Electronic Supplementary Material (ESI) for Physical Chemistry Chemical Physics.

Supporting Information

Machine-learning-assisted molecular design of phenylnaphthylamine-type antioxidants

Shanda Du, [a] Xiujuan Wang, [b] Runguo Wang, [a] Ling Lu, [a] Yanlong Luo, [c] Guohua You *[d] and Sizhu Wu *[a]

[a] State Key Laboratory of Organic–Inorganic Composites, Beijing University of Chemical Technology, Beijing 100029, China. E-mail: wusz@mail.buct.edu.cn

[b] Key Laboratory of Rubber–Plastics, Ministry of Education/Shandong Provincial Key Laboratory of Rubber–Plastics, Qingdao University of Science & Technology, Qingdao 266042, China.

[c] College of Science, Nanjing Forestry University, Nanjing 210037, China.

[d] College of Information Science and Technology, Beijing University of Chemical Technology, Beijing 100029, China. E-mail: yough@mail.buct.edu.cn

S1. Quantum mechanics simulation

The complete quantum mechanical computation process could be regarded as solving the Kohn-Sham (KS) equation.[1] The generalized gradient approximation (GGA) class with Perdew-Burke-Ernzerhof (PBE) specific functional was employed as exchange-correlation potential type, and the triple numerical plus polarization (TNP) with high accuracy was applied to the atomic orbital basis set.[2, 3] Concerning free radicals, the multiplicity was doublet with spin unrestricted. In order to accelerate computation convergence, the self-consistent field with the threshold convergence of 10^{-6} au on the energy and 0.005 Hartree of smearing in orbital occupancy were selected.

All structures, including antioxidant (AH), antioxidants free radical (A·), and hydrogen atom (H·), as shown in Fig. S1, were optimized first by the geometry optimization task with convergence tolerance of 10^{-5} au for energy, 0.002 Hartree/Å for maximum force, and 0.005 Å for maximum displacement. The calculated energy using the DMol³ module is electronic energy at 0 K. A thermodynamic cycle is needed to determine the energy at other temperatures (*T*), and Fig. S1 shows the cycle with the example of N-phenyl-1-naphthylamine (PANA).[4]

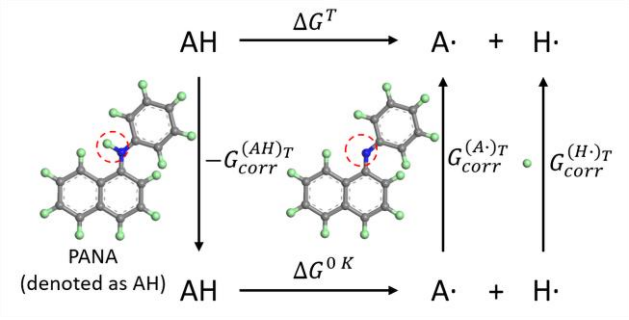

Fig. S1 Thermodynamic cycle for N-phenyl-1-naphthylamine (PANA). The gray, blue, and green spheres represent C, N, and H atoms, respectively.

Thus, the *BDE* value can be obtained from following Equation S1:

$$\Delta G^T = \left[E(\mathrm{A}\cdot) + G_{\mathrm{corr}}^{(\mathrm{A}\cdot)_T}\right] + \left[E(\mathrm{H}\cdot) + G_{\mathrm{corr}}^{(\mathrm{H}\cdot)_T}\right] - \left[E(\mathrm{AH}) + G_{\mathrm{corr}}^{(\mathrm{AH})_T}\right] \quad \text{(S1)}$$

where ΔG^{T} is the *BDE* value at a certain temperature *T*. $E(\mathrm{A}\cdot)$, $E(\mathrm{H}\cdot)$, and $E(\mathrm{AH})$ represent the energy of A·, H·, and AH at 0 K, respectively. $G_{\mathrm{corr}}^{(\mathrm{A}\cdot)_T}$, $G_{\mathrm{corr}}^{(\mathrm{H}\cdot)_T}$, and $G_{\mathrm{corr}}^{(\mathrm{AH})_T}$ are energy corrections at specific temperature *T*.

S2. Molecular dynamics simulation

Molecular dynamics (MD) simulation can calculate the future velocities and positions of atoms in a specific forcefield based on the initial state by solving the Newton equation.[5] For this reason, MD simulations were performed using Materials Studio software to investigate antioxidants activity in this study. Before implementing the MD process, the AH and diisooctyl sebacate (DIOS)/AH amorphous models were generated at 298 K, and each DIOS/AH model contains five antioxidants and 27 DIOS molecules. Pursuant to the previous section, 302 kinds of antioxidant molecule models have been constructed, which means 302 * 2 = 604 amorphous cells would be simulated.

After constructing the amorphous cell model, the simulation was carried out in two phases, i.e., the optimization and equilibration phases.[6] The geometry optimization task was performed with fine quality (i.e., the cutoff distance was set to 15.5 Å) and smart algorithm in the optimization phase. Then, the optimized amorphous models were used for annealing under the NVE ensemble. The anneal task outputs the low energy structures by periodically increasing and decreasing temperature between 300 K and 500 K.[7] The annealed amorphous models were equilibrated via 500 ps of NVT and 1000 ps of NPT ensemble simulation at 298 K with a time step of 1 fs in the equilibration phase. In the above description, *N* represents constant number of particles, *V* represents constant volume, *E* represents constant energy, *T* represents constant temperature, *P* represents constant pressure. The *ab* into Condensed-phase Optimized Molecular Potentials for Atomistic Simulation Studies (COMPASS) force field with high quality was utilized for simulation.[8] The Berendsen barostat and Anderson thermostat were set to maintain the pressure and temperature, respectively.[6, 9] In addition, the electrostatic interactions were evaluated by the Ewald summation method with an accuracy of 0.0001 kcal·mol^{-1}, and the non-bond van der Waals interactions were truncated using an atom based cutoff distance of 15.5 Å. The equilibrated models could be obtained after these processes have been completed.

S3. Molecular connectivity index

The *m*-th order valence molecular connectivity index (MCI) ${}^{m}\chi_{\mathrm{p}}^{\mathrm{v}}$ is defined as following equation:[10, 11]

$$ {}^{m}\chi_{\mathrm{p}}^{\mathrm{v}} = \sum_{l=1}^{n} \prod_{i=1}^{m+1} (\delta_i^{\mathrm{v}})_l^{-0.5} \tag{S2} $$

where *m* is the order of the connectivity index, p is paths type, *n* is the number of *m*-th paths type fragments, δ_i^{v} is the valence connectivity degree of atoms, which could be expressed as follows:

$$ \delta_i^{\mathrm{v}} = \frac{Z^{\mathrm{v}} - h}{Z - Z^{\mathrm{v}} - 1} \tag{S3} $$

where Z^{v} is the number of valence electrons in the *i*-th atom, Z is the atomic number, *h* is the number of hydrogen atoms connecting to the *i*-th atom. From the above description, ${}^{m}\chi_{\mathrm{p}}^{\mathrm{v}}$ is related to the type and connection way of atoms. Hence, third, fourth, and fifth-order MCI (${}^{3}\chi_{\mathrm{p}}^{\mathrm{v}}$, ${}^{4}\chi_{\mathrm{p}}^{\mathrm{v}}$, ${}^{5}\chi_{\mathrm{p}}^{\mathrm{v}}$) could be employed to describe various connections of group descriptors in the skeleton of phenylnaphthylamine antioxidants.

S4. Data sets normalization

The normalization not only eliminates the dimension influence of features but also accelerates the convergence of models.[12] The original data sets were normalized into [0,1] using the following equation:

$$ x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{S4} $$

where x' is normalized value, x is the original value of input (output) value, $x_{\min}$ and $x_{\max}$ are minimum and maximum of input (output) value, respectively. The obtained normalized data sets would be used for machine learning model construction.

S5. Determination of hidden layer neurons

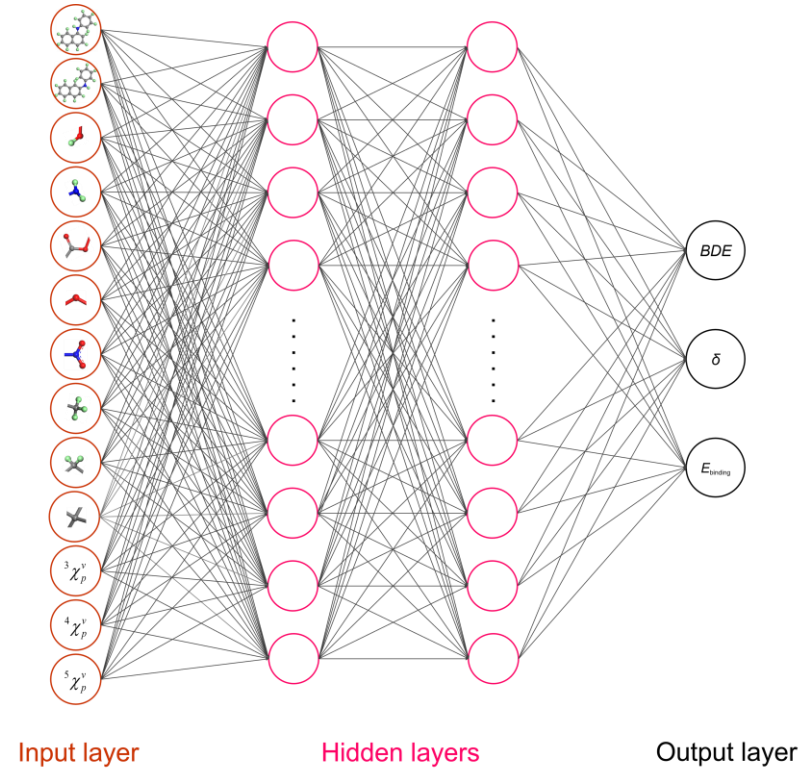

Fig. S2 Structure diagram of the fully connected artificial neural network (ANN) model. Every circle represents a neuron, and each neuron is connected to the neurons of the next layer while there is no link between neurons in the same layer. That means the information is only transmitted between layers instead of exchanging information within the layer.

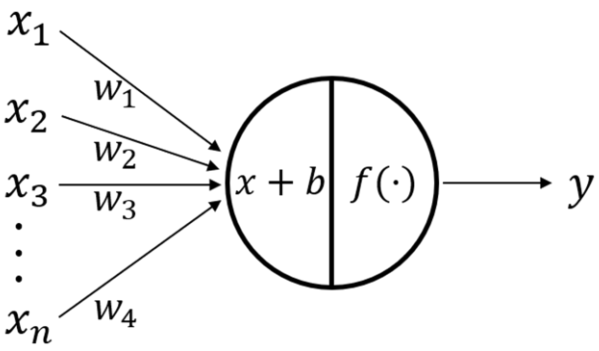

Fig. S3 Structure diagram of the typical neuron. Neurons will receive the output signal from the antecedent layer. Its net input contains the sum of weighted output ($x = \sum_{i=1}^{n} w_i x_i$, w_i is the weight matrix, x_i is the output from the previous layer) of the previous layer and the neuron bias (b). Then a new output ($y = f(x + b)$) is generated by the activation function ($f(\cdot)$), and y is passed to the next layer of neurons as a new x_i. The activation function of hidden neurons is the hyperbolic tangent function (tanh) which is widely used in machine learning.

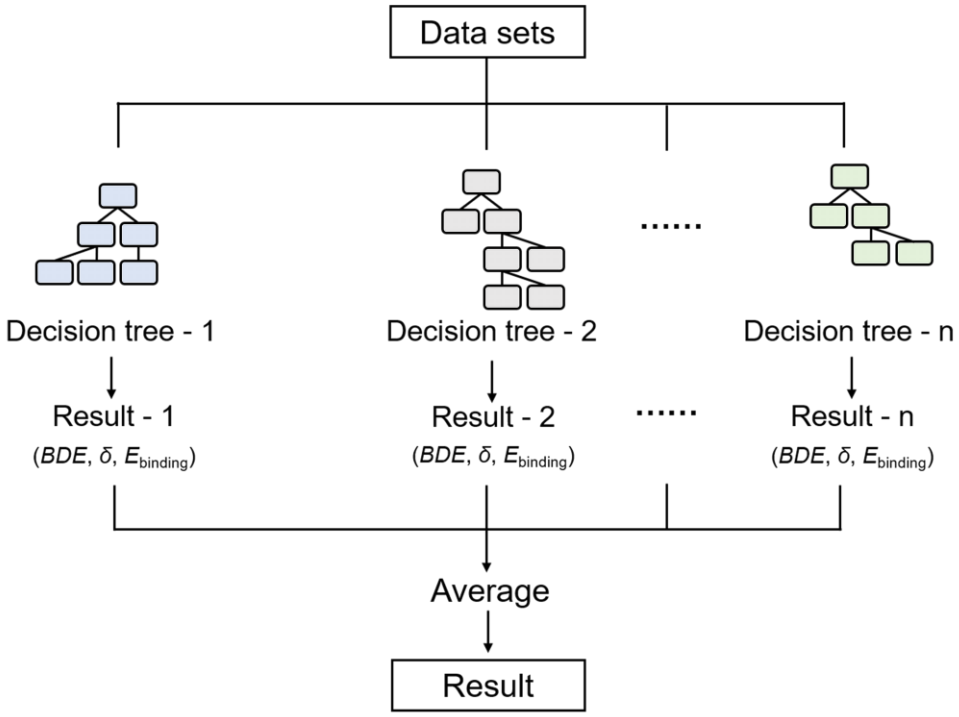

Fig. S4 Structure diagram of the random forest model. The final output value is based on the average values estimated by many decision trees.

The maximum number of hidden neurons (H) could calculate by the following two empirical equations:[13, 14]

$$H = \sqrt{n + l} + a \tag{S5}$$

$$H < \frac{D - 1}{n - 2} \tag{S6}$$

where n is the number of input neurons, l is the number of output neurons, a is the tuning parameter (often located within 1–10), and D is the number of data points for artificial neural network (ANN) calibration. Here, the calculated H corresponding to Equation S5 is 5–14, and H is less than 22 by Equation S6. Accordingly, the ANN models containing 1–20 hidden neurons were established, and the

average relative error (*ARE*) was used to estimate the predicting accuracy.

When two hidden layers keep the same number of neurons, Fig. S5 shows the relationship between *ARE* and the number of neurons in hidden layers. Note that all sets refer to the statistics of the training, validation, and test set together. It can be seen from these curves that *ARE* of each data set shows a decreasing trend with the increase of the number of hidden neurons. When the number reaches 12, the change of *ARE* tends to be flat, which demonstrates the number of hidden neurons is gradually saturated, and the excessive hidden neurons would cause overfitting. Combined with the mean square error (*MSE*) curve (Fig. S6) in ANN model, we further analyzed the models with more than 12 hidden neurons. When there are 15 hidden neurons, the validation and test set show overfitting, which is undesirable. When it is reduced to 13, the overfitting phenomenon is inconspicuous, and the *MSE* difference among the training, validation, and test sets is also decreased. Therefore, each hidden layer containing 13 neurons was determined.

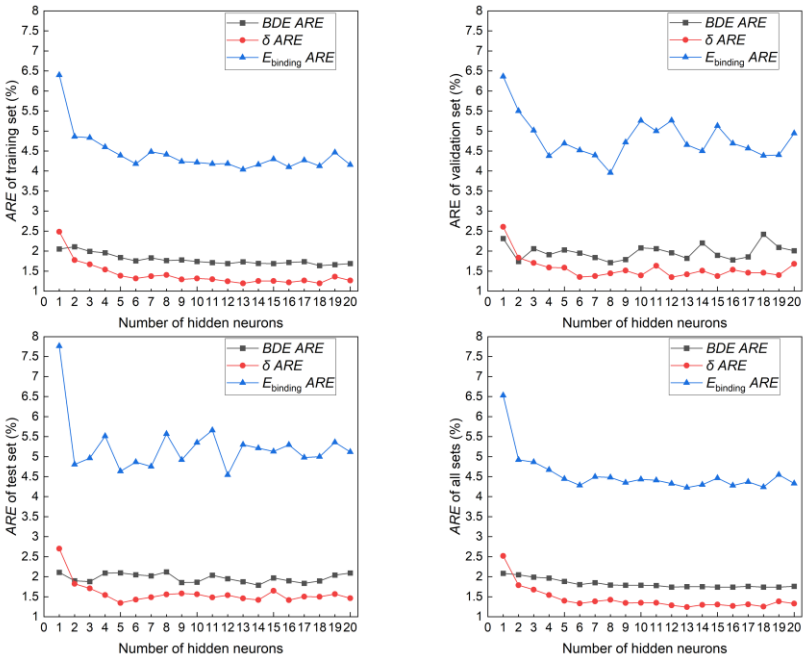

Fig. S5 The average relative error (*ARE*) curves of antioxidant parameters in (a) Training set, (b) Validation set, (c) Test set, and (d) All sets. Here, the bond dissociation energy, solubility parameter, and binding energy are denoted as *BDE*, *δ*, and $E_{binding}$, respectively. *BDE ARE*, *δ ARE*, and $E_{binding}$ *ARE* represent the *ARE* of *BDE*, *δ*, and $E_{binding}$, respectively.

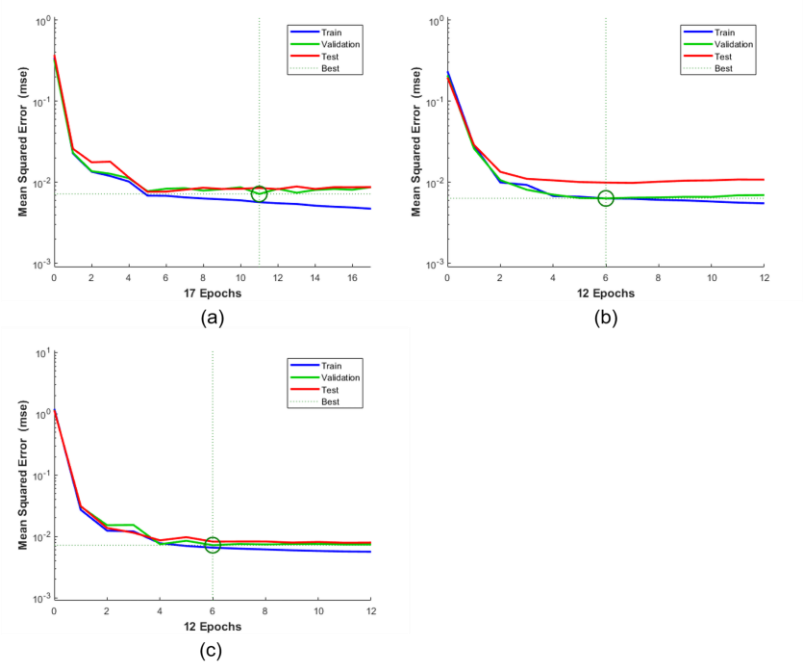

Fig. S6 The mean square error (*MSE*) curves of ANN models with (a) 15, (b) 14, and (c) 13 hidden neurons.

References

1. İ. Temizer, *Comput. Methods Appl. Mech. Engrg.*, 2021, **386**, 114094.
2. H. Jin, J. Zhang, W. Zhang, Y. Zhang, S. Ma, Y. Du, J. Qin and Q. Wang, *J. Phys. Chem. Solids*, 2022, **161**, 110453.
3. Y. Maekawa, K. Sasaoka and T. Yamamoto, *Appl. Phys. Express*, 2019, **12**, 115001.
4. W. Zheng, Y. Wu, W. Yang, Z. Zhang, L. Zhang and S. Wu, *J. Phys. Chem. B*, 2017, **121**, 1413-1425.
5. J. Xu, X. Chen, G. Yang, X. Niu, F. Chang and G. Lacidogna, *Constr. Build. Mater.*, 2021, **312**, 125389.
6. J. Gupta, C. Nunes, S. Vyas and S. Jonnalagadda, *J. Phys. Chem. B*, 2011, **115**, 2014-2023.
7. S. Sharma, S. K. Tiwari and S. Shakya, *Def. Technol.*, 2021, **17**, 234-244.
8. H. Sun, P. Ren and J. R. Fried, *Comput. Theor. Polym. Sci.*, 1998, **8**, 229-246.
9. M. Khalili, A. Liwo, A. Jagielska and H. A. Scheraga, *J. Phys. Chem. B*, 2005, **109**, 13798-13810.
10. J. R. Baker, J. R. Mihelcic and A. Sabljic, *Chemosphere*, 2001, **45**, 213-221.
11. A. Mozrzymas, *J. Solution Chem.*, 2013, **42**, 2187-2199.
12. J. Cai, X. Chu, K. Xu, H. Li and J. Wei, *Nanoscale Advances*, 2020, **2**, 3115-3130.
13. T. A. Albahri, *Fluid Phase Equilib.*, 2014, **379**, 96-103.
14. S. Pang, J. Luo and Y. Wu, *Macromol. Theory Simul.*, 2020, **29**, 1900063.