

多组件学习器实现有机分子沸点的精准预测

刘雨泽^a 李昆华^a 黄佳兴^a 于曦^{*,a,b} 胡文平^{*,a,b}

(^a 天津大学化学系 天津 300072)

(^b 天津大学天津市分子光电科学重点实验室 天津 300072)

摘要 沸点(BP)是有机分子液体的基本物理化学量,也是化学工业生产中的重要参数。有机分子的沸点由分子结构决定,呈现复杂的结构-沸点关系,函数法(Function Method)、基团贡献法(Group Contribution Method)等传统方法无法应对复杂多样有机分子结构的预测,应用范围狭窄,预测精度低。本研究中,我们利用基于人工神经网络(ANN)和支持向量机(SVM)的多组件学习器实现有机分子沸点的精准预测。我们构建了基于可解释性描述符的 ANN、基于相关性描述符的 ANN 及基于复合分子指纹的 SVM 三个异质模型,并通过包含 4550 个各种类别的有机分子沸点的数据集进行训练得到了三个异质性学习器,最后集成三个学习器对有机分子沸点进行预测。相比于传统方法和此前的定量结构性质关系(QSPR)模型,多组件模型结合了三模型的优点,展现出很好的预测精度和泛化能力以及低的过拟合,实现了对多种类型有机分子的沸点的有效预测。

关键词 化学信息学; 机器学习; 异质学习模型; 人工神经网络; 支持向量机; 集成学习; 有机分子沸点

Accurate Prediction of the Boiling Point of Organic Molecules by Multi-Component Heterogeneous Learning Model

Liu, Yuze^a Li, Kunhua^a Huang, Jiaxing^a Yu, Xi^{*,a,b} Hu, Wenping^{*,a,b}

(^a Department of Chemistry, Tianjin University, Tianjin 300072)

(^b Tianjin Key Laboratory of Molecular Optoelectronic Sciences, Tianjin University, Tianjin 300072)

Abstract Boiling point (BP) is the basic physical chemistry quantity of organic molecular liquids, and an important parameter in the chemical industry. The boiling point of organic molecules is determined by the molecular structure, showing a complex structure-property relationship. Traditional methods such as function method and group contribution method cannot cope with the prediction of BP of organic molecules of complex and diverse structures. In this study, we developed an ensemble machine learning method, combining artificial neural networks (ANN) and support vector machines (SVM), to build a multi-component model to realize the boiling point prediction of high accuracy in wide structure space. We developed three heterogeneous models: ANN based on interpretable descriptors, ANN based on correlated descriptors, and SVM based on hybrid molecular fingerprints. The three heterogeneous models were evaluated and optimized using a boiling points data set containing 4550 organic molecules of various categories. The first model (Model A) is an interpretable ANN with selected descriptors possessing physical properties such as molecular composition and molecular chain branches, and with an ANN structure of 19-14-1. The second model (Model B) is an ANN with 35-30-1 structure, of which the descriptors were obtained by using correlation coefficient selecting and bidirectional stepwise regression method. The third model (Model C) is a fingerprint SVM, which had integrated molecular fingerprint information of 1679 bit. Among the three component models, Model A had the best generalization ability and Model B had the best prediction accuracy. The heterogeneous multi-component learner combined with the three models showed a better prediction accuracy and generalization property with low overfitting than that of a single model or homogeneous model. The mean square error on test set of the multi-component model is 12 K by a 5-fold cross validation on 4550 organic molecular data sets, better than all the previous BP prediction methods. The ensemble machine learning strategy we developed here based on heterogeneous models can also be well generalized to the prediction of physical chemical properties beyond BP.

Keywords chemical informatics; machine learning; heterogeneous learning model; artificial neural network; supported

* E-mail: xi.yu@tju.edu.cn; wphu@tju.edu.cn

Received January 10, 2022; published February 25, 2022.

Supporting information for this article is available free of charge via the Internet at <http://sioc-journal.cn>.

Project supported by the National Natural Science Foundation of China (Nos. 21973069, 21773169), the PEIYANG Young Scholars Program of Tianjin University (No. 2018XRX-0007), the Tianjin University-Qinghai Nationalities University Innovation Cooperation Fund (No. 2021XZC-0064), the Qinghai University for Nationalities and Qinghai-Tibet Plateau Resource Chemistry and Ecological Environmental Protection Key Laboratory of National Civil Affairs Commission Open Fund (No. 2021GKF-1002), and the College Student Innovation and Entrepreneurship Training Program of Tianjin University (No. 201910056451).

国家自然科学基金(Nos. 21973069, 21773169)、天津大学北洋青年学者计划(No. 2018XRX-0007)、天津大学-青海民族大学自主创新合作基金(No. 2021XZC-0064)、青海民族大学青藏高原资源化学与生态环境保护国家民委重点实验室开放基金(No. 2021GKF-1002)和天津大学大学生创新创业训练计划(No. 201910056451)资助项目。

1 引言

沸点是有机分子最基本的物性数据,也是化学化工生产和科学实验中的重要参考参数。尽管沸点的实验获取通常相对容易,但在许多条件下沸点的测量仍存在困难,如易燃易爆分子、有毒有害的分子、易解离分子、昂贵不易得量少的分子,以及新的未获得的分子,都需要理论方法预测沸点。沸点是微观复杂分子作用的宏观体现,分子结构与其沸点是唯一对应的,因而利用分子结构预测其沸点从原则上是可能的。然而由于有机分子结构的巨大的多样性和分子间相互作用的复杂性,基于统计力学的方法对分子沸点的预测是不现实的,因此寻求局部经验预测或近似模型成为可行替代方案。

目前已存在多种预测有机分子沸点的方法。早期的经验性方法提出一系列规则和公式用于预测同系物的沸点。如函数法(Function Method)对同系物的沸点、闪点等性质进行近似预测,利用原子数目、分子摩尔质量、摩尔折射率构造拟合公式。Mills 构造的预测公式能够准确预测简单烷烃的沸点^[1],但随着烷烃链增长预测偏差逐渐增大。为实现对更多有机化合物的预测,函数法引入更多参数,提出更加复杂的公式使得预测问题难以求解。基团贡献法不再停留于提出数学公式,基于有机分子沸点主要取决于不同官能团间相互作用的原则,首先将分子基团按种类和所处化学环境分成不同的功能基团,通过包含大量分子的数据库拟合计算出每个功能基团参数(被称为“基团贡献”),分子的沸点为分子内功能基团参数即基团贡献的全体加和。基团贡献法使通过单模型预测多官能团化合物成为可能。如 Joback 等^[2]利用包含 7000 有机分子的数据集,计算近 80 种基团的基团贡献,得到预测集平均绝对误差(MAE)为 26 的预测结果。

随着构效关系的研究和计算机科学的辅助,通过数理统计分析计算化合物性质的定量结构性质关系法(QSPR)快速发展。QSPR 使用定量描述符表征有机分子结构参数,除拓扑描述符,实验描述符外,其沿袭基团贡献法,以分子功能基团数目作为基团描述符。基于各种方式生成的大量描述符几乎涵盖有机分子所包含的所有结构信息,因此可以有效利用计算机拟合建立描述符和沸点的对映关系。Wiener^[3]提出烷烃拓扑结构 p 因子,使用摩尔分子量和 p 因子对烷烃的沸点进行二元线性拟合,首次将 QSPR 运用于预测沸点。近年来,随着机器学习算法趋于成熟,以人工神经网络(ANN)和支持向量机(SVM)等为代表的机器学习算法逐渐被应用于化学中构效关系的研究,成为 QSPR 预测分子性质更受欢迎的算法^[4-9]。Fissa 等^[10]利用 ANN 对 223 个纯碳氢化合物的沸点,相对流体密度进行预测,得到训练集均方

根误差(RMSE)为 2.15 K,远优于使用 MLR 等传统方案所得结果。但由于小样本数据集和狭窄的预测范围,模型只能提供有参照意义的统计参数,无法应对实际预测任务。Goll 等^[11]基于三个数据集分别包含 298 个碳氢化合物+含氮杂原子化合物、277 个不含氮杂原子化合物、104 个含氮卤代烃,利用遗传算法和浅层神经网络,得到三个数据集的预测标准偏差分别为 8.47, 11.03, 18.78 K。由于使用的数据集分子范围狭窄,三个不同的模型无法用于预测各类化合物的沸点。Beck 等^[12]及 Chalk 等^[13]以分子间相互作用为出发点,结合分子结构和量子力学半经验法计算得到 18 个包含分子极化率、分子静电势、氢键给受体数等描述符,使用神经网络模型预测 6000 多个化合物的沸点,由于模型选取描述符具有清晰物理意义,预测结果被证实具有优秀的泛化能力。Gharagheizi 等^[14]使用包含 17768 个化合物更大的沸点数据库,使用序列前向选择(Sequential Forward Selection, SFS)算法筛选出 44 个描述符,利用人工神经网络训练模型,在测试集上得到平均绝对偏差为 3.2%。由于其采用的描述符物理意义不明确,过拟合问题及泛化能力都有待进一步考察。分子沸点是分子结构影响的微观作用力的宏观体现,这种细致的差异一方面需要对分子结构提出更加细致的描述符,然而以上的方法在进一步提高精度上存在诸多困难,甚至会带来如过拟合等风险,限制预测模型实用化的发展。

本工作提出一种多组件异质学习模型策略用于更精确有效的预测有机分子的沸点。使用不同的描述符和算法描述沸点信息构建了三个异质性模型,分别为解释性 ANN、回归分析 ANN 和分子指纹 SVM。解释性 ANN 以分子结构、氢键数等具备可解释性的基本结构参数作为输入量建立 ANN 模型,在三个模型中具有最强的物理解释性,其预测呈现一般物理规律,具有较强的泛化能力;回归分析 ANN 以包含众多各异描述符的参数池为起点,经由逐步回归算法及相关性筛选得到和沸点高度关联的描述符子集作为输入量建立 ANN 模型;分子指纹 SVM 以复合指纹作为输入参数,基于分子特征唯一性与结构相似性,将分子编码为结构性指纹和哈希指纹,利用 SVM 建立复合指纹中各种结构信息和沸点的对应关系。我们在一个包含 4550 个有机分子的沸点数据库上对模型进行训练,得到三个独立异质性模型。进一步通过三个组件模型预测结果的平均得到最终模型预测结果。这一方法得到的预测结果优于目前文献报道的所有单体模型,展现出很好的预测精度和泛化能力以及较低的过拟合程度,实现了有机分子沸点的有效预测。

2 方法

2.1 数据集

Yaws' Dataset^[15]包含各种材料、流体的实验热力学性质。其中包含数万种有机化合物的沸点数据以及相应的系统命名及CAS号。其数据集的分子沸点分布于100~700 K, 呈现相对全面的分布。数据集分子量分布于16~590 Da, 包含最小的有机分子甲烷至多碘烷烃等重原子。数据集包含多种不同类别化合物, 可按分子所含杂原子和共轭结构大致分为脂肪烃、芳香烃、卤代烃、含氧化合物、含氮化合物和含硫化合物等。分子属性的合理分布有利于提升模型的稳健性, 而部分化合物具有多基团的复杂特征, 为模型训练提出了更高的挑战。

进一步对数据进行筛选, 除去样本容量有限的金属有机化合物, 以及非常压下的沸点数据, 最后剩下共计4550个分子(如图1所示)。这些分子按照4:1的比例随机分成3640分子训练集和910分子测试集用于后续模型学习与预测。

2.2 分子特征提取

我们利用PubChem网络数据库^[16]补充了数据集中每个分子对应的简化分子线性描述规范(Simplified molecular input line entry system, SMILES), 再通过AlvaDesc^[17], RDkit, PaDEL等计算化学软件生成近4200种涵盖结构特征、拓扑描述、电学性质的分子描述符和MACSS、PubChem、拓展连通性等指纹。依据不同的策略, 三个模型从所生成的描述符库中分别选取不同的输入参数构建机器学习模型。

2.2.1 可解释性描述符

解释性描述符依据沸点的分子间相互作用物理来源选取分子基本结构参数, 如分子组成描述符(碳原子数、氧原子数等), 物化参数描述符(分子质量 M_w , 油水

分布系数 $\log P$, 折射率 R 等)和结构描述符(氢键给受体数目)。Kubic等^[18]的工作中总结有关含氧化合物的性质预测复杂并且精度较低, 对于含有不同功能基团, 含氮氧原子数相同的分子, 其分子极性, 范德华半径均不同。使用氧原子数氮原子数两个简单结构参数描述包含氧原子和氮原子的杂环化合物、酰基化合物存在明显局限。因此依据基团贡献法的思路, 我们进一步将分子内氧和氮原子数细化为不同的含氧或者氧的官能团数作为描述符。Kier等^[19]提出Kappa3、Chi2n拓扑因子等基于分子图描述分子结构, 并预测支链烷烃沸点, 依据其计算步长的大小可以对不同大小分子的分支度进行区分。AromaticRatio为分子中芳香体系原子质量占分子中重原子(非氢原子)质量的比例, 用于描述分子芳香性。依据相应的解释性描述符, 可以构成一个易于计算, 生成模型泛化稳定的描述符集(表1)。

2.2.2 回归法筛选描述符

描述符回归法筛选法以包含大量描述符的数据集作为初始池, 通过数学筛选得到最终描述符集。描述符集的筛选存在三种常规方案^[20]。

(1)过滤式, 通过描述符间关系或描述符与预测性质的关联直接进行筛选; (2)包裹式, 组合不同描述符生成模型进行穷举学习, 以最终模型质量选取特定组合; (3)嵌入式, 通过引入范数正则化误差函数, 使模型获得稀疏解。

包裹学习为取得最优模型和模型描述符组合, 需要进行枚举或使用遗传算法(GA), 枚举或GA均需要大量的运算容量和时间开销; 嵌入法以数据集直接作为初始输入, 通过稀疏化求解, 不适用于输入参数各异的化学预测问题。最终使用相关系数筛选, 双向逐步回归的过滤法得到描述符集^[21]。

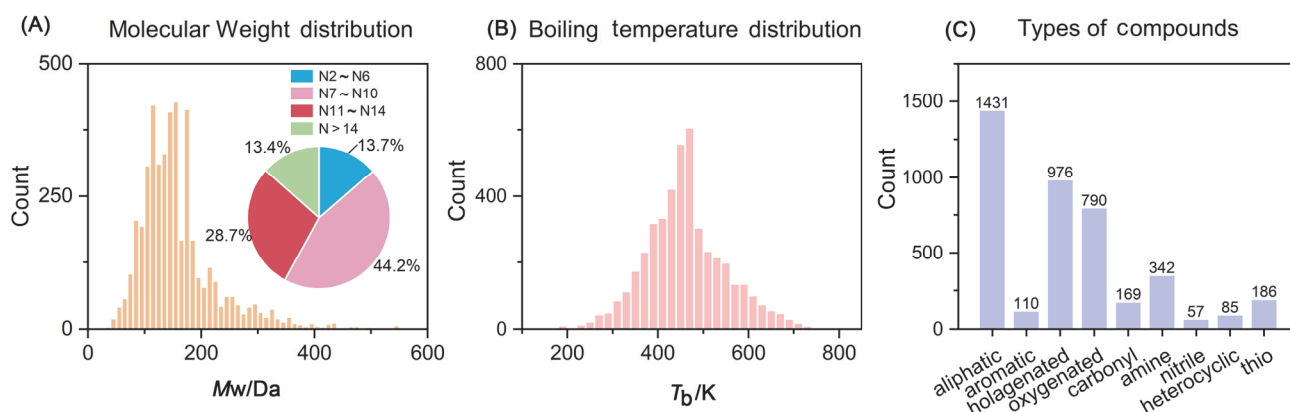


图1 (A)有机分子数据集的分子量分布, 按照重原子数(N)分类, 大部分化合物具有7~10个重原子; (B)数据集分子的沸点分布; (C)数据集分子的种类分布, 按照官能团特征分为脂肪烃、芳香烃等9类

Figure 1 (A) Shows the molecular weight distribution of the organic molecular data set. According to the number of heavy atoms (N), most compounds have 7~10 heavy atoms; (B) shows the boiling point distribution of molecules in the data set; (C) shows the species distribution of molecules in the data set, which is divided into 9 categories according to the characteristics of functional groups

表1 解释性描述符集

Table 1 Interpretable descriptor set

No.	Name	Type	Detail
1	Log <i>P</i>	Molecular properties	Octanol-water partition coeff.
2	nROH	Functional group	Number of hydroxyl groups
3	nROR	Functional group	Number of ethers
4	nR=O	Functional group	Number of carbonyl
5	nRNH ₂	Functional group	Number of 1st amine
6	nRNHR	Functional group	Number of 2st amine
7	nRNR ₂	Functional Group	Number of 3st amine
8	nRCN	Functional Group	Number of nitriles
9	nSatom	Constitutional indices	Number of Sulfur atom
10	nFatom	Constitutional indices	Number of Fluorine atom
11	nClatom	Constitutional indices	Number of Chlorine atom
12	nBratom	Constitutional indices	Number of Bromine atom
13	nIatom	Constitutional indices	Number of Iodine atom
14	nCatom	Constitutional indices	Number of carbon atom
15	HA	Structural Parameter	Number of acceptor atom for H-bond
16	HD	Structural Parameter	Number of donor atom for H-bond
17	nCharge	Constitutional indices	Total charge
18	Chi2n	Topological indices	2-Path kier shape index
19	AromR	Ring descriptors	Aromatic Ratio

相关系数筛选基于 Spearman 秩系数选取和沸点高相关度的描述符, 基于该子集利用双向逐步回归进行筛选. 双向逐步回归利用 F 检验引入描述符, t 检验剔除引入后非显著的描述符, 通过增删取得和沸点关系显著的描述符集.

最终得到的描述符集如表 2 所示, 共 35 个, 包含组成因子, 信息因子, 二维自相关等描述符. 相较于物理意义清晰的解释性描述符, 利用数学方法筛选出的回归描述符大部分意义抽象.

表2 回归描述符集

Table 2 Descriptor set got by regression

No.	Descriptor ^a	Type ^b	Detail
1	Sv	Constitutional indices	Sum of VDW volumes
2	Sp	Constitutional indices	Sum of polarizabilities
3	nBO	Constitutional indices	Number of non-H bonds
4	SCBO	Constitutional indices	Sum of conventional bond orders
5	nH	Constitutional indices	Numbers of Hydrogen atoms
6	piID	Walk and path counts	Conventional bond order ID number
7	IDET	Information indices	Total information content on the distance equality
8	HVcpx	Information indices	graph vertex complexity index
9	TIC1	Information indices	Total Information Content index
10	CIC0	Information indices	Complementary Information Content index
11	ATS4m	2D autocorrelations	B-M autocorrelation of lag 4 (log function) weighted by mass
12	ATS2p	2D autocorrelations	B-M autocorrelation of lag 2 (log function) weighted by polarizability
13	ATS1li	2D autocorrelations	B-M autocorrelation of lag 1 (log function) weighted by ionization potential
14	ATS3i	2D autocorrelations	B-M autocorrelation of lag 3 (log function) weighted by ionization potential
15	ATSC1i	2D autocorrelations	Centred B-M autocorrelation of lag 1 weighted by ionization potential
16	MATS1i	2D autocorrelations	Moran autocorrelation of lag 1 weighted by ionization potential
17	GGI10	2D autocorrelations	topological charge index of order 10
18	SpMin3_Bh(m)	Burden eigenvalues	smallest eigenvalue n. 3 of Burden matrix weighted by mass
19	SpMin4_Bh(m)	Burden eigenvalues	smallest eigenvalue n. 4 of Burden matrix weighted by mass
20	P_VSA_m_2	P_VSA-like descriptors	P_VSA-like on mass, bin 2
21	P_VSA_e_2	P_VSA-like descriptors	P_VSA-like on Sanderson electronegativity, bin 2
22	P_VSA_p_1	P_VSA-like descriptors	P_VSA-like on polarizability, bin 1

续表

No.	Descriptor ^a	Type ^b	Detail
23	P_VSA_ppp_L	P_VSA-like descriptors	P_VSA-like on potential pharmacophore points, <i>L</i> -lipophilic
24	P_VSA_ppp_ar	P_VSA-like descriptors	P_VSA-like on potential pharmacophore points, ar-aromatic atoms
25	P_VSA_charge_6	P_VSA-like descriptors	P_VSA-like on partial charges, bin 6
26	nCbH	Functional group counts	number of unsubstituted benzene C(sp ²)
27	CATS2D_03_LL	Pharmacophore descriptors	CATS2D Lipophilic-Lipophilic at lag 03
28	Qpos	Charge descriptors	total positive charge
29	RPCG	Charge descriptors	relative positive charge
30	MLOGP2	Molecular properties	squared logP
31	LOGPcons	Molecular properties	Octanol-water partition coeff. (consensus)
32	PDI	Molecular properties	packing density index
33	BLTD48	Molecular properties	Verhaar Daphnia base-line toxicity from MLOGP (mmol/L)
34	DLS_cons	Drug-like indices	DRAGON consensus drug-like score
35	MDEC-23	MDE descriptors	molecular distance edge between all secondary and tertiary carbons

2.2.3 分子指纹

分子指纹 MACSS、PubChem、拓展连通性指纹、路径指纹等, 可按照生成算法分为结构指纹和拓扑指纹. 结构指纹包裹分子内子结构的信息, 拓扑指纹将路径参数哈希化至特定长度布尔字符串. 我们使用如下规则过滤了分子指纹中的无效信息: (1)删除分子指纹全 0 值的 bit 位. (2)删除分子指纹中值线性相关的 bit 位. 将对模型训练无益的 bit 位过滤后我们将 MACSS 指纹、PubChem 指纹和拓展连通性指纹首尾拼接构成 1679 bit

复合指纹, 包含相对而言最为全面的结构信息. 经过过滤操作, 复合指纹满足高密度的结构信息分布, 可以有效避免训练过程中无效信息对模型的干扰.

2.3 异质性模型

多组件异质模型预测方法流程如图 2 所示. 三个异质模型分别选取上述不同类型的输入参数并依据参数特征选用相应的机器学习算法训练. 神经网络作为合适的方法适应于相比输入参数较大规模的样本数量. 解释性描述符模型和回归分析描述符模型的样本数目和输入参数的比值(*S-P ratio*)分别为 217 和 130, 充足的样本数量适合构建神经网络模型. 分子指纹输入参数为布尔比特串, 其中每一个 bit 位均是输入参数, 其 *S-P*

ratio 为 4, 适合构建支持向量机模型以避免因样本过少导致的欠拟合.

2.3.1 解释性 ANN(模型 A)

采取解释性描述符及全连通前馈神经网络构建模型 A, 三层全连通神经网络分别包含 19, 14, 1 个节点, 分别代表 19 个输入参数, 14 个隐含层节点, 1 个输出参数. 输入层和隐含层间, 隐含层和输出层间分别采用双曲正切函数 *tansig*, 线性函数作为层间激活函数. 使用贝叶斯正则化算法^[22] 作为全局优化函数, 均方误差 (MSE) 为损失函数或全局优化目标. 每个神经元节点使用带动量因子的梯度下降 (Gdm) 作为学习函数, 并设置学习率为 0.01, 最大迭代次数为 1000. 模型会在 MSE 变化率低于 10e-4 或达到最大迭代次数时停止学习.

2.3.2 回归分析 ANN(模型 B)

采用 35:30:1 节点数构成的三层神经网络, 与解释性 ANN 所使用的训练参数不同, 回归模型神经网络所设置的学习率为 0.005, 最大迭代次数为 80, 马夸尔特因子变化率为 $\sqrt{10}$. 由于选用的输入参数接近解释性模型的 2 倍, 回归分析 ANN 在训练过程中从单个数据中学习更多的信息, 经过过多次数的迭代将导致神经网络对训练集过度学习, 产生过拟合现象, 对预测集的预测

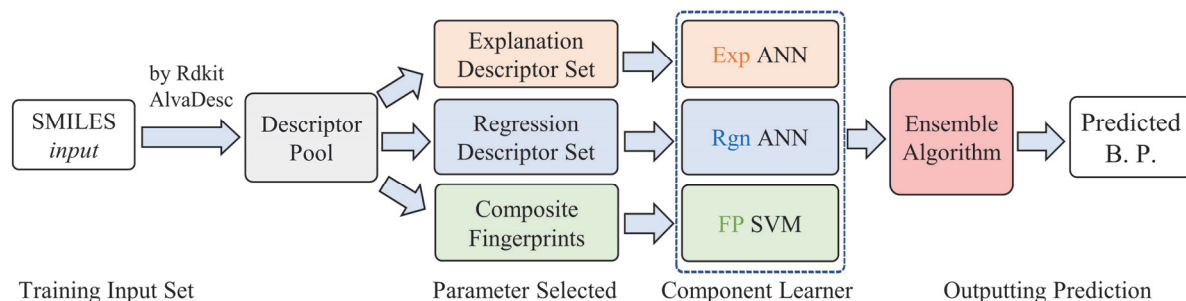


图 2 利用多组分异质模型预测沸点的流程. 经过 SMILES 输入计算得到描述符集和分子指纹集, 而后构建基于可解释性描述符的 ANN、基于相关性描述符的 ANN 及基于复合分子指纹的 SVM 三组分模型, 最后集成三个模型共同预测有机分子的沸点

Figure 2 A process for predicting boiling points using a multicomponent heterogeneity model. Descriptor sets and molecular fingerprint sets are obtained through SMILES input calculation. Then three-component models: ANN based on interpretability descriptors, ANN based on correlation descriptors and SVM based on compound molecular fingerprints, are constructed. Finally, the three models are integrated to predict Boiling point of organic molecules

精度低于少量迭代的结果. 使用提前结束法避免过拟合发生, 限制神经网络的迭代次数, 降低马夸尔特因子的变化率使神经网络在有限迭代次数内找到近似最优解.

2.3.3 分子指纹 SVM(模型 C)

使用支持向量拟合 SVR 方法, 1679 bit 布尔字符串分别对应 1679 个输入参数. 采用径向基函数 RBF 作为模型核函数, 径向基函数将输入参数变换至高维空间, 可以有效求解高参非线性拟合. 首先使用自适应核尺寸 (AutoScale) 进行变换, 发现核尺寸约为 10 时, 可以取得全局最优解. 支持向量拟合采用支持向量作为拟合的锚点, 模型训练过程使用自适应支持向量数往往取得最好的结果.

2.4 多组件异质模型的集成

构建集成模型是数理统计常用的提高对具体问题预测精度的方法. 利用多个模型, 采取合适的集成策略常常能取得比单一模型性更好的效果^[23]. 集成模型用于预测物化性质已有先例, Chalk 等使用 10-折交叉验证构建 10 个模型用于预测沸点, 降低均方误差 2~3 K.

考虑到三个异质模型对沸点的预测具有相对独立性, 利用平均法对三个异质模型的预测值求平均是最直观的集成方法:

$$P = \frac{(A+B+C)}{3} \quad (1)$$

3 结果与讨论

3.1 组件模型

3.1.1 模型 A

使用三次平行交叉验证的平均值作为模型预测值, 并结合模型总均方误差(MSE)、平均绝对误差(MAE)、相关系数 R^2 , 评价选取模型的综合性能, 用于与前人工作进行比较.

解释性模型的取得依赖于分子的物理性质和模型最终预测参数. 遵循优先选取分子最基本组成的描述符, 排除量化描述符、派生描述符等的标准, 并以具体任务即拟合沸点能力作为评价模型标准. 有机分子沸点取决于分子间的色散力, 取向力和诱导力, 选取对分子间作用力产生直接影响的 12 种基本结构参数, 包括分子量、油水系数、不同原子数, 派生出若干解释性模型: 包含基本描述符集构成的 12-8-1 模型取得 5-折交叉验证的 MSE 为 515, 训练集和测试集 MAE 分别为 11 和 12.3, 较高的 MAE 表明数据的整体预测精度不高. 我们进一步引入分子电荷 nCharge 和拓扑因子 Chi2n, Kappa3. nCharge 表示分子内的电荷或路易斯式的电荷, 描述分子诱导力; 拓扑因子描述分子长链的分支化程度, 影响分子色散力. 基本描述符和引入描述符构成 15-9-1 模型, MSE 为 530, 训练集和测试集 MAE 降低至 10 和 11.8; 预测精度在含氮氧化合物中表现较差, 参照基团

贡献法, 将氮氧原子数描述为含氮氧的不同基团个数, 分别扩充氮氧描述符构建 17-11 模型和 22-14 模型, 取得 MSE 为 800, MAE 为 9.4 和 12, 8.6 和 12. 加入氮氧描述符后模型质量显著降低, 经分析, 由于芳香化合物上的氮氧基团对沸点的贡献异于脂肪族化合物, 产生对 MSE 值贡献更大的离群值, 加入芳香比例系数降低 MSE, 获得 23-14 模型 MSE 为 750; 同时, 模型的预测精度也与数据集分子特征显著相关. 如 4550 分子共包含 79 个含有氰基官能团的分子, 其余 nCN 值均为 0, 尝试排除该描述符, 获得 MSE 为 689, 表明放弃对 79 个分子的拟合能够增强依赖于数据库的模型性能. 为使解释性模型能够具有更强的泛化性能, 应该保证描述符集不依赖于数据库. 因此保留氰基, 通过压缩含氮描述符即仅使用一级胺, 二级胺, 三级胺和氰基, 取得 21-14 模型 MSE 为 674, MAE 分别为 8.3 和 11.5.

一般认为, 不同描述符对模型的贡献可通过取舍描述符的两个对照模型判断, 为避免描述符添加过程中出现赘余, 我们基于该思路对描述符集中的描述符进行取舍. 物理性质描述符 M_w , Log P 分别从模型中排除, 获得 MSE 为 700 和 1000; 拓扑描述符 Kappa3 和 Chi2n 分别排除获得 MSE 为 541 和 700. 当 Kappa3 从模型中被排除时, MSE 值显著降低, 测试集 MAE 降低至 11, 预测值的系统误差降低近 5 K. 由于 Kappa3 和 Chi2n 显著的描述重叠, 我们最终排除 M_w 和 Kappa3 得到 19-14 模型取得 MSE 为 485, 为基于解释性描述符派生出的最优描述符集(如图 3 所示).

3.1.2 模型 B

大量数学算法被用于构建模型的最适输入参数, 从通过模型结果迭代生成输入集的遗传算法^[24], 退火算法^[27], 到使用嵌入学习的 LASSO^[25]等算法, 大多适用于解决特定问题, 并面临复杂度过高, 难以收敛等问题. 回归分析模型利用逐步回归从描述符池中选取描述符构建模型, 通过逐步回归能够快速获得次优输入集, 广泛用于解决简单机器学习输入集构建. 对于包含 4179 个不同种类描述符的参数池, 首先进行预筛选以获得缩略的描述符池可以有效提高选取效率. 通过计算每个描述符和沸点的 Spearman 系数, 从 4179 个描述符中选取共计 436 个 Spearman 系数绝对值大于 0.8 的描述符. 使用 p 因子为 $10e-6$ 的双向逐步回归最终获得 35 个描述符参数输入, 采用经验公式(2)^[26]计算隐含层节点数为 30, 最终构造的 35-30-1 模型, MSE 为 323, MAE 分别为 5.6 和 8.7. 利用回归选取的描述符在经过训练时, 随训练循环次数增大而呈现出明显的过拟合, 通过调整马尔夸特因子为 3.16, 固定训练循环 60 次, 能够同时保留模型 B 的精度并提升整体泛化能力.

$$N_h = \frac{N_s}{(\alpha * (N_i + N_o))} \quad (2)$$

其中 N_h 为隐含层节点数, N_s 为训练集样本数, N_i 和 N_o 为

输入输出参数数目, α 为调节因子, 取值 1~10.

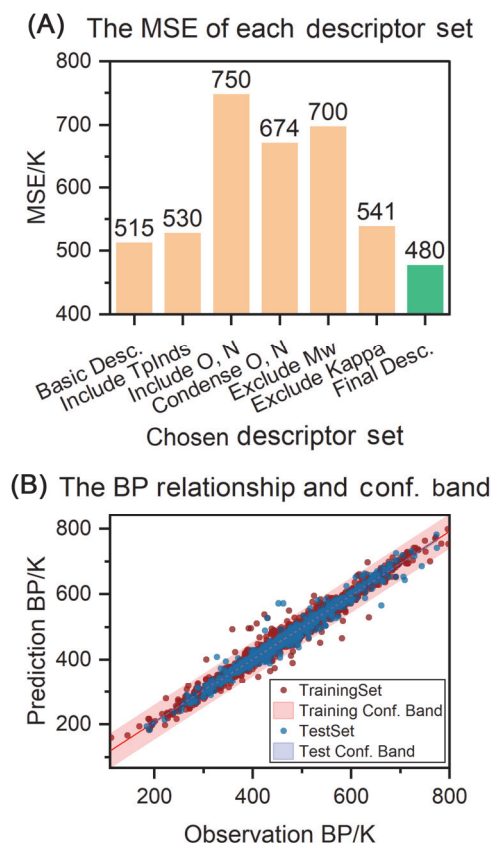


图3 (A)构建解释性模型 A 时选取不同描述符集进行 5 折交叉验证获得的平均均方误差; (B)最终选取描述符集下模型 A 的预测值-实际值关系与置信带

Figure 3 (A) Shows the average MSE obtained by 5-fold cross validation of different descriptor sets selected when constructing Interpretable model A; (B) shows the predicted value actual value relationship and confidence band of model a under the final selected descriptor set

3.1.3 模型 C

分子指纹预测沸点取决于模型是否能够对分子指纹字符串进行聚类分析. 生成 166MACCS 指纹, 构建 166 输入模型, 得到 MSE 为 1950; 将 PubChem 指纹和 MACCS 指纹连接在一起, 筛去指纹中结构缺失项, 构建 655 mix 指纹模型, 得到 MSE 为 813; 随着结构性指纹的信息增长, 模型 MSE 逐渐降低, 趋于稳定. 由于结构性指纹稀疏性较大, 聚类效果一般, 生成 1024-2-2 的拓展连通性指纹, 得到 MSE 为 378, MAE 为 11.2 和 6.4. 基于不同的参数, 生成若干不同的拓扑分子指纹. 尝试了通过增长哈希指纹的映射长度或增大指纹算法搜索半径的方法, 模型预测效果并未呈现明显提升, 说明采用 1024 bit 扩展连通性指纹通过 SVM 能够有效实现聚类. 将结构性指纹和连通性指纹拼接得到 1024+655 mix 指纹模型, 得到 MSE 为 345, MAE 为 10.4 和 6, 此即为基于分子指纹派生出的最优输入集(如表 3 所示).

表 3 利用不同拓扑指纹作为参数的 SVM 模型交叉验证结果

Table 3 MSE results of SVM model using different topological fingerprint as parameter

Fingerprint	Length	Radical	MSE
ECFP ^a	1024	2-2	378
ECFP	2048	2-2	398
ECFP	2048	2-4	420
PFP ^b	1024	2-6	685

^a ECFP is extended connectivity fingerprints; ^b PFP is path fingerprints.

3.2 多组件异质模型的集成

综合上述三个异质模型可以得到更好的预测效果. 考虑到每次训练和评估中模型的结果并不完全一致, 对 3 个模型以及集成模型做共计三次 5 折交叉验证以得到稳定的平均效果.

尝试了不同的模型集成方法, 包括简单平均法和投票法. 简单平均法集成策略直接将三个模型的预测值平均得到组合模型的预测值, 是除复杂逻辑外最简单有效的方法, 由于三个模型均在组合模型中存在不可取代的影响, 应用平均法避免任意单模型的贡献被忽视. 投票法依据三个模型的预测值进行表决, 两个相近的预测值相比较远的另一个预测值将会占有上风(投票法的细节见支持信息).

最终的模型表现如表 4 和图 4 所示. 可以看到简单平均法增强了沸点的预测效果. 组合模型的评估参数 MSE 得到显著降低, 在 3×5 折交叉验证中取得的平均 MSE_{train} 由模型 A 的 186.7, 模型 B 的 81.4, 模型 C 的 83.2 降至组合模型的 77.8; 平均 MSE_{test} 由模型 A 的 333.9, 模型 B 的 270.1, 模型 C 的 350.1 降至组合模型的 209.8. 对模型在不同数据集上的统计参数作差 $P_{\text{train}} - P_{\text{test}}$, 也可以发现, 模型 B 和模型 C 的训练测试集的 DMSE (Difference of MSE) 较大, 存在一定过拟合现象, 而模型 A 则较小. 这是由于模型 A 是可解释模型, 所选取的特征描述符具有明确的物理意义, 从而原则上具有更好的泛化能力. 但同时, 模型 A 的预测精度相对其他两个模型较低. 当把 3 种模型集成以后, 组合模型保持了 B 模型预测精度, 同时也获得了 A 的优点, DMSE 为 132, 相较模型 B 中 DMSE 为 189, 异质组合模型的 DMSE 明显降低, 显示异质模型在训练集和测试集上的差异更小, 模型泛化能力增强.

为证明组合模型的异质性对于模型预测性能的提升有不可忽视的作用, 同时使用三个相同训练测试集生成的模型 B 构建同质组合模型(Homogeneous Model). 由于已知模型 B 性能统计参数在三个单组件模型中最优, 所生成的同质组合模型应具有更高的预测性能. 但经过 3×5 折交叉验证, 发现同质 B 组合模型取得 MSE_{test} 为 245.6, 异质模型的 MSE_{test} 为 209.8, 同质模型劣于异质模型. 可见, 异质模型可以综合不同模型的优点, 保证高预测精度和较低的过拟合风险.

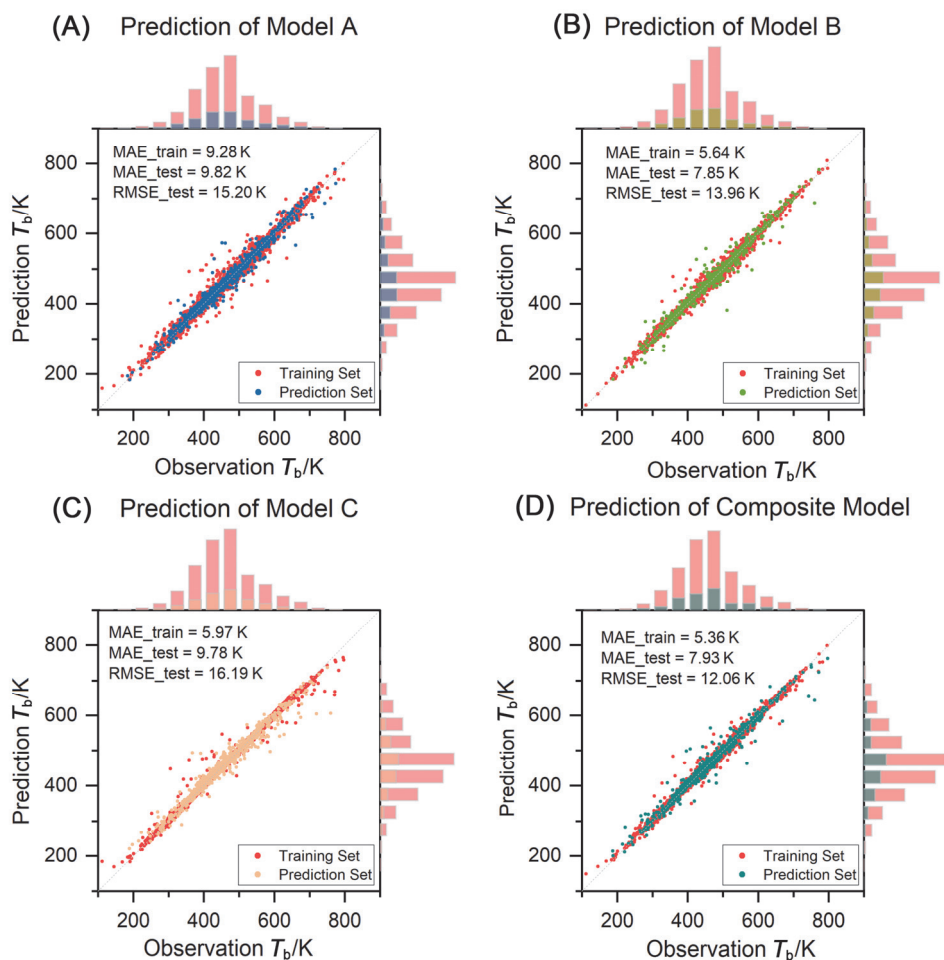


图4 图A, B, C分别为模型A, B, C的沸点实际值-预测值关系; 图D为异质组合模型的沸点实际值-预测值关系; 模型均为最优性能的参数下所取得的模型

Figure 4 Figures A, B and C show the Observation-Prediction Boiling point relationship of models A, B and C respectively; Figure D shows the O-P BP relationship of heterogeneous combination model; All of these models are under the parameters of the optimal performance

表4 单模型和不同多组件模型的预测性能

Table 4 Prediction performance of single model and different composite model

Model composition	R^2_{train}	R^2_{test}	MAE _{train}	MAE _{test}	MSE _{train}	MSE _{test}
Model A	0.9885	0.9793	9.23	10.84	186.7	333.9
Model B	0.995	0.9833	5.61	8.53	81.4	270.1
Model C	0.9951	0.9789	6.01	10.45	83.2	350.1
Heterogenous ^a	0.9953	0.9872	5.52	7.95	77.8	209.8
Heterovoted ^b	0.9953	0.9864	5.55	8.1	73.8	214.6
Homogenous ^c	0.9955	0.9848	5.16	7.96	79.9	245.6

^aHeterogenous Model is generated by average method from Model A, B, C; ^bheterovoted model is generated by the voted method from Model A, B, C; ^chomogenous Model is generated from three parallel Model B and compared with heterogenous Model.

3.3 模型预测性能比较

回顾沸点预测的诸多工作, 对于有机分子沸点的预测由简单烷烃展开发展为对各类化合物的沸点预测, 所建立的预测模型也随化合物结构的变化由简单的多元线性分析发展为机器学习方法。

Wiener, Needham, Balaban 发展出一系列拓扑因子应对简单烷烃(C2~C7)的预测, 预测方根误差降低至 4 以内。Stanton 提出拓扑电性描述符(CPSA)以解决对五

元杂环(四氢呋喃、呋喃、噻吩)的预测, 取得预测均方根误差为 11.2。Randic 对 58 个脂肪醇进行预测, 所得预测方根误差仅为 3.64。

随后的多数工作聚焦于对各类化合物的研究。Katrisky 采用 DDDR 库中的 298 个分子, 采用多种描述符包括(CPSA), 取得预测平方根误差为 16。Varamesh 采用径向基神经网络对各类化合物进行预测。Gharagheizi 引入 14216 个分子, 基于回归分析选取描述符。

表5 利用 QSPR 法预测有机分子沸点相关工作的总结

Table 5 Summary of related work on predicting boiling point of organic molecules by QSPR method

Type of compound	N	Descriptor	Method	R ²	S	Reference
Alkanes	94	Tls(W, P)	MLR		0.97	[3]
Alkanes	74	Tls(5)	MLR	0.999	1.86	[28]
Alkanes(C2~C7)	72	Tls(LOVI's)	RA	0.994	3.9	[29]
Furans, thiophenestrahdrofuran	209	Tls, electronic, geometrical	RA	0.969	11.2	[30]
Alkanes, alcohols	245	Atom-type E	MLR		8	[31]
Alcohols	58	Weighted path related Tls	MRA	0.978	3.64	[32]
Diverse organic compounds	298	Constitutional, topological, geometrical and CPSA(8)	MLR	0.954	16.15	[33]
Hydrocarbons	143	Tls	MLR	0.9821	7.2549	[34]
Diverse organic compounds	612	Constitutional, topological, geometrical and CPSA(8)	MLR	0.965	15.5	[35]
(contain N, O, F, Cl, Br and I)						
Diverse organic compounds	450	Molecular weight and specific gravity	RBF		18.78	[36]
Single ring, fused rings, halogens						
OH, COOH, COOR, CON, CN, NH ₂ and NO ₂	155	Tls	BMLR	0.9864	9.1	[37]
Diverse organic compounds	14216	Mathematical Selection	ANN	0.943	22	[14]
Diverse organic compounds	4550	Tls, Constitutional, Fingerprints	Composite ML	0.9953	10.0873	Present work

本工作使用三个异质独立模型的组合, 实现对 Yaws 数据库中 4550 个分子预测, 所得预测均方根误差为 12.

模型的结果与前人模型结果的比较见表 5. 可以看出, 多组件异质模型有着更好的表现力和预测力. 同时多组件模型独立预测结果彼此间的差别也提供了模型置信程度, 从而使得我们的方法更具有实用价值.

4 结论

本工作提出一种多组件异质学习模型策略用于预测有机分子的沸点. 构建了三个异质性模型, 即基于可解释性描述符的 ANN、基于相关性描述符的 ANN 及基于复合分子指纹的 SVM, 并集成三个异质学习器对有机分子沸点进行预测. 在包含 4550 个有机分子沸点的数据集上进行训练和性能测试, 发现相比于传统方法和此前的定量结构性质关系(QSPR)模型, 多组件集成模型可以结合三种异质模型的优点, 展现出很好的预测精度和泛化能力以及较低的过拟合程度, 预测结果优于目前文献报道的所有单体模型, 实现了对有机分子沸点的有效预测. 通过描述符优化和集成方法的改进以及数据库的扩大, 这一方法的预测精度和泛化能力可以进一步提升. 同时, 多组件异质学习器策略也有望应用于其他常见的有机分子理化性质的预测, 成为一种有效的通用的策略.

References

- [1] Walker, J. *J. Chem. Soc.* **1894**, 65, 193.
- [2] Joback, K. G.; Reid, R. C. *Chem. Eng. Commun.* **1987**, 57, 230.
- [3] Wiener, H. *J. Am. Chem. Soc.* **1947**, 69, 17.
- [4] Katritzky, A. R.; Kuanar, M.; Slavov, S.; Hall, C. D.; Karelson, M.; Kahn, I.; Dobchev, D. A. *Chem. Rev.* **2010**, 110, 5714.
- [5] Zhu, B.; Wu, R.; Yu, X. *Acta Chim. Sinica* **2020**, 78, 1366 (in Chinese).
- [6] 朱博阳, 吴睿龙, 于曦, 化学学报, **2020**, 78, 1366.
- [7] Wei, J.; Chu, X.; Sun, X.; Xu, K.; Deng, H.; Chen, J.; Wei, Z.; Lei, M. *InfoMat* **2019**, 1, 338.
- [8] Song, Z.; Chen, X.; Meng, F.; Cheng, G.; Wang, C.; Sun, Z.; Yin, W. *J. Chin. Phys. B* **2020**, 29, 68.
- [9] Wu, W.; Sun, Q. *Scientia Sinica: Physica, Mechanica et Astronomica* **2018**, 48, 58 (in Chinese).
- [10] 吴炜, 孙强, 中国科学: 物理学, 力学, 天文学, **2018**, 48, 58.
- [11] Liu, Y. D.; Yang, Q.; Li, Y.; Zhang, L.; Luo, S. Z. *Chin. J. Org. Chem.* **2020**, 40, 3812 (in Chinese).
- [12] 刘伊迪, 杨骐, 李遥, 张龙, 罗三中, 有机化学, **2020**, 40, 3812.
- [13] Fissa, M. R. *J. Mol. Graph. Model.* **2019**, 87, 109.
- [14] Goll, E. S.; Jurs, P. C. *J. Chem. Inf. Model.* **1999**, 39, 974.
- [15] Beck, B.; Breindl, A.; Clark, T. *J. Chem. Infor. Comp. Sci.* **2000**, 40, 1046.
- [16] Chalk, A. J.; Beck, B.; Clark, T. *J. Chem. Infor. Comp. Sci.* **2001**, 41, 457.
- [17] Gharagheizi, F.; Mirkhani, S. A.; Ilani-Kashkouli, P.; Mohammadi, A. H.; Ramjugernath, D.; Richon, D. *Fluid Phase Equilib.* **2013**, 354, 250.
- [18] Yaws, C. L. *Yaws' Critical Property Data for Chemical Engineers and Chemists*, **2012**, <http://app.knovel.com/hotlink/toc/id:kpYCPDCECD/yaws-critical-property/yaws-critical-property>
- [19] PubChem-<https://pubchem.ncbi.nlm.nih.gov/>
- [20] Mauri, A. *Ecotoxicological QSARs, Methods in Pharmacology and Toxicology*, Roy, K., New York, **2020**, pp. 801~820.
- [21] Kubic, W. L.; Jenkins, R. W.; Moore, C. M.; Semelsberger, T. A.; & Sutton, A. D. *Ind. Eng. Chem. Res.* **2017**, 56, 12236.
- [22] Kier, L. B.; Hall, L. H. In *Molecular Connectivity in Chemistry and Drug Research*, New York, **1976**, pp. 27~39.
- [23] Dash, M.; Liu, H. *Intell. Data Anal.*, **1997**, 1, 131.
- [24] Roy, P. P.; Leonard, J. T.; Roy, K. *Chemometr. Intell. Lab. Syst.* **2008**, 90, 31.
- [25] MacKay, J. C. *Neural Comput.* **1992**, 4, 415
- [26] Zhou, Z. H. *Ensemble Methods: Foundations and Algorithms*, Chapman and Hall/CRC, **2012**, p. 129
- [27] Zhou, L.; Wang, B.; Jiang, J.; Pan, Y.; & Wang, Q. *Chemometr. Intell. Lab. Syst.* **2017**, 167, 190.
- [28] Robert, T. *J. R. Stat. Soc. Ser. A* **1996**, 58, 267.
- [29] Sheela, K.; Deepa, S. N. *Math. Probl. Eng.* **2013**, 2013, 11.
- [30] Eriksson, L.; Jaworska, J.; Worth, A. P.; Cronin, M. T. D.; McDowell, R. M.; Gramatica, P. *Environ. Health Perspect.* **2003**, 111, 1361.
- [31] Needham, D. E.; Wei, I. C.; Seybold, P. G. *J. Am. Chem. Soc.* **1988**, 110, 4186.
- [32] Balaban, A. T.; Ciubotariu, D.; Medeleanu, M. *J. Chem. Infor. Comp. Sci.* **1991**, 313, 517.
- [33] Stanton, D. T. *J. Chem. Infor. Comp. Sci.* **2000**, 40, 81.
- [34] Hall, L. H.; Kier, L. B. *J. Chem. Infor. Comp. Sci.* **1995**, 35, 1039.
- [35] Randić, M.; Balaban, A. T.; Basak, S. *J. Chem. Infor. Comp. Sci.* **2001**, 41, 593.
- [36] Katritzky, A. R.; Mu, L.; Lobanov, V. S.; Karelson, M. *J. Phys. Chem.* **1996**, 100, 10400.
- [37] Zhou, C. Y.; Nie, C. M.; Li, S.; Li, Z. H. *J. Comput. Chem.* **2007**, 28,

- 2413.
- [35] Katritzky, A. R.; Lobanov, V. S.; Karelson, M. *J. Chem. Infor. Comp. Sci.* **1998**, *38*, 28.
- [36] Varamesh, A.; Hemmati-Sarapardeh, A.; Dabir, B.; Mohammadi, A. *H. J. Mol. Liq.* **2017**, *242*, 59.
- [37] Sola, D.; Ferri, A.; Banchero, M.; Manna, L.; Sicardi, S. *Fluid Phase Equilib.* **2008**, *263*, 33.

(Lu, Y.)