PAPER • OPEN ACCESS

Information bottleneck in peptide conformation determination by x-ray absorption spectroscopy

Journal of Physics Communications

PAPER

Information bottleneck in peptide conformation determination by x-ray absorption spectroscopy

Eemeli A Eronen[1], **Anton Vladyka**[1], **Florent Gerbon**[2], **Christoph J Sahle**[2] and **Johannes Niskanen**[1,*]

[1] University of Turku, Department of Physics and Astronomy, FI-20014 Turun yliopisto, Finland
[2] ESRF, The European Synchrotron, 71 Avenue des Martyrs, CS40220, 38043 Grenoble Cedex 9, France
[*] Author to whom any correspondence should be addressed.

E-mail: eemeli.a.eronen@utu.fi and johannes.niskanen@utu.fi

Keywords: x-ray spectroscopy, machine learning, statistical analysis, secondary structure, peptide, molecular dynamics, density functional theory

Supplementary material for this article is available online

OPEN ACCESS

RECEIVED
6 October 2023

REVISED
28 December 2023

ACCEPTED FOR PUBLICATION
17 January 2024

PUBLISHED
7 February 2024

Abstract

We apply a recently developed technique utilizing machine learning for statistical analysis of computational nitrogen K-edge spectra of aqueous triglycine. This method, the emulator-based component analysis, identifies spectrally relevant structural degrees of freedom from a data set filtering irrelevant ones out. Thus tremendous reduction in the dimensionality of the ill-posed nonlinear inverse problem of spectrum interpretation is achieved. Structural and spectral variation across the sampled phase space is notable. Using these data, we train a neural network to predict the intensities of spectral regions of interest from the structure. These regions are defined by the temperature-difference profile of the simulated spectra, and the analysis yields a structural interpretation for their behavior. Even though the utilized local many-body tensor representation implicitly encodes the secondary structure of the peptide, our approach proves that this information is irrecoverable from the spectra. A hard x-ray Raman scattering experiment confirms the overall sensibility of the simulated spectra, but the predicted temperature-dependent effects therein remain beyond the achieved statistical confidence level.

1. Introduction

Proteins are formed of amino acids, often cited as the building blocks of life. The backbone of the amino acid chain consists of a -CO-NH-C_α—sequence, where adjacent residues are bound by the peptide bond [1]. The geometry of the backbone contains pairs of flexible dihedral angles, the Ramachandran angles, which together define the macroscopic structure and the function of the protein. The related folding of proteins is a complex and complicated question [2], affected by both intramolecular and intermolecular interactions. The process is ultimately dependent on the primary order of the amino acids [1] and the surrounding solvent [3, 4].

While diffraction experiments allow for structure determination of proteins, they require a crystalline sample. Owing to its localized mechanism, x-ray spectroscopy maintains its sensitivity to atomistic structure also in the soft condensed phase, although statistical variation of the individual spectra in such environments has been found to be huge [5–13]. This variation can be accounted for by evaluation of the spectra at numerous structures, sampled from the respective statistical ensemble, and averaging over them. Furthermore, x-ray spectroscopy is sensitive to the local structure of amino acids and peptides [6, 14–20]. For example, different protonation forms of glycine yield different resonant inelastic x-ray scattering spectra [19] and x-ray photoelectron spectra [6, 21]. It has also been found that near-edge x-ray absorption fine structure is sensitive to a few neighbouring amino acids in poly-Gly peptides [15, 16, 20].

To predict the native form of proteins in their biological environment, classical molecular dynamics (MD) is often used [22]. In this technique, the atomistic composition of the system is accounted for, but the forces between the atoms are computed from a prior parametrization rather than from the electronic-nuclear system

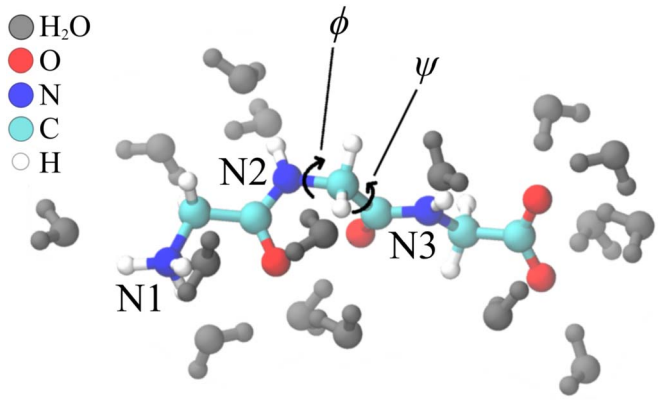

Figure 1. A schematic illustration of aqueous triglycine. The absorption sites are denoted as N1, N2 and N3. The backbone dihedrals known as the Ramachandran angles ϕ and ψ are shown for the central residue. The molecular plot was prepared using the VMD software [66].

that ultimately defines them. Being computationally much lighter, and allowing for time and size scales not otherwise accessible, this makes classical MD subject to the quality of the parametrization—the force field—used in the calculations [23, 24]. For the development of these models in the solution environment, their performance assessment in the atomistic level would be valuable.

Interpretation of x-ray spectra is typically not straightforward. This is due to the complicated relation between spectra and structures, originating from quantum nature of the electronic system. In addition, the statistical aspects of the liquid state cause an experiment to probe only the ensemble average. To tackle these problems, machine learning and related emulator-based component analysis (ECA) [25] may prove useful. The ECA algorithm carries out dimensionality reduction in structural space to identify structural variations having the strongest influence on the spectra. Thus the method identifies recoverable (and irrecoverable) structural information without a prior hypothesis.

In this work we report a theoretical study of the smallest tripeptide triglycine in aqueous solution (see figure 1) at temperatures of 300 K and 350 K, expecting a change in the distribution of the Ramachandran angles. We calculate an extensive set of N K-edge x-ray absorption spectra from MD trajectories with three different force fields, observing notable structural and spectral variation. Next, we study the dependency between intensities of spectral regions of interest (ROI) and the corresponding structures using a neural network (NN) and ECA. The results show that while x-ray absorption ROIs are sensitive to the nearest neighboring atoms of the absorption site, they are not sensitive to the secondary structure *i.e.* the Ramachandran angles of the system, in agreement with Schwartz *et al* [26]. The simulated spectra are in good agreement with an x-ray Raman scattering (XRS) experiment we performed. However, the predicted spectral difference profile as a function of temperature remains unconfirmed with the achieved statistical uncertainty.

2. Methods

2.1. Simulations

We simulated the NPT ensemble by classical MD at 300 K and at 350 K and 1 bar using three different force fields: AMBER-03 [27], Charmm27 [28, 29] and OPLS-AA [30]. For the first two force fields we applied the TIP3P [31] water model and for the latter the TIP4P [31] model. The cubic simulation cell had an edge length of ~50 Å. The input files were prepared using a desktop installation of GROMACS [32] package (version 2020.1), and the simulations were performed using the version 2020.5 [33] on a computing cluster. Rigid water molecules were used whereas no constraints were applied on the triglycine molecule. After 1 ns of initial thermalization with Berendsen thermostat and barostat [34], we ran MD for 51 ns (timestep 0.5 fs) using the Nosé-Hoover thermostatting [35, 36] (time constant $\tau_T = 0.5$ps) and Isotropic Parrinello-Rahman barostatting [37, 38] (time constant $\tau_P = 5.0$ ps, compressibility 4.5×10^{-5} bar^{-1}). To avoid the 'hot-solvent/cold-solute' problem, we used separate thermostats for the solute and for the solvent in the latter run, with conserved energy drift of the order 1 kJ mol^{-1}ns^{-1}atom^{-1}. The last 50 ns of the run were sampled with 10 ps spacing for spectrum calculations.

From the MD snapshots we calculated the N K-edge x-ray absorption spectra of 30 006 structures using the projector-augmented-wave (PAW) method [39] with plane wave basis and density functional theory (DFT), as implemented in GPAW version 22.1.0 [39–41]. As the simulation cell of MD is too large to be used in the

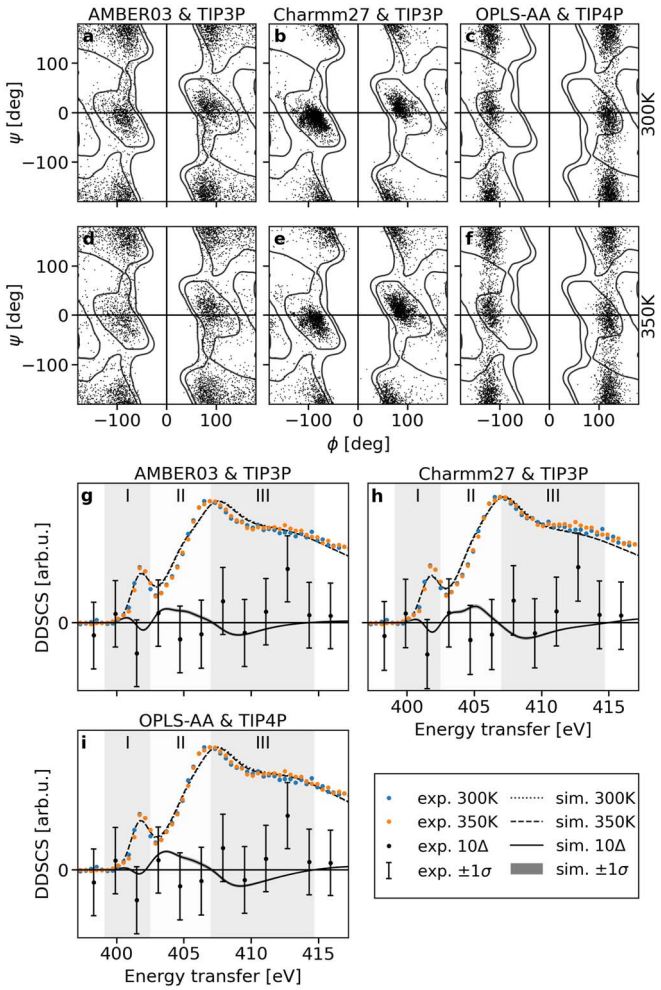

Figure 2. The calculated and experimental results for aqueous triglycine. (a)–(f): Scatter plot of the Ramachandran angles of the central residues from the simulated trajectories and the definitions [58] of the allowed (99.95%) and favored (99.8%) regions of glycine residue according to the Top8000 data set [59]. (g)–(i): Computational ensemble mean and experimental spectra (background removed) with the respective temperature difference profiles Δ and spectral regions of interest. For the experiment, the difference profile has been 4-fold binned from the spectra. The error bars and shading of the simulated curve indicate the statistical uncertainty σ (confidence level 68%). The computational spectra have been shifted by −2.9 eV in all cases for the pre peak to match with the experiment. The experiment is presented scaled for the same main peak height as the respective 300 K simulation.

quantum mechanical electron structure calculations, we used cut structures including only the water molecules within 3.0 Å of the solute. A vacuum with a radius of 3.0 Å was added around each structure. Excitations to the lowest 1500 valence single-electron states were evaluated in the transition potential half-hole (TP-HH) approximation [42] for each nitrogen site (N_{ex}). The spectrum onset was corrected using Δ-DFT method for the lowest core-excited state. The calculations utilized plane wave basis with the energy the cutoff of 350 eV and the Perdew–Burke–Ernzerhof (PBE) functional [43]. To aid convergence, occupation smearing by the Fermi–Dirac distribution (width 0.25 eV) was used.

The simulations resulted in energy–intensity pairs for the transitions of each absorption site (denoted as N1, N2 and N3; see figure 1), from which the spectra were obtained by convolution with Gaussian functions of increasing width analogously to the procedure presented in [44]. The full width at half maximum of the function was obtained by a numerical grid search for the best match with the experiment; we used 1.4 eV for the lowest state and increased the value linearly in energy to 4.5 eV for states 5.5 eV above it, or higher. An alternative set of convolution parameters was tested (from 0.2 eV to 4.25 eV for states 10 eV above the lowest state, or higher) without qualitative changes in the results (see Supplementary Information). The spectrum of a single snapshot was evaluated as the sum of the spectra from the three nitrogen atoms and the resulting ensemble mean spectra of all the systems matched the experiment well as seen in figure 2. In this set, 11 spectra were omitted in the subsequent analysis due to obviously nonsensical results.

To validate the computational results, sets of a few hundred to a few thousand spectra were calculated while varying one of the following simulation parameters: the number of water molecules included (both with a 6.0 Å cutoff, and also without any water), the plane wave energy cutoff (600 eV) or the functional (RPBE [45]). Some of

these calculations were troubled with convergence issues, but the successful ones show none of the parameter changes to have a significant effect on the temperature difference profile (see Supplementary Information). In total, the spectrum calculations took approximately 300 000 CPU hours on Intel Xeon Gold 6148 processors.

2.2. Data analysis

We divided the spectra into three ROIs, the selection of which was based on mean zero-passing location in the T-difference profiles Δ of the three models. These areas roughly correspond to pre peak (I), main edge (II), and continuum (III) in the spectrum, and each of them has a consistent T-dependence in the simulations. Using the full spectrum instead of ROIs would probably allow more information to be captured, but the risk of overinterpretation also grows with the grid tightness, as simulations are always erroneous in reproducing the experiment [46]. To conclude, we analyze the dependency between the ROI intensities **S** and the corresponding structure **R**.

We apply emulator-based component analysis [25] to this data for its interpretation. The algorithm searches for a few orthogonal basis vectors of structural space by maximizing the generalized covered variance (R^2-score) of the prediction for projections of the data onto the spanned subspace. This procedure utilizes the known target values of the original data points in the evaluation of the score. The components of the ECA basis vectors indicate dominant structural features in terms of the variation of the resulting output (in this work the spectral ROIs). The method needs a suitable machine learning (ML) based emulator $\mathbf{S}_{\mathrm{emu}}$ to predict the intensities of the three ROIs of any given structure **R**. We used an emulator consisting of nine separate NNs, one for each nitrogen atom and ROI, implemented by scikit-learn [47].

The individual structures **R** were encoded using the local many-body tensor representation (LMBTR) [48] implemented in the DScribe package [49]. The system is represented as distances between pairs and angles between triples of each element combination that includes the central atom—the absorption site. The internal hyperparameters of the LMBTR descriptor and the emulator were chosen by an alternating search, for best average ROI prediction (see Supplementary Information). In the search for the best descriptor, some LMBTR features were zero depending on the according set of hyperparameters. These features were manually removed before training the final model, without a significant effect on performance. The total dimensionality of the final LMBTR vector **D**(**R**) describing the local neighborhoods of the three nitrogen atoms was 1140. The ECA algorithm aims at dimensionality reduction, an optimization task complicated by the notably large number of dimensions of this space. We applied an alternating partial optimization with respect to 10% of the ECA vector components at a time. The routine was completed with a full optimization of all components at once. For these tasks we used the optimization toolbox of the SciPy [50] Python library and the trust-region interior point method [51] therein.

For the analysis of general spectrum–structure relationships, we combined all the data from the three force fields and the two temperatures into one data set, which was randomly divided for model selection and training (80%) and for testing and application (20%) of the emulator.

2.3. Experiment

We measured XRS spectra [52] at the N K-edge of aqueous triglycine using the multi-element XRS endstation [53] at beamline ID20 at the European Synchrotron Radiation Facility (ESRF). In the experiment scattering signal proportional to the double differential scattering cross section (DDSCS) [54] $\mathrm{d}^2\sigma/\mathrm{d}\Omega\mathrm{d}\omega_2$ at finite angle element $\mathrm{d}\Omega$ and outgoing photon energy element $\mathrm{d}\omega_2$ is recorded with a monochromatic and collimated incident beam of hard x-rays. When carried out at small scattering angles (forward direction) the XRS core-level signal is reduced to the dipole spectrum, and thus yields a spectrum equivalent to XAS [55]. Four analyzer modules, 12 spherically bent Si(660) crystals (bending radius $R_b = 1$ m) in each, were used to detect scattering in near-forward direction. We used the spacial imaging property of the bent analyzer crystals to include signal only from the interaction region by selecting the corresponding pixels of the CCD detector in the analysis. We observed different signal from the bulk liquid region and from the capillary walls and therefore excluded the latter pixels from the analysis.

To minimize the effect from radiation damage, a heatable liquid flow cell (modified version of a cell described elsewhere [56]) was used. In the system, a magnetically driven pump rotor is used to circulate ~ 5 ml of sample liquid through a capillary (2 mm outer diameter, 0.01 mm wall thickness), where inelastic x-ray scattering takes place. Moreover, the sample was replaced regularly to further reduce possible degradation. The sample solution of molality 0.3 mol/kg of aqueous triglycine was prepared by dissolving the powder sample (Sigma–Aldrich, purity $\geqslant 99.0\%$, lot # BCCB1590) into de-ionized water ($\rho \approx 18.2$ MΩ cm, TOC ≈ 2 ppb).

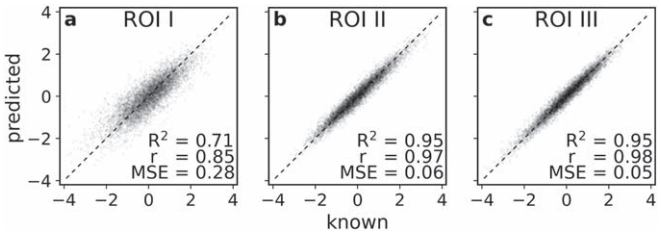

Figure 3. Emulator evaluation with standardized data. (a)–(c): Predicted intensities for ROIs I–III, respectively, plotted against the known ones. Covered variance R^2, Pearson's r, and the mean squared error MSE are given for each panel. Overall covered variance of the three ROIs is 0.874.

Table 1. Covered ROI intensity variances as a function of the rank of the ECA decomposition.

Rank	ECA fit	ECA validation
1	0.536	0.519
2	0.753	0.712
3	0.858	0.813
4	0.865	0.817
5	0.877	0.817

3. Results

The Ramachandran angles of the central residue, ϕ and ψ (see figure 1), can be visualized in a corresponding scatter plot. Depending on these two dihedrals, the residue can either exhibit an α-helix or β-sheet secondary structure. Our MD simulations show triglycine appears in both of these structural classes (figures 2(a)–(f)), but the distributions are different between the different force fields. As the molecule has no restrictive side chains, the plot is symmetric for the left-handed and right-handed isomers. The conformational variation of the liquid system gives rise to significant spectral variation. Using simulations allows studying the dependency between the shape of a spectrum and the underlying structure.

Figures 2(g)–(i) present the experimental spectra as well as the simulated ensemble-mean spectra for the corresponding force fields and the two temperatures. The temperature difference profile Δ is also shown. The simulated curves were shifted by -2.9 eV to match the pre peak of the experiment, and the experiment was scaled for the same main peak height with the respective 300 K simulation. Due to the huge Compton scattering background, approximately subtracted from the spectra, our experiment suffers from a rather large statistical uncertainty. This prohibits further conclusions as the predicted difference profiles lie within. However, an impressive match between the simulated and experimental spectra is obtained.

We did not observe a significant linear correlation between any of the structural features (internal coordinates and water related parameters) and the ROI intensities (see Supplementary Information). However, a well-functioning emulator will enable the use of ECA decomposition as a more advanced analysis tool. In this case, the apparent collaborative action of the structural degrees of freedom is, indeed, captured by an NN-based emulator as depicted in figure 3. The emulator achieved covered variances (R^2 score) of 0.714, 0.945, and 0.953 for each region I–III with the test data. The prediction accuracy is systematically worse for region I, the origin of which is the N1s$\rightarrow\pi^*$ resonance of the N atoms of the peptide bonds. Combining the regions together, the total covered variance stood at 0.874 for z-score standardized ROIs and 0.942 for the absolute intensity ROIs. The difference can be attributed to much larger total areas of ROIs II and III dominating the absolute score. Going forward, a score of 0.874 is the upper limit for covered spectral variance for the standardized ROI intensities by ECA decomposition, as even with complete structural coverage the respective ML-emulator-induced error will remain.

We split the test data further into two parts of the same size for the ECA. The first part was used for the optimization of the standardized structural descriptor space basis vectors (fit) and the latter was used for an independent test (validation). The covered variance as a function of the rank of the ECA expansion are given in table 1 for both data parts. While the covered ROI intensity variance improves up to rank five (emulator limit is achieved), convergence for the validation data is practically reached with three components.

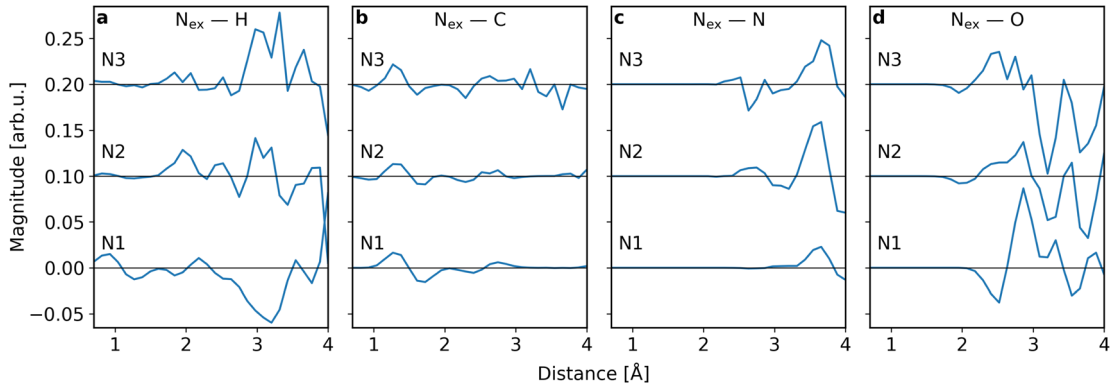

Figure 4. The interatomic distance distributions for each of the nitrogen atoms deduced from the first component vector transformed into the descriptor space. (a)–(d): Distances from the absorption site (denoted as N_{ex}) to the neighboring hydrogen, carbon, nitrogen and oxygen atoms, respectively. The ROIs of the total N K-edge spectrum is sensitive to the hydrogen and oxygen atoms at around 3 Å, which can mainly be contributed to the water molecules. In addition, the ROIs are sensitive to the nearest carbon and nitrogen atoms of each absorption site. Finally, the ROIs are also sensitive to the three nearest hydrogen atoms of N1.

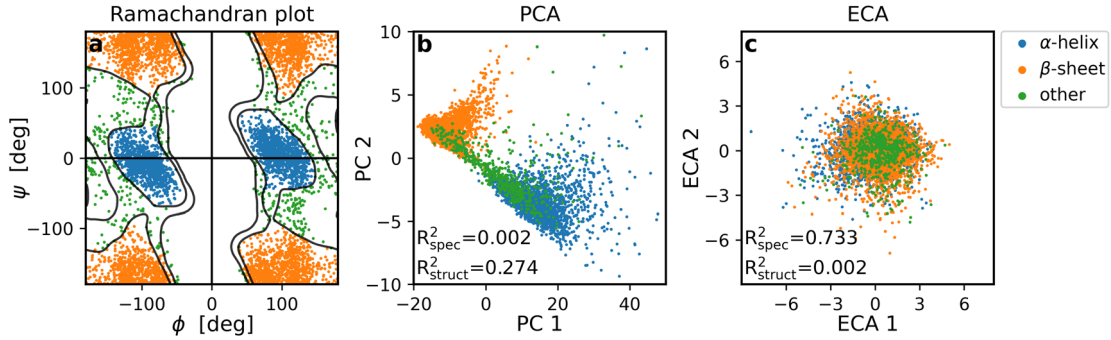

Figure 5. Analysis of the structural classes of the test data with PCA of LMBTR features and ECA reconstruction. (a): Data points classified to structural classes. The allowed and favored regions according to the Top8000 data set [59] are shown as contours. (b): A 2-component PCA decomposition of the LMBTR features of the data identifies the structural classes. (c): A 2-component ECA-coordinate reconstruction, based on spectral ROIs, covers drastically more spectral variation but does not distinguish the structural classes. For details, see text.

The ECA components can be used for structural interpretation [25, 57]. We analyze the first component (figure 4) in terms of convoluted distributions of interatomic distances, a subset of the entire LMBTR descriptor. This component shows the predominant structural changes associated with spectral ROI variation: increase in ROIs I and III, and decrease in ROI II (see Supplementary Information). There is a striking shift in N1–H, N_{ex}–C and N_{ex}–O interatomic distances, and N_{ex}–N distances, which however remain somewhat inconclusive due to the limited distance range of the descriptor. The surrounding water affects the spectrum as for both N_{ex}–H and N_{ex}–O there is a shift at around 3 Å near the first H_2O solvation shell. Interpretation of the angular structural features from the LMBTR is more complicated because exponential distance weighting was found to be required for best emulation performance. The weighting mode was one of the numerous hyperparameters varied in the model selection phase.

We next turn our focus on the features representing the secondary structure, the Ramachandran angles. We assigned each simulated structure in the test set (includes ECA fit and ECA validation data) to one of three classes: α-helix, β-sheet or 'other' based on the contours [58] from the Top8000 data set [59]. Borderline cases were assigned manually as shown in figure 5(a). The fractions of α-helices and β-sheets are given in table 2. We then used principal component analysis (PCA) to reduce the dimensionality of the LMBTR-encoded structures to two (spectral $R^2 = 0.002$) as shown in figure 5(b). This plot shows clear clusters of α-helices and β-sheets with 'others' found in between the two. The LMBTR can thus encode this information, even though the dihedrals ϕ and ψ are not directly included in it. Correspondingly, we used the ECA approach as a 2D dimensionality reduction tool for the same data (spectral $R^2 = 0.733$), as depicted in figure 5(c), without clearly observable clusters. This difference can be explained by PCA focusing only on covered structural variance, whereas ECA is

Table 2. The relative occurence (in %) of α-helices and β-sheets at 300 K and 350 K with the three different force fields.

	300 K		350 K	
Force field	α	β	α	β
AMBER03	31	61	27	62
Charmm27	66	30	58	34
OPLS-AA	27	64	24	63

guided by the spectral one. The result therefore indicates that, for reasonable structures, XAS (or XRS) is insensitive to the Ramachandran angles, which in turn cannot be reconstructed from the spectral ROIs.

Feature importance is a metric used for the significance of an input feature with respect to the output of an ML model [60]. We define an importance score for each LMBTR feature as its absolute value in the first ECA vector. Next, we trained a new emulator with the same architecture as the original one (except for the input layer) using a given number of LMBTR features of the highest importance score. We tested the according models using the ECA validation set and find that most LMBTR groups (e.g. a pair-wise distance distribution) contain highly meaningful features in terms of performance which, in turn, indicates strong interplay of the structural features of all kinds. Improvement of covered spectral ROI variance is monotonous with respect to the importance score and strongly saturates already with 300 features (see Supplementary Information). This shows that magnitude of a feature in the ECA vector directly measures its spectral significance.

4. Discussion

Although the different force fields give different structural distributions (including the those of the Ramachandran angles) the spectral effect predicted upon change of temperature is similar for all of them. This complicates the validation of the force fields by core-level excitation spectra. Moreover, the predicted difference profiles are small and cannot be confirmed or rejected by the current experiment, as all predicted changes are within the error bars that also include the possibility of zero effect. However, the experiment supports the validity of the spectrum simulations, together with the respective convergence checks.

Schwartz and co-workers used a liquid jet to measure the nitrogen K-edge x-ray absorption spectrum of triglycine(aq) in ambient temperature [26] by total electron yield. They obtained a spectrum with roughly similar features, but notably different pre peak to main edge ratio. Interestingly, our study is in agreement with x-ray spectroscopy of solid triglycine [15, 16, 20]. The bump at the post edge seen only in our 350K spectrum seems to be present in the other studies [15, 16, 20, 26] and is also present in the N K-edge spectrum of the similar molecule diglycine [15]. The cause of this feature is not explained by our simulations or the measured 300 K spectrum, but based on these references the possibility of some solid triglycine in our 350 K heated sample cannot be excluded. The aforementioned results have been obtained with numerous yield techniques not necessarily equivalent to XRS used here, or the definition of XAS. Although the XRS has a rather low count rate, it is known to be bulk sensitive and extremely stable.

The descriptor used in an ML work sets the 'language' for the resulting scientific discussion. Although several descriptors for the atomistic structure of molecules have been developed [48, 49, 61–65], we see three general requirements for the spectrum analysis by ECA:

1. The descriptor must allow for accurate ML with the available data.
2. The descriptor must allow for ECA decomposition of high spectral variance coverage with only a few dominant components.
3. The descriptor must allow for interpretation in terms of structure; preferably it is translatable into simple structural information.

The architecture of a well-performing descriptor gives insight into the physical system, as it can encode relevant structural information. With LMBTR, the distance-based weighting of angles improved the ML prediction accuracy, which is understandable as the nearby atoms should, by intuition, affect the spectrum more than the distant ones. Unfortunately, the weighting also makes the angular features of the descriptor harder to interpret and (3) is not reached. Thus, effective encoding of physical information for condition (1) may render some of it unrecoverable, and moreover, a trade-off between all three conditions can be expected. Previously, we have

applied a similar method for glassy GeO_2 [57], where we found a variant of a Coulomb matrix [63] (similar to the Bag of Bonds [64]) a suitable descriptor for ML and ECA. Arguably the well-defined covalent bond topology renders LMBTR well-suited for aqueous triglycine.

The ECA algorithm applies projection of the structural descriptor vector onto a limited number of basis vectors and relies on an emulator to predict the spectral information for these new points much faster than the corresponding electron structure calculation would do. This allows for using iterative algorithms for the search of the the optimal basis vectors. The emulator must be complex enough of a function (likely non-linear) yet generalizable to sufficiently capture the relation between any relevant structure **R** and its spectrum **S**. The selection of this mapping $\mathbf{S}_{\mathrm{emu}}(\mathbf{D}(\mathbf{R}))$ is a complicated problem with plenty of tunable hyperparameters, including those of the descriptor $\mathbf{D}(\mathbf{R})$. In this work we settled for nine independent neural networks, one for each atomic site and ROI with a combined output, as they worked best for the LMBTR-encoded triglycine system with points (1)–(3) in mind. As often the case in machine learning, the choice cannot be shown to be the global optimum, but only the best in the particular hyperparameter search. Up to date there are no generally accepted performance criteria in the x-ray spectroscopic community.

The results of the importance score analysis show that almost all groups of distance and angular distributions are necessary to maximise the covered spectral ROI variance. On the other hand, a significant number of features from each group could be ignored, as only 300 of them in total (originally 1140) are necessary to cover nearly all the accessible variance, which itself rises quickly and monotonously with respect to the number of selected features. This shows that the first ECA vector can indeed find the most relevant features of the descriptor in the order of their spectral significance and could serve as a tool for feature selection. Using the ECA, the dependence of ROI intensities, as captured by the descriptor–emulator–ROI mapping $\mathbf{S}_{\mathrm{emu}}(\mathbf{D}(\mathbf{R}))$, can practically be condensed into *three* structural degrees of freedom when a 1140-dimensional LMBTR is used.

The quantitative analysis of our simulations proves that while information about the Ramachandran angles is present in the LMBTR-encoded structural space, it is largely lost in the subspace covering a significant portion of the spectral ROI variance. Therefore, our results support the conclusion of Schwartz and co-workers [26] in that XAS is insensitive to the secondary structure of triglycine (within the reasonably expected structural space). Furthermore, our results are at least partly in contradiction with that of Gordon and co-workers [15] as we see a small dependency between the most relevant part of the backbone conformation and its spectrum (figure 5(c)) at most. This conclusion is manifested by also taking the mean spectra of all α-helices and all β-sheets in the data (see Supplementary Information), which show a deviating difference profile from those presented in figure 2. The advantage of K-edge spectra, locality, seems to be a limitation when it comes to secondary structure of proteins, for which there is an information bottleneck for reasonably expected structures. The ECA shows the ROIs of the total spectrum to mainly be sensitive to the nearest C and N atoms from the solute together with the O and H atoms mostly from the water; and additionally, to the nearest H atoms in the case of N-terminus of the peptide.

5. Conclusions

Experimental N K-edge x-ray Raman scattering spectra of aqueous triglycine can be modelled by classical MD and TP-DFT calculations for a good match. However, the experiment is not able to confirm or reject the predicted temperature difference effect. A machine learning emulator can predict the intensities of spectral regions of interest from the corresponding local many-body tensor representation encoded structures. This enabled the application of the emulator-based component analysis, which was able to condense the structure–spectrum dependency practically into three degrees of freedom. Moreover, the spectral significance of structural features was found to be ordered along the magnitude of the according component in the first obtained basis vector. From the analysis we conclude that the details of the secondary structure of aqueous triglycine, expressed by the Ramachandran angles, are lost in x-ray absorption spectral regions of interest due to an information bottleneck. Instead, in the structural information available for analysis from the used descriptor, the distance distributions between the absorption site and its nearest neighbors significantly account for the spectral region variance. Each of the tried force fields results in a similar temperature-difference profile, and therefore distinguishing between them by x-ray absorption seems improbable. On the other hand, the spectrum can be predicted equally well with all of them.

In this work we have pushed structural decomposition by the ECA algorithm to maximal size scales that x-ray spectra can be hoped to have sensitivity to. The method identifies spectrally relevant structural subspace in a complicated system without prior knowledge. Future prospects of this approach include reconstructing the maximum obtainable structural information from the spectra, the first steps of which have already been taken [57]. We also note that the method is not limited to x-ray spectroscopy, but can be applied to a wide variety of inverse problems for which an emulator for the forward problem is known.

Acknowledgments

E A E acknowledges Jenny and Antti Wihuri Foundation for funding. E A E , A V and J N acknowledge Academy of Finland for funding via project 331 234. The authors acknowledge CSC—IT Center for Science, Finland, and the FGCI—Finnish Grid and Cloud Infrastructure for computational resources. We acknowledge the European Synchrotron Radiation Facility (ESRF) for provision of synchrotron radiation facilities. Prof M Metsä-Ketelä is thanked for sharing his insight in sample environment materials used in biochemical experiments. Dr H Müller at the ESRF is thanked for help regarding chemistry lab and sample preparation during the experiment. Dr Y Watier at the ESRF is thanked for support with the sample environment.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://doi.org/10.5281/zenodo.8239300. Data will be available from 11 September 2023.

Author contributions

E.A.E. Data analysis, machine learning, experiment, simulations and writing manuscript. A.V. Data analysis, experiment and writing manuscript. F.G. experiment. C.J.S Experiment, data analysis and writing manuscript. J. N. Research design, experiment, simulations, writing manuscript and funding.

Conflicts of interest

There are no conflicts to declare.

ORCID iDs

Eemeli A Eronen https://orcid.org/0000-0003-0027-2199
Anton Vladyka https://orcid.org/0000-0003-1770-8139
Christoph J Sahle https://orcid.org/0000-0001-8645-3163
Johannes Niskanen https://orcid.org/0000-0002-5501-3082

References

[1] Liljas A, Liljas L, Piskur J, Lindblom G, Nissen P and Kjeldgaard M 2009 *Textbook of structural biology* (World Scientific Publishing)
[2] Dill K A and MacCallum J L 2012 *Science* **338** 1042–6
[3] Ball P 2017 Water is an active matrix of life for cell and molecular biology *PNAS* **114** 13327–35
[4] Canchi D R and García A E 2013 *Annu. Rev. Phys. Chem.* **64** 273–93
[5] Wernet P *et al* 2004 *Science* **304** 995–9
[6] Ottosson N, Børve K J, SpÅngberg D, Bergersen H, Sæthre L J, Faubel M, Pokapanich W, Öhrwall G, Björneholm O and Winter B 2011 JACS **133** 3120–30
[7] Niskanen J, Sahle C J, Ruotsalainen K O, Müller H, Kavčič M, Žitnik M, Bučar K, Petric M, Hakala M and Huotari S 2016 *Sci. Rep.* **6** 21012
[8] Niskanen J *et al* 2016 *Phys. Chem. Chem. Phys.* **18** 26026–32
[9] Niskanen J, Sahle C J, Gilmore K, Uhlig F, Smiatek J and Föhlisch A 2017 *Phys. Rev.* E **96** 013319
[10] Sahle C, Niskanen J, Gilmore K and Jahn S 2018 *J. Electron. Spectrosc. Relat. Phenom.* **222** 57–62
[11] Niskanen J *et al* 2019 Compatibility of quantitative X-ray spectroscopy with continuous distribution models of water at ambient conditions *PNAS* **116** 4058–63
[12] Vaz da Cruz V *et al* 2019 *Nat. Commun.* **10** 1013
[13] Pietzsch A *et al* 2022 Cuts through the manifold of molecular H2O potential energy surfaces in liquid water at ambient conditions *PNAS* **119** e2118101119
[14] Kaznacheyev K, Osanna A, Jacobsen C, Plashkevych O, Vahtras O, Ågren H, Carravetta V and Hitchcock A P 2002 *The Journal of Physical Chemistry*A **106** 3153–68
[15] Gordon M L, Cooper G, Morin C, Araki T, Turci C C, Kaznatcheev K and Hitchcock A P 2003 *The Journal of Physical Chemistry* A **107** 6144–59
[16] Zubavichus Y, Zharnikov M, Schaporenko A and Grunze M 2004 *J. Electron. Spectrosc. Relat. Phenom.* **134** 25–33
[17] Zubavichus Y, Shaporenko A, Grunze M and Zharnikov M 2005 *The Journal of Physical Chemistry* A **109** 6998–7000
[18] Messer B M, Cappa C D, Smith J D, Drisdell W S, Schwartz C P, Cohen R C and Saykally R J 2005 *The Journal of Physical Chemistry* B **109** 21640–6
[19] Blum M, Odelius M, Weinhardt L, Pookpanratana S, Bär M, Zhang Y, Fuchs O, Yang W, Umbach E and Heske C 2012 *The Journal of Physical Chemistry* B **116** 13757–64
[20] Weinhardt L *et al* 2019 *Phys. Chem. Chem. Phys.* **21** 13207–14

[21] Niskanen J, Arul Murugan N, Rinkevicius Z, Vahtras O, Li C, Monti S, Carravetta V and gren H 2013 *Phys. Chem. Chem. Phys.* **15** 244–54
[22] Hollingsworth S A and Dror R O 2018 *Neuron* **99** 1129–43
[23] Lindorff-Larsen K, Maragakis P, Piana S, Eastwood M P, Dror R O and Shaw D E 2012 *PLoS One* **7** 1–6
[24] Martín-García F, Papaleo E, Gomez-Puertas P, Boomsma W and Lindorff-Larsen K 2015 *PLoS One* **10** 1–16
[25] Niskanen J, Vladyka A, Niemi J and Sahle C 2022 *Royal Society Open Science* **9** 220093
[26] Schwartz C P, Uejio J S, Duffin A M, England A H, Kelly D N, Prendergast D and Saykally R J 2010 Investigation of protein conformation and interactions with salts via X-ray absorption spectroscopy *PNAS* **107** 14008–13
[27] Duan Y *et al* 2003 *J. Comput. Chem.* **24** 1999–2012
[28] MacKerell A D *et al* 1998 *The Journal of Physical Chemistry* B **102** 3586–616
[29] Mackerell A D Jr, Feig M and Brooks C L III 2004 *J. Comput. Chem.* **25** 1400–15
[30] Kaminski G A, Friesner R A, Tirado-Rives J and Jorgensen W L 2001 *The Journal of Physical Chemistry* B **105** 6474–87
[31] Jorgensen W L, Chandrasekhar J, Madura J D, Impey R W and Klein M L 1983 *J. Chem. Phys.* **79** 926–35
[32] Abraham M J, Murtola T, Schulz R, Páll S, Smith J C, Hess B and Lindahl E 2015 *SoftwareX* **1-2** 19–25
[33] Lindahl, Abraham, Hess and van der Spoel 2021 Gromacs 2020.5 manual
[34] Berendsen H J C, Postma J P M, van Gunsteren W F, DiNola A and Haak J R 1984 *J. Chem. Phys.* **81** 3684–90
[35] Nosé S 1984 *Mol. Phys.* **52** 255–68
[36] Hoover W G 1985 *Phys. Rev.* A **31** 1695–7
[37] Andersen H C 1980 *J. Chem. Phys.* **72** 2384–93
[38] Parrinello M and Rahman A 1980 *Phys. Rev. Lett.* **45** 1196–9
[39] Enkovaara J *et al* 2010 *J. Phys. Condens. Matter* **22** 253202
[40] Mortensen J J, Hansen L B and Jacobsen K W 2005 *Phys. Rev.* B **71** 035109
[41] Larsen A H *et al* 2017 *J. Phys. Condens. Matter* **29** 273002
[42] Triguero L, Pettersson L G M and Ågren H 1998 *Phys. Rev.* B **58** 8097–110
[43] Perdew J P, Burke K and Ernzerhof M 1996 *Phys. Rev. Lett.* **77** 3865–8
[44] Leetmaa M, Ljungberg M P, Lyubartsev A, Nilsson A and Pettersson L G 2010 *J. Electron. Spectrosc. Relat. Phenom.* **177** 135–57
[45] Hammer B, Hansen L B and Nørskov J K 1999 *Phys. Rev.* B **59** 7413–21
[46] Niskanen J, Vladyka A, Kettunen J A and Sahle C J 2022 *J. Electron. Spectrosc. Relat. Phenom.* **260** 147243
[47] Pedregosa F *et al* 2011 *Journal of Machine Learning Research* **12** 2825–30 http://jmlr.org/papers/v12/pedregosa11a.html
[48] Huo H and Rupp M 2022 *Machine Learning: Science and Technology* **3** 045017
[49] Himanen L, Jäger M O J, Morooka E V, Federici Canova F, Ranawat Y S, Gao D Z, Rinke P and Foster A S 2020 *Comput. Phys. Commun.* **247** 106949
[50] Virtanen P *et al* 2020 SciPy 10 Contributors *Nat. Methods* **17** 261–72
[51] Byrd R H, Hribar M E and Nocedal J 1999 *SIAM J. Optim.* **9** 877–900
[52] Schülke W 2007 *Electron dynamics by inelastic X-ray scattering* vol 7 (Oxford University Press)
[53] Huotari S *et al* 2017 *J. Synchrotron Radiat.* **24** 521–30
[54] Sahle C, Mirone A, Niskanen J, Inkinen J, Krisch M and Huotari S 2015 *J. Synchrotron Radiat.* **22** 400–9
[55] Mizuno Y and Ohmura Y 1967 *J. Phys. Soc. Jpn.* **22** 445–9
[56] Sahle C, Henriquet C, Schroer M, Juurinen I, Niskanen J and Krisch M 2015 *J. Synchrotron Radiat.* **22** 1555–8
[57] Vladyka A, Sahle C J and Niskanen J 2023 *Phys. Chem. Chem. Phys.* **25** 6707–13
[58] Lovell S C, Davis I W, Arendall W B III, de Bakker P I W, Word J M, Prisant M G, Richardson J S and Richardson D C 2003 *Proteins Struct. Funct. Bioinf.* **50** 437–50
[59] Hintze B J, Lewis S M, Richardson J S and Richardson D C 2016 *Proteins Struct. Funct. Bioinf.* **84** 1177–89
[60] Géron A 2019 *Hands-on machine learning with Scikit-Learn, Keras and TensorFlow: concepts, tools, and techniques to build intelligent systems* II edn (O'Reilly Media, Inc.)
[61] Behler J 2011 *J. Chem. Phys.* **134** 074106
[62] Bartók A P, Kondor R and Csányi G 2013 *Phys. Rev.* B **87** 184115
[63] Rupp M, Tkatchenko A, Müller K R and von Lilienfeld O A 2012 *Phys. Rev. Lett.* **108** 058301
[64] Hansen K, Biegler F, Ramakrishnan R, Pronobis W, von Lilienfeld O A, Müller K R and Tkatchenko A 2015 *J. Phys. Chem. Lett.* **6** 2326–31
[65] Chandrasekaran A, Kamal D, Batra R, Kim C, Chen L and Ramprasad R 2019 *NPJ Comput. Mater.* **5** 22
[66] Humphrey W, Dalke A and Schulten K 1996 *J. Mol. Graphics* **14** 33–8