https://doi.org/10.17952/38EPS.2026.P268

Machine Learning–driven Prediction of Circular Dichroism Spectra for Peptides through Multi-scale Structural Descriptors

Michele Casoria[1,2,4], Marina Macchiagodena[1,3], Paolo Rovero[2,3], Claudia Andreini[1,4], Anna Maria Papini[2,3], Marco Pagliai[1,3]

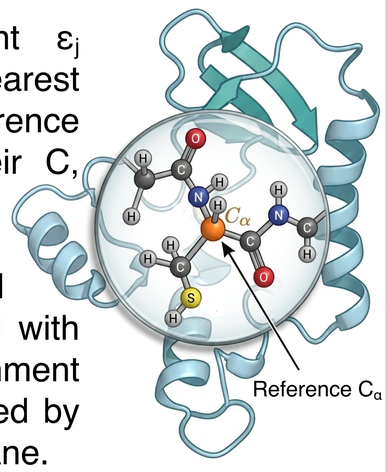
1 Research Unit of Computational Chemistry, University of Florence, 50019 Sesto Fiorentino, Italy
2 Interdepartmental Research Unit of Peptide and Protein Chemistry and Biology, University of Florence, 50019 Sesto Fiorentino, Italy
3 Department of Chemistry “Ugo Schiff”, University of Florence, 50019 Sesto Fiorentino, Italy
4 Magnetic Resonance Center - University of Florence, 50019 Sesto Fiorentino, Italy
e-mail: michele.casoria@unifi.it

Introduction

Circular dichroism (CD) spectroscopy provides rapid characterization of protein secondary structure, but limited scalability and high-purity requirements hinder high-throughput applications. Experimental CD spectra were retrieved from the Protein Circular Dichroism Data Bank (PCDDB), curated by removing duplicates and retaining room-temperature measurements associated with experimentally determined structures. Here, we develop a structure-based machine learning framework using a rotationally invariant Smooth Overlap of Atomic Positions (SOAP) representation and tensor neural network to predict CD spectra from 3D structures.

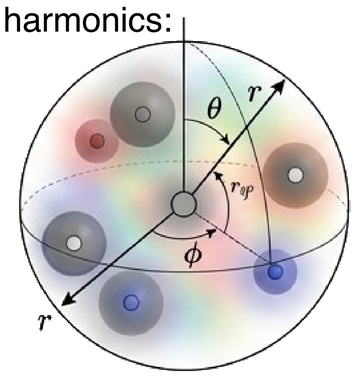
Local Environment in Reference Frame

Each local environment ε_j comprises the six nearest residues around the reference Cα, represented by their C, Cα, N, and O atoms.
A canonical frame is then constructed from the local geometry, with z aligned with the mean environment direction and x determined by PCA in the orthogonal plane.



Atom-Type Resolved Spherical Expansion

The local atomic density is expanded over Gaussian radial functions and complex spherical harmonics:



$$\rho_a(r) \rightarrow C_{nlm}^{(a)}$$

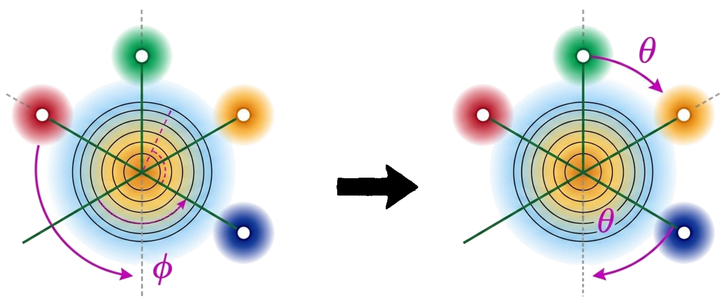
Radial + Angular chemical information

$$C_{anlm}^{(i)} = \sum_{j \in \mathcal{E}_i} \delta_{a,a_j}\, g_n(r_j)\, Y_l^m(\theta_j, \phi_j)$$

Rotationally Invariant Power Spectrum

Rotational invariance is obtained by contracting over m: for four atom types, the 10 unique atom-type pairs are retained.

$$P_{abnn'l}^{(i)} = \sum_{m=-l}^{l} C_{anlm}^{(i)} C_{bn'lm}^{(i)*}$$



Residue-level descriptors are averaged over the protein:

$$\langle P \rangle = \frac{1}{N_{\text{res}}} \sum_{i=1}^{N_{\text{res}}} P^{(i)} \in \mathbb{R}^{10\times13\times13\times9}$$



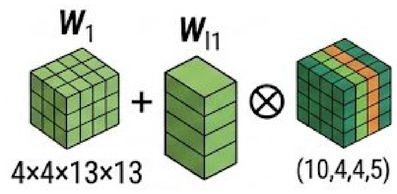
Spectral Target Compression

Experimental CD spectra over the selected wavelength range are compressed using PCA.
The first five principal components are retained as regression targets.

$$\mathbf{y}_{\text{CD}} \in \mathbb{R}^{N_\lambda} \xrightarrow{\text{PCA}} \mathbf{y}_{\text{PCA}} \in \mathbb{R}^5$$

Tensor Neural Network

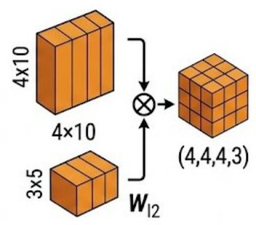
The protein-level descriptor $\langle P \rangle$ retains the structure of atom-type, radial (n, n′), and angular (l) correlations.
Instead of flattening the descriptor, the network performs learnable tensor contractions along the physical dimensions.
The first layer simultaneously contracts the two radial dimensions and mixes the angular components using:



$$W_1 \in \mathbb{R}^{4\times4\times13\times13}$$

$$W_{l1} \in \mathbb{R}^{5\times9}$$

The second layer contracts the atom-type and angular dimensions while preserving the two learned radial dimensionss:

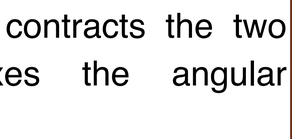


$$W_2 \in \mathbb{R}^{4\times10}$$

$$W_{l2} \in \mathbb{R}^{3\times5}$$

The resulting latent tensor is contracted with a learnable readout tensor to predict the five principal components of the CD spectrum.

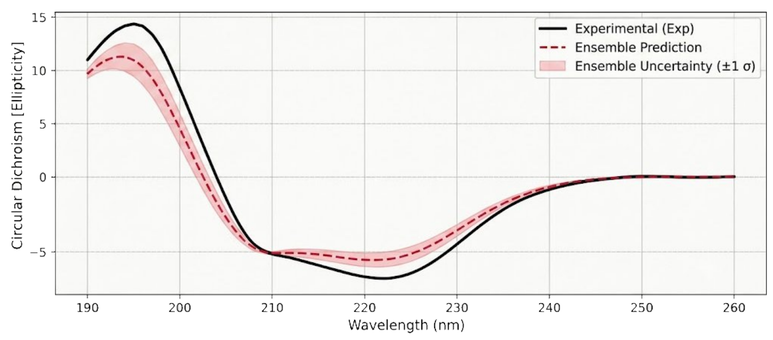
$$W_3 \in \mathbb{R}^{5\times4\times4\times4\times3}$$



The training loss is computed as the mean squared error (MSE) between the five predicted PCA components and the corresponding target components. The architecture preserves the physical organization of the descriptor while learning separate transformations along atom-type, radial, and angular dimensions.

Model Training & Validation

The dataset was randomly partitioned into five folds, using four folds for training and one for validation in each iteration. The network was optimized by minimizing the mean squared error (MSE) between the predicted and target five-dimensional PCA representations.
For validation, the predicted PCA components were transformed back into the original CD spectral space using the inverse PCA transformation. The coefficient of determination R^2 was then computed between the reconstructed and experimental CD spectra.
The model parameters yielding the highest validation R^2 were retained for each fold, resulting in a five-fold ensemble.



Metric	Value
Mean R^2	0.85
Mean RMSD	0.66
Pearson Correlation	0.96
Sign Accuracy	87.8%
Ensamble Uncertainty	0.42
Spectral Diversity	0.85

Conclusions

The proposed tensor-based neural network accurately reproduces experimental CD spectra while preserving most of their spectral diversity, with low ensemble uncertainty, and will be further validated on independent experimental peptide spectra to assess its generalizability and robustness.

Acknowledgments

European Peptide Society is kindly acknowledged for the awarded registration grant to Dr. Michele Casoria.