Supplementary Information: Machine learning modeling of electronic spectra and thermodynamic stability for a comprehensive chemical space of melanin

Arpan Choudhury* and Debashree Ghosh*

School of Chemical Sciences, Indian Association for the Cultivation of Science, Kolkata 700032, India

E-mail: arpanchoudhury29@gmail.com; pcdg@iacs.res.in

Contents

Page No.

Figure S1: Linear, branched and cyclic tetramer structures. . . . 4

Section S1: Fingerprint generation method. . . . 5

Table S1: Optimized hyperparameter values of multi-output KRR for learning excitation energies of lowest 60 excited states. . . . 7

Table S2: Optimized hyperparameter values of multi-output KRR for learning oscillator strengths of lowest 60 excited states. . . . 7

Table S3: Optimized hyperparameter values of multi-output KRR for learning bin intensities using fingerprint input. . . . 7

Table S4: Optimized hyperparameter values of single-output KRR for learning relative stability using fingerprint input. . . . 7

Figure S3: Learning excitation energies and oscillator strengths of the lowest 60 excited states using SLATM, Coulomb matrix and fingerprint ML-input. . . . 8

Table S5: Nature of the excited states in each bin across the UV-visible wavelength. . . . 9

Figure S4: Learning curves showing the overlap metric for 50 nm bin resolution using fingerprint, SLATM and Coulomb matrix input. . . . 10

Figure S5: ML-predicted UV-visible spectra. . . . 11

Figure S6: Comparison between DLPNO-CCSD(T), B3LYP and ML-predicted relative energies. . . . 12

Figure S7: Scatter plots of DFT vs. ML-predicted relative energies (in kcal/mol) for different proportions of reduced and oxidized monomers in the tetramer structures. . . . 12

Figure S8: The Boltzmann-weighted average spectrum of DHI-melanin containing linear, branched and cyclic tetramers.....13

Figure S9: Electronic absorption spectrum of DHI monomer at CAM-B3LYP/6-31G(d) level of theory.....13

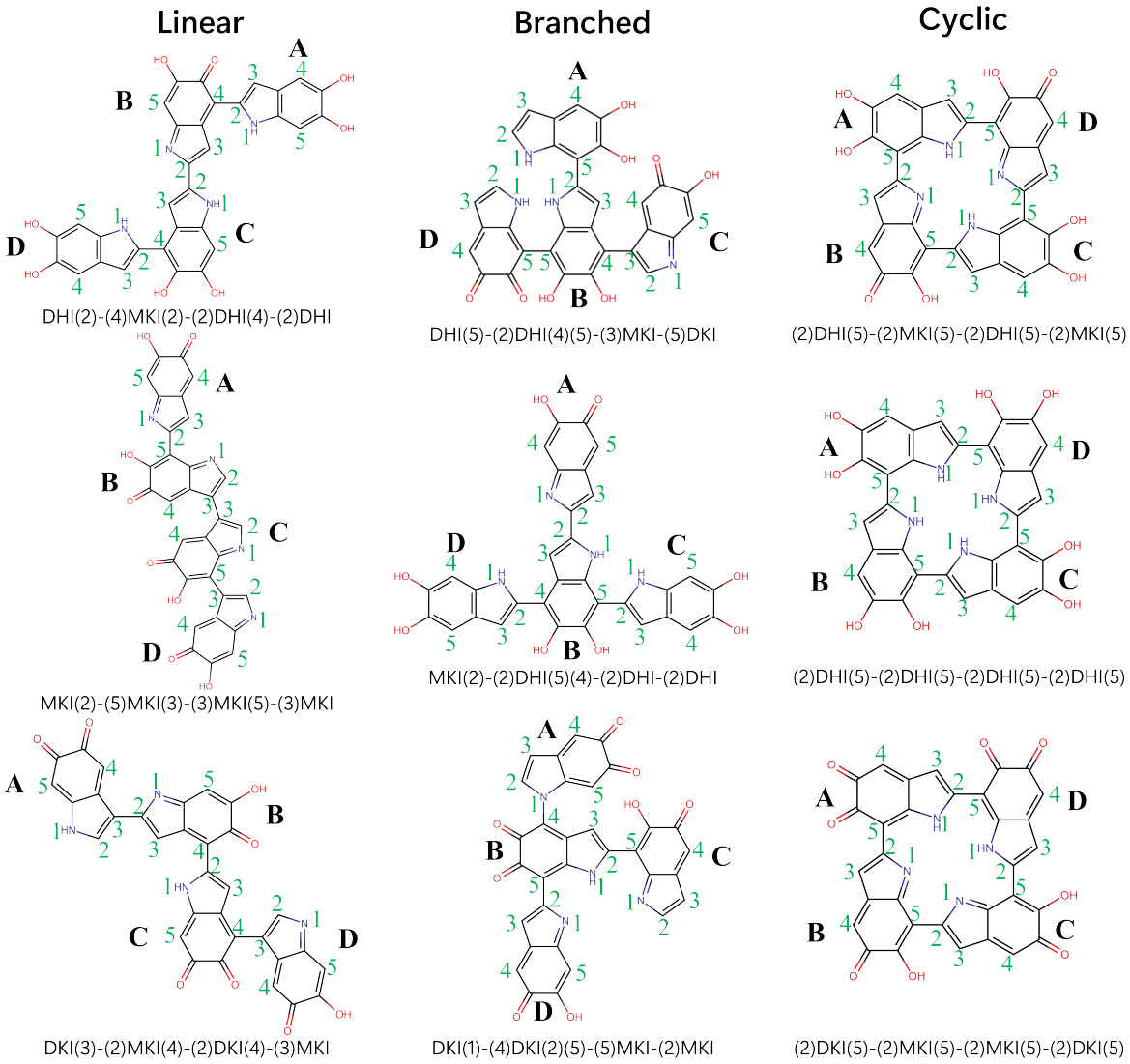


Figure S1: Examples of linear, branched and cyclic tetramer structures. Each of them is assigned to unique name which includes the position numbers through which adjacent monomers are connected and the oxidation states of the oxygen. DHI (dihydroxyindole), MKI (monoketoindole) and DKI (diketoindole) refers to the different oxidation states.

Section S1: Fingerprint Generation Method:

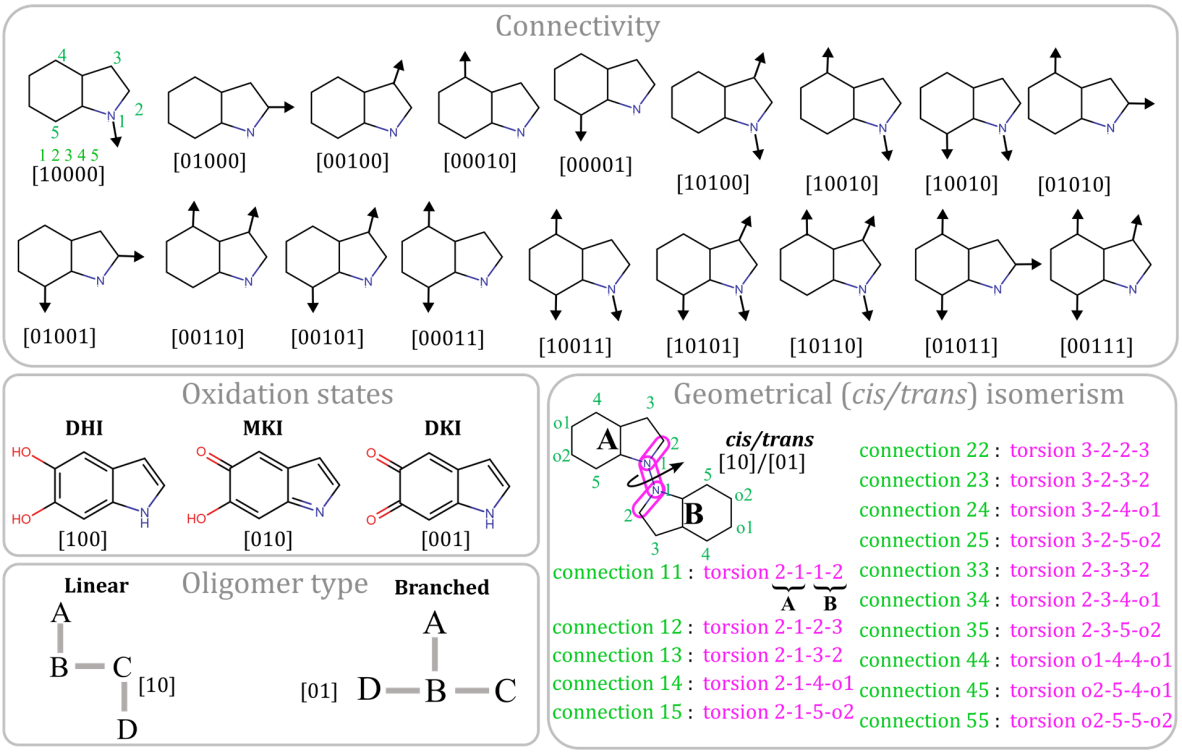


Figure S2: Molecular fingerprints generation. Bit strings to codify the connectivity patterns, oxidation state of the monomer oxygens, oligomer type and geometrical isomerism are shown here (see text for explanations).

The fingerprints are generated based on the following structural information: (i) connectivity pattern between the adjacent monomers; (ii) oxidation states of the hydroxyl oxygens; (iii) *cis/trans* isomerism about the connecting bonds; and (iv) nature of the tetramer. These structural information are encoded sequentially into bit strings to produce the final fingerprint as illustrated in Fig. S2.

Connectivity: Each monomer unit has five positions available for oligomerization. We encode these positions by '1' if the monomer is connected to another monomer through this position (represented with arrows in Fig. S2), otherwise it is encoded by '0'. For each monomer it produces a bit string of length 5 and thus for a tetramer it is a bit string of length 20.

Oxidation states: Experiments have identified three types of monomer units in melanin,

namley dihydroxyindole (DHI), monoketoindole (MKI) and diketoindole (DKI) due to different oxidation states of the hydroxyl oxygens. We encode DHI as '100', MKI as '010' and DKI as '001'.

Oligomer type: In our dataset there are three types of tetramer molecules: linear, branched and cyclic. As there are only 16 cyclic tetramers, we excluded them from our machine learning modeling. As shown in the figure, there are connections such as AB and BC which are present for linear and branched both tetramers. However, the only difference between linear and branched structure is whether there is CD or BD connection. Thus based on the 'AB,BC,CD,BD' connection string, linear structure can be characterized as '1110' and branched structure can be characterized as '1101'. Here the presence/absence of a connection is encoded by '1'/'0'. Since the first two bits in '1110' and ''1101 are '11', we removed them to simply the representation as '10' and '01', respectively.

Geometrical isomerism: The *cis/trans* isomerism about specific torsional angles between two adjacent monomers are also encoded in the fingerprint. The torsional angles along which we infer the *cis/trans* isomerism are mentioned in Fig. S2 for all possible connections between two adjacent monomers (one example torsional angle is also drawn). Along these torsional angles, *cis* geometry is encoded by '10' and *trans* geometry is encoded by '01'.

Table S1: Optimized hyperparameter values of multi-output KRR for learning excitation energies of lowest 60 excited states.

Input	**Kernel function**	**Hyperparameter**
Fingerprint	gaussian	$\sigma = 10, \lambda = 10^{-3}$
Fingerprint	laplacian	$\sigma = 10^2, \lambda = 10^{-2}$
SLATM	gaussian	$\sigma = 10^2, \lambda = 10^{-6}$
SLATM	laplacian	$\sigma = 10^3, \lambda = 10^{-3}$
Coulomb matrix	gaussian	$\sigma = 10^5, \lambda = 10^{-7}$
Coulomb matrix	laplacian	$\sigma = 10^6, \lambda = 10^{-4}$

Table S2: Optimized hyperparameter values of multi-output KRR for learning oscillator strengths of lowest 60 excited states.

Input	**Kernel function**	**Hyperparameter**
Fingerprint	gaussian	$\sigma = 1, \lambda = 1$
Fingerprint	laplacian	$\sigma = 10^2, \lambda = 10^3$
SLATM	gaussian	$\sigma = 10^2, \lambda = 10^3$
SLATM	laplacian	$\sigma = 10, \lambda = 1$
Coulomb matrix	gaussian	$\sigma = 10^3, \lambda = 10^3$
Coulomb matrix	laplacian	$\sigma = 10^4, \lambda = 10^3$

Table S3: Optimized hyperparameter values of multi-output KRR for learning bin intensities using fingerprint input.

Bin resolution	**Kernel function**	**Hyperparameter**
25 nm	gaussian	$\sigma = 10, \lambda = 10^{-2}$
	laplacian	$\sigma = 10, \lambda = 1$
50 nm	gaussian	$\sigma = 10, \lambda = 10^{-2}$
	laplacian	$\sigma = 10, \lambda = 1$
100 nm	gaussian	$\sigma = 10, \lambda = 10^{-2}$
	laplacian	$\sigma = 10, \lambda = 10^{-1}$

Table S4: Optimized hyperparameter values of single-output KRR for learning relative stability using fingerprint input.

Input	**Kernel function**	**Hyperparameter**
Fingerprint	gaussian	$\sigma = 1, \lambda = 10^{-3}$
Fingerprint	laplacian	$\sigma = 1, \lambda = 10^{-1}$

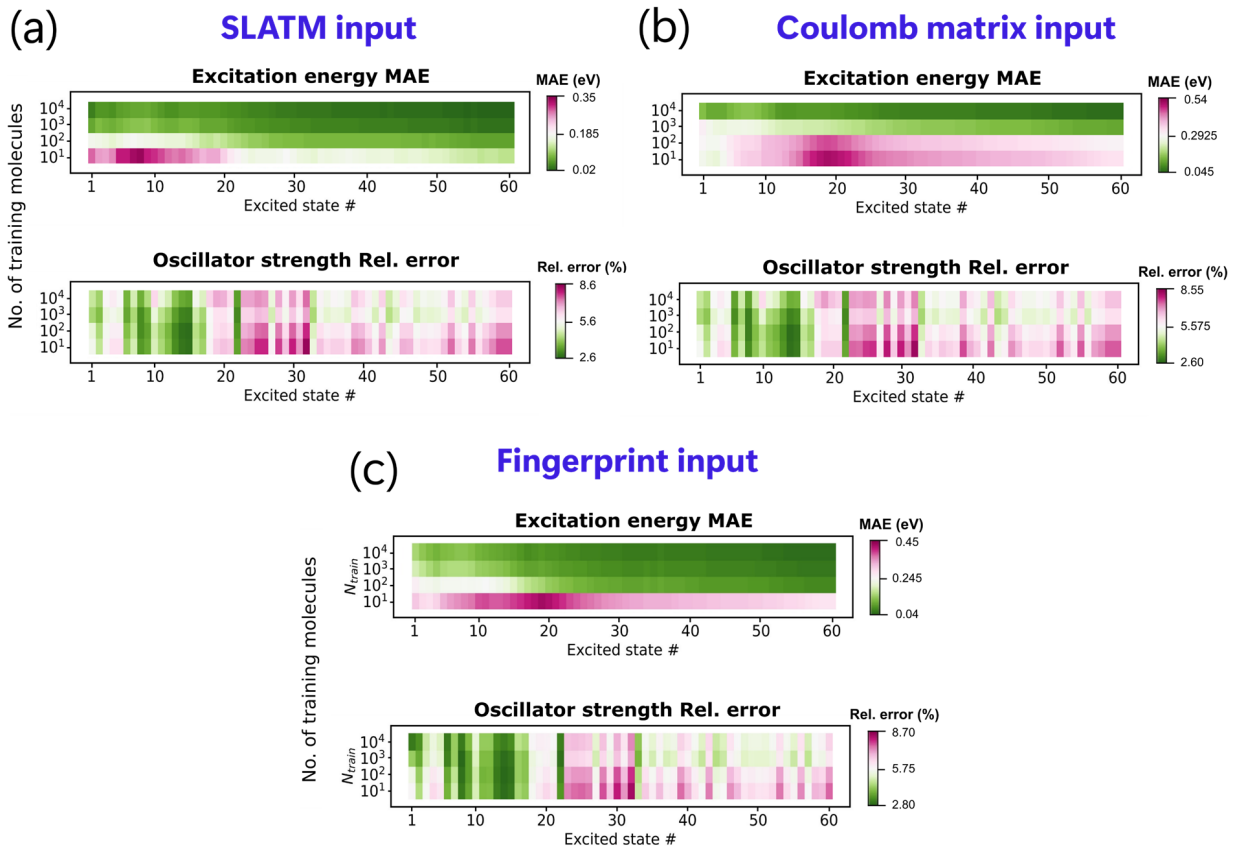
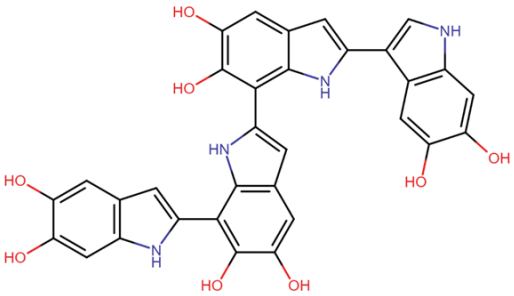
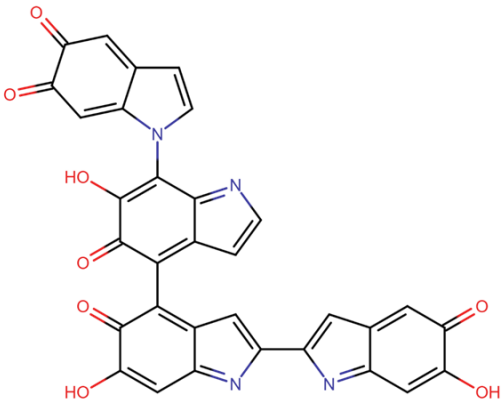
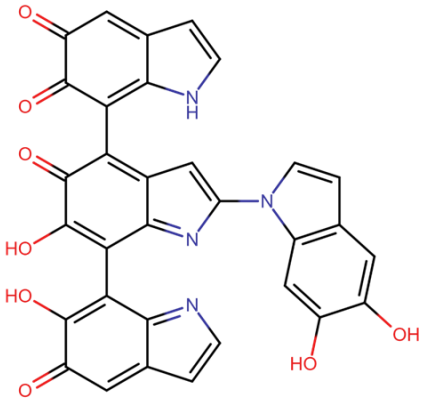


Figure S3: Learning excitation energies and oscillator strengths of the lowest 60 excited states. The test errors calculated over a hold-out dataset are shown by color bars. The X-axis represents the individual excited states and the Y-axis represents the training dataset size. Learning is shown for 3D geometry-based ML input (a) SLATM and (b) Coulomb matrix which are derived from B3LYP/6-31G(d) optimized geometries. (c) Learning for fingerprint-based input.

Table S5: Nature of the excited states within each bin for three representative tetramers. LE: local excitation; CT: charge transfer; Mixed: combination of LE and CT.

Molecule	Absorption range	Nature of the states
	200-250 nm	LE, CT, Mixed
	250-300 nm	LE, Mixed
	300-350 nm	LE
	200-250 nm	LE, CT, Mixed
	250-300 nm	LE, CT, Mixed
	300-350 nm	LE, CT, Mixed
	350-400 nm	LE, CT, Mixed
	400-450 nm	LE, CT
	450-500 nm	LE
	500-550 nm	CT
	550-600 nm	LE, Mixed
	650-700 nm	LE
	700-750 nm	CT
	200-250 nm	LE, CT, Mixed
	250-300 nm	LE, CT, Mixed
	300-350 nm	CT, Mixed
	350-400 nm	LE, CT, Mixed
	400-450 nm	CT
	450-500 nm	CT
	500-550 nm	LE
	550-600 nm	LE
	600-650 nm	LE
	750-800 nm	LE

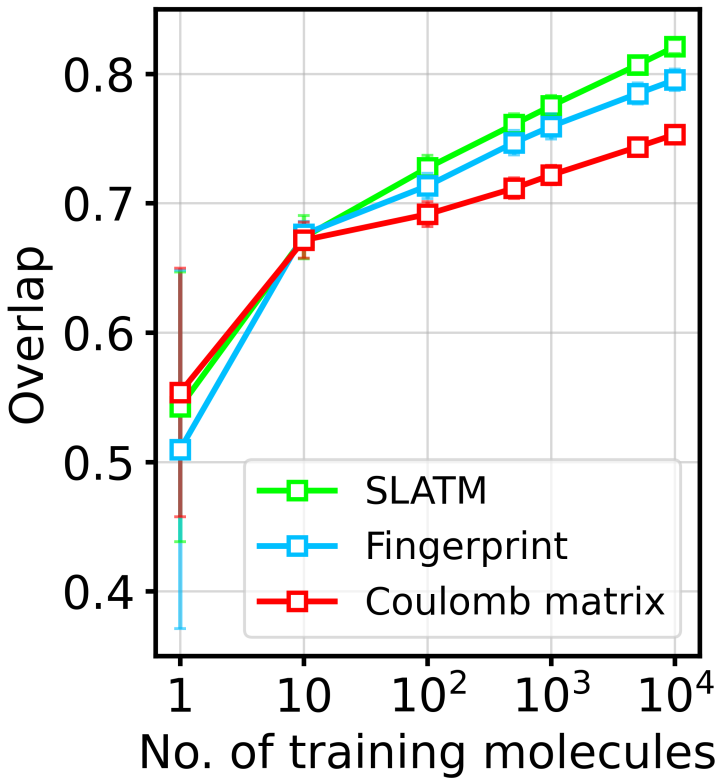


Figure S4: Learning curves showing the overlap metric for 50 nm bin resolution using fingerprint, SLATM and Coulomb matrix input. The vertical error bars correspond to the uncertainty over 20 independent runs.

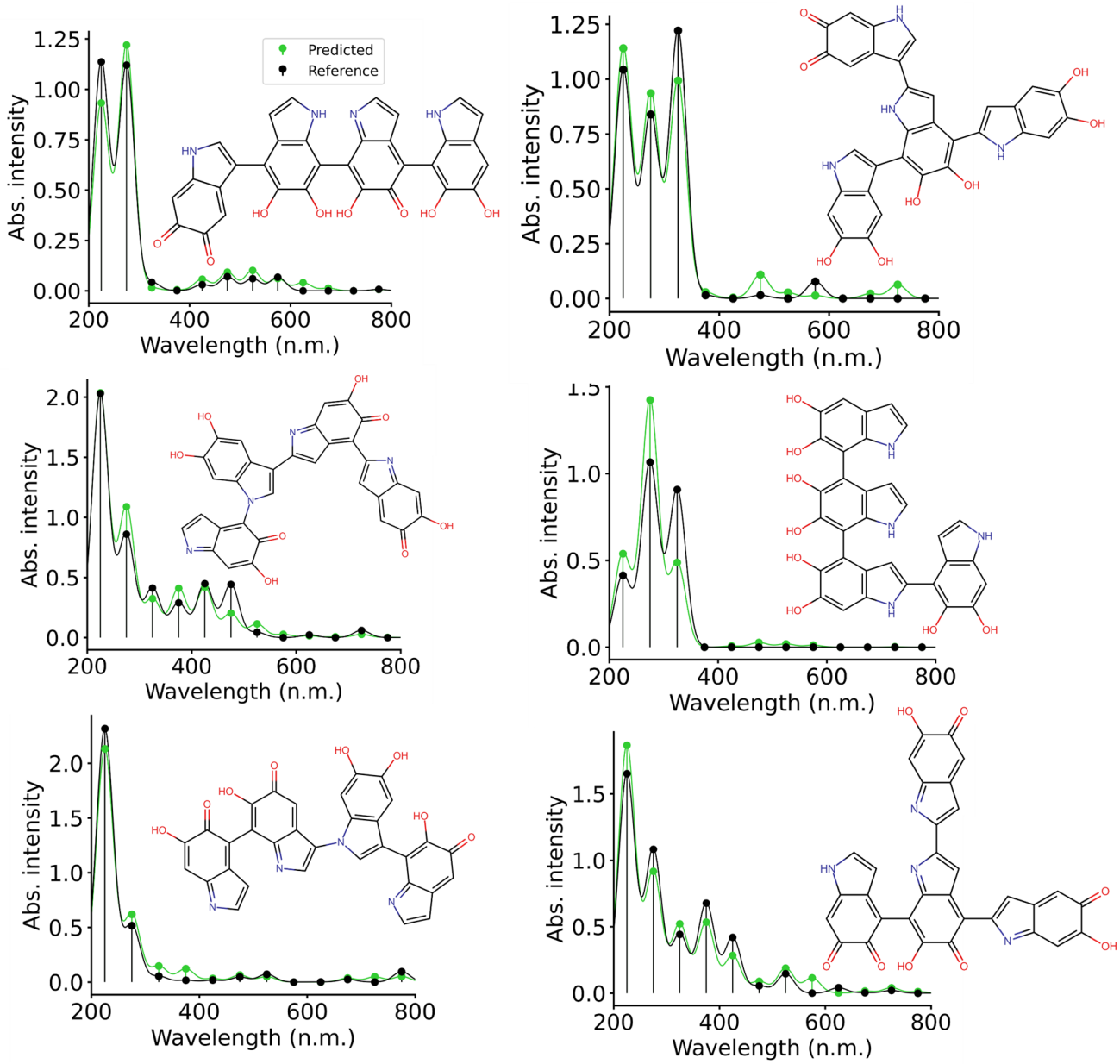


Figure S5: Predicted UV-visible absorption spectra of molecules not present in the training dataset using multi-output KRR-ML model trained on 10k dataset with fingerprint input. The vertical lines with circles on the top refer to the intensity value of the bins and the curves are Gaussian broadening with FWHM equal to the bin resolution (i.e. 50 nm)

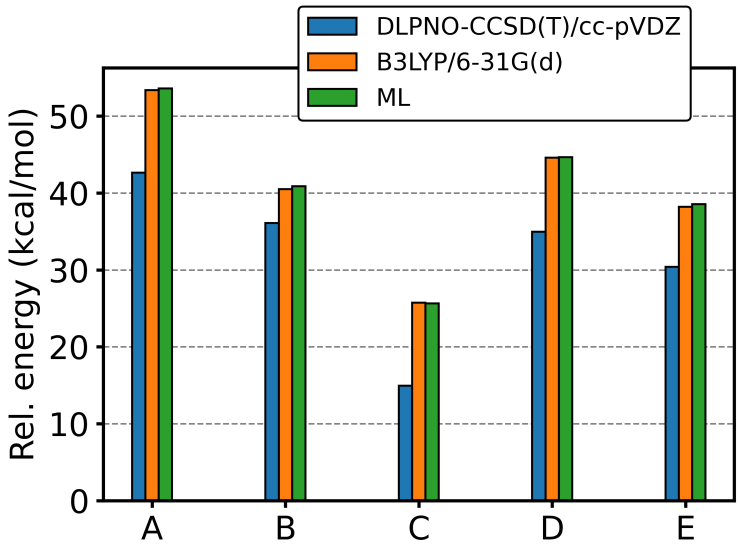


Figure S6: Relative energies (in kcal/mol) for five random molecules using DLPNO-CCSD(T)/cc-pVDZ, B3LYP/6-31G(d) and ML prediction.

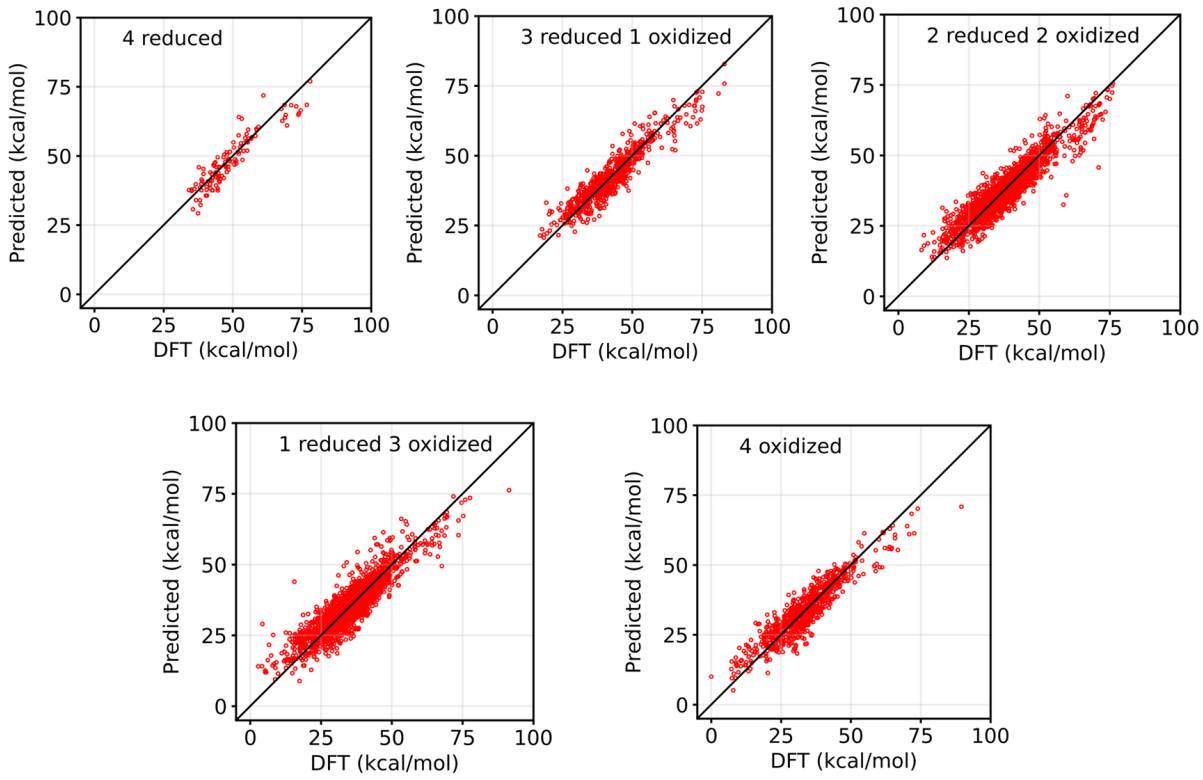


Figure S7: Scatter plots of DFT vs. ML-predicted relative energies (in kcal/mol) for different proportions of reduced and oxidized monomers in the tetramer structures.

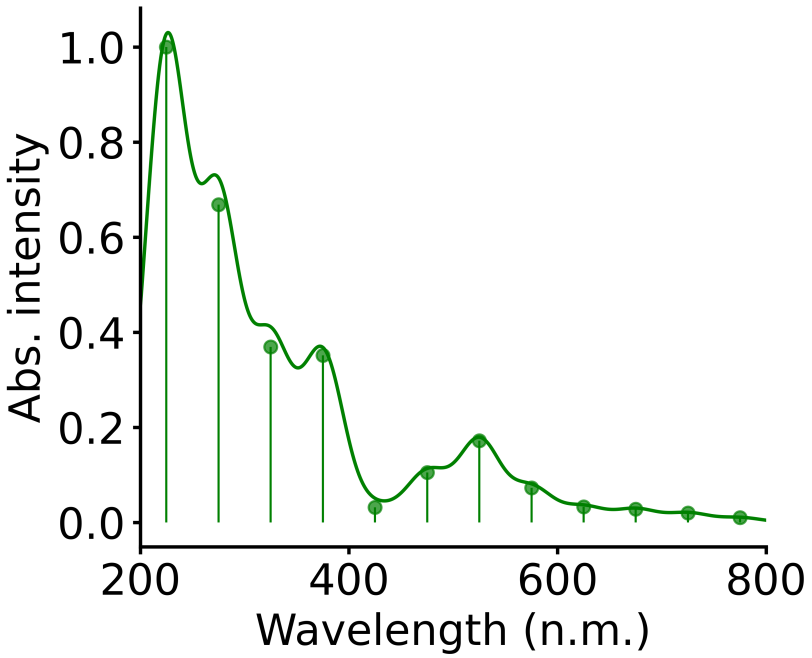


Figure S8: The Boltzmann-weighted average spectrum of DHI-melanin containing linear, branched and cyclic tetramers.

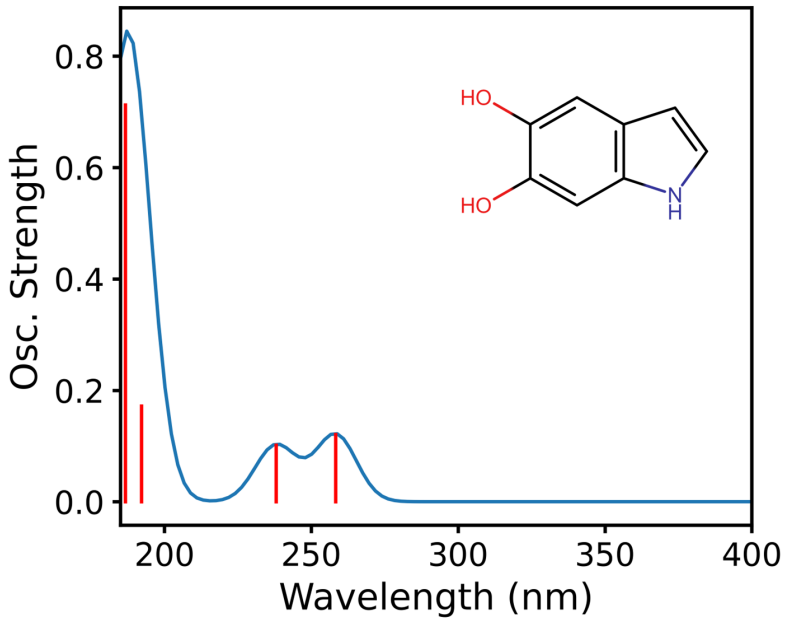


Figure S9: Electronic absorption spectrum of DHI monomer at CAM-B3LYP/6-31G(d) level of theory.