Electronic Supplementary Information (ESI)

A novel unambiguous strategy of molecular feature extraction in machine learning assisted predictive models for environmental properties

Zihao Wang,[a] Yang Su,[a] Saimeng Jin,[a] Weifeng Shen,*[a] Jingzheng Ren,[b] Xiangping Zhang[c] and James H. Clark[d]

[a] *School of Chemistry and Chemical Engineering, Chongqing University, Chongqing 400044, People's Republic of China*

[b] *Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University, Hong Kong SAR, People's Republic of China*

[c] *Beijing Key Laboratory of Ionic Liquids Clean Process, CAS Key Laboratory of Green Process and Engineering, Institute of Process Engineering, Chinese Academy of Sciences, Beijing, 100190, People's Republic of China*

[d] *Green Chemistry Centre of Excellence, University of York, York YO105DD, UK*

Corresponding Author: *(W.S) E-mail: shenweifeng@cqu.edu.cn

Table of Contents

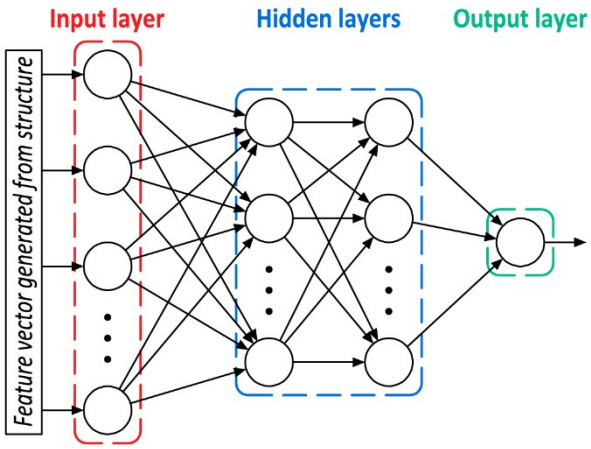

Fig. S1 The structure of the four-layer neural network.

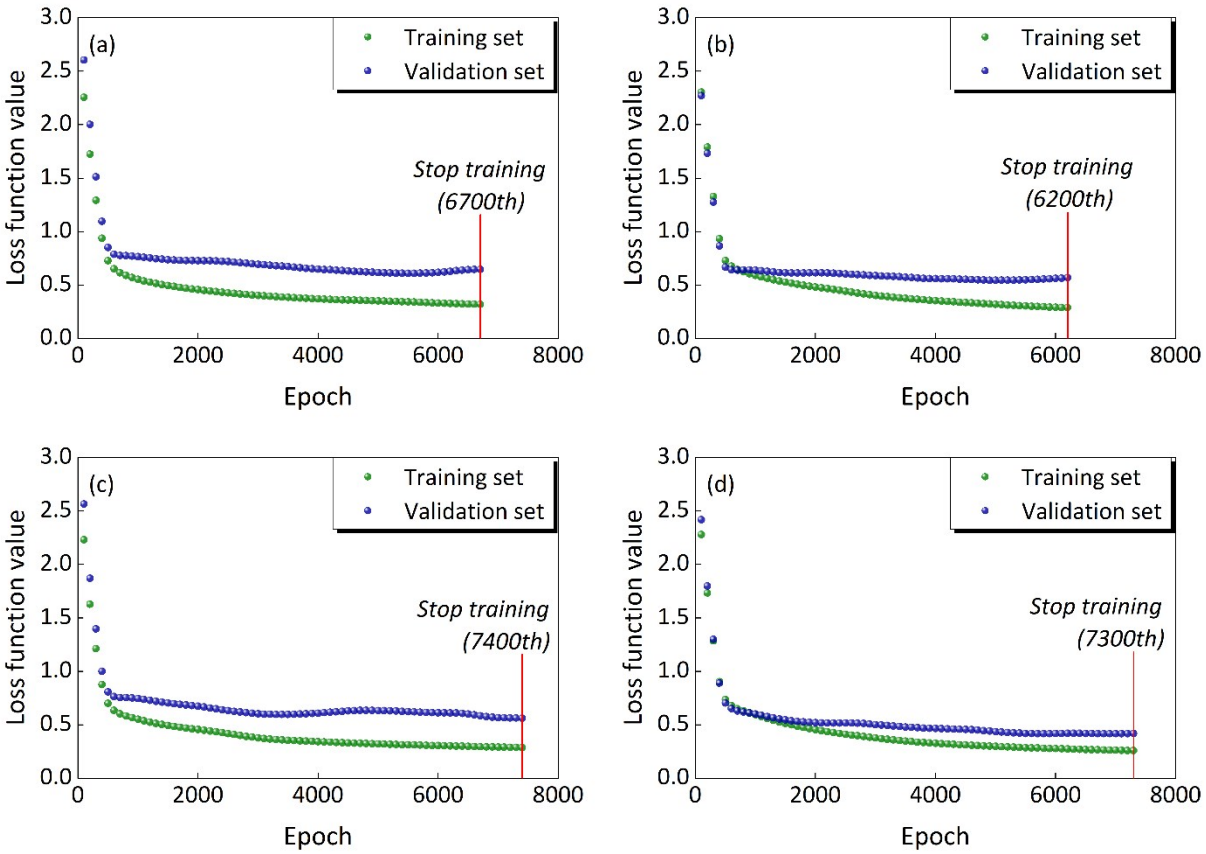

Fig. S2 The learning curves for training model using (a) feature vector under random sampling (**Scheme 1**); (b) feature vector under cluster sampling (**Scheme 2**); (c) feature vector supplemented with PBF under random sampling (**Scheme 3**); (d) feature vector supplemented with PBF under cluster sampling (**Scheme 4**).

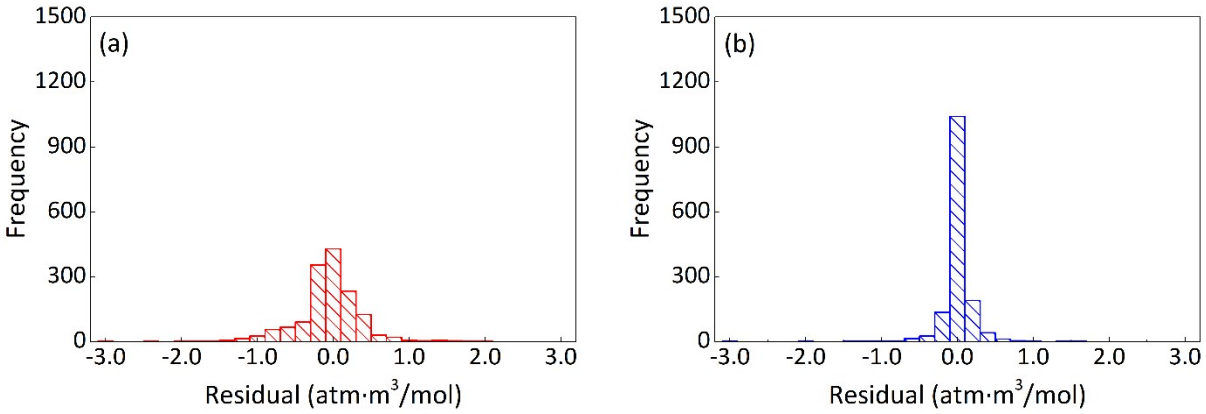

Fig. S3 The residual distributions of log*HLC* values estimated with (a) ER model and (b) NN model.

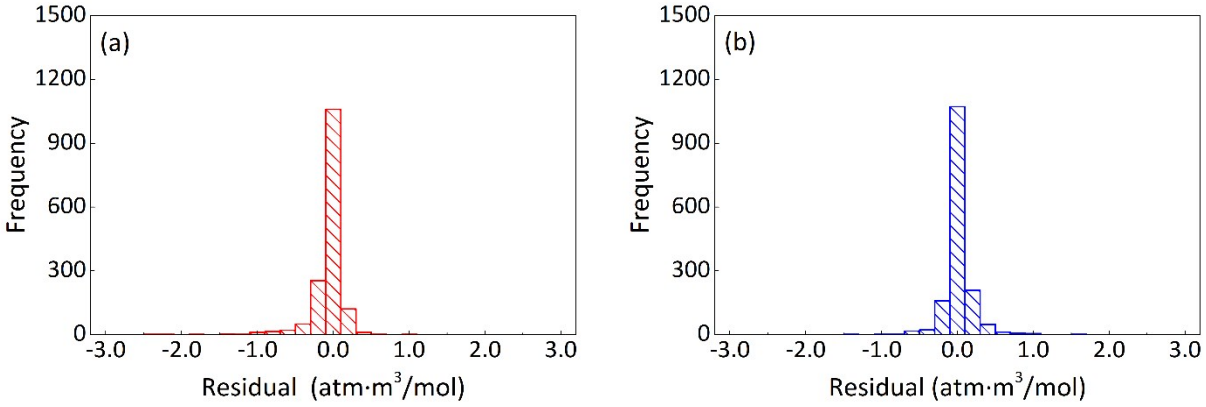

Fig. S4 The residual distributions of log*HLC* values estimated with (a) GN model and (b) NN model.

Table S1 The explanations and examples for the characters involved in identifiers.

Symbol	Explanation	Example
A	Atom in aliphatic compound	"[C]" Carbon atom in aliphatic compound
a	Atom in aromatic compound	"[c]" Carbon atom in aromatic compound
H*n*	Atom with n hydrogen atoms	"[CH3]" Carbon atom with three hydrogen atoms
+ (inside [ ])	Atom with a positive charge	"[N+]" Nitrogen atom with a positive charge
– (inside [ ])	Atom with a negative charge	"[N–]" Nitrogen atom with a negative charge
– (outside [ ])	Single bond not in rings	"[C]–" Carbon atom with a single bond not in rings
.	Single bond within rings	"[C]." Carbon atom with a single bond within rings
=	Double bond not in rings	"[C]=" Carbon atom with a double bond not in rings
:	Double bond within rings	"[C]:" Carbon atom with a double bond within rings
#	Triple bond not in rings	"[C]#" Carbon atom with a triple bond not in rings
*	Aromatic bond	"[c]*" Carbon atom with an aromatic bond
@	Atom is a R-chirality centre	"[C@]" Carbon atom is a R-chirality centre
@@	Atom is a S-chirality centre	"[C@@]" Carbon atom is a S-chirality centre
|	Separator	-

Table S2 The equations for activation functions in the ANN model.

Activation function	Equation
sigmoid	$f(x)=\frac{1}{1+e^{-x}}$
softplus	$f(x)=\ln(1+e^{-x})$

Table S3 The molecular features represented by identifiers for describing molecular structures.

Chemical element	Molecular features expressed with identifiers
Carbon (C)	[CH0]-#; [CH0]----; [CH0]--..; [CH0]--=; [CH0]-...; [CH0]-.:; [CH0]=..; [CH0|]=..; [CH0]==; [CH0|@]--..; [CH0|@]-...; [CH1]#; [CH1]---; [CH1]-..; [CH1]-=; [CH1]...; [CH1].:; [CH1|@@]---; [CH1|@@]-..; [CH1|@]---; [CH1|@]-..; [CH1|@]...; [CH2]--; [CH2]..; [CH2|]=; [CH3]-; [CH4]; [cH0]***; [cH0]**.; [cH0]-**; [cH0]=**; [cH1]**
Oxygen (O)	[OH0|-]-; [OH0]--; [OH0]..; [OH0]=; [OH1]-; [oH0]**
Nitrogen (N)	[NH0|+]--=; [NH0]#; [NH0]---; [NH0]-=; [NH1]--; [NH1]..; [NH2]-; [nH0|]**; [nH1]**
Phosphorus (P)	[PH0]---=
Sulfur (S)	[SH0]--; [SH0]--==; [SH0]..; [SH0]=; [SH1]-; [sH0]**
Halogen	[FH0]-; [ClH0]-; [BrH0]-; [IH0]-

Table S4 The weight and bias matrixes of the predictive model.

Layer	Item	Matrix
Hidden layer 1	Weight [15*59]	[[-0.5551, -0.1243, -0.1166, -0.5558, 0.3310, -0.1169, 0.1428, -0.3357, 0.1969, -0.0548, -0.3600, -2.1200, 0.5740, 0.4849, 0.4082, -0.1036, 0.0099, -0.3805, 0.4664, -0.3595, 0.1756, -0.1791, 0.4811, 0.3586, 0.2290, 0.0292, -0.0973, 0.1650, -0.2006, -1.2431, -0.8260, -0.1069, -0.7904, -0.5551, 0.2309, 0.4153, -0.3225, 0.9519, 0.0653, 0.3494, -0.7509, -0.3488, 0.6242, -1.3429, -0.4255, -2.4164, -0.1656, -0.3714, 0.2089, 0.0738, -0.6320, -0.2532, 0.9213, -0.0461, -0.8731, -0.8667, 0.2339, -0.7594, -0.1683], [-1.1445, 0.7524, 0.6763, 2.1176, 1.1718, 0.9026, 0.8898, 0.0768, 0.6349, 0.6817, 0.2186, 0.1175, 0.1142, 0.5652, 0.2433, 0.4463, 0.7342, -1.8364, 0.0221, 0.2782, 0.1124, -0.0040, 0.1612, -0.3896, 0.4234, -0.0886, -0.4974, -0.1524, -1.2312, -1.2951, -0.3075, -0.2186, 1.0099, 0.6045, -0.7052, 1.0142, 0.4080, 0.9281, -0.4038, -1.0711, -0.8105, -0.4205, 0.6156, 1.6647, 0.8063, 0.1876, 0.1640, 0.7760, -0.4733, -0.8753, -0.3390, -0.7104, -0.6790, -0.1183, -0.6014, 1.9527, 0.7430, 0.4879, 0.2809], [-0.2627, 0.4235, 0.5918, -0.2294, 0.4176, 0.3365, 0.6405, 0.0755, 0.5571, 0.3744, 0.0032, -0.9396, 0.1632, -0.1640, -0.0928, -0.0580, 0.5227, 0.6318, 0.0442, -0.1895, 0.7351, -0.0343, 0.3366, -0.0715, -0.0968, 0.0161, 0.0148, -0.1321, 0.4513, 0.2418, 0.0866, 0.5171, -0.8276, 1.3685, 0.5662, 0.5331, 0.4220, 0.5517, 0.4325, -0.2188, -0.6198, -1.1576, 0.1476, 0.6086, 0.0700, -2.9789, 0.1630, -1.8153, 1.6312, -0.0590, -0.3019, -0.0790, -0.8150, 0.1590, -1.7631, -0.4307, -0.2310, -0.3035, -0.2887], [-0.0404, -0.2396, -0.3320, 0.9721, -0.1078, 0.1860, 0.4755, -0.0737, 0.2183, -0.3358, -0.0721, 1.1476, -0.3615, -0.0289, -0.3090, 0.2757, -0.3534, 0.5233, -0.3922, 0.6416, 0.1409, 0.0135, 0.3476, -0.3701, 0.2537, -0.6910, -0.1489, -0.0012, 0.0427, 0.3067, 0.2349, -0.2702, -1.0385, -1.1897, -1.0251, -1.1899, -0.2675, -0.3693, -0.2120, 0.8304, 0.2073, 0.0292, 0.2765, 0.1960, -0.0646, 2.3980, -0.0043, 1.9410, 0.0985, 0.1996, 0.4118, 0.0508, 0.8238, 0.4151, 1.7575, -1.7015, -0.1624, 0.2443, -0.2081], [-0.1021, 0.4834, 0.5327, 0.0847, 0.4024, 0.6320, -0.3237, 0.1111, 0.6100, 0.6859, -0.1004, -1.8436, 0.2495, -0.1387, -0.1928, -0.2406, 0.4783, 0.4969, 0.2313, 0.0330, -0.1494, 0.0723, 0.1011, 0.0521, -0.1068, 0.5969, -0.0652, -0.1454, -0.6852, -0.7167, -0.2115, 0.4746, -0.2392,

0.8679, -0.4110, 0.7407, 0.5316, -0.1730, 0.2888, 0.5219, -0.7025, 0.5011, 0.2885, -1.1689, 0.2881, 2.6972, 0.0870, -1.7867, 0.1572, 0.2115, -0.2453, 1.6128, -1.1208, 0.0208, 2.8755, 1.1858, -0.4252, -0.5613, 0.0316], [0.5346, -0.5644, -0.6866, 0.8249, -0.2065, 0.7422, 0.4439, -0.0180, 0.1128, -0.5536, -0.5030, -1.5883, 0.6868, -0.6718, -0.6468, -0.1098, -0.6980, -0.0767, 0.1750, -0.0489, -0.0374, 0.2012, 0.0154, -0.0546, -0.1679, 0.8877, -0.2237, 0.0968, 0.9115, 0.4850, -1.2501, -0.7286, -0.2516, -0.4399, -0.3929, -0.4775, -0.3900, -1.2794, -0.8103, -0.2022, -0.9891, -1.1330, 0.2496, -0.7981, -0.0723, -1.5741, -0.1795, -0.5349, 0.4203, -1.4506, 0.1580, 0.0268, 1.0863, -0.3693, 0.3255, -1.2440, 0.5154, 0.4345, -0.7041], [-0.5225, -0.5618, -0.7030, 0.8656, -0.1465, 0.9507, -0.6585, -0.1280, -0.2756, -0.5665, -0.3250, -1.0291, -0.3102, -0.3938, -0.4328, -0.3916, -0.6076, 1.6398, -0.7688, 0.0049, -0.4531, -0.0792, -0.0282, -1.1256, -0.2402, 0.8741, -0.7416, 0.1687, 0.9929, 0.7413, -2.2218, -2.1078, 0.7739, 0.1715, -0.2899, -0.4112, -0.5721, 1.0882, -1.9529, -0.4880, -1.4426, -1.4040, -1.8052, -2.3323, -0.3236, -0.7727, -0.1985, -0.7354, 0.0348, -1.4994, 0.0247, 0.2599, 0.6627, 0.8779, -2.8069, -1.4799, -0.4597, -0.5257, -0.3509], [-0.1823, -0.6104, -0.5902, -1.0433, 0.1854, -1.2936, 0.7351, -0.0700, -0.3373, -0.6544, -0.4590, -0.5952, -0.1619, 0.0885, -0.7355, 0.0522, -0.6571, 0.2418, 0.5306, 0.3537, 0.7576, -0.1558, -0.0668, 0.1552, -0.3325, 0.8494, -0.0681, 0.2262, -1.0509, -0.4190, 0.3358, 0.0760, -1.0562, -2.5233, 0.2815, -0.6333, -0.3256, -0.9827, 0.0924, -0.9850, -0.3811, -0.5731, -0.2089, -1.1051, -0.2211, 2.0265, -0.1295, -0.6091, 0.3520, -2.2550, -0.3800, 0.2316, 0.9124, 0.0516, 1.4810, -1.2615, -0.5774, 0.5311, 0.4195], [0.3297, -0.6270, -0.5184, -0.0001, -0.1344, -0.2602, 0.7228, -0.1293, -0.0492, -0.4949, -0.0068, 1.4706, -0.0569, 0.0216, 0.4051, -0.1639, -0.4694, -0.6655, -0.4070, -0.3767, 1.1551, 0.1881, -0.1311, -0.1761, -0.2408, -0.5054, 0.2797, 0.1910, 0.0092, -0.4521, 0.0262, -1.4069, 0.4183, -0.1883, -1.7836, -0.1853, -0.3275, 0.0477, -1.5211, 0.5493, -0.0980, 1.1636, -1.6350, 0.3507, -0.1037, -0.6095, -0.1995, -1.1813, -0.5232, -0.0838, -0.0824, 0.4686, 1.0359, -0.0386, 2.1823, 1.1994, -0.3637, -0.5591, 0.4144], [0.9658, 0.7169, 0.7267, -0.5248, -0.0334, -0.7806, 0.4960, 0.1686, 0.0065, 0.5741, 0.5074, 0.8497, 0.3339, -0.1080, 0.4714, -0.1064, 0.7668, 1.7776, 0.4220, 0.2435, 0.0023, -0.1212, 0.2060, 0.3327, -0.2409, -0.8629, -0.3322, -

0.2047, 1.4291, 0.4418, 0.4233, -0.7371, -0.3256, -0.1920, -0.2943, 0.6833, 0.5674, 0.6882, -0.8769, 1.7714, 1.8984, 2.0298, 1.1647, 0.8849, 0.9083, 0.5671, 0.0745, 0.7590, 0.4507, 0.5049, 0.2421, 0.1654, -1.0934, 0.7304, 0.5843, 1.1869, 0.2714, 0.5947, -0.1383], [0.2617, 0.6616, 0.7318, 1.1551, 0.5321, 1.1331, 1.6102, 0.1766, 0.5186, 0.6500, 0.4423, -0.5550, 0.7559, 0.6829, 1.0318, 0.8967, 0.7114, 0.0769, 0.0627, -0.0085, 1.3713, -0.0722, 0.4316, -0.2189, 0.1574, -0.4041, -0.1659, -0.3109, 1.2799, 1.0121, 0.7968, -0.0813, 0.2605, -0.3193, -1.5810, 1.2700, 0.5897, -0.1661, -0.2012, 2.7981, 2.1122, 2.4654, 1.3117, 0.7278, 0.7463, 0.0840, 0.0507, -2.5999, 0.9077, -0.4482, 0.3615, -0.0118, -1.1613, 0.2723, 1.6361, -0.1620, -0.3608, -0.5333, 0.6124], [-1.0274, -0.6220, -0.5760, -1.4399, 0.2951, -0.6712, 0.7101, -0.2802, 0.0481, -0.6213, -0.4712, -1.5923, -0.5857, -0.6066, 0.0713, -0.1812, -0.7511, -0.1017, -0.2065, -0.3670, -0.3517, -0.0424, 0.1196, 0.0477, -0.1896, -0.3620, 0.0324, 0.1231, 0.1490, 0.7312, -1.0898, -0.7231, -1.7364, -1.0782, 1.3815, -0.7864, -0.5625, -0.5381, -0.7928, -0.9974, -1.7343, -0.3056, -0.5720, -1.4705, -0.5455, -2.0937, -0.1039, -0.3672, -0.2063, 0.8215, -0.2877, 0.1770, 1.0974, -0.5266, -0.7838, -0.1224, 0.0540, -0.4689, 0.0733], [0.0617, 0.3720, 0.4163, -1.0678, 0.2520, -0.0619, 1.0311, 0.0393, -0.1530, 0.3998, 0.0025, -1.1511, -0.3312, -0.6319, 0.2637, 0.2115, 0.2174, -0.5546, 0.0680, 0.3705, 0.3461, 0.0003, -0.3743, 0.2349, -0.7693, -0.9561, -0.0221, -0.1356, 0.2975, -0.0103, -0.0802, -0.3659, 1.5652, 0.4389, 1.9141, 0.4440, 0.5573, -0.0965, -0.2766, 0.7411, -0.3417, -0.3733, 0.1247, 1.0812, 0.3428, 0.2053, 0.2235, -1.7108, 0.4129, -0.0103, -0.3739, 0.4919, -0.8348, -0.6363, -1.2862, 1.8661, 0.3991, -0.1274, 0.6001], [-0.1142, -0.2331, -0.1668, 0.5074, 0.5585, 0.2867, 0.1174, -0.1263, 0.0794, -0.1952, -0.7356, -0.3561, -0.2185, 0.4505, 0.1469, 0.1583, -0.2380, -0.0811, 0.2966, 0.4116, 0.2081, 0.0563, 0.0049, 0.2757, 0.4135, -0.7708, 0.2071, 0.2006, -0.1642, 0.2116, 0.3038, 1.4067, 1.1782, 1.4135, 2.0640, 1.1072, -0.3753, -1.0436, 1.1823, -0.9329, -1.6763, -0.0891, -0.9871, -0.4069, 0.5713, 1.3180, -0.0822, -0.1678, 0.6441, -0.0980, -0.3655, 0.0923, 0.6055, 0.0550, -0.2838, 0.1430, 0.2209, -0.1314, 0.3711], [0.1721, -0.2072, -0.3255, 0.2517, 1.0057, 1.2301, 0.6759, -0.0604, -0.4366, -0.4091, -0.4851, -1.3021, 0.0131, 1.3435, 0.4329, 0.7540, -0.2247, -1.3158, 0.8149, 0.7685, -0.0619, -

		0.1239, 0.1380, 0.2862, 0.1765, -0.2681, -0.3243, -0.2189, -0.7128, -0.3304, -0.9921, -1.2587, 0.1638, 0.0790, 0.7907, 1.0461, 0.0927, 0.7727, -1.0874, -2.1761, -0.7496, 1.7541, -1.1920, 1.2847, 0.4046, -5.4185, 0.0050, -1.7528, -1.3572, -1.0772, -0.2070, 0.0327, -0.0527, 0.1734, 2.5980, 2.1890, 0.3602, -0.0865, -0.3807]]
	Bias [15*1]	[[-1.0093], [-1.8459], [-1.1276], [0.4659], [-1.0576], [-0.8017], [-2.4100], [-1.3136], [0.9413], [1.7379], [1.5329], [-0.6604], [-1.6602], [-1.0659], [-1.9738]]
Hidden layer 2	Weight [6*15]	[[0.7745, -0.2409, -0.2465, 0.4620, -0.1624, 0.7049, 0.5933, 0.6288, 0.7126, -0.3740, -0.2241, 0.2953, 0.0666, 0.7033, -0.2707], [-0.2959, 0.9810, 0.7755, -0.5564, 0.7002, -0.3882, -0.1177, -0.4274, -0.5900, 0.5160, 0.3990, -0.2966, 0.7109, -0.4268, 0.9236], [-0.6980, 0.9719, 0.8850, -0.3788, 0.8093, -0.1369, -0.1718, -0.3038, -0.4573, 0.2724, 0.4433, -0.2234, 0.9396, -0.6361, 1.0899], [0.6278, -0.1884, -0.3386, 0.5629, 0.0467, 0.4708, 0.4545, 0.3323, 0.5761, -0.3871, -0.5625, 0.4270, -0.3494, 1.0080, 0.1378], [-0.6941, 1.5144, 1.2823, -0.8733, 0.6350, -0.4231, -0.1760, -0.0216, -0.7807, 0.5474, 0.4918, -0.5700, 0.9954, -0.4693, 1.1910], [1.0177, -0.3520, -0.0647, 0.6185, -0.1277, 0.6333, 0.3160, 0.5120, 0.5422, -0.6076, -0.7118, 0.3580, -0.1280, 1.1257, -0.1203]]
	Bias [6*1]	[[0.2088], [0.3252], [0.3543], [0.4004], [0.2958], [0.6652]]
Output layer	Weight [1*6]	[[0.6656, -1.0053, -0.9710, 0.8382, -1.0497, 0.7874]]
	Bias [1*1]	[[-0.2849]]

The application of the developed tool in making predictions for new compounds is illustrated in Fig. S5. The molecular structure (SMILES notation) of a new compound is used to generate the molecular feature vector. Relying on the parameters and hyper-parameters of the ANN framework determined in the model development, the predictive model (predictive tool) makes a numeric prediction and outputs a predicted value for the log*HLC* of the compound.

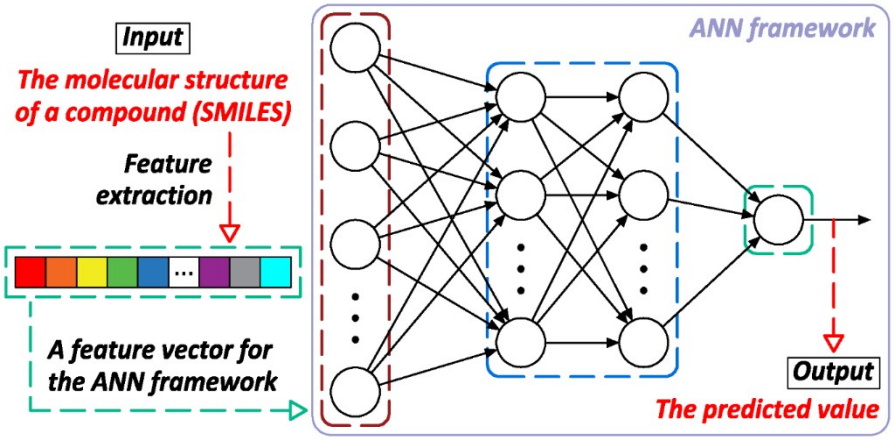

Fig. S5 The application of the developed tool in making predictions for new compounds.

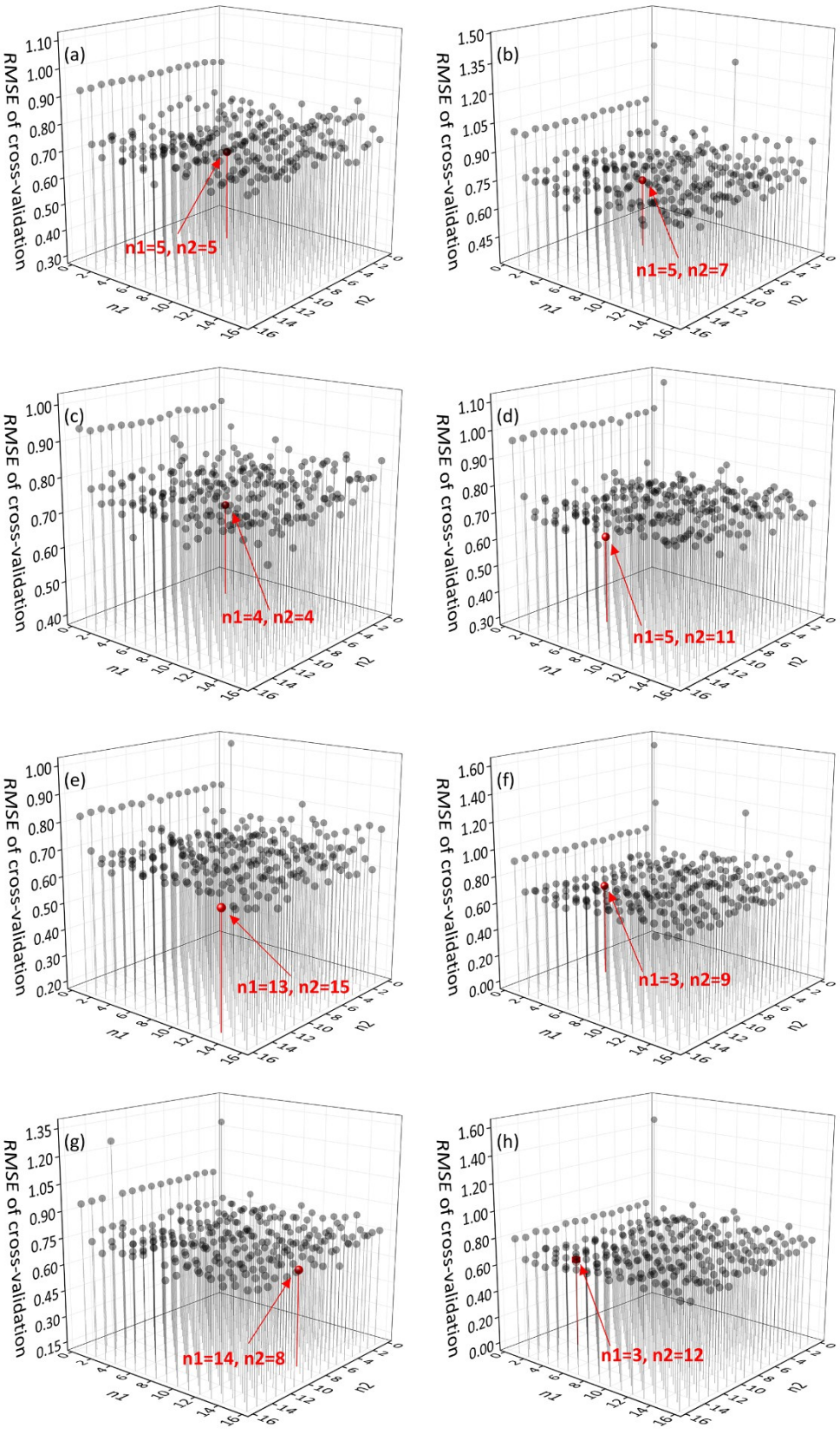

Fig. S6 The optimization and determination for the numbers of neurons in two hidden layers (n1 and n2) based on Scheme 4 trained with different numbers of cluster centres: (a) two; (b) three; (c) five; (d) six; (e) seven; (f) eight; (g) nine; (h) ten.

In order to directly compare model performance with the number of clusters, different numbers of cluster centres are adopted in the cluster sampling of model development based on Scheme 4. Analogously, the structures of these predictive models are optimized with the five-fold cross-validation as shown in Fig. S6.

Relying on the optimized ANN structures, the predictive models are developed using different number of clusters in sampling for data partitioning, and the model performance is provided in Table S5.

Table S5 The statistical analysis for the whole dataset in log*HLC* prediction adopting different numbers of clusters in Scheme 4.

Number of clusters	*N*	*RMSE*	*MAE*	R^2
2	2566	0.4009	0.2000	0.9740
3	2566	0.3952	0.2167	0.9747
4	2566	0.2981	0.1544	0.9856
5	2566	0.4274	0.2360	0.9704
6	2566	0.4016	0.2182	0.9739
7	2566	0.4465	0.2297	0.9677
8	2566	0.4576	0.2331	0.9661
9	2566	0.4454	0.2334	0.9679
10	2566	0.4470	0.2329	0.9676

Moreover, the predictive performance (indicated with RMSE and MAE) is directly compared with the number of clusters as shown in Fig. S7, along with the CH index which is used to measure the performance of clustering. It is observed that the cluster performance is better (indicated with a greater CH index) when four cluster centres are assigned in sampling. Meanwhile, the model accuracy is also better (indicated with a smaller RMSE or MAE) in the case of four cluster centres. As the number of cluster increases, the variation of CH index is basically contrary to that of RMSE or MAE, demonstrating that the optimization criterion (CH index) used is also optimal for the model performance.

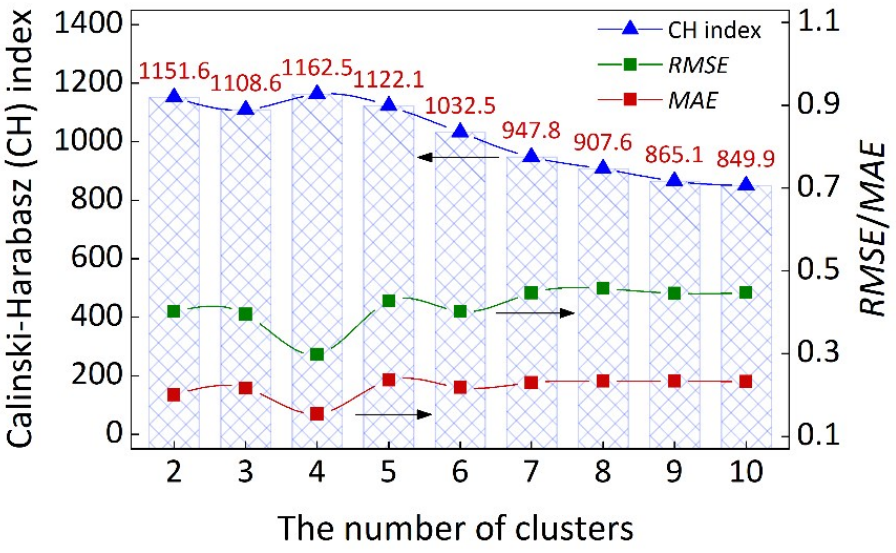

Fig. S7 The optimization for the number of cluster centres in cluster sampling and the effect of cluster centres on model performance.

373 compounds that are not included in the employed dataset are gathered from the literature[1] to validate the performance of the predictive model. Despite experimental HLC values and chemical names are provided, the necessary SMILES notations of compounds are unavailable. Therefore, we acquire the SMILES notations in the PubChem database by recognizing the names of compounds, and finally, a total of 233 compounds are assigned with correct SMILES notations and they are used to evaluate the performance of the predictive model. Hereinto, 89 compounds contain no less than ten carbon atoms (C10-C17) and 144 compounds contain less than ten carbon atoms (C3-C9). The predictive model in Scheme 4 is used to perform predictions on these compounds, and the predictive performance is presented in Fig. S8. It is observed that the predictions on relatively complex compounds (C10-C17) present a bit larger deviations than those on relatively simple compounds (C3-C9). Overall, the predictive model reveals decent and acceptable accuracy on these compounds outside the employed dataset in the model development, demonstrating the satisfactory predictability of the developed predictive model.

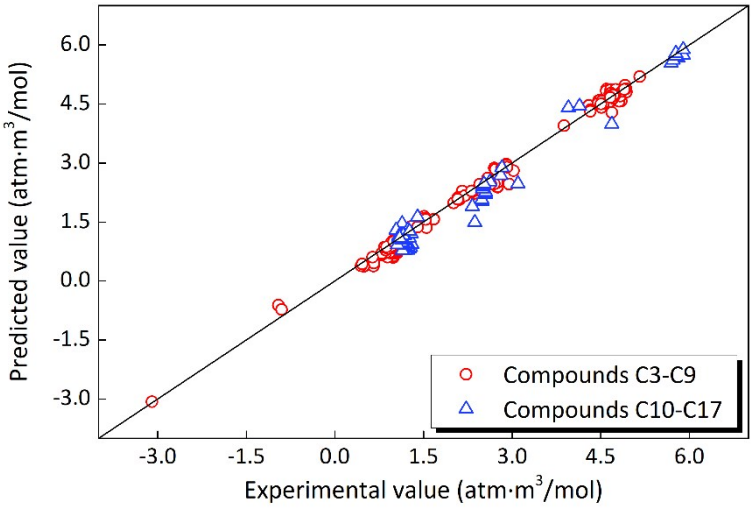

Fig. S8 The scatter plot of experimental and predicted log*HLC* values for the collected compounds.

Reference

1 F. Gharagheizi, A. Eslamimanesh, A. H. Mohammadi and D. Richon, *J. Chem. Thermodyn.*, 2012, **47**, 295-299.