ELECTRONIC SUPPORTING INFORMATION

Fingerprint-based deep neural networks can model thermodynamic and optical properties of eumelanin DHI dimers

Daniel Bosch,[a] Jun Wang[b] and Lluís Blancafort*

[a] Institut de Química Computacional i Catàlisi and Departament de Química, Universitat de Girona. Facultat de Ciències, C/M. A. Capmany 69, 17003 Girona (Spain)

[b] Jiangsu Key Laboratory for Chemistry of Low-Dimensional Materials, Jiangsu Engineering Laboratory for Environment Functional Materials, Huaiyin Normal University. No. 111 West Changjiang Road, Huaian 223300, Jiangsu Province (P. R. China).

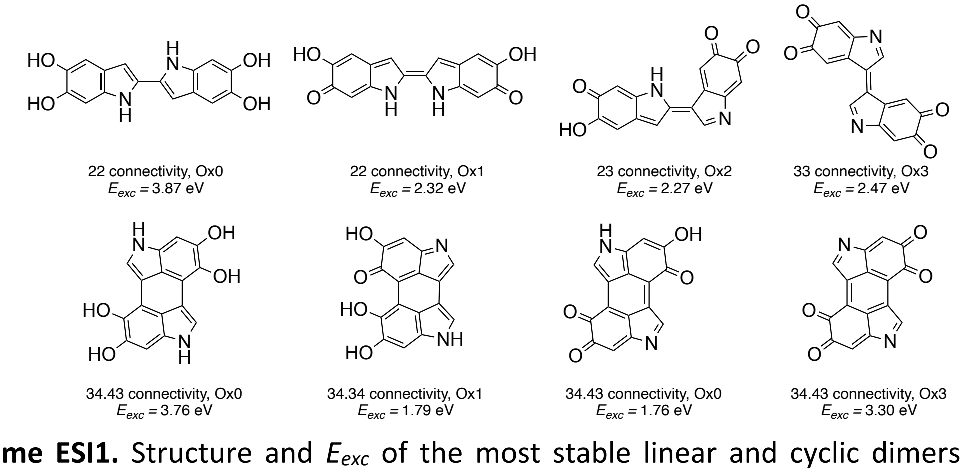
TABLE OF CONTENTS

ESI1. Data set composition and density functional theory details

The data set introduced in Ref. [1] covers the relative free energy of formation, G_{rel}, and the lowest vertical electronic absorption, E_{exc}, of 830 DHI dimers with different connectivity and oxidation state. The compound library contains only closed-shell dimers (*ie* with a total even number of oxidized sites) and has two main groups of compounds, linear and cyclic dimers. There are 538 linear dimers with a single interfragment CC bond. The set includes both possible *cis*/*trans* stereoisomers at the connecting bond. There are also 292 polycyclic dimers formed by a pair of DHI molecules connected at two sites, giving rise to a new interfragment ring. These dimers are called cyclic dimers here for simplicity. The cyclic dimers have 5- or 6-membered interfragment rings containing up to two nitrogen or oxygen atoms. For oxygen-containing rings, only 6-membered rings are considered. This compound set includes all possible dimers with the described characteristics except for six linear isomers that where discarded due to instabilities in the electronic wave function, indicating a possible open-shell character, or spontaneous tautomerization to a different isomer (see Ref. [1] for details). The total compound distribution in terms of linear and polycyclic structures and oxidation state is shown in Table SI1. Ox*n* refers to an oxidation state with *n* going from 0 - 3 (elimination of *n* H_2 molecules). For instance, Ox0 stands for a fully reduced structure (all heteroatoms present as OH or NH), and Ox1 corresponds to elimination of 1 H_2 molecule, *ie* 2 oxidized heteroatoms. An overview of the most stable compounds of each type is shown in Scheme SI1.

Table ESI1. Number of linear and cyclic dimers of each oxidation state.

	Ox0	Ox1	Ox2	Ox3	Total
Linear	20	246	252	20	538
Cyclic	18	168	102	4	292
Total	38	414	354	24	830



Scheme ESI1. Structure and E_{exc} of the most stable linear and cyclic dimers of each oxidation state.

G_{rel} of the dimer set was calculated in Ref. [1] at the CAM-B3LYP/6-311G(d,p) level of theory (SMD bulk solvation in water[2]). It is calculated relative to the most stable dimer, using a thermodynamic scheme based on the oxidation of the dimers with oxygen to estimate the energy of compounds with different stochiometry (see Ref. [1]). G_{rel} spans a range of 0-183 kcal·mol^{-1}, with cyclic and more oxidized dimers being more stable than linear and more reduced ones, respectively. The majority of compounds lies in a range of 50 - 150 kcal·mol^{-1} (see Figure SI1a). E_{exc} is the vertical absorption of the lowest excited state, calculated at the time-dependent (TD) CAM-B3LYP/6-311G(d,p) level (see Ref. [1]). E_{exc} of the set covers a range of 0.46 - 4.71 eV, with Ox0 structures covering the 3.3 - 4.7 eV range, and the remaining ones the lower energy range.

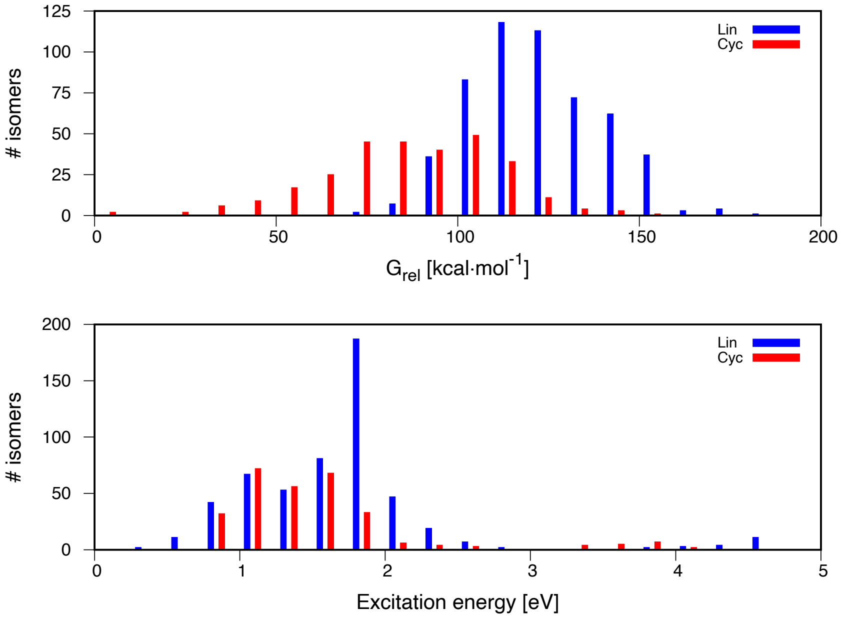


Figure ESI1. G_{rel} and E_{exc} distribution histograms (panels a and b, respectively), grouping the dimers into linear and cyclic structures.

ESI2. Generation of molecular string representations

ESI2.1 Rules for molecular string representations

This section provides the rules for generation of the different molecular representation strings. It is supported by Table ESI2, which gives the possible values of every digit for every fingerprint type.

ESI2.1.1 Quasi-binary (QB) strings. These strings have a variable number of connectivity digits (see following paragraphs), six oxidation digits, and one stereo-/regiochemistry digit. The oxidation digits correspond to the three heteroatoms of each fragment, and they have a value of '0' if the heteroatom is reduced and a value of '1' if it is reduced. The connectivity digits have all a value of '0' except for the connectivity that is present in the represented molecule. The rules for these digits depend on the specific fingerprint type and are as follows:

ESI2.1.1.1 QBB strings. There are 19 connectivity digits: 10 codify bonds in the linear or cyclic dimers and 9 codify bonds that are only present in the cyclic dimers. The digits get a value of '1' for the interfragment bonds, *ie* there are one and two digits equal to '1' in linear and cyclic dimers, respectively. The exception is cyclic dimers between equivalent fragments, *eg* 2 (43) fragments connected such that there are two 34 bonds. In that case, the cyclic dimer has a single non-zero digit which is set equal to '2'. The stereochemistry digit has a value '0' for cyclic dimers, '1' and '2' for linear *cis* and *trans* dimers.

ESI2.1.1.2 QBF strings. There are 10 connectivity digits, 4 for the linear fragments and 6 for the cyclic fragments. These digits get a value of '1' for the fragments that form the dimer. If the dimer is symmetric (*ie* formed by a single type of fragment), the corresponding digit gets a value of '2'. The stereo-/regiochemistry digit is used to distinguish the two possible stereoisomers of the linear dimers and the two regioisomers of the cyclic dimers (*eg* distinguish 23.43 from 23.34 connectivity).

ESI2.1.1.3 QBS strings. There are 12 connectivity strings, 5 for the sites on the first fragment and 7 for the sites on the second fragment. The different number of sites for each fragment is due to the fact that we do not consider dimers where two oxygen atoms take part in the formation of the interfragment ring, and site (5) and (6) connectivity is only considered for one fragment. The value of the connectivity digits is set to '1' if it corresponds to the connection site in a linear dimer, or 2 for a cyclic dimer. The stereo-/regiochemistry digit is used to distinguish the two possible stereoisomers of the linear dimers and the two regioisomers of the cyclic dimers.

ESI2.1.2 Multi-valued (MV) strings.

ESI2.1.2.1 MV reduced (MVR) strings. Here we use a single string to code connectivity and another one to code oxidation pattern, with 28 and 32 possible values, respectively. The third string codifies interfragment stereochemistry and has possible values of '1' and '2' for linear cis and trans dimers, and '0' for cyclic ones.

ESI2.1.2.2 MV site-based (MVS) strings. 7 strings represent the possible connection sites and 3 the oxidation sites, and every string represents equivalent positions in the two fragments. The possible values of each connectivity node are 0 if the site is not connected to any fragment, 1 if it is connected in the first fragment, 2 if it is connected in the second fragment and 3 if it is connected in both fragments. Similarly, the oxidation nodes can have values of 0 if the site is reduced in the two fragments, 1 if it is only oxidized in the first fragment, 2 if it is only oxidized in the second fragment and 3 if it is oxidized in both fragments. The stereo-/regiochemistry digit is used to distinguish the two possible stereoisomers of the linear dimers and the two regioisomers of the cyclic dimers.

Table ESI2. Total number of possible values for each type of nodes for all input strings.

Input strings	Connectivity node values	Oxidation node values	Stereochemistry node values
QBB	4	2	3
QBF	2	2	2
QBS	3	2	2
MVR	28	32	3
MVS	4	4	2

ESI2.2 Application to dimer #339

The molecular string representations for dimer #339 of Figure 2 are given in Figure ESI2. This structure has bonds between carbons 2 of the first fragment and 4 of the second fragment, and between 3 of the first fragment and 3 of the second fragment. The (24) and (33) digits are set to '1' in the QBB representation. It is formed by fragments (23) and (34), which are the digits set to '1' in the QBF representation. In the QBS representation, sites 2 and 3 on the first fragment and 3 and 4 on the second fragment are set to '1'.

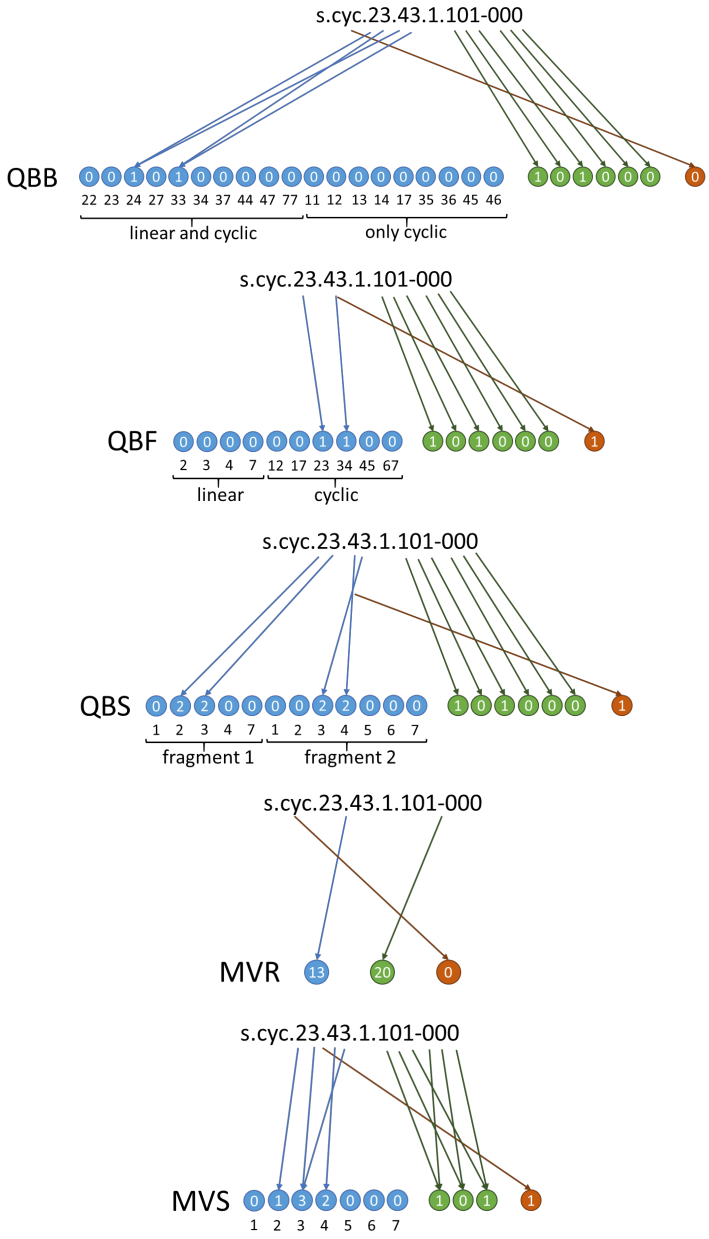


Figure ESI2. Generation of molecular string representations for cyclic dimer #339 of Figure 2.

ESI3. Neural network set up

We have trained deep neural networks with a varying number of input nodes corresponding to the digits of the molecular representation strings introduced in the previous section, and a single output node (G_{rel} or E_{exc}). Independent NN are trained for G_{rel} and E_{exc}. Our total data set is divided into 60% training, 20% validation and 20% test sets. We have tested 1 to 3 hidden layers with varying number of nodes. The training is carried out with a learning rate of $1{\cdot}10^{-2}$ except where noted, using the root mean square deviation with respect to the DFT values as loss function. We use the rectified linear unit (ReLU) activation function and Glorot initialization.[3] The training is done for 200 epochs in the initial tests, and 5000 epochs for the prediction runs. For the E_{exc}

predictions we use transfer learning with two phases over 500 and 5000 epochs with one and two hidden layers. The training and prediction cycles are repeated 10 times for every setup, and the results provided are average and standard deviation (SD) of the errors for the 10 runs (initial runs were carried out with a smaller number of runs, see Tables ESI2 - ESI4). In the initial tests, for every training cycle the data set is split into different training, validation and test subsets, whereas in the prediction runs we have worked with the same data set in the 10 runs to simulate a prediction situation where one works with a fixed training and validation set, and the DFT values of the predicted compounds are unknown. Training curves (plots of mean training and validation loss over each epoch) for one prediction run with every fingerprint model are provided in the section ESI4.

To condition the data and set up the neural network we follow the workflow of Ref. [4] and use the following Python 3.0 modules: Pandas[5] to read in the external values, Scikit-Learn[6] to handle data, and TensorFlow[7] to train the network and do the predictions. We use Matplotlib,[8] Pyplot and gnuplot[9] for plotting.

ESI4. Neural network training results

ESI4.1 Initial tests

A set of tests with a small number of 200 epochs was carried out to compare the performance of multiple-valued and quasi-binary inputs. We tested MVR and QBB with one or two hidden layers with up to 16 nodes per layer (see details in Tables ESI2 and ESI3). QBB performs better than MVR. For G_{rel} (stability), with the QBB inputs we have errors of approximately 12 kcal/mol with no big changes between validation, training and test error, whereas the errors with MVR are in the range 24 - 34 kcal/mol, respectively. A similar conclusion can be drawn from the results for E_{exc}, where the test errors with QBB and MDR are 0.36 - 0.40 and 0.39 - 0.56 eV. Therefore, the multiple-valued based inputs were discarded for the rest of the work. These tests were carried out with learning rates of $1{\cdot}10^{-2}$ for G_{rel} and $5{\cdot}10^{-3}$ for E_{exc}. Single-run tests with different learn-rates showed that the variation of the results with respect to the learning rate is small for both endpoints, and these learning rates were kept for the rest of the work.

Given that the relative errors for E_{exc} prediction are higher than for G_{rel} (9% of the total range vs 5-7%, approximately), we tested further alternatives for E_{exc}, including transfer learning, batch normalization,[10] and He initialization[11] (see Table ESI4). Batch normalization and He initialization did not provide any improvements. Following these tests, we have set up for the predictions a transfer learning approach using 7 hidden nodes over 500 epochs and adding a second hidden layer of 7 hidden nodes for 5000 epochs, and a learning rate of $5{\cdot}10^{-3}$.

Table ESI2. Results of preliminary tests for G_{rel} modeling with the MVR and QBB strings. Results are averages and standard deviation in kcal·mol^{-1} over 5 training runs.[a]

Input strings	Nodes per layer[b]	N_{Pars}[c]	Training error	Validation error	Test error	Over-training[d]
MVR	3,12,1	48	26.4 ± 1.7	32.6 ± 8.7	33.6 ± 8.9	6.2
	3,16,1	64	27.7 ± 1.0	25.4 ± 2.9	25.3 ± 3.3	-2.3
	3,6,1	24	28.0 ± 2.3	26.1 ± 4.8	26.2 ± 3.6	-1.9
	3,10,4,1	74	24.1 ± 1.0	24.8 ± 1.7	23.9 ± 1.2	0.7
	3,7,7,1	77	24.8 ± 1.5	26.2 ± 2.8	27.4 ± 2.8	1.4
	3,4,4,1	32	24.6 ± 1.9	24.4 ± 2.1	24.3 ± 2.7	-0.2
QBB	26,12,1	324	11.2 ± 0.2	11.7 ± 0.4	12.0 ± 0.6	0.5
	26,16,1	432	11.2 ± 0.6	11.1 ± 0.5	11.8 ± 1.0	-0.1
	26,6,1	162	11.4 ± 0.3	11.8 ± 0.5	12.4 ± 1.2	0.4
	26,10,4,1	304	11.2 ± 0.8	12.4 ± 2.1	12.4 ± 2.2	1.2
	26,7,7,1	238	11.0 ± 0.6	11.8 ± 1.7	11.9 ± 2.2	0.8
	26,4,4,1	124	11.1 ± 0.3	12.2 ± 1.5	12.4 ± 1.9	1.1

[a] Training over 200 epochs, learning rate 1·10^{-2}. [b] Number of nodes per each layer. [c] Number of parameters of the model, equal to the number of edges in the NN graph. [d] Difference between validation and training error.

Table ESI3. Results of preliminary tests for E_{exc} modeling with the MVR and QBB strings. Results are averages and standard deviation in eV over 5 training runs.

Input strings	Nodes per layer[b]	N_{Pars}[c]	Training error	Validation error	Test error	Over-training[d]
MVR	3,10,1	40	0.39 ± 0.01	0.39 ± 0.02	0.39 ± 0.04	0.00
	3,16,1	64	0.38 ± 0.01	0.41 ± 0.04	0.40 ± 0.01	0.03
	3,6,1	24	0.54 ± 0.14	0.58 ± 0.13	0.55 ± 0.12	0.04
	3,10,4,1	74	0.39 ± 0.01	0.40 ± 0.04	0.39 ± 0.02	0.01
	3,7,7,1	77	0.40 ± 0.02	0.41 ± 0.03	0.41 ± 0.02	0.01
	3,4,4,1	32	0.54 ± 0.14	0.57 ± 0.16	0.56 ± 0.13	0.03
QBB	26,12,1	324	0.32 ± 0.02	0.37 ± 0.01	0.40 ± 0.02	0.05
	26,16,1	432	0.30 ± 0.01	0.36 ± 0.01	0.39 ± 0.01	0.06
	26,6,1	162	0.34 ± 0.01	0.41 ± 0.05	0.40 ± 0.03	0.07
	26,10,4,1	304	0.29 ± 0.02	0.39 ± 0.03	0.36 ± 0.03	0.10
	26,7,7,1	238	0.31 ± 0.04	0.37 ± 0.01	0.39 ± 0.02	0.06
	26,4,4,1	124	0.35 ± 0.02	0.40 ± 0.02	0.39 ± 0.03	0.05

[a] Training over 500 epochs, learning rate 5·10^{-3}. [b] Number of nodes per each layer. [c] Number of parameters of the model, equal to the number of edges in the NN graph. [d] Difference between validation and training error.

Table ESI4. Training results [eV] for E_{exc} using **QBB** input string.[a]

Approach	Training error	Validation error	Test error	Overtraining[b]
Transfer learning (TL)	0.25 ± 0.05	0.37 ± 0.04	0.38 ± 0.02	0.12
TL + He initialization	0.24 ± 0.04	0.38 ± 0.05	0.38 ± 0.03	0.14
TL + Batch normalization	0.28 ± 0.02	0.40 ± 0.05	0.36 ± 0.03	0.12

[a] Training with transfer learning over 500+5000 epochs, learning rate 5·10^{-3}, using a hidden layer of 7 nodes in the first phase and an additional layer of 7 nodes in the second phase. Results are averages and standard deviation over 10 runs. [b] Difference between validation and training error.

ESI4.2 Prediction run results

Detailed results of the prediction runs are provided in Tables ESI5 and ESI6.

Table ESI5. Training and prediction results for G_{rel} using four different quasi-binary input strings. Results are averages and standard deviation in kcal·mol^{-1}, for 10 training/prediction runs.[a]

Input strings	N_{In}[b]	N_{Pars}[c]	Training error	Validation error	Test error	Over-training[d]	Prediction error[e]	MAE[f]
QBB	26	238	5.8 ± 0.7	8.8 ± 0.9	9.2 ± 1.0	3.0	9.1 ± 0.9	25.0
QBF	17	175	6.2 ± 0.7	8.6 ± 0.8	8.3 ± 0.8	2.4	8.5 ± 0.8	21.6
QBS	19	189	6.6 ± 0.8	8.9 ± 1.6	8.9 ± 1.7	2.3	9.1 ± 1.6	27.2
MVS	11	133	8.7 ± 1.0	11.2 ± 2.0	11.2 ± 2.1	2.5	11.3 ± 2.1	29.2

[a]Training over 5000 epochs using two intermediate layers of 7 nodes each except where noted, learning rate 1·10^{-2}. [b]Number of input nodes. [c]Number of parameters of the model, equal to the number of edges in the NN graph. [d]Difference between validation and training error. [e]Calculated as RMSD of the joint validation and test sets. [f]Maximum absolute prediction error.

Table ESI6. Training and prediction results for E_{exc} using four different quasi-binary input strings. Results are averages and standard deviation in eV, for 10 training/prediction runs.[a]

Input strings	N_{In}[b]	N_{Pars}[c]	Training error	Validation error	Test error	Over-training[d]	Prediction error[e]	MAE[f]
QBB	26	238	0.25 ± 0.05	0.37 ± 0.04	0.38 ± 0.02	0.12	0.38 ± 0.02	1.05
QBF	17	175	0.24 ± 0.03	0.32 ± 0.03	0.34 ± 0.03	0.08	0.33 ± 0.03	0.83
QBS	19	189	0.24 ± 0.02	0.35 ± 0.04	0.37 ± 0.03	0.11	0.36 ± 0.03	1.04
MVS	11	133	0.30 ± 0.03	0.36 ± 0.02	0.38 ± 0.02	0.06	0.37 ± 0.02	1.29

[a]Training with transfer learning over 500+5000 epochs using a hidden layer of 7 nodes in the first phase and an additional layer of 7 nodes in the second phase; learning rate 5·10^{-3}. [b]Number of input nodes. [c]Number of parameters. [d]Difference between validation and training error. [e]Calculated as RMSD of the joint validation and test sets. [f]Maximum absolute prediction error.

ESI4.3. Training curves for representative prediction runs

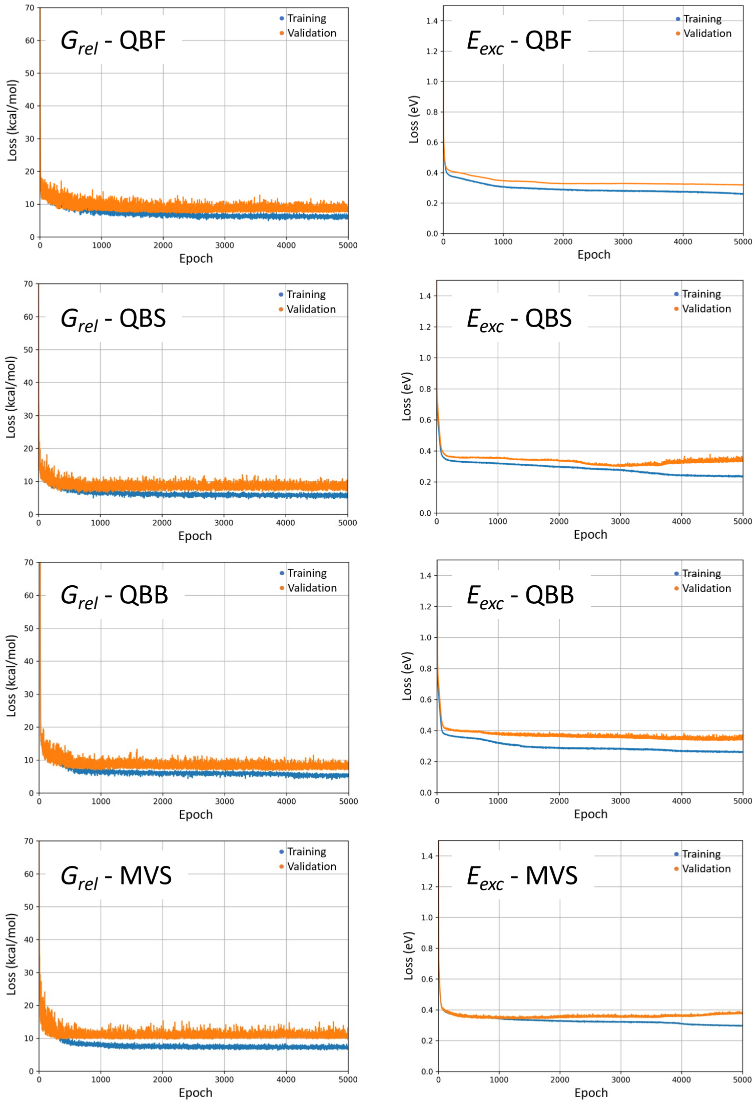


Figure ESI3. Training curves (mean training and validation loss over each epoch) for one representative prediction run with every fingerprint model.

ESI5. Code and data availability

The data set and code for fingerprint generation and NN training prediction are provided in the Electronic Supporting Information and are publicly available in GitHub (https://github.com/llblancafort/dhi_dimers_nn). The data set is contained in file dimer_summary.dat, a three-column file containing the dimer names following the nomenclature from Ref. [1], and the G_{rel} and E_{exc} endpoints for each dimer. Python scripts are provided for fingerprint generation (*input_layer_generator.py*) and for NN training and prediction for each endpoint (*NN_G_rel.py* and *NN_E_exc.py*).

Fingerprint generation. The *input_layer_generation.py* script generates five files with the MVR, MVS, QBF, QBB and QBS input sets.

Training and prediction. Run the *NN_G_rel.py* or *NN_E_exc.py* script passing the name of the input set file as argument, *eg* `python NN_G_rel.py MVS.out`. The *NN_G_rel.py* script generates the following files:

- *NN_G_rel_names.txt*: Names of dimers in the validation and tests sets.
- *NN_G_rel.txt*: Final value of training, validation and test loss.

- *NN_G_rel_DFT_energies.txt*: DFT energies of the dimers in the validation and test sets, following the order of *NN_G_rel_names.txt*.
- *NN_G_rel_predict.txt*: Predicted dimer energies for the validation and test sets, following the order of *NN_G_rel_names.txt*.
- *NN_G_rel.csv*: History of training and validation losses for each epoch.
- *NN_G_rel.png*: Plot of *NN_G_rel.csv* contained in the `*images/NN_G_rel*` directory.

The *NN_E_exc.py* script generates analogous files with `*G_rel*` replaced by `*E_exc*`.

ESI6. Supporting Information references

1. J. Wang and L. Blancafort, *Angew. Chem. Int. Ed.*, 2021, **60**, 18800-18809.
2. A. V. Marenich, C. J. Cramer and D. G. Truhlar, *The Journal of Physical Chemistry B*, 2009, **113**, 6378-6396.
3. X. Glorot and Y. Bengio, presented at the Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, Proceedings of Machine Learning Research, 2010.
4. A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow : Concepts, Tools, and Techniques to Build Intelligent Systems*, O'Reilly Media, Inc., Sebastopol, United States, 2019.
5. W. McKinney, *Proceedings of the 9th Python in Science Conference*, 2010, DOI: 10.25080/Majora-92bf1922-00a, 56-61.
6. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, *J. Mach. Learn Res.*, 2011, **12**, 2825-2830.
7. M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Craig, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenber, M. Wicke, Y. Yu and X. Zheng TensorFlow: Large-scale machine learning on heterogeneous systems, 2015.
8. J. D. Hunter, *Comput. Sci. Eng.*, 2007, **9**, 90-95.
9. T. Williams and C. Kelley gnuplot 5.4, 2021.
10. S. Ioffe and C. Szegedy, presented at the Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37, Lille, France, 2015.
11. K. He, X. Zhang, S. Ren and J. Sun, presented at the 2015 IEEE International Conference on Computer Vision (ICCV), 7-13 Dec. 2015, 2015.