DOI: 10.1016/S0076-6879(04)83013-1.

(15) Berova, N.; Polavarapu, P. L.; Nakanishi, K.; Woody, R. W. *Comprehensive Chiroptical Spectroscopy: Instrumentation, Methodologies, and Theoretical Simulations*; John Wiley & Sons: Hoboken, NJ, USA, 2012; Vol. 1.

(16) Jurinovich, S.; Pescitelli, G.; Di Bari, L.; Mennucci, B. A TDDFT/MMPol/PCM model for the simulation of exciton-coupled circular dichroism spectra. *Phys. Chem. Chem. Phys.* **2014**, *16*, 16407.

(17) Segatta, F.; Cupellini, L.; Garavelli, M.; Mennucci, B. Quantum Chemical Modeling of the Photoinduced Activity of Multichromophoric Biosystems. *Chem. Rev.* **2019**, *119*, 9361–9380, DOI: 10.1021/acs.chemrev.9b00135.

(18) Morrison, A. F.; You, Z.-Q.; Herbert, J. M. Ab Initio Implementation of the Frenkel–Davydov Exciton Model: A Naturally Parallelizable Approach to Computing Collective Excitations in Crystals and Aggregates. *J. Chem. Theory Comput.* **2014**, *10*, 5366–5376, DOI: 10.1021/ct500765m.

(19) Loco, D.; Jurinovich, S.; Di Bari, L.; Mennucci, B. A fast but accurate excitonic simulation of the electronic circular dichroism of nucleic acids: how can it be achieved? *Phys. Chem. Chem. Phys.* **2016**, *18*, 866–877.

(20) Padula, D.; Jurinovich, S.; Di Bari, L.; Mennucci, B. Simulation of Electronic Circular Dichroism of Nucleic Acids: From the Structure to the Spectrum. *Chem. Eur. J.* **2016**, *22*, 17011–17019.

(21) Ianeselli, A.; Orioli, S.; Spagnolli, G.; Faccioli, P.; Cupellini, L.; Jurinovich, S.; Mennucci, B. Atomic Detail of Protein Folding Revealed by an Ab Initio Reappraisal of Circular Dichroism. *J. Am. Chem. Soc.* **2018**, *140*, 3674–3682, DOI: 10.1021/jacs.7b12399.

(22) Bondanza, M.; Nottoli, M.; Cupellini, L.; Lipparini, F.; Mennucci, B. Polarizable embedding QM/MM: the future gold standard for complex (bio)systems? *Phys. Chem. Chem. Phys.* **2020**, *19*, DOI: 10.1039/D0CP02119A.

(23) Chai, J.-D.; Head-Gordon, M. Long-range corrected hybrid density functionals with damped atom–atom dispersion corrections. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6615, DOI: 10.1039/b810189b.

(24) Lipparini, F.; Mennucci, B. Hybrid QM/classical models: Methodological advances and new applications. *Chem. Phys. Rev.* **2021**, *2*, 041303, DOI: 10.1063/5.0064075.

(25) Curutchet, C.; Muñoz-Losa, A.; Monti, S.; Kongsted, J.; Scholes, G. D.; Mennucci, B. Electronic Energy Transfer in Condensed Phase Studied by a Polarizable QM/MM Model. *J. Chem. Theory Comput.* **2009**, *5*, 1838–1848, DOI: 10.1021/ct9001366.

(26) Frisch, M. J. et al. Gaussian 16 Revision A.03. 2016; Gaussian Inc. Wallingford CT.

(27) Lipparini, F. General Linear Scaling Implementation of Polarizable Embedding Schemes. *J. Chem. Theory Comput.* **2019**, *15*, 4312–4317, DOI: 10.1021/acs.jctc.9b00585.

(28) Jurinovich, S.; Guido, C.; Bruhn, T.; Pescitelli, G.; Mennucci, B. The role of magnetic-electric coupling in exciton-coupled ECD spectra: the case of bis-phenanthrenes. *Chem. Commun.* **2015**, *51*, 10498–10501, DOI: 10.1039/c5cc03167b.

(29) Jurinovich, S.; Cupellini, L.; Guido, C. A.; Mennucci, B. EXAT: EXcitonic analysis tool. *J. Comput. Chem.* **2018**, *39*, 279–286, DOI: 10.1002/jcc.25118.

(30) Bouř, P.; Keiderling, T. A. Partial optimization of molecular geometry in normal coordinates and use as a tool for simulation of vibrational spectra. *The Journal of Chemical Physics* **2002**, *117*, 4126–4132, DOI: 10.1063/1.1498468.

(31) Mensch, C.; Barron, L. D.; Johannessen, C. Ramachandran mapping of peptide conformation using a large database of computed Raman and Raman optical activity spectra. *Phys. Chem. Chem. Phys.* **2016**, *18*, 31757–31768, DOI: 10.1039/C6CP05862K.

(32) Cheeseman, J. R.; Frisch, M. J. Basis Set Dependence of Vibrational Raman and Raman Optical Activity Intensities. *Journal of Chemical Theory and Computation* **2011**, *7*, 3323–3334, DOI: 10.1021/ct200507e, PMID: 26598166.

(33) Niederhafner, P.; Šafařík, M.; Neburková, J.; Keiderling, T. A.; Bouř, P.; Šebestík, J. Monitoring peptide tyrosine nitration by spectroscopic methods. *Amino acids* **2021**, *53*, 517–532, DOI: 10.1007/s00726-020-02911-7.

(34) Debie, E.; De Gussem, E.; Dukor, R. K.; Herrebout, W.; Nafie, L. A.; Bultinck, P. A confidence level algorithm for the determination of absolute configuration using vibrational circular dichroism or Raman optical activity. *ChemPhysChem* **2011**, *12*, 1542–1549, DOI: 10.1002/cphc.201100050.

(35) Mensch, C.; Johannessen, C. The influence of the amino acid side chains on the Raman optical activity spectra of proteins. *ChemPhysChem* **2019**, *20*, 42–54.

(36) Myshakina, N. S.; Ahmed, Z.; Asher, S. A. Dependence of amide vibrations on hydrogen bonding. *The Journal of Physical Chemistry B* **2008**, *112*, 11873–11877, DOI: 10.1021/jp8057355.

(37) Loeffen, P.; Pettifer, R.; Fillaux, F.; Kearley, G. Vibrational force field of solid imidazole from inelastic neutron scattering. *The Journal of chemical physics* **1995**, *103*, 8444–8455, DOI: 10.1063/1.470155.