Electronic Supplementary Material (ESI) for ChemComm.

Electronic Supplementary Information: Structure Prediction from Spectra amidst Dynamical Heterogeneity in Melanin

Arpan Choudhury,[†] Raghunathan Ramakrishnan,[*,‡] and Debashree Ghosh[*,†]

[†]*School of Chemical Sciences, Indian Association for the Cultivation of Science, Kolkata 700032, India*

[‡]*Tata Institute of Fundamental Research Hyderabad, Hyderabad 500046, India*

E-mail: ramakrishnan@tifrh.res.in; pcdg@iacs.res.in

Methods

In this Section, we outline the methodologies used for the ML-guided assignment of the valence excitation spectra of DHICA melanin. The steps involved in this process are depicted in the workflow diagram (Fig 2) in the main paper.

Data generation. Molecular dynamics simulations of DHICA, MKICA, and DKICA tetramers (see Supplementary Fig S1-S3) were carried out with a general AMBER force field (GAFF) as implemented in AMBER16.[1] The equilibrations in a canonical (NVT) ensemble were done with 0.5 fs of integration time step at 300K using the Langevin thermostat. The trajectories were run for a total of 1 ns, with the first 100 ps discarded, and configurations were sampled every 1 ps. Electronic excited state calculations for these configurations were carried out at the TDDFT level of theory using the CAM-B3LYP functional and 6-31G(d) basis set, utilizing the QChem software package.[2] The excitation energies and oscillator strengths for all singlet excitations were calculated within the wavelength range of 200-800 nm. Solvent effects have also been evaluated via continuum solvation models and we notice only small changes in the spectra upon solvation[3] (see supplementary Fig S7).

Data clustering. The K-means algorithm was utilized to cluster the structures based on their three inter-monomer dihedral angles (θ), with a requirement of a standard deviation of $\sim 5^\circ$ within each cluster. Further sub-clustering was performed within each cluster based on the values of 8/7/6 ring-OH dihedral angles (ϕ) in DHICA/MKICA/DKICA melanin. These clusters are the atropisomers and their sub-clusters are the dynamic structures of statistically dominant conformational macrostates which are responsible for the population of each cluster. Finally, we take the mean of the structures and the spectra within each sub-cluster to generate unique ML training data for each cluster.

Machine learning architecture. In this study, we have developed a ML model using kernel-ridge regression (KRR)[4,5] to autonomously assign the DHICA melanin spectra. KRR-based ML approach has been shown to result in asymptotically vanishing errors for predicting molecular ground[6] and excited state[7] energies across the chemical space using structural

descriptors as the input. However, in this study, we have applied KRR for the inverse problem of structure assignment using the valence-excitation spectrum as the input, and the dihedral angles as the output. A novel aspect of our study is that our models are applicable for assigning the individual structure in a mixture of melanin.

For training the ML models, we obtained our data by randomly sampling a mixture of clusters (i.e., atropisomers) for each DHICA, MKICA and DKICA. In experiments, the objective is to identify the composition and structure of melanin that correspond to a particular spectral fingerprint within a limited wavelength range. To support these investigations, the full spectrum (200-800 nm) was divided into small 10 nm windows. Then the ML models were trained for each of these windows, with the intensities of all clusters in the mixture being combined into M bins. In other words, we train ML models for the entire UV-vis spectrum. Irrespective of the number of clusters in a mixture, the input to the ML models is a summed-intensity vector, $\mathbf{p}$, of length M.

The prediction of an angle for an individual configuration in a query mixture is computed by using the kernel-Ansatz

$$\theta^{pred.} = \sum_{i=1}^{N} \alpha_i k(\mathbf{p}_q, \mathbf{p}_i). \tag{1}$$

The equation states that the predicted angle, $\theta^{pred.}$, for a query mixture (q) is the sum of all α_i multiplied by the kernel function, $k(\mathbf{p}_q, \mathbf{p}_i)$, for each element ($i$) in the training set of size N. The kernel function $k(\cdot)$ measures the similarity between the query and training spectra, $\mathbf{p}_q$ and $\mathbf{p}_i$. After evaluating the performance of different kernel functions, the Laplacian kernel was selected that is defined as $k(\mathbf{p}_q, \mathbf{p}_i) = \exp(-|\mathbf{p}_q - \mathbf{p}_i|_1/\sigma)$, where σ is a hyperparameter that determines the width of the kernel and $|\cdot|_1$ denotes the L_1 norm (see Supplementary Fig S8). We trained separate ML models for each dihedral angle by setting σ for all models constant according to the single-kernel formalism suitable for multi-property modeling.[8] This value of σ obtained was found to be similar to the value obtained from a 5-fold cross-validated model (see Supplementary Fig S9). The regression coefficients in Eq.(1) are determined as

the vector, α, for each angle by solving the linear equation

$$[\mathbf{K} + \lambda \mathbf{I}]\boldsymbol{\alpha} = \mathbf{x}, \tag{2}$$

where $\mathbf{x}$ is a vector containing the angle in question for all training instances, and $\mathbf{K}$ is a square matrix that comprises all the pairwise kernel functions between the training instances. Due to the absence of outliers in our dataset, we have set the regularization strength (λ) to a constant value of 10^{-4} that is sufficient to make the kernel matrix positive definite.

DHICA

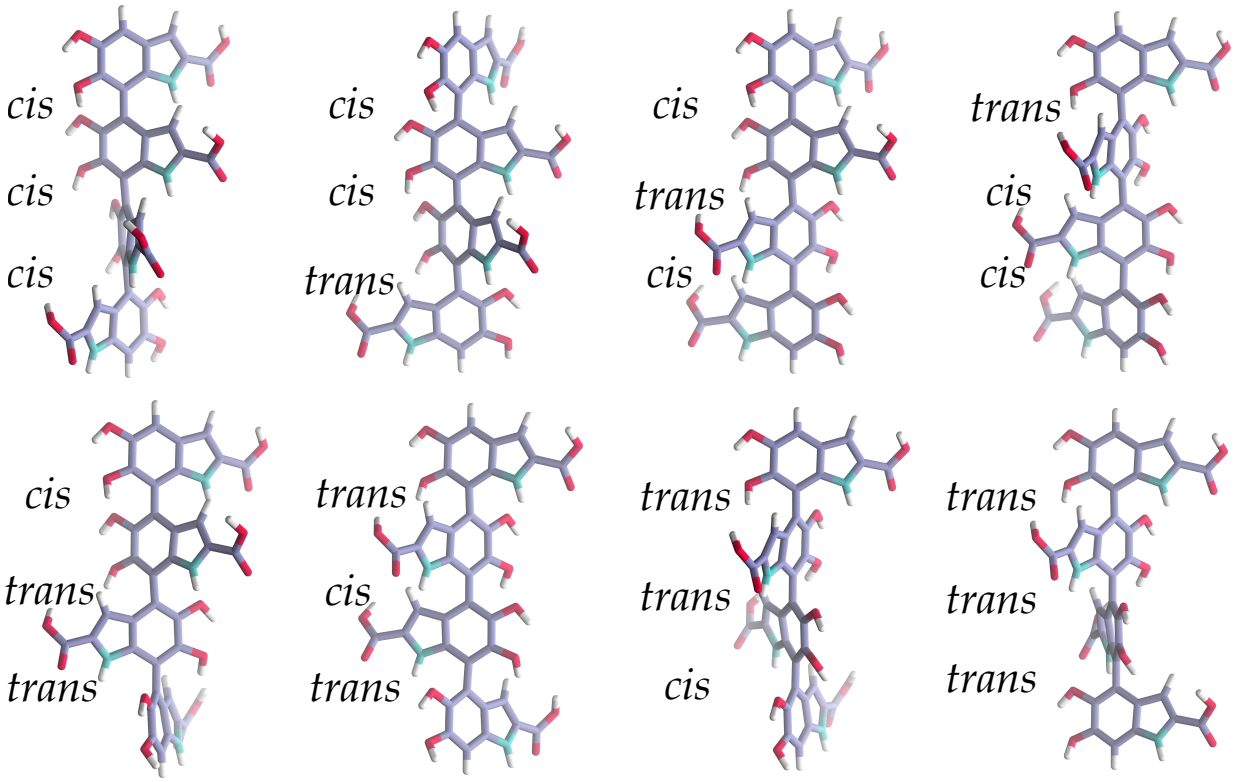


Fig S1. Initial structures of DHICA tetramers which are taken for MD simulations.

MKICA

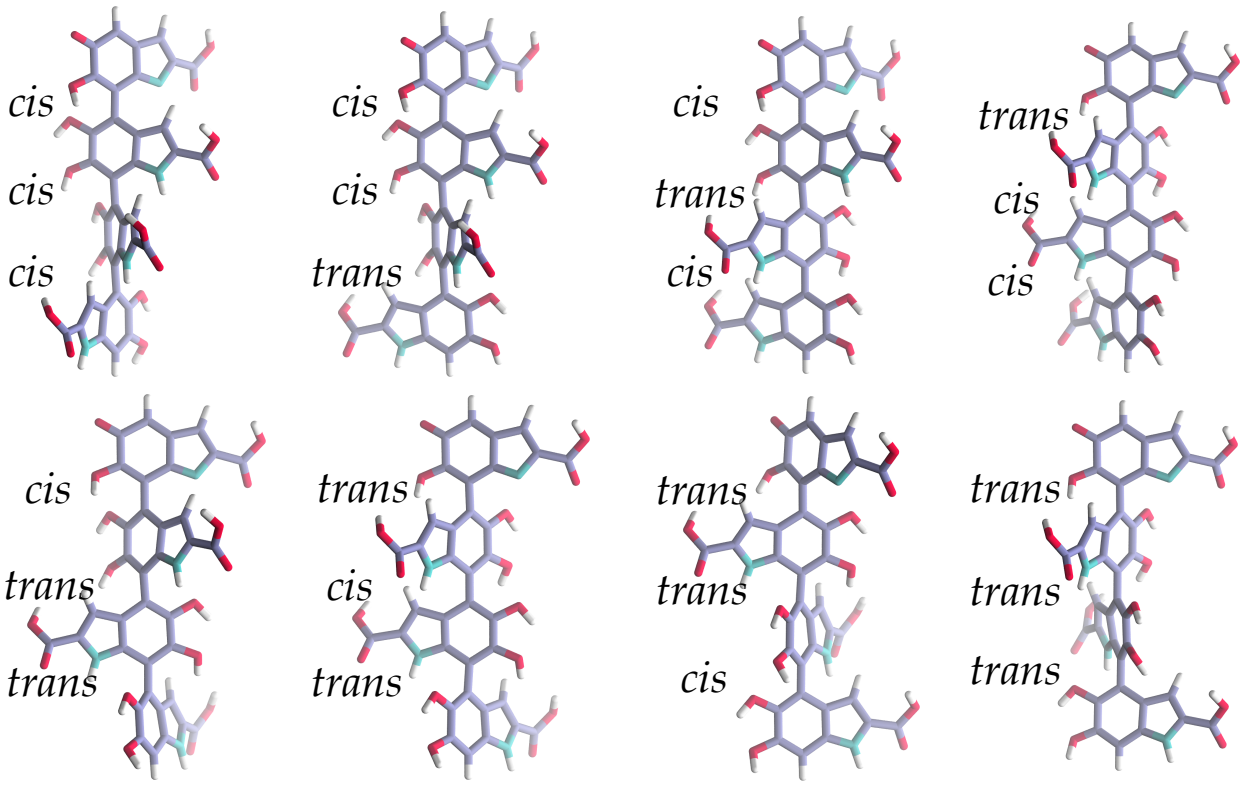


Fig S2. Initial structures of MKICA tetramers which are taken for MD simulations.

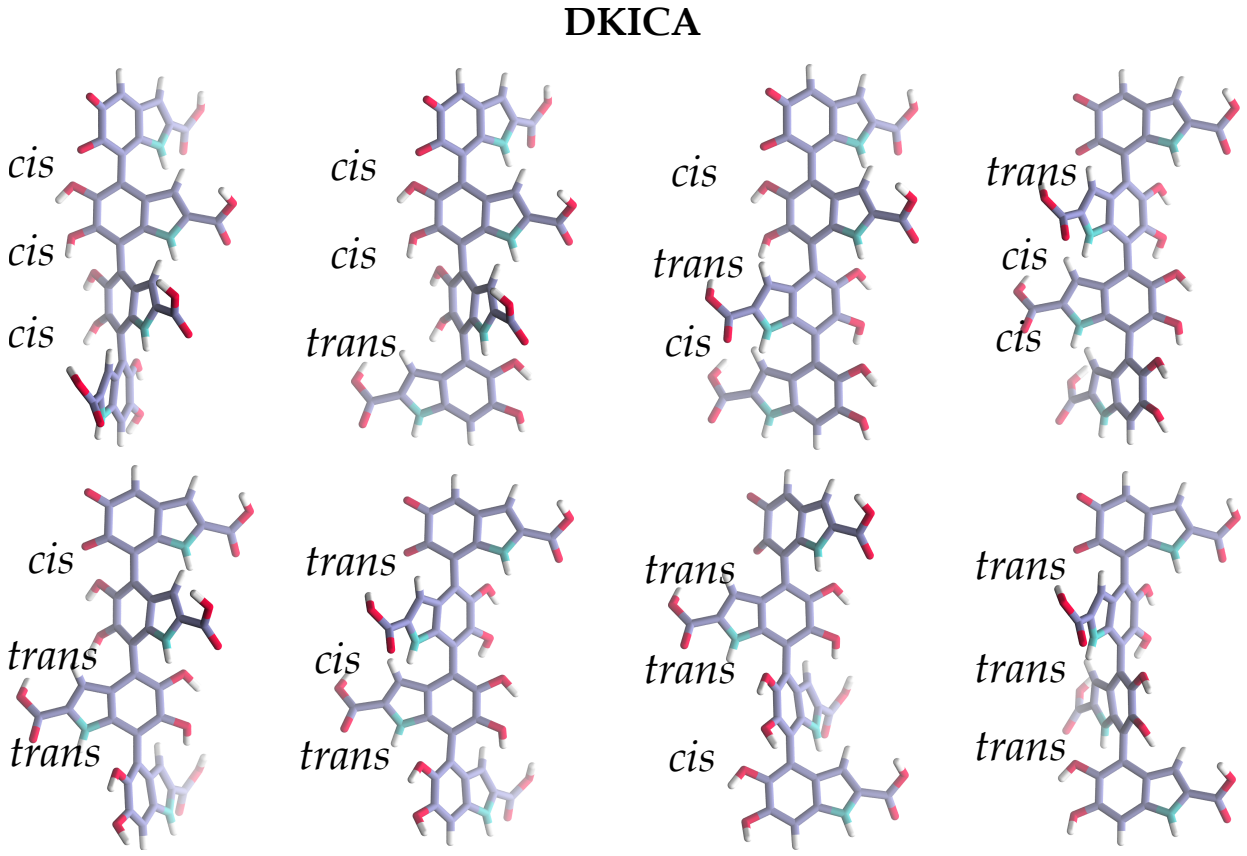
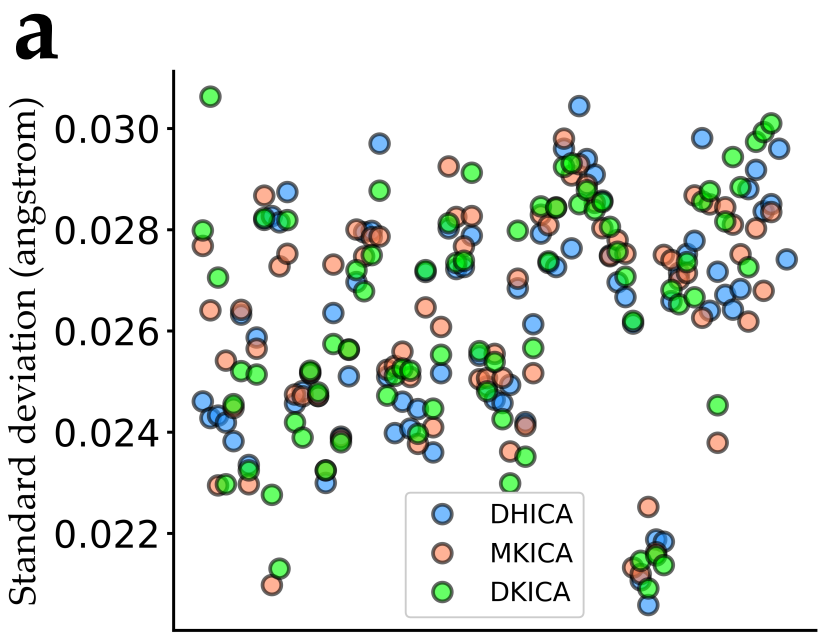


Fig S3. Initial structures of DKICA tetramers which are taken for MD simulations.



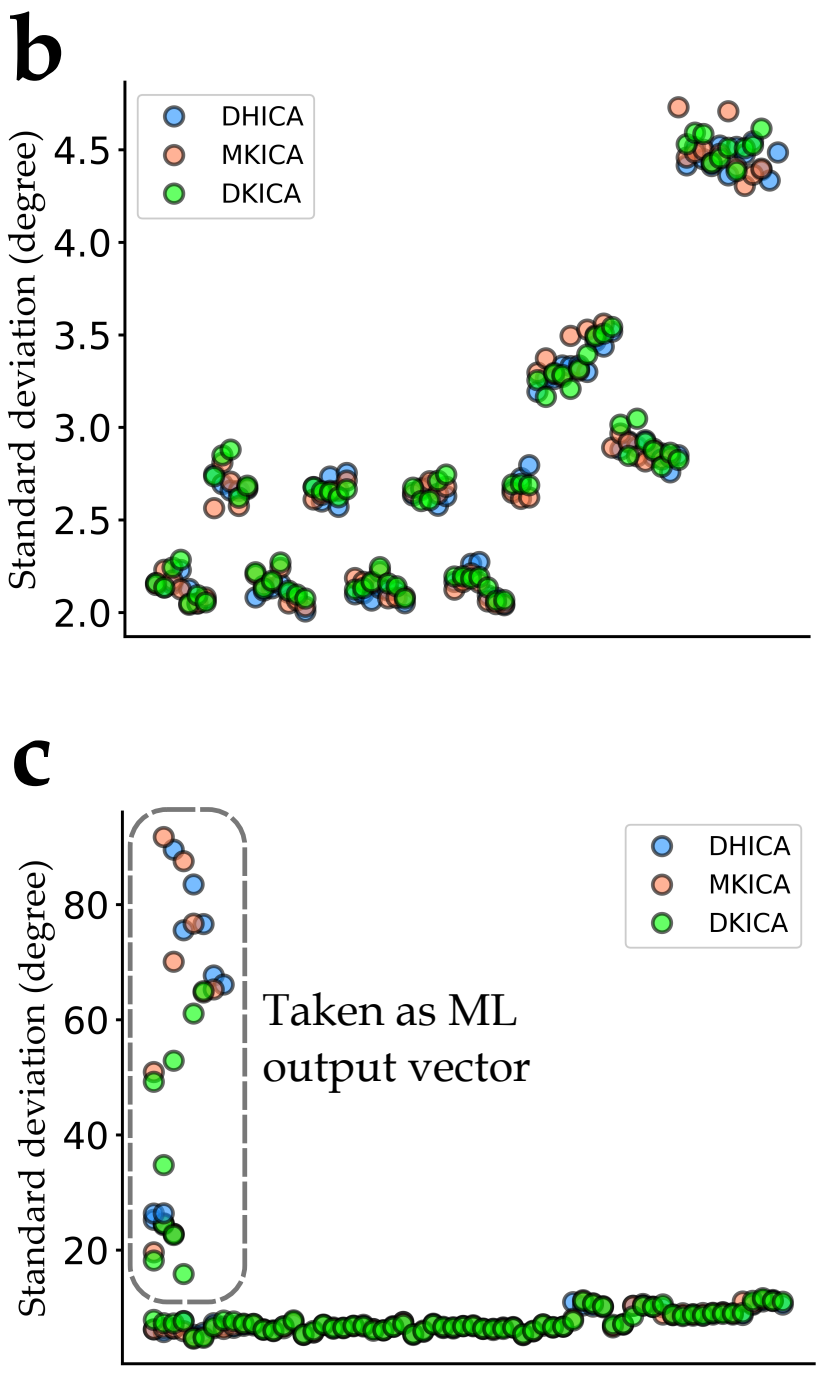


Fig S4. Standard deviation of all the structural degrees of freedoms along the MD trajectory are shown for DHICA, MKICA and DKICA melanin tetramers. The standard deviation of bond distances (**a**) and bond angles (**b**) are very low. But for the dihedral angles (**c**), only a few important parameters (i.e. inter-monomer and ring-OH dihedral angles) show significant standard deviation (marked by the dashed box). These are chosen as ML output vectors. All other degrees of freedom are considered as rigid.

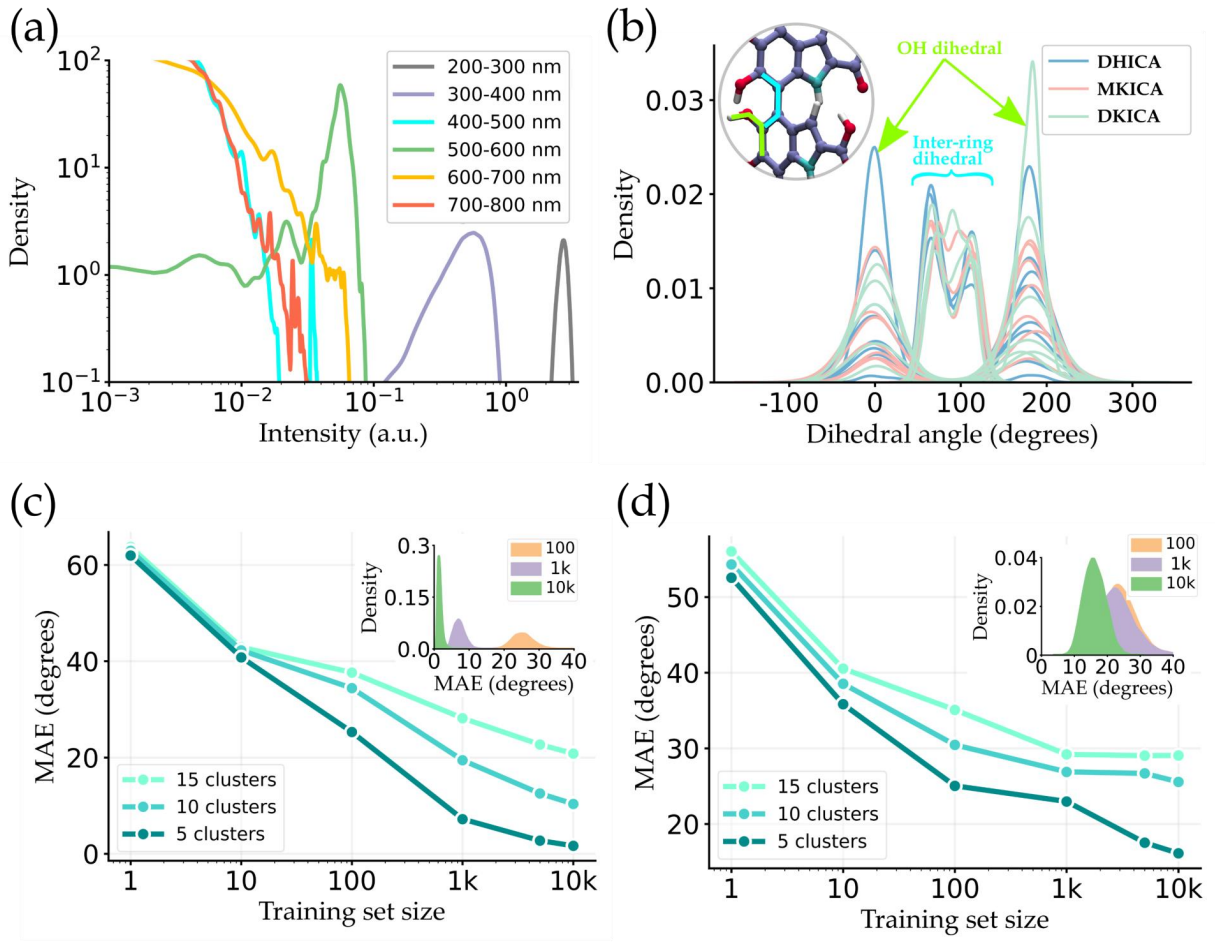


Fig S5. Data distribution and performance of ML models for predicting the structures of different melanin configurations using an ensemble-averaged resultant spectrum as the input. **(a)** Distribution of cumulative spectral intensities for unique configurations of DHICA, MKICA, and DKICA melanin tetramers across different wavelength ranges. **(b)** Distribution of the dihedral angles $\{\theta, \phi\}$ for unique configurations of melanin. Learning curves for predicting structures with 5, 10, and 15 most important clusters for each DHICA, MKICA and DKICA using spectral intensities in the **(c)** 290-300 nm and **(d)** 560-570 nm wavelength range as the input. In **(c)** and **(d)**, the mean absolute error (MAE) of the angles in the ML-reconstructed structures is plotted for various training set sizes. Predictions were made on a 10k holdout set. The inset displays error distribution for various ML models.

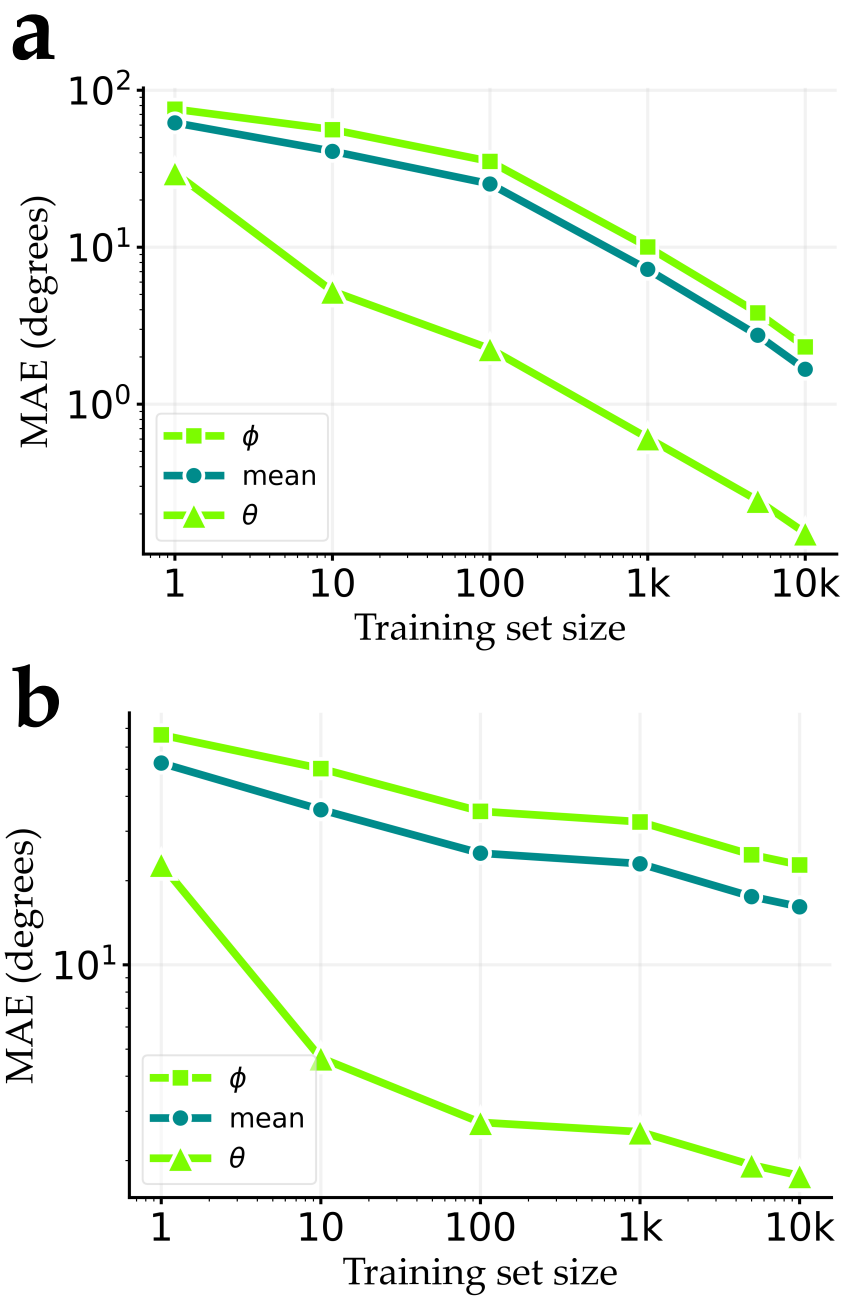
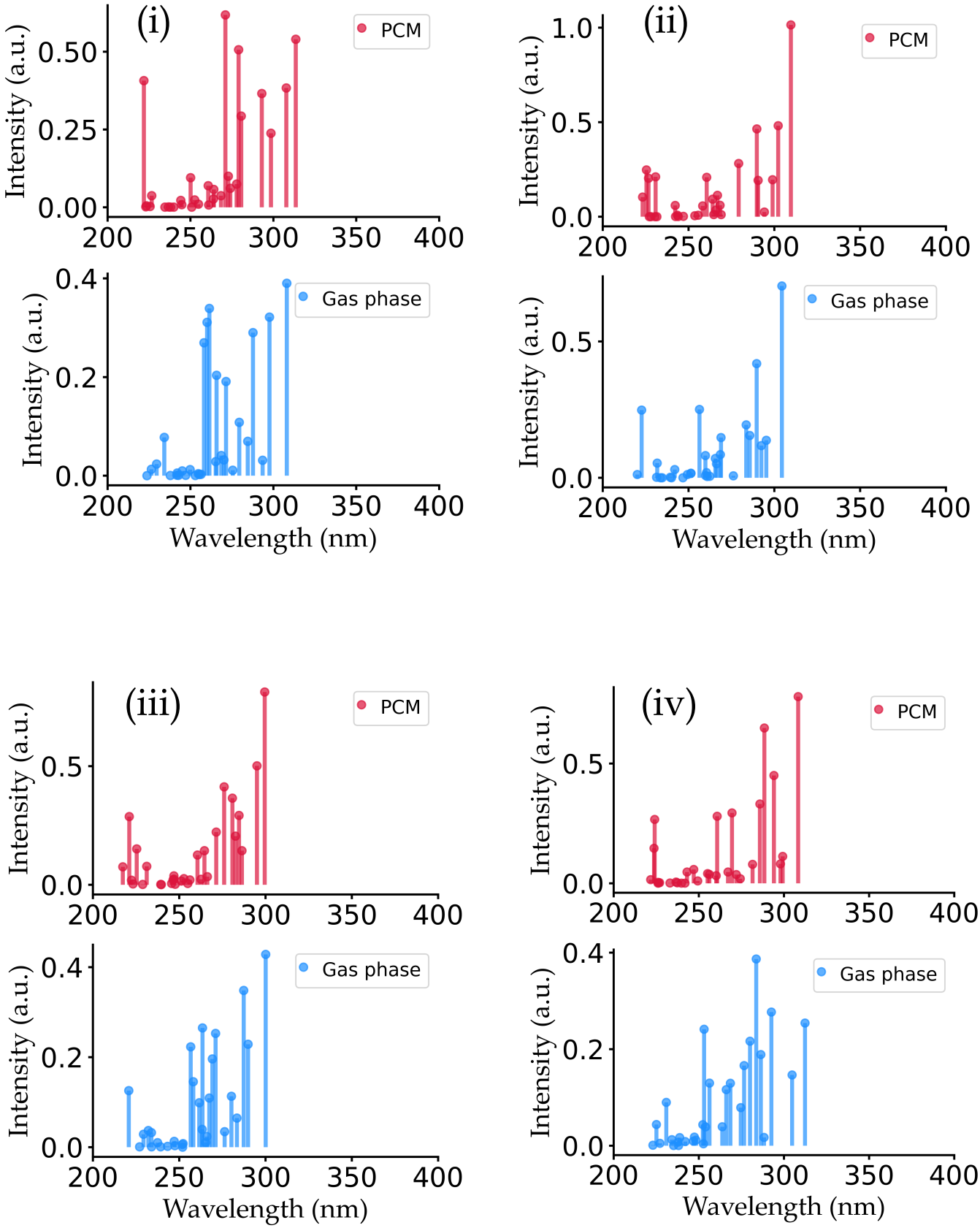
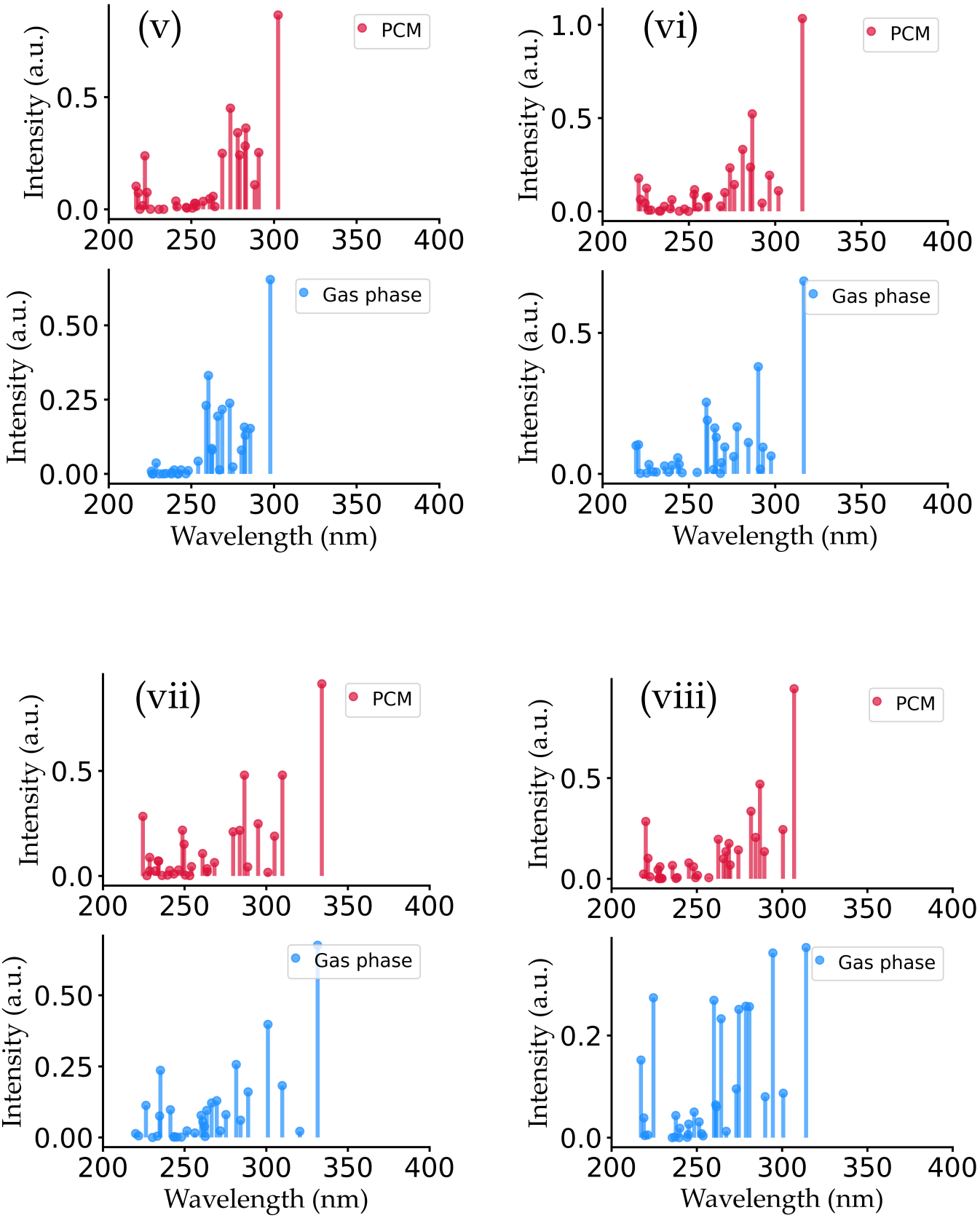


Fig S6. Error breakdown. Separate mean absolute error (MAE) for inter-monomer dihedral angle (θ) and ring-OH dihedral angle (ϕ) in comparison to the 'mean' of them are given for the intensities in the (**a**) 290-300 nm and (**b**) 560-570 nm wavelength range. It shows that the learning curves show much better convergence when predicting only the inter-monomer dihedral angles (i.e., the atropisomers).


(i)
PCM
Intensity (a.u.)
0.00
0.25
0.50
200
250
300
350
400
Gas phase
Intensity (a.u.)
0.0
0.2
0.4
Wavelength (nm)
(ii)
PCM
Intensity (a.u.)
0.0
0.5
1.0
Gas phase
Intensity (a.u.)
0.0
0.5
Wavelength (nm)
(iii)
PCM
Intensity (a.u.)
0.0
0.5
Gas phase
Intensity (a.u.)
0.0
0.2
0.4
Wavelength (nm)
(iv)
PCM
Intensity (a.u.)
0.0
0.5
Gas phase
Intensity (a.u.)
0.0
0.2
0.4
Wavelength (nm)


(v)
PCM
Gas phase
Intensity (a.u.)
Wavelength (nm)
(vi)
PCM
Gas phase
Intensity (a.u.)
Wavelength (nm)
(vii)
PCM
Gas phase
Intensity (a.u.)
Wavelength (nm)
(viii)
PCM
Gas phase
Intensity (a.u.)
Wavelength (nm)

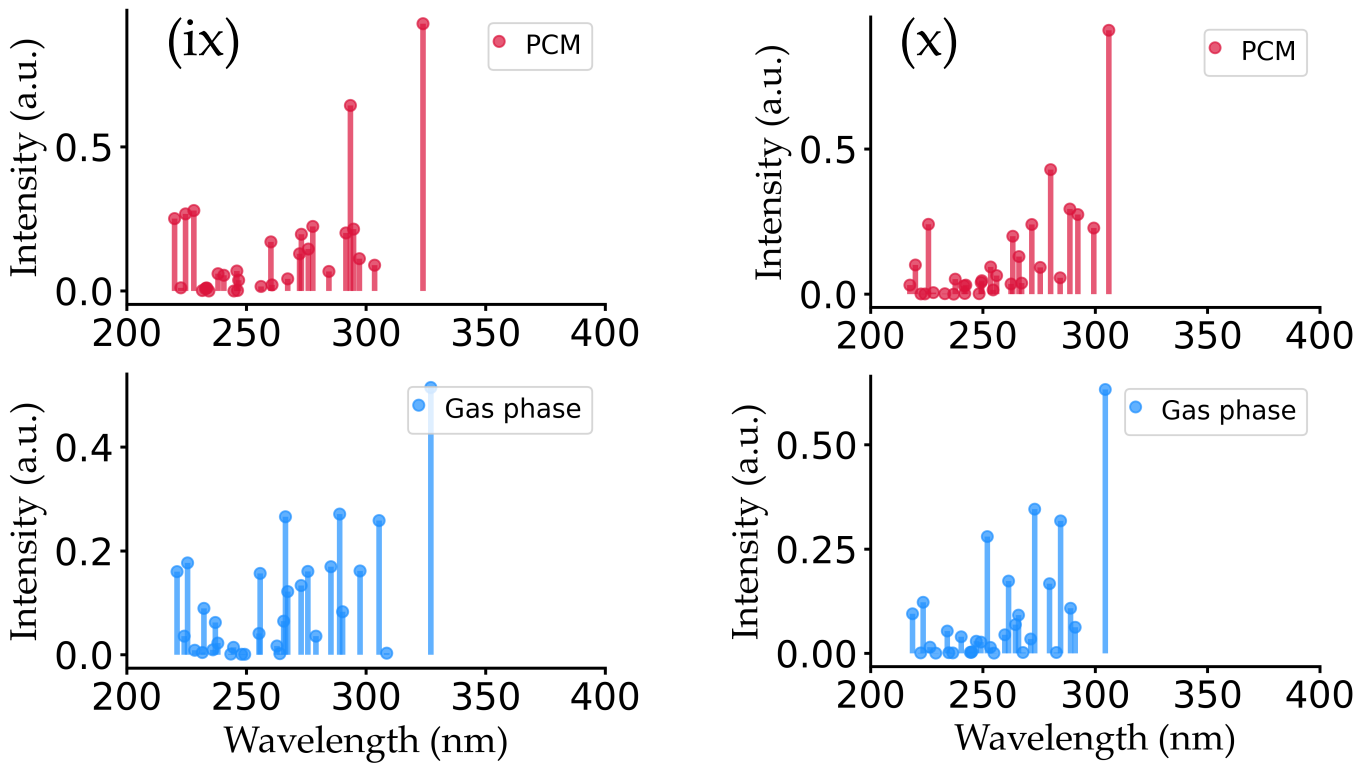


Fig S7. Effect of solvent. Comparison of TDDFT stick spectra in solvent phase (PCM with dielectric constant of water) and gas phase for ten random molecules in the dataset.

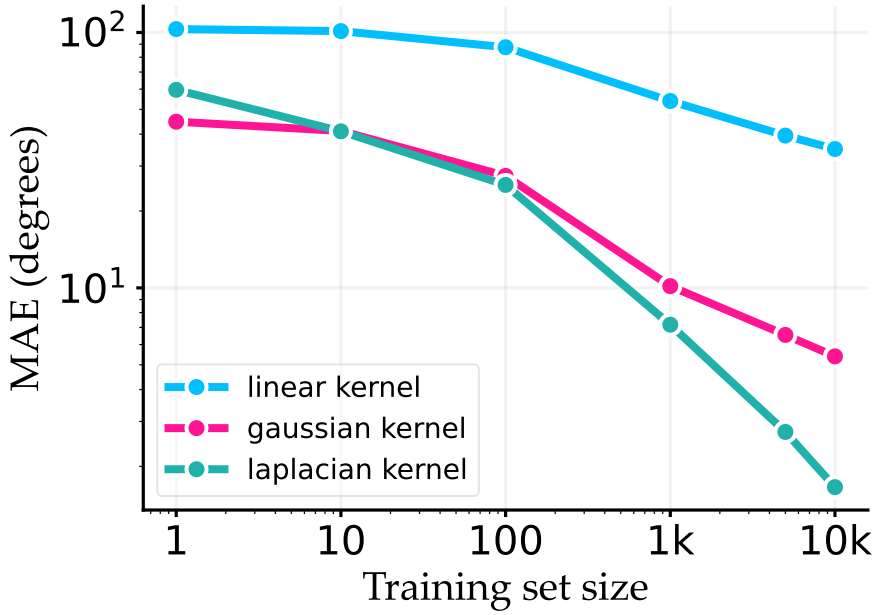


Fig S8. Kernel function benchmarking. The learning curves for linear, gaussian and laplacian kernel functions are shown in the 290-300 nm wavelength region.

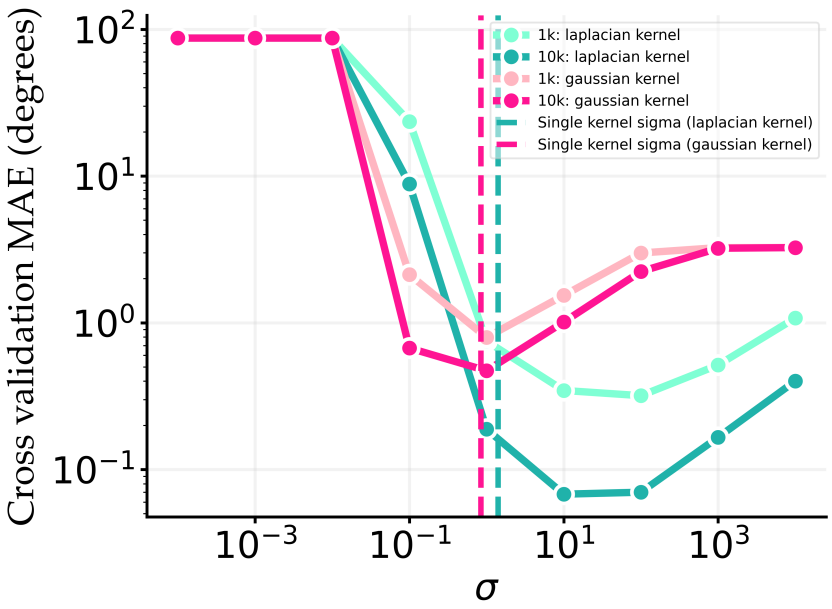


Fig S9. Hyperparameter benchmarking. Kernel width (σ) calculated with single kernel ansatz (vertical dashed lines) are compared with 5-fold cross-validation model (solid lines). Two different kernel functions for two training set sizes are used in the cross-validation. Single kernel σ are calculated with random 500 input representations from the training data.

Section S1. Validation of the ML models on an arbitrary artificial spectrum

To further validate our approach, we evaluated our ML models for an artificial model spectrum with a bimodal distribution (Fig S10). Our goal is to assign absorptions near the peak maxima in the wavelength ranges 290-300 nm and 690-700 nm. We found that for the 690-700 nm range, the only possible species was DKICA tetramers, while for the 290-300 nm range, all three forms were possible. The characteristic structures for the 5 most important clusters in these regions were reconstructed. They are shown in Fig S10 (blue denotes structures that absorb in the range 290-300 nm and green denotes those that absorb in the 690-700 nm range). The percentage contributions of each cluster to the resultant spectrum are also displayed next to the corresponding structure. The predicted inter-monomer dihedral angle values and standard deviation of the clusters are also shown in the figure. The full list of

predicted dihedral angles is shown in the supplementary Table S2 and S3.

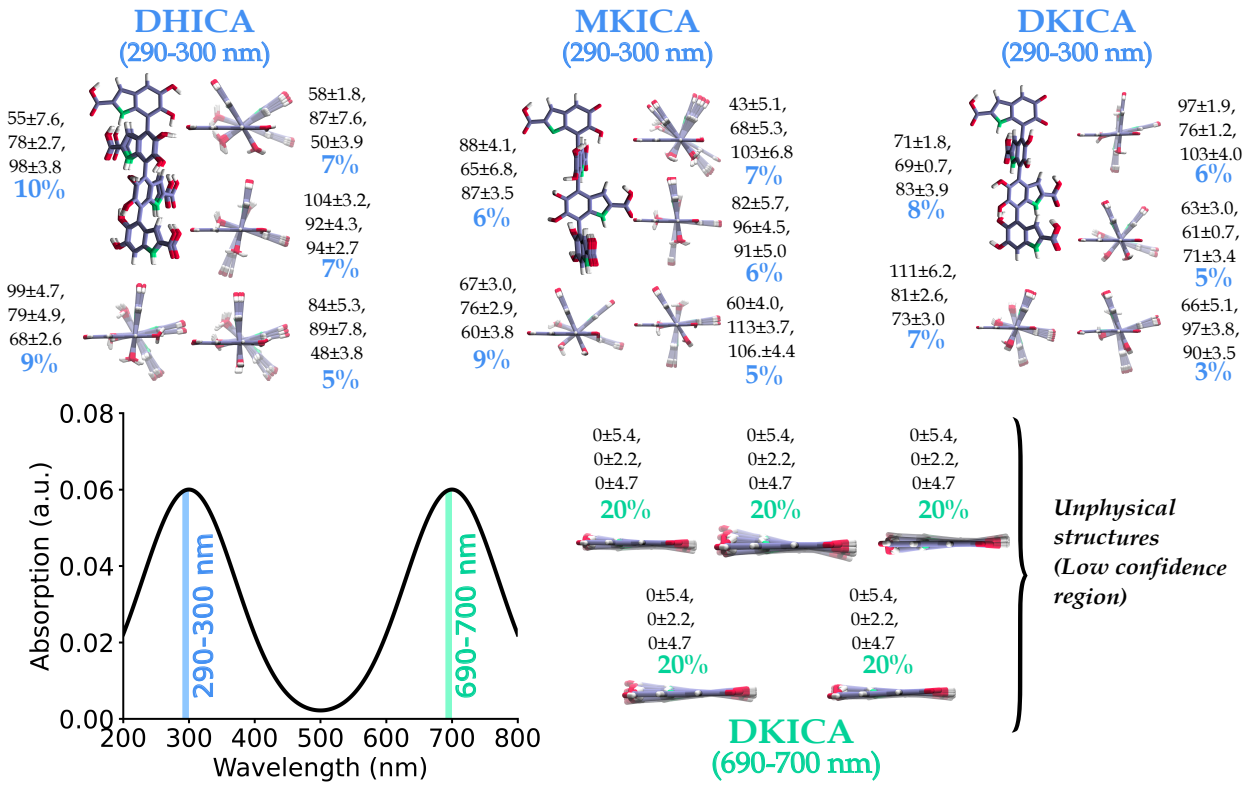


Fig S10. Analysis of the predicted melanin structures for an artificial bimodal spectrum with peak intensities at 300 nm and 700 nm. Models were trained on mixtures of 5 clusters for each DHICA, MKICA and DKICA using inter-monomer $\{\theta\}$ and ring-OH $\{\phi\}$ dihedral angles as the targets. ML predictions were made for the spectral intensities in 290-300 nm and 690-700 nm. Reconstructed structures with their percentage contribution to the input spectrum are shown for each cluster. The individual ML-predicted angles $\{\theta\}$ and their uncertainty determined by the K-means clustering of the training set are also provided.

In the spectral range 690-700nm, the predicted structures display inter-monomer dihedral angles close to 0 degrees, resulting in unphysical planar structures. This is attributed to the low-intensity region of the tetramer spectra in the database (see supplementary Fig S5a). The low intensity in this range is indicative of the absence of low energy (stable or most probable) conformers that can absorb significantly in this wavelength region.

Overall, we have tested our model on a holdout validation set as well as on an artificial spectrum. The prediction on the artificial spectrum has shown expected outcomes of stable conformers in the high confidence region of the spectral window where the melanin absorbs

significantly and null values for the structural parameters in the low confidence region where melanin does not absorb significantly. We have further confirmed the validity of our predictions in the 290-300 nm range by again calculating the TDDFT spectra of the ML predicted structures and their intensities are within the correct wavelength range (supplementary Table S4).

Table S1. Training set bias check. Inter-monomer dihedral angle (θ) predictions on a query spectrum for 5 different randomly chosen training sets of size 10k in the 290-300 nm wavelength range. All 5 training sets predict the same cluster for all the configurations in DHICA, MKICA and DKICA which indicates that there is no training set bias.

Training set 1	**Training set 2**	**Training set 3**	**Training set 4**	**Training set 5**
DHICA	DHICA	DHICA	DHICA	DHICA
$\theta_1 = 54.82$ $\theta_2 = 78.13$ $\theta_3 = 98.51$ (Cluster id = 59)	$\theta_1 = 54.60$ $\theta_2 = 77.92$ $\theta_3 = 98.29$ (Cluster id = 59)	$\theta_1 = 54.41$ $\theta_2 = 77.80$ $\theta_3 = 98.16$ (Cluster id = 59)	$\theta_1 = 54.57$ $\theta_2 = 77.91$ $\theta_3 = 98.26$ (Cluster id = 59)	$\theta_1 = 54.49$ $\theta_2 = 77.79$ $\theta_3 = 98.15$ (Cluster id = 59)
$\theta_1 = 84.42$ $\theta_2 = 88.97$ $\theta_3 = 47.78$ (Cluster id = 50)	$\theta_1 = 84.24$ $\theta_2 = 88.79$ $\theta_3 = 47.66$ (Cluster id = 50)	$\theta_1 = 83.86$ $\theta_2 = 88.56$ $\theta_3 = 47.64$ (Cluster id = 50)	$\theta_1 = 84.34$ $\theta_2 = 88.76$ $\theta_3 = 47.51$ (Cluster id = 50)	$\theta_1 = 83.90$ $\theta_2 = 88.47$ $\theta_3 = 47.65$ (Cluster id = 50)
$\theta_1 = 103.93$ $\theta_2 = 91.81$ $\theta_3 = 93.98$ (Cluster id = 39)	$\theta_1 = 103.66$ $\theta_2 = 91.62$ $\theta_3 = 93.80$ (Cluster id = 39)	$\theta_1 = 103.55$ $\theta_2 = 91.49$ $\theta_3 = 93.65$ (Cluster id = 39)	$\theta_1 = 103.60$ $\theta_2 = 91.62$ $\theta_3 = 93.78$ (Cluster id = 39)	$\theta_1 = 103.54$ $\theta_2 = 91.46$ $\theta_3 = 93.67$ (Cluster id = 39)
$\theta_1 = 99.29$ $\theta_2 = 78.82$ $\theta_3 = 68.07$ (Cluster id = 44)	$\theta_1 = 99.80$ $\theta_2 = 80.13$ $\theta_3 = 68.47$ (Cluster id = 44)	$\theta_1 = 99.01$ $\theta_2 = 78.87$ $\theta_3 = 67.87$ (Cluster id = 44)	$\theta_1 = 99.93$ $\theta_2 = 80.32$ $\theta_3 = 68.43$ (Cluster id = 44)	$\theta_1 = 98.55$ $\theta_2 = 77.96$ $\theta_3 = 67.72$ (Cluster id = 44)
$\theta_1 = 58.34$ $\theta_2 = 87.13$ $\theta_3 = 50.24$ (Cluster id = 25)	$\theta_1 = 58.29$ $\theta_2 = 86.88$ $\theta_3 = 50.00$ (Cluster id = 25)	$\theta_1 = 58.20$ $\theta_2 = 86.72$ $\theta_3 = 50.00$ (Cluster id = 25)	$\theta_1 = 58.22$ $\theta_2 = 86.87$ $\theta_3 = 50.05$ (Cluster id = 25)	$\theta_1 = 58.19$ $\theta_2 = 86.66$ $\theta_3 = 49.88$ (Cluster id = 25)
MKICA	MKICA	MKICA	MKICA	MKICA
$\theta_1 = 88.41$ $\theta_2 = 64.73$ $\theta_3 = 86.86$ (Cluster id = 53)	$\theta_1 = 88.20$ $\theta_2 = 64.61$ $\theta_3 = 86.74$ (Cluster id = 53)	$\theta_1 = 88.08$ $\theta_2 = 64.47$ $\theta_3 = 86.51$ (Cluster id = 53)	$\theta_1 = 88.10$ $\theta_2 = 64.53$ $\theta_3 = 86.61$ (Cluster id = 53)	$\theta_1 = 87.98$ $\theta_2 = 64.47$ $\theta_3 = 86.51$ (Cluster id = 53)
$\theta_1 = 81.84$ $\theta_2 = 95.95$ $\theta_3 = 90.57$ (Cluster id = 54)	$\theta_1 = 81.60$ $\theta_2 = 95.68$ $\theta_3 = 90.35$ (Cluster id = 54)	$\theta_1 = 81.53$ $\theta_2 = 95.58$ $\theta_3 = 90.24$ (Cluster id = 54)	$\theta_1 = 81.54$ $\theta_2 = 95.56$ $\theta_3 = 90.40$ (Cluster id = 54)	$\theta_1 = 81.46$ $\theta_2 = 95.49$ $\theta_3 = 90.22$ (Cluster id = 54)

$\theta_1 = 67.16$ $\theta_2 = 75.89$ $\theta_3 = 60.08$ (Cluster id = 24) $\theta_1 = 42.89$ $\theta_2 = 68.29$ $\theta_3 = 102.84$ (Cluster id = 21) $\theta_1 = 60.29$ $\theta_2 = 113.46$ $\theta_3 = 105.70$ (Cluster id = 33)	$\theta_1 = 67.08$ $\theta_2 = 75.81$ $\theta_3 =$ (Cluster id = 24) $\theta_1 = 42.51$ $\theta_2 = 68.11$ $\theta_3 = 102.53$ (Cluster id = 21) $\theta_1 = 60.21$ $\theta_2 = 113.16$ $\theta_3 = 105.43$ (Cluster id = 33)	$\theta_1 = 66.96$ $\theta_2 = 75.69$ $59.96\theta_3 = 59.90$ (Cluster id = 24) $\theta_1 = 42.79$ $\theta_2 = 68.01$ $\theta_3 = 102.47$ (Cluster id = 21) $\theta_1 = 60.08$ $\theta_2 = 113.02$ $\theta_3 = 105.21$ (Cluster id = 33)	$\theta_1 = 67.05$ $\theta_2 = 75.78$ $\theta_3 = 59.88$ (Cluster id = 24) $\theta_1 = 42.53$ $\theta_2 = 68.11$ $\theta_3 = 102.47$ (Cluster id = 21) $\theta_1 = 60.13$ $\theta_2 = 113.07$ $\theta_3 = 105.49$ (Cluster id = 33)	$\theta_1 = 66.91$ $\theta_2 = 75.59$ $\theta_3 = 59.85$ (Cluster id = 24) $\theta_1 = 42.24$ $\theta_2 = 67.97$ $\theta_3 = 102.30$ (Cluster id = 21) $\theta_1 = 60.10$ $\theta_2 = 112.87$ $\theta_3 = 105.26$ (Cluster id = 33)
DKICA	DKICA	DKICA	DKICA	DKICA
$\theta_1 = 70.96$ $\theta_2 = 68.69$ $\theta_3 = 83.30$ (Cluster id = 20) $\theta_1 = 63.52$ $\theta_2 = 61.00$ $\theta_3 = 71.25$ (Cluster id = 46) $\theta_1 = 96.72$ $\theta_2 = 75.72$ $\theta_3 = 102.84$ (Cluster id = 27) $\theta_1 = 65.92$ $\theta_2 = 97.51$ $\theta_3 = 90.18$ (Cluster id = 23) $\theta_1 = 111.07$ $\theta_2 = 81.12$ $\theta_3 = 72.59$ (Cluster id = 2)	$\theta_1 = 70.79$ $\theta_2 = 68.53$ $\theta_3 = 83.09$ (Cluster id = 20) $\theta_1 = 63.37$ $\theta_2 = 60.85$ $\theta_3 = 71.08$ (Cluster id = 46) $\theta_1 = 96.46$ $\theta_2 = 75.53$ $\theta_3 = 102.62$ (Cluster id = 27) $\theta_1 = 65.66$ $\theta_2 = 97.31$ $\theta_3 = 89.99$ (Cluster id = 23) $\theta_1 = 111.31$ $\theta_2 = 78.19$ $\theta_3 = 74.80$ (Cluster id = 2)	$\theta_1 = 70.70$ $\theta_2 = 68.45$ $\theta_3 = 83.00$ (Cluster id = 20) $\theta_1 = 63.27$ $\theta_2 = 60.77$ $\theta_3 = 70.99$ (Cluster id = 46) $\theta_1 = 96.36$ $\theta_2 = 75.43$ $\theta_3 = 102.45$ (Cluster id = 27) $\theta_1 = 65.80$ $\theta_2 = 97.16$ $\theta_3 = 89.78$ (Cluster id = 23) $\theta_1 = 110.87$ $\theta_2 = 80.05$ $\theta_3 = 73.05$ (Cluster id = 2)	$\theta_1 = 70.77$ $\theta_2 = 68.54$ $\theta_3 = 83.06$ (Cluster id = 20) $\theta_1 = 63.35$ $\theta_2 = 60.84$ $\theta_3 = 71.05$ (Cluster id = 46) $\theta_1 = 96.40$ $\theta_2 = 75.50$ $\theta_3 = 102.63$ (Cluster id = 27) $\theta_1 = 65.64$ $\theta_2 = 97.22$ $\theta_3 = 89.93$ (Cluster id = 23) $\theta_1 = 111.17$ $\theta_2 = 79.01$ $\theta_3 = 74.08$ (Cluster id = 2)	$\theta_1 = 70.66$ $\theta_2 = 68.42$ $\theta_3 = 82.95$ (Cluster id = 20) $\theta_1 = 63.26$ $\theta_2 = 60.75$ $\theta_3 = 70.94$ (Cluster id = 46) $\theta_1 = 96.34$ $\theta_2 = 75.38$ $\theta_3 = 102.45$ (Cluster id = 27) $\theta_1 = 65.87$ $\theta_2 = 97.10$ $\theta_3 = 89.69$ (Cluster id = 23) $\theta_1 = 110.86$ $\theta_2 = 79.81$ $\theta_3 = 73.18$ (Cluster id = 2)

Table S2. All predicted inter-monomer $\{\theta\}$ and ring-OH $\{\phi\}$ dihedral angles for the artificial spectrum in the 290-300 nm wavelength range.

Predicted dihedral angles (in deg.) of DHICA				
Configuration 1	Configuration 2	Configuration 3	Configuration 4	Configuration 5
θ_1 = 54.82	θ_1 = 84.42	θ_1 = 103.93	θ_1 = 99.29	θ_1 = 58.34
θ_2 = 78.13	θ_2 = 88.97	θ_2 = 91.81	θ_2 = 78.82	θ_2 = 87.13
θ_3 = 98.51	θ_3 = 47.78	θ_3 = 93.98	θ_3 = 68.07	θ_3 = 50.24
ϕ_1 = -4.82	ϕ_1 = 25.38	ϕ_1 = -7.72	ϕ_1 = -313.63	ϕ_1 = 8.00
ϕ_2 = 160.28	ϕ_2 = 176.95	ϕ_2 = 164.21	ϕ_2 = 495.45	ϕ_2 = 168.38
ϕ_3 = 71.89	ϕ_3 = 183.97	ϕ_3 = 93.17	ϕ_3 = 72.95	ϕ_3 = 28.16
ϕ_4 = 148.06	ϕ_4 = 13.70	ϕ_4 = 86.94	ϕ_4 = 87.61	ϕ_4 = 139.05
ϕ_5 = 70.45	ϕ_5 = -85.11	ϕ_5 = 121.93	ϕ_5 = -173.72	ϕ_5 = 72.42
ϕ_6 = 152.26	ϕ_6 = 216.35	ϕ_6 = 57.90	ϕ_6 = 278.49	ϕ_6 = 106.17
ϕ_7 = 159.84	ϕ_7 = 102.77	ϕ_7 = 139.41	ϕ_7 = 150.34	ϕ_7 = 76.45
ϕ_8 = 6.65	ϕ_8 = 60.23	ϕ_8 = 20.92	ϕ_8 = -12.53	ϕ_8 = 77.98
Predicted dihedral angles (in deg.) of MKICA				
Configuration 1	Configuration 2	Configuration 3	Configuration 4	Configuration 5
θ_1 = 88.41	θ_1 = 81.84	θ_1 = 67.16	θ_1 = 42.89	θ_1 = 60.29
θ_2 = 64.73	θ_2 = 95.95	θ_2 = 75.89	θ_2 = 68.29	θ_2 = 113.46
θ_3 = 86.86	θ_3 = 90.57	θ_3 = 60.08	θ_3 = 102.84	θ_3 = 105.70
ϕ_1 = 155.50	ϕ_1 = 169.33	ϕ_1 = 134.17	ϕ_1 = 201.59	ϕ_1 = 160.55
ϕ_2 = 121.67	ϕ_2 = 65.23	ϕ_2 = -166.22	ϕ_2 = 500.60	ϕ_2 = 6.09
ϕ_3 = 84.68	ϕ_3 = 97.48	ϕ_3 = 177.83	ϕ_3 = -321.00	ϕ_3 = 164.56
ϕ_4 = 99.12	ϕ_4 = 81.89	ϕ_4 = 170.79	ϕ_4 = 170.64	ϕ_4 = 113.69
ϕ_5 = 75.48	ϕ_5 = 90.49	ϕ_5 = 265.65	ϕ_5 = 521.45	ϕ_5 = 153.18
ϕ_6 = 160.81	ϕ_6 = 162.98	ϕ_6 = -151.34	ϕ_6 = 179.04	ϕ_6 = 143.46
ϕ_7 = -4.80	ϕ_7 = 3.02	ϕ_7 = 321.42	ϕ_7 = -42.59	ϕ_7 = 45.24
Predicted dihedral angles (in deg.) of DKICA				
Configuration 1	Configuration 2	Configuration 3	Configuration 4	Configuration 5
θ_1 = 70.96	θ_1 = 63.52	θ_1 = 96.72	θ_1 = 65.92	θ_1 = 111.07
θ_2 = 68.69	θ_2 = 61.00	θ_2 = 75.72	θ_2 = 97.51	θ_2 = 81.12
θ_3 = 83.30	θ_3 = 71.25	θ_3 = 102.84	θ_3 = 90.18	θ_3 = 72.59
ϕ_1 = 164.09	ϕ_1 = 165.58	ϕ_1 = 127.68	ϕ_1 = 9.51	ϕ_1 = -5.71
ϕ_2 = 2.86	ϕ_2 = 153.71	ϕ_2 = 155.68	ϕ_2 = 198.27	ϕ_2 = 132.62
ϕ_3 = -26.20	ϕ_3 = -4.59	ϕ_3 = 33.28	ϕ_3 = 259.94	ϕ_3 = -22.82
ϕ_4 = 165.45	ϕ_4 = 165.62	ϕ_4 = 160.24	ϕ_4 = 197.30	ϕ_4 = 142.56
ϕ_5 = 222.96	ϕ_5 = 147.02	ϕ_5 = 149.21	ϕ_5 = 225.04	ϕ_5 = -12.37
ϕ_6 = -29.54	ϕ_6 = 9.75	ϕ_6 = 4.15	ϕ_6 = -40.28	ϕ_6 = 194.45

Table S3. All predicted inter-monomer $\{\theta\}$ and ring-OH $\{\phi\}$ dihedral angles for the artificial spectrum in the 690-700 nm wavelength range.

Predicted dihedral angles (in deg.) of DKICA				
Configuration 1	Configuration 2	Configuration 3	Configuration 4	Configuration 5
$\theta_1 = 0.027$	$\theta_1 = 0.033$	$\theta_1 = 0.041$	$\theta_1 = 0.039$	$\theta_1 = 0.039$
$\theta_2 = 0.028$	$\theta_2 = 0.029$	$\theta_2 = 0.042$	$\theta_2 = 0.041$	$\theta_2 = 0.026$
$\theta_3 = 0.023$	$\theta_3 = 0.031$	$\theta_3 = 0.040$	$\theta_3 = 0.029$	$\theta_3 = 0.040$
$\phi_1 = 0.022$	$\phi_1 = 0.002$	$\phi_1 = 0.028$	$\phi_1 =$ -0.002	$\phi_1 = 0.002$
$\phi_2 = 0.060$	$\phi_2 = 0.063$	$\phi_2 = 0.066$	$\phi_2 = 0.064$	$\phi_2 = 0.063$
$\phi_3 = 0.033$	$\phi_3 = 0.028$	$\phi_3 = 0.039$	$\phi_3 = 0.002$	$\phi_3 = 0.018$
$\phi_4 = 0.065$	$\phi_4 = 0.065$	$\phi_4 = 0.064$	$\phi_4 = 0.055$	$\phi_4 = 0.068$
$\phi_5 = 0.044$	$\phi_5 = 0.031$	$\phi_5 = 0.024$	$\phi_5 = 0.039$	$\phi_5 = 0.036$
$\phi_6 = 0.031$	$\phi_6 = 0.027$	$\phi_6 = 0.034$	$\phi_6 = 0.062$	$\phi_6 = 0.033$

Table S4. TDDFT calculations of the ML predicted structures in the 290-300 nm wavelength range of the artificial spectrum.

DHICA		**MKICA**		**DKICA**	
Wavelength (nm)	Intensity (a.u.)	Wavelength (nm)	Intensity (a.u.)	Wavelength (nm)	Intensity (a.u.)
296	0.100	290	0.116	291	0.674
285	0.081	285	0.599		
295	0.832	290	0.081	301.27	0.591
				301.16	0.088
				295	0.163
295	0.236	302	0.489	288	0.567
288	0.057	291	0.070		
		288	0.170		
296	0.620	304	0.487	286.61	0.388
288	0.027	293	0.137	286.28	0.053
		287	0.188		
303	0.649	304	0.553	291	0.664
293	0.111	292	0.190		
		286	0.119		

Table S5. All predicted inter-monomer $\{\theta\}$ and ring-OH $\{\phi\}$ dihedral angles for the experimental spectrum in the 280-290 nm wavelength range.

Predicted dihedral angles (in deg.) of DHICA				
Configuration 1	Configuration 2	Configuration 3	Configuration 4	Configuration 5
θ_1 = 84.49	θ_1 = 56.67	θ_1 = 102.37	θ_1 = 110.14	θ_1 = 84.17
θ_2 = 91.44	θ_2 = 69.75	θ_2 = 91.00	θ_2 = 89.41	θ_2 = 60.40
θ_3 = 109.04	θ_3 = 64.55	θ_3 = 74.40	θ_3 = 111.56	θ_3 = 69.77
ϕ_1 = -2.40	ϕ_1 = -6.47	ϕ_1 = 88.55	ϕ_1 = 6.72	ϕ_1 = -5.35
ϕ_2 = 184.38	ϕ_2 = 170.31	ϕ_2 = 91.04	ϕ_2 = 167.35	ϕ_2 = 180.38
ϕ_3 = 123.60	ϕ_3 = 125.28	ϕ_3 = 26.81	ϕ_3 = 158.86	ϕ_3 = 26.74
ϕ_4 = 50.46	ϕ_4 = 86.77	ϕ_4 = 192.58	ϕ_4 = 35.76	ϕ_4 = 45.07
ϕ_5 = 7.46	ϕ_5 = 83.32	ϕ_5 = 87.25	ϕ_5 = 144.33	ϕ_5 = -0.37
ϕ_6 = 153.55	ϕ_6 = 91.66	ϕ_6 = -39.57	ϕ_6 = 33.92	ϕ_6 = 219.89
ϕ_7 = 159.93	ϕ_7 = 167.42	ϕ_7 = 265.89	ϕ_7 = 157.19	ϕ_7 = -33.53
ϕ_8 = -8.41	ϕ_8 = 6.46	ϕ_8 = -70.13	ϕ_8 = 4.25	ϕ_8 = 141.50
Predicted dihedral angles (in deg.) of MKICA				
Configuration 1	Configuration 2	Configuration 3	Configuration 4	Configuration 5
θ_1 = 73.71	θ_1 = 100.15	θ_1 = 85.95	θ_1 = 95.32	θ_1 = 77.62
θ_2 = 87.03	θ_2 = 85.36	θ_2 = 113.50	θ_2 = 62.56	θ_2 = 99.24
θ_3 = 89.01	θ_3 = 92.38	θ_3 = 77.84	θ_3 = 107.96	θ_3 = 87.00
ϕ_1 = 135.83	ϕ_1 = 172.82	ϕ_1 = 100.76	ϕ_1 = -45.46	ϕ_1 = 62.24
ϕ_2 = 111.78	ϕ_2 = 182.22	ϕ_2 = 2.41	ϕ_2 = 90.21	ϕ_2 = 35.96
ϕ_3 = 203.20	ϕ_3 = 91.24	ϕ_3 = 173.97	ϕ_3 = 192.35	ϕ_3 = 137.99
ϕ_4 = 108.26	ϕ_4 = 30.34	ϕ_4 = -82.39	ϕ_4 = 323.72	ϕ_4 = -2.98
ϕ_5 = 112.53	ϕ_5 = 182.55	ϕ_5 = 100.54	ϕ_5 = -183.77	ϕ_5 = 94.46
ϕ_6 = 155.67	ϕ_6 = 169.66	ϕ_6 = 128.65	ϕ_6 = 63.95	ϕ_6 = 175.25
ϕ_7 = 24.04	ϕ_7 = 2.88	ϕ_7 = 131.21	ϕ_7 = 89.74	ϕ_7 = 7.69
Predicted dihedral angles (in deg.) of DKICA				
Configuration 1	Configuration 2	Configuration 3	Configuration 4	Configuration 5
θ_1 = 65.95	θ_1 = 69.50	θ_1 = 84.13	θ_1 = 82.41	θ_1 = 105.88
θ_2 = 80.19	θ_2 = 90.62	θ_2 = 94.37	θ_2 = 69.18	θ_2 = 83.83
θ_3 = 100.09	θ_3 = 70.66	θ_3 = 100.92	θ_3 = 96.90	θ_3 = 66.08
ϕ_1 = -4.83	ϕ_1 = 0.47	ϕ_1 = 5.69	ϕ_1 = -5.05	ϕ_1 = -41.32
ϕ_2 = 201.58	ϕ_2 = 176.57	ϕ_2 = 177.28	ϕ_2 = 157.23	ϕ_2 = 186.02
ϕ_3 = 255.72	ϕ_3 = 65.91	ϕ_3 = 269.98	ϕ_3 = -3.51	ϕ_3 = 193.82
ϕ_4 = 153.85	ϕ_4 = 166.14	ϕ_4 = 173.79	ϕ_4 = 171.53	ϕ_4 = 163.81
ϕ_5 = 289.77	ϕ_5 = 123.24	ϕ_5 = 145.94	ϕ_5 = 165.04	ϕ_5 = -83.45
ϕ_6 = -96.50	ϕ_6 = 100.83	ϕ_6 = 0.98	ϕ_6 = 3.01	ϕ_6 = 259.15

Table S6. All predicted inter-monomer $\{\theta\}$ and ring-OH $\{\phi\}$ dihedral angles for the experimental spectrum in the 560-570 nm wavelength range.

Predicted dihedral angles (in deg.) of MKICA				
Configuration 1	Configuration 2	Configuration 3	Configuration 4	Configuration 5
$\theta_1 = 103.45$	$\theta_1 = 49.17$	$\theta_1 = 106.07$	$\theta_1 = 106.00$	$\theta_1 = 53.96$
$\theta_2 = 92.05$	$\theta_2 = 107.63$	$\theta_2 = 48.50$	$\theta_2 = 73.02$	$\theta_2 = 106.34$
$\theta_3 = 56.97$	$\theta_3 = 50.99$	$\theta_3 = 90.01$	$\theta_3 = 93.77$	$\theta_3 = 100.31$
$\phi_1 = 142.78$	$\phi_1 = 133.99$	$\phi_1 = 158.81$	$\phi_1 = 159.10$	$\phi_1 = 106.39$
$\phi_2 = 89.25$	$\phi_2 = 30.90$	$\phi_2 = 75.95$	$\phi_2 = 133.91$	$\phi_2 = -3.09$
$\phi_3 = 102.63$	$\phi_3 = 116.26$	$\phi_3 = 145.01$	$\phi_3 = 94.75$	$\phi_3 = 152.62$
$\phi_4 = 116.35$	$\phi_4 = 81.97$	$\phi_4 = 42.07$	$\phi_4 = 131.68$	$\phi_4 = -4.22$
$\phi_5 = 98.99$	$\phi_5 = 115.89$	$\phi_5 = 130.41$	$\phi_5 = 136.12$	$\phi_5 = 150.84$
$\phi_6 = 161.02$	$\phi_6 = 119.63$	$\phi_6 = 102.10$	$\phi_6 = 153.61$	$\phi_6 = 118.58$
$\phi_7 = -2.30$	$\phi_7 = 29.87$	$\phi_7 = 59.28$	$\phi_7 = 9.64$	$\phi_7 = 92.12$
Predicted dihedral angles (in deg.) of DKICA				
Configuration 1	Configuration 2	Configuration 3	Configuration 4	Configuration 5
$\theta_1 = 65.45$	$\theta_1 = 61.11$	$\theta_1 = 81.83$	$\theta_1 = 82.27$	$\theta_1 = 65.55$
$\theta_2 = 79.03$	$\theta_2 = 58.86$	$\theta_2 = 90.90$	$\theta_2 = 109.16$	$\theta_2 = 123.50$
$\theta_3 = 86.23$	$\theta_3 = 69.04$	$\theta_3 = 69.34$	$\theta_3 = 96.98$	$\theta_3 = 73.23$
$\phi_1 = 15.49$	$\phi_1 = 133.13$	$\phi_1 = 2.12$	$\phi_1 = 3.08$	$\phi_1 = 8.18$
$\phi_2 = 158.89$	$\phi_2 = 148.44$	$\phi_2 = 161.31$	$\phi_2 = 159.12$	$\phi_2 = 131.77$
$\phi_3 = 70.57$	$\phi_3 = -0.71$	$\phi_3 = -0.55$	$\phi_3 = 14.26$	$\phi_3 = 2.70$
$\phi_4 = 147.13$	$\phi_4 = 161.21$	$\phi_4 = 166.67$	$\phi_4 = 157.94$	$\phi_4 = 179.86$
$\phi_5 = -3.63$	$\phi_5 = 145.89$	$\phi_5 = 92.01$	$\phi_5 = 148.88$	$\phi_5 = -3.43$
$\phi_6 = 186.07$	$\phi_6 = 8.51$	$\phi_6 = 161.73$	$\phi_6 = 5.46$	$\phi_6 = 168.04$

Table S7. MAE (in degree) with variable spectral wavelength range for 10k training set. Below results show that the 10 nm wavelength range offers the best error. MAEs are calculated on a 10k hold-out set.

Wavelength range (nm)	**MAE in inter-monomer dihedral angle (θ)**	**MAE in ring-OH dihedral angle (ϕ)**
299-301	1.57	21.23
298-302	0.30	5.61
295-305	0.09	1.85
290-310	0.11	1.95
275-325	0.23	3.76

References

(1) Case, D.; Berryman, J.; Betz, R.; Cerutti, D.; Cheatham III, T.; Darden, T.; Duke, R.; Giese, T.; Gohlke, H.; Goetz, A.; others *AMBER 2016; University of California: San Francisco*; 2016.

(2) Shao, Y.; Gan, Z.; Epifanovsky, E.; Gilbert, A. T.; Wormit, M.; Kussmann, J.; Lange, A. W.; Behn, A.; Deng, J.; Feng, X.; others Advances in molecular quantum chemistry contained in the Q-Chem 4 program package. *Molecular Physics* **2015**, *113*, 184–215.

(3) Gauden, M.; Pezzella, A.; Panzella, L.; Neves-Petersen, M.; Skovsen, E.; Petersen, S. B.; Mullen, K.; Napolitano, A.; d'Ischia, M.; Sundstrom, V. Role of solvent, pH, and molecular size in excited-state deactivation of key eumelanin building blocks: implications for melanin pigment photostability. *Journal of the American Chemical Society* **2008**, *130*, 17038–17043.

(4) Murphy, K. P. *Machine learning: a probabilistic perspective*; MIT press, 2012.

(5) Schölkopf, B.; Tsuda, K.; Vert, J.-P.; others *Kernel methods in computational biology*; MIT press, 2004.

(6) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Physical review letters* **2012**, *108*, 058301.

(7) Ramakrishnan, R.; Hartmann, M.; Tapavicza, E.; von Lilienfeld, O. A. Electronic spectra from TDDFT and machine learning in chemical space. *The Journal of chemical physics* **2015**, *143*, 084111.

(8) Ramakrishnan, R.; von Lilienfeld, O. A. Many molecular properties from one kernel in chemical space. *CHIMIA International Journal for Chemistry* **2015**, *69*, 182–186.