计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(9):1−11 [doi: 10.15888/j.cnki.csa.008076]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 ANN 的新型 MOFs 性能预测①

赖　欣[1], 卢　罡[2], 王　磊[3], 毕志远[1], 阳庆元[3], 俞度立[1,4]

[1](北京化工大学 信息科学与技术学院, 北京 100029)
[2](北京化工大学 信息科学与技术学院 智能无人系统研究中心, 北京 100029)
[3](北京化工大学 有机无机复合材料国家重点实验室, 北京 100029)
[4](北京化工大学 软物质科学与工程高精尖创新中心, 北京 100029)
通讯作者: 卢　罡, E-mail: lugang@mail.buct.edu.cn

摘　要: 在 MOFs 研究领域, 探寻新型 MOFs 仍然是非常困难的研究问题. 将 MOFs 进行“材料基因编码”后, 应用遗传算法 (Genetic Algorithm, GA) 可以快速探索新型 MOFs, 但其性能依赖于设定的个体适应度函数, 且对新生成的 MOFs 个体的有效评估也影响了该方法的效果. 机器学习方法可以对 MOFs 的构效关系进行评估与预测, 人工神经网络 (Artificial Neural Network, ANN) 是众多机器学习方法中具有代表性的一个, 可以发掘非线性的构效关系. 本文提出将神经网络用于预测遗传算法生成的新型 MOFs 个体对 CH_4 气体的吸附能力, 从而帮助遗传算法搜索新型 MOFs. 实验结果表明, 神经网络可以有效评估新型 MOFs 材料, 证明了将神经网络与遗传算法相结合用于新型 MOFs 搜索和筛选的可行性.

关键词: 机器学习; 遗传算法 (GA); 神经网络; 材料基因编码; MOFs

引用格式: 赖欣,卢罡,王磊,毕志远,阳庆元,俞度立.基于 ANN 的新型 MOFs 性能预测.计算机系统应用,2021,30(9):1–11. http://www.c-s-a.org.cn/1003-3254/8076.html

ANN-Based Prediction about Performance of Novel MOFs

LAI Xin[1], LU Gang[2], WANG Lei[3], BI Zhi-Yuan[1], YANG Qing-Yuan[3], YU Du-Li[1,4]

[1](College of Information Science and Technology, Beijing University of Chemical Technology, Beijing 100029, China)
[2](Research Center for Intelligent Unmanned Systems, College of Information Science and Technology, Beijing University of Chemical Technology, Beijing 100029, China)
[3](State Key Laboratory of Organic-Inorganic Composites, Beijing University of Chemical Technology, Beijing 100029, China)
[4](Beijing Advanced Innovation Center for Soft Matter Science and Engineering, Beijing University of Chemical Technology, Beijing 100029, China)

Abstract: In the field of MOFs research, searching for novel MOFs is still a complicated problem. After MOFs are processed by “material genetic encoding”, the Genetic Algorithm (GA) can be used to rapidly explore novel MOFs, but their performance depends on the setting of individual fitness functions, and the effective evaluation of the novel MOFs also contributes to the effectiveness of this method. As one of the representative methods of machine learning, the Artificial Neural Network (ANN) can uncover the non-linear constitutive relationships. In this paper, the neural network is introduced to predict the adsorption capacity for CH_4 gas by the novel MOFs generated by GA, thereby facilitating the search for novel MOFs by GA. The experimental results show that the neural network can thoroughly evaluate the novel MOFs materials, demonstrating the feasibility of combining the neural network and GA for the search and screening of the novel MOFs.

Key words: machine learning; Genetic Algorithm (GA); neural network; material genetic encoding; MOFs

① 基金项目: 国家自然科学基金 (22078004); 中央高校基础研究基金 (buctrc201727); 北京化工大学大科学项目 (XK180301)
Foundation item: National Natural Science Foundation of China (22078004); Foundation of Basic Research for the Central Universities of China (buctrc201727); Big Science Project of Beijing University of Chemical Technology (XK180301)
收稿时间: 2020-12-07; 修改时间: 2021-01-08; 采用时间: 2021-01-20; csa 在线出版时间: 2021-09-02

近年来, 以机器学习、深度学习为代表的人工智能理论和方法受到人们的广泛关注. 尤其是谷歌 DeepMind 团队开发的 AlphaGo, 在围棋领域中的精彩表现令人印象深刻[1,2]. 在此之后, DeepMind 又迅速对计算机视觉等领域做出了可喜成果[3]. 如今, 机器学习已被广泛应用于自然语言处理[4,5]、数据挖掘[6]、证券市场分析[7]、机器人应用[8,9]、医学诊断[10,11] 等领域.

在材料科学领域, 材料的各种反应、合成中会产生海量的数据, 而将善于从海量数据中发掘规律的机器学习方法应用于材料科学领域便顺理成章[12,13]. 实验研究发现, 由于 MOFs 具有较高的孔隙率和具有规律性、可组合性、多元性等特点, 能够高效地通过计算机模拟预测 MOFs 材料的物理化学性质[14]. 通过 GCMC (Grand Canonical Monte Carlo) 分子模拟方法对 MOFs 进行高通量筛选已经被证实是一种有效的实验手段[15,16]. 目前应用的分子模拟方法主要有分子动力学、蒙特卡罗、密度泛函理论等. 在探寻物理化学性能优秀的 MOFs 材料过程中, 需要对材料的结构特性、物理性质、化学性质等进行搜索分析, 通常可以应用 GCMC 分子模拟方法. 然而, 可能存在的 MOFs 结构存在于一个近乎无穷大的样本空间, 要将所有 MOFs 材料逐一进行分子模拟计算, 从而挑选出性能出众的材料, 其计算成本是无法估量的. 近年来, 人们已经开始关注如何在准确预测 MOFs 性能基础上, 提高计算效率. Simon 研究组将少量 MOFs 吸附材料放入综合数据库中, 对其进行 GCMC 模拟, 找出吸附材料的物理结构特性与其对 CH_4 吸附能力之间的关系[17]. 其中用到的 MOFs 数据有 Zeolites[18]、hypothetical MOFs (hMOFs)[19]、Porous Polymer Networks (PPNs)[20]、hypothetical Zeolitic Imidazolate Frameworks (hZIFs)[21] 以及 Computation-Ready Experimental (CoRE) MOF[22] 等 MOFs 材料数据. 材料数据包含多种性质特征, 利用机器学习挖掘其定量构效关系 (Quantitative Structure–Property Relationship, QSPR) [23], 可将这些结构性质作为参数, 对材料分子的气体吸附能力进行回归分析和预测. Fernandez 等通过晶体学的 RDF 分析方法, 利用 RDF 得分评估 MOF, 同时利用多元线性回归、支持向量机等方法, 构建了处于不同压力环境下, MOFs 材料针对 CO_2、N_2 与 CH_4 的气体吸附与 RDF得分的 QSPR 模型[24]. 之后, Fernandez 小组利用孔隙率和孔径等物理结构变量, 预测 MOFs 对 CH_4 的吸收, 并在实验中得到 $R^2 = 0.85$ 的结果[25]. Fernandez 等还应用了分类方法, 基于 QSPR预测表现最佳的 CO_2 吸附 MOFs 材料, 达到 94.5% 的准确率[26]. Sezginel 等经过 QSPR 分析, 提出一种多变量线性模型, 利用该模型与 MOFs 的结构特性, 包括表面积、晶体密度、孔隙率、孔径以及等量热吸附 (Qst), 预测出 MOFs 吸附剂对 CH_4 的吸收能力, 实验结果发现, 孔隙率与等量热吸附是影响 MOFs 气体吸附能力的关键因素[27]. Chung 等利用遗传算法, 对捕获 CO_2 的 MOFs进行筛选, 在计算效率上获得了 50 倍左右的提升[28]. 这些工作在材料筛选效率上有着出色的表现.

本文工作受到 Chung 等 2016 年关于遗传算法 (Genetic Algorithm, GA) 方面工作[28] 的启发. 他们在材料数据库中通过遗传算法进行材料筛选, 但是对于遗传算法生成的库中没有的新个体并未进行进一步的评估. 本文用原始 MOFs 数据集训练人工神经网络 (Artificial Neural Network, ANN), 并用 ANN 对遗传算法生成的新型 MOFs 个体的性能进行预测评估, 从而搜索对 CH_4 气体具有较高吸附性的 MOFs. 我们首先通过 GCMC 模拟计算文献 [28] 中数据集的每个 MOFs 在一定条件下对于 CH_4 气体的吸附性能, 然后用该结果训练一个 ANN, 使其能够评估和预测 MOFs 基因与 CH_4 气体吸附性之间的构效关系. 实验结果表明, 基于 ANN 搜索并预测的材料吸附性能平均表现优于原始材料数据库中的最优材料, 证实了该方法的可行性和有效性.

1 面向 GA 和 ANN 的 MOFs 数据集

1.1 针对 MOFs 材料数据进行基因编码

为了通过 GA 搜索新型 MOFs, 需要根据 MOF 的特征设计 GA 所需的基因编码. 为 MOFs 进行基因编码的方式没有特定的规则, 但应能够尽量反映 MOFs 的结构特征及各组分、配体之间的组合特征, 从而在 GA 运行过程中, 基因编码的变化能够反映出 MOFs 组合结构的变化.

本文的原始数据来自于 WLLFHS hMOF 数据集[19]. 该数据集中 MOFs 的参数由 Wilmer 研究组汇编和验证, 具有丰富多样的 MOFs 材料结构, 适合进行分子模拟筛选与机器学习分析. 文献 [28] 将该数据集中的 MOFs 进行基因编码, 该编码利用 6 个整数作为“基因”, 每个“基因”都表示了一种分子的特性或者功能[28].

本文沿用该基因编码, 具体设定如下:

第 1 位基因, 表示潜在互穿能力, 共 4 种, 用 0 至 3 的整数表达; 第 2 位基因, 表示实际互穿能力, 共 4 种, 用 0 至 3 的整数表达; 第 3 位基因, 表示无机配体, 共 5 种, 用 0 至 4 的整数表达; 第 4 位基因, 表示主要有机连接单元, 共 40 种, 用 0 至 39 的整数表达; 第 5 位基因, 表示次要有机连接单元, 共 40 种, 用 0 至 39 的整数表达; 第 6 位基因, 表示化学官能团, 共 15 种, 用 0 至 14 的整数表达.

根据上述设定, MOFs 材料的搜索空间大小为: $4\times4\times5\times40\times40\times15=1920\,000$. 在这种编码方式下, 构象异构体之间以及只有官能团定位不同的 MOFs 之间具有相同的基因编码. Chung 等[28] 分析发现, 构象异构体之间、只有官能团定位不同的 MOFs 之间不仅结构类似, 化学性能也相当. 因此, 他们从类似的 MOFs 中选择一个作为代表, 缩减数据集的规模. 最终, 文献 [28] 整理了具有 51 163 个 MOF 基因编码的数据集.

1.2 计算对 CH_4 气体的吸附值

在文献 [28] 整理的数据集基础上, 我们进一步采用自主开发的力场参数和自主研发的模拟计算软件, 通过 GCMC 模拟计算其中每个 MOFs 在 298 K (K 为开尔文, 热力学温度单位, 下同) 条件下对 CH_4 气体的吸附能力.

MOF 材料和气体分子之间的相互作用采用范德华力 (vdW) 和库仑势的组合来表示[29]. 其中范德华力采用 Lennard-Jones (LJ) 方程描述. LJ 势能参数取自 UFF 力场, CH_4 分子势能参数取自 TraPPE 力场. 不同原子之间的 LJ 势能参数采用 Lorentz-Berthelot 混合规则计算.

在前期的研究工作中, 我们利用量子密度泛函和 Monte Carlo 模拟相结合的跨尺度手段开发出了新的力场[30], 其中基于量子力学层次的密度泛函理论 (Density Functional Theory, DFT)[31] 计算被用于确定材料与气体分子之间的精确相互作用参数. DFT 计算基于 Materials Studio 软件中的 Dmol3 模块, 采用 GGA 交换泛函 Perdew-Burke-Ernzerhof (PBE) 和含轨道极化函数的双数值轨道基组 (DNP), 并结合 Grimme 的色散校正作用 (DFT-D2), 对 MOF 中获取的模型簇进行优化, 并计算出不同距离下无机单元与 CH_4 之间的相互作用能. 在此基础上, 通过 Monte Carlo 模拟实现了 MOFs 对 CH_4 气体在 298 K 条件下吸附量的量化, 计算得到了数据集中每个 MOF 对 CH_4 气体的吸附值.

本文基于自主研发的模拟计算软件 HT-CADSS (http://jshx.buct.edu.cn/yjcg/bzxcg/86799.htm), 采用 GCMC 方法研究了 298 K 条件下, 文献 [28] 的数据集中 51 163 个 MOFs 对 CH_4 气体的吸附能力. 在 GCMC 模拟中, 采用 Peng-Robinson (PR) 方程将压力转换为逸度作为计算的输入值. 所有的 MOFs 均视为刚性材料, 并在三维尺度上采用周期性边界条件. 计算范德华作用的截断半径 (cut-off) 设置为 1.4 nm. 对于每一个吸附模拟过程, 模拟总步数为 3000 万步, 前 1500 万步用于系统平衡, 后 1500 万步用于获得热力学性质的统计平均值. 对 CH_4 单组分吸附模拟, 涉及分子的平移、插入和删除. 在无限稀释条件下, MOF 骨架与气体分子之间相互作用力的相对强弱采用无限稀释吸附热进行表征. 无限稀释吸附热采用基于正则系综 (NVT) 的 Widom 测试粒子方法[32] 计算.

经过以上的计算, 最终得到 51 163 个经过基因编码的 MOFs 对 CH_4 气体的吸附值, 最大为 528, 其基因编码为 2-0-0-29-29-12. 数据示例如表 1 所示.

表 1 数据示例

基因编码	对CH_4气体吸附值
2-0-0-9-9-0	245.5
1-1-0-13-0-3	49.7
1-0-0-13-0-7	170.6
1-1-0-13-13-7	140.2
0-0-0-13-0-7	258.7
2-0-0-13-13-7	328.7
2-1-0-13-13-7	149.7
2-2-0-13-13-7	62.7
0-0-0-13-13-10	361.5
⋮	⋮

1.3 单特征分析

1.1 节中, 一个 MOF 被编码成了一个具有 6 个基因的染色体, 6 个基因分别代表了它的 6 个结构特征. 本节中, 我们分别分析了这 6 个特征与 MOFs 的 CH_4 气体吸附能力之间的构效关系, 结果如图 1 所示.

图 1(a) 为 MOFs 潜在互穿能力与 MOFs 对 CH_4 气体吸附能力的关系. WLLFHS 数据库以材料结构的多样性著称, 因此, 具有不同潜在互穿能力的 MOFs 在 CH_4 气体吸附能力上分布较为均匀, 体现了该数据集中样本的多样性和完整性. 图 1(b) 显示了在实际互穿能力的维度上, 数据集中 MOFs 对 CH_4 气体吸附能力的分布. 可以看到, 实际互穿能力越高的 MOFs 对 CH_4 气体吸附能力相对越差. 这是由于互穿较多的 MOFs

稳定性较高, 一定程度上阻碍了气体分子的吸附[33]. 对于图 1(c) 中的无机节点而言, 带有锌或铜桨轮与对位连接的 MOFs, 在分析结果中表现出更强的 CH_4 气体吸附性能, 这是由于部分 MOFs 材料在活化的过程中, 遇金属簇配位溶剂分子或水分子易脱落, 形成不饱和金属位点, 从而增强了对 CH_4 气体的吸附作用. 另外, 研究表明, 当 MOFs 材料与水接触时, 结构的结晶性会在一定时间内消失. 大多数情况下, 水的存在是不可避免的, 具有二价金属离子 (例如 Zn^{2+}和 Cu^{2+}) 的 MOF 在有水的情况下极易出现这种不稳定性[34,35]. 主有机连接单元与次有机连接单元对 MOFs 的 CH_4 气体吸附性能的影响分别如图 1(d) 和图 1(e). 可以看到, 表现良好的有机连接单元主要集中在 12–30 号区间内, 而 31–39 号有机连接单元在低性能 MOFs 中缺失. 图 1(f) 显示, 含有 0 号、7 号、10 号、12 号化学官能团的 MOFs 材料对 CH_4 气体吸附能力突出, 其中 0 号表示不考虑官能团影响, 其余 3 种官能团分别对应甲基、乙基、丙基[28]. 我们认为, 这是因为这类烃基官能团与 CH_4 有相似的机构和化学性质.

对于单特征的分析表明, MOFs 对 CH_4 气体吸附能力受多种因素共同作用影响, 包括 MOFs 材料的拓扑结构、有机配体和无机单元的结构、官能团的选择等. 单纯的针对其中某一方面进行修改, 并不能保证有效提升 MOFs 材料对于 CH_4 气体的吸附能力. 这也进一步体现了应用包括 ANN 在内的机器学习方法发掘这种非线性构效关系的意义.

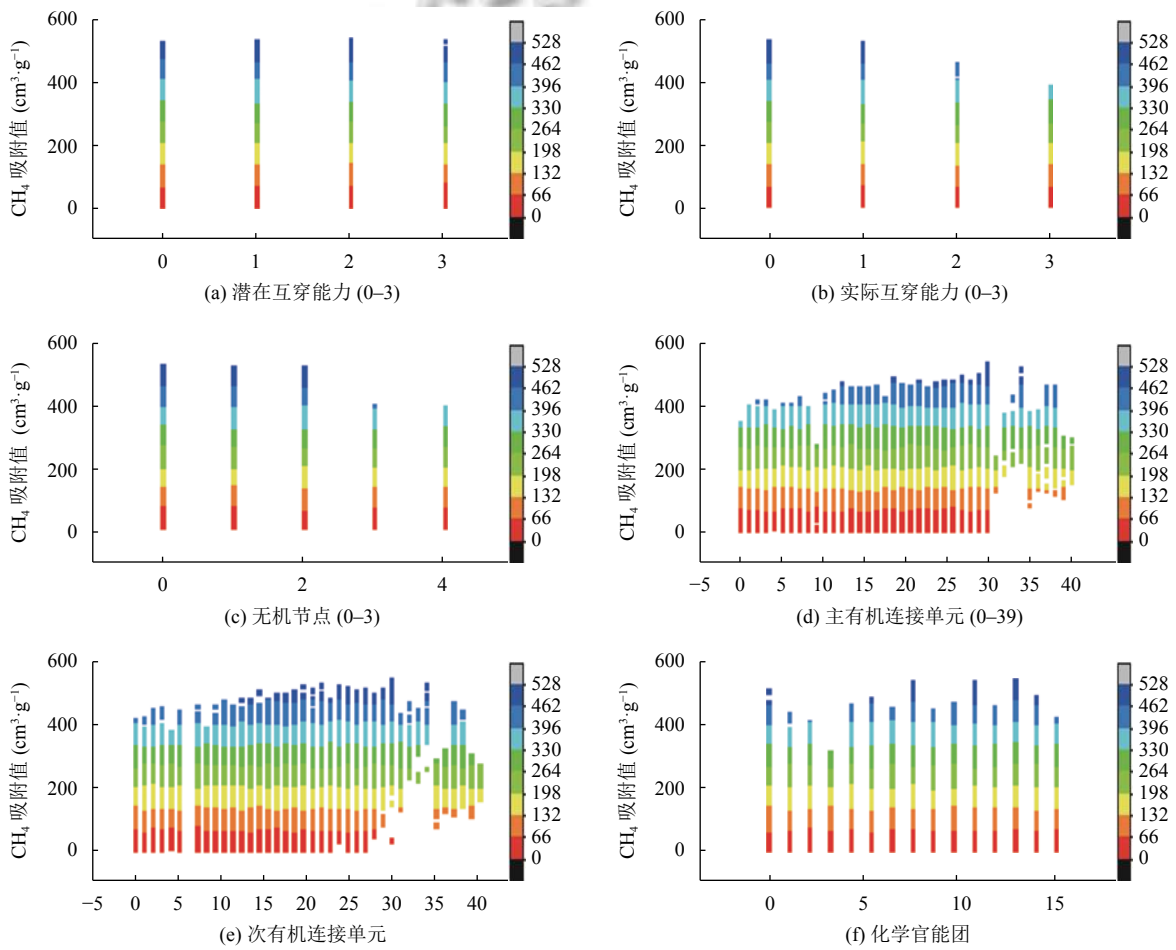


图 1 各基因编码与 CH_4 气体吸附能力的构效关系

2 ANN 模型的训练

对于 GA 产生的新型 MOFs 个体, 从仅有的 6 个基因位点的值构建 MOF 结构, 再生成相应的数据进行 GCMC 模拟计算, 从而进行性能评估, 将是一个非常繁琐及耗时的过程. 因此我们提出将 MOFs 的基因编码作为输入, GCMC 模拟计算的目标性能作为输出,

训练 ANN 作为挖掘 MOFs 构效关系的机器学习模型, 从而能够对 GA 生成的新的 MOFs 个体进行性能预测评估.

2.1 ANN 模型评价指标

ANN 通过模仿人类大脑的思维方式, 进行大规模高维数据处理和分析. 一个 ANN 包含输入层、隐含层和输出层, 其中隐含层可以有多层. ANN 的本质是非线性函数映射, 通过对高维数据的低维非线性映射, 转变为人类可理解的结果输出. 由于需要预测 MOFs 对 CH_4 气体的吸附值, 因此我们将 ANN 构建为输出层只有一个神经元的回归神经网络, 从而输出一个实数值. 作为预测具体数值的回归 ANN, 其评价指标 R^2 的值越接近 1, 模型的预测性能越好, 其定义如下:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i^{\text{true}} - y_i^{\text{predict}}\right)^2}{\sum_{i=1}^{n}\left(y_i^{\text{true}} - \overline{y_i^{\text{predict}}}\right)^2} \tag{1}$$

其中, n 是测试集中 MOFs 个体的数量, y_i^{predict} 为第 i 个 MOFs 结果的预测值, y_i^{true} 是通过 GCMC 模拟得到的结果. $\overline{y_i^{\text{predict}}}$ 是所有 y_i^{predict} 的平均值. 另一个评价指标均方误差 (Mean Square Error, *MSE*), 是预测值和真实值之间误差的平方和, 其定义为:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i^{\text{true}} - y_i^{\text{predict}}\right)^2 \tag{2}$$

2.2 数据集准备

对 1.2 节生成的数据集中 51 163 条 MOFs 数据的 CH_4 气体吸附值以 20 为长度进行区间划分, 进而对数据的分布情况进行初步统计, 结果如图 2 所示. 统计结果显示, 在该数据集中, 存在极少数 CH_4 气体吸附值大于 480 的 MOFs. 这种数据分布的倾斜, 会影响模型的学习和预测性能. 因此, 我们从吸附值大于 280 的 MOFs 样本中随机重复抽取一定数量的样本, 然后对每一个样本的吸附值引入以该吸附值为均值、方差为 1 的高斯随机误差. 经过这样的随机上采样后, 数据集扩充到 67 878 条, 其分布如图 3 所示.

对经过上采样后的 67 878 条数据的各特征值进行最大最小标准化预处理, 以消除数据集不同特征取值范围不同对模型训练的影响, 并加快模型训练的收敛速度. 最大最小标准化的方法如式 (3) 所示:

$$x^* = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{3}$$

其中, $x_{\max}$ 是样本数据的最大值, $x_{\min}$ 是样本数据的最小值.

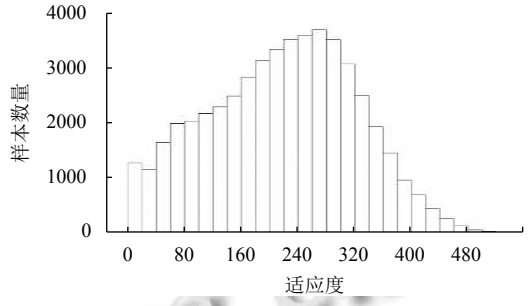


图 2 原始数据集分布直方图

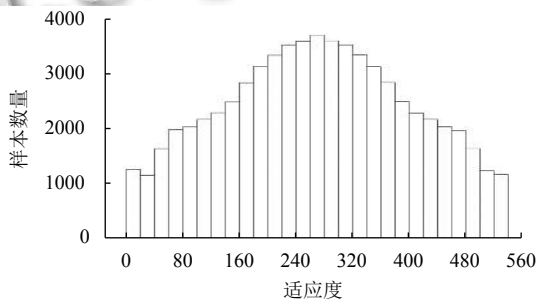


图 3 上采样后数据集分布直方图

经过以上预处理后, 我们将所有 67 878 条数据随机抽取 80% 作为训练集, 剩余 20% 作为测试集.

2.3 BPNN 和 RBFNN

本文分别采用 BP 神经网络 (Back Propagation Neural Network, BPNN) 和径向基函数神经网络 (Radial Basis Function Neural Network, RBFNN) 对 GA 生成的新型 MOFs 个体进行了针对 CH_4 气体吸附性能的预测评估实验. 利用 BP 神经网络与径向基函数神经网络进行针对 CH_4 气体吸附性能预测评估的优点主要有:

(1) BP 神经网络拥有高容错性, 并行计算, 自适应和可学习等优点, 在针对 MOFs 材料吸附 CH_4 气体能力这类非线性关系的预测方面具有显著的优势.

(2) 径向基神经网络在本文中设置为 BP 神经网络的对照, 作为一种性能优秀的前馈型神经网络, 理论上可以逼近任意非线性函数, 具有全局逼近能力, 从根本上解决了 BP 神经网络由于梯度下降所导致的局部最优问题, 且由于其整体网络结构紧凑, 收敛速度快. 而 BP 神经网络中权值调节采用负梯度下降法, 收敛速度递减而较慢.

(3) BP 神经网络学习速率是固定的, 因此对于一些复杂问题, BP 算法需要的训练时间可能非常长, 这主要是由于学习速率太小造成的. 而径向基神经网络是高效的前馈式网络, 它训练速度相对较快的同时, 具有 BP 神经网络所不具有的最佳逼近性能和全局最优特性.

2.3.1 BPNN 的构建和训练

BPNN 是 ANN 中经典且常见的一种神经网络. 结构上, BPNN 包含输入层、隐含层和输出层. 其本质是通过对高维数据的低维非线性映射, 转变为人类可理解的输出结果. 由于 1.1 节中将 MOFs 结构编码为具有 6 个基因的染色体, 故本文 BPNN 的输入层相应地设置为 6 个神经元, 对应 6 个基因的输入, 而输出为 1 个神经元, 用于预测吸附值. 在保持较为简单的网络结构的前提下, 经过多次调整、实验, 最终确定了 2 个隐层后增加 1 个批归一化层的基本结构. 具体地, 隐藏层的激活函数使用 ReLU 函数, 输出层以 Sigmoid 函数作为激活函数. 训练的 epoch 设为 100, batch_size 设为 128, 采用随机梯度下降 (Stochastic Gradient Descent, SGD) 迭代训练模型, 并采用 Adam 方法进行优化, 学习率设为 0.002. 其中, 激活函数是一种神经网络常用的非线性函数, 用于实现对上一层神经元输出的线性组合进行非线性变换. 批归一化 (Batch Normalization, BN) 层的 Scale and Shift 操作, 可以加速训练过程的收敛、控制过拟合、降低网络对初始化权重的敏感程度, 并允许使用比较大的学习率.

通过用训练集进行 5 折交叉验证, 最终网络结构调整实验的结果如表 2 所示. 表中结果按照 R^2 降序排列. 可以看到, 2 个隐层神经元个数分别为 32 和 15 时的模型准确度最高. BPNN 训练过程需要调节的参数个数为$6\times32\times15+15=2895$.

2.3.2 RBFNN 的构建和训练

作为比较, 我们还构建了另外一种常见的人工神经网络——径向基函数神经网络. RBFNN是一种前馈型的 3 层神经网络, 激励函数使用径向基函数. 其隐含层中神经元与输入、输出层的神经元之间的关系不再是全连接, 而是用径向基函数代替. 本文中, 采用常用的高斯函数作为径向基函数. 与 BPNN 相比, RBFNN 通常泛化能力更强, 能够避免 BPNN 可能出现的局部最优问题, 理论上能够在充分训练的情况下完全逼近要拟合的数据. 该网络可以方便地增加神经元进行训练, 直到满足精度要求为止, 这样的网络结构拥有更为突出的定向信息处理能力. 本文通过调整神经网络的结构, 将隐含层神经元个数从 100 个开始, 逐次递增, 每次调节增加 100 个神经元, 观察评价指标 R^2 与 *MSE* 的数值变化情况, 从而确定最优的网络结构.

表 2 BPNN 结构调节实验结果

各隐层神经元个数(不含BN层)	R^2	*MSE*
32, 15	**0.87**	**2072.54**
32, 10	0.86	2277.64
32, 16	0.86	2197.68
32, 4	0.86	2286.26
16, 8	0.85	2391.52
32, 2	0.83	2656.54
9, 3	0.78	2556.24
8, 2	0.69	3312.34
9, 4	0.63	3893.6
10, 4	0.62	4014.69
9, 2	0.56	4637.91
10, 3	0.53	4953.52
8, 3	0.47	5543.52
8, 4	0.26	7845.51

使用与 2.3.1 节相同的训练集进行 5 折交叉验证, 比较训练结果, 得到当隐层节点设置为 600 个时, 其 *MSE*=2264.83、R^2=0.854 为最优, 即 RBFNN 的结构确定为 6-600-1. 调节隐含层神经元个数的比较结果如表 3 所示. 从中可以看到, 隐层节点数为 800 时, 结果与隐层节点数为 600 的相差无几, 但从模型复杂度、参数数量等角度综合考虑, 最终, 选取隐层节点数为 600 为最合适的网络隐层节点个数. 由于径向基函数神经网络是局部逼近网络, 其对于输入空间的某个局部区域只有少数几个连接权值影响输出, 故而该网络实际需要调节的参数数量大大小于 BP 神经网络.

表 3 RBFNN 结构调节实验结果

各隐层神经元个数	R^2	*MSE*
100	0.815	2824.41
200	0.825	2612.60
300	0.842	2395.55
400	0.846	2358.59
500	0.846	2366.56
600	**0.854**	**2264.83**
700	0.852	2281.90
800	0.852	2265.85
900	0.841	2364.48

2.3.3 BPNN 和 RBFNN 的性能比较

通过上述对 BPNN 和 RBFNN 结构的优化, 我们分别训练并确定了 BPNN 和 RBFNN 的结构和参数.

该过程已经初步显示了二者的性能. 图 4 和图 5 分别展示了二者在测试集上回归预测的具体性能表现.

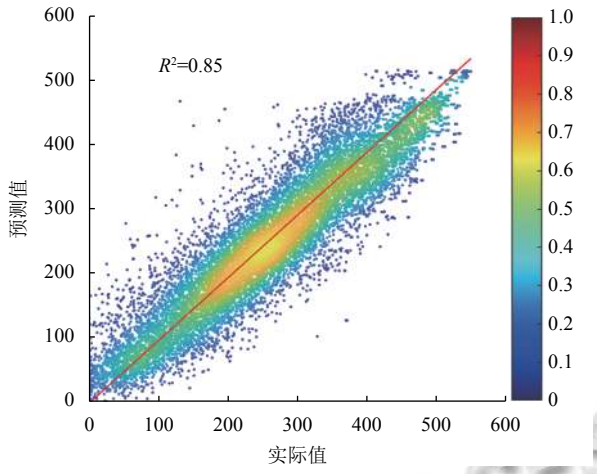


图 4 BPNN 网络模型在测试集上的回归散点图

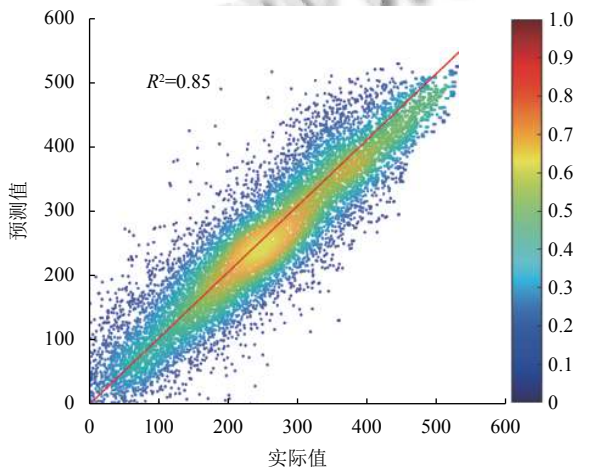


图 5 RBFN 网络模型在测试集上的回归散点图

上述模型回归结果图中, 颜色表明数据点的密集程度, 颜色越接近红色, 数据分布越密集. 当实际值与预测值接近时, 数据点会均匀分布在红色标识实线及两侧. 从图 4 中可以直观地看出, 采用 2.3.1 节所得的 6-32-15-1 结构的 BPNN 时, 在测试集上的实验结果 R^2=0.85; 类似地, RBFN 以 2.3.2 节所得的隐层节点数为 600 时, 其在测试集上得到 R^2=0.85. 可见, 两种模型在测试集上均可以实现较为准确的回归. 这些结果表明基于训练数据构建的非线性模型具有可以预测新型 MOFs 材料气体吸附性能的能力.

3 实验结果

基于 1.2 节构建的基础数据集, 以及 2.2 节对于数据集的处理和划分, 我们应用 GA 实现了对 MOFs 个体的演化搜索, 并应用 ANN 对搜索到的新型 MOFs 进行了基于 CH_4 吸附值的性能预测评估实验.

3.1 基于 GA 的 MOFs 搜索

GA 是一种具有生存与检测特征、不断进行迭代过程的一种全局优化搜索算法. 在迭代过程中, 会产生大量通过基因编码表示的个体, 每个个体的基因特征会随着进化的进行, 根据优胜劣汰的基本原则进行代际遗传, 从而产生优秀个体, 实现在搜索空间中对最优解的搜索.

GA 的主要参数包括种群规模 M、进化代数 N、遗传交叉率 a, 以及遗传变异率 b. 其中, a 决定了两个个体进行交叉操作从而产生新子代的概率, b 为一个个体的某个随机基因发生变异的概率. GA 中的种群规模, 代表着数据域内数据点的密度, 密度越大, 其覆盖最优解的可能性越高, 即对求解最优解越有利. 但相应的, 其计算量也将会快速增加; 对于 GA 进化代数的限制, 是为了让算法能够在合理的实验时间内完成迭代搜索; GA 中的变异概率与交叉概率设定, 是为了让数据域内的数据点保持相对分散的分布, 避免陷入局部最优的困境, 文献 [28] 中的设定为<M=100, N=100, a=0.65, b=0.05>. 此外, 适应度函数也是 GA 的一项重要设定. 通过适应度函数, GA 计算个体的适应度, 评估个体的性能. 适应度越高, 种群越朝着有利于发展的方向进化. 文献 [28] 采取的操作是, 在原数据集中查找新生成的个体, 若找到, 则通过 GCMC 模拟方法计算它对 CO_2 的工作容量、CO_2/H_2 的选择性, 以及对 CO_2 的吸附值, 分别以这 3 个指标作为个体的适应度值以评估个体性能; 若新生成的个体不在原数据集中, 则重新进行基因操作, 直到生成数据集中存在的个体. 也就是说, 文献 [28] 中并未对原数据集中不存在的新个体进行评估并加入新的子代.

本文以 1.2 节计算的 MOFs 对 CH_4 气体的吸附值 F 作为适应度, 参照文献 [28] 中的参数设定, 以产生新型 MOFs 个体的数量 X、搜索到最优个体所进化的代数 G, 以及搜索到的最优个体吸附值 F 为评价指标, 针对 M 和 N 两个参数进行了 6 组参数设定的实验. 具体步骤为:

(1) 在原始数据集上构建初始种群. 初始种群中的个体可从数据集中进行多次随机选择并择优, 也可加入一些人为设定的策略. 例如, 文献 [28] 中人工选择

100 个 MOFs 个体构建初始种群, 从而保证所设计的每个基因都至少出现一次, 个体演化过程中不会有基因的缺失.

(2) 执行遗传算法, 开始种群的演化. 这个过程包含了遗传算法中的经典操作, 例如从种群中进行个体的择优、交叉、变异, 从而产生下一代种群, 不断迭代, 直到算法停止

(3) 在 GA 迭代演化过程中, 对于产生的 MOF 个体, 如果存在于原数据集中, 则直接使用其已经计算得到的目标性能指标值作为个体性能的评估结果, 并加入下一代种群; 否则将新个体暂存.

(4) 算法的停止条件, 可以为指定的演化迭代次数、指定的搜索到新的优秀个体数量等.

按照上述实验步骤不断循环迭代, 本文依据实际实验条件, 综合考虑遗传算法的计算效果与计算周期, 调整实验参数, 经过 6 组实验最终将遗传算法参数设定为<M=50, N=200, a=0.65, b=0.05>, 如表 4 所示.

最终, GA 算法搜索到 907 个原数据集中不存在的新 MOFs 个体. 同时我们观察到, 原始数据集中, 第 1 个基因 (潜在互穿能力) 的值均不小于第 2 个基因 (实际互穿能力) 的值. 这是因为, 潜在互穿能力表示理论上可能的互穿能力, 所以实际互穿能力不会超过它. 因此, 我们将 907 个新型 MOFs 个体中不符合该条件的 144 个删除, 剩余 763 个新型 MOFs 个体作为实验对象.

表 4 GA 参数组合实验结果

M	N	a	b	X	G/F
50	50	0.65	0.05	368	34/528
50	100	0.65	0.05	615	34/528
50	**200**	**0.65**	**0.05**	**907**	**34/528**
100	50	0.65	0.05	577	50/512
100*	100*	0.65	0.05	942	100/522
100	200	0.65	0.05	1426	200/523

*注: 本组M=100, N=100的实验参数来自文献[28].

3.2 ANN 对新型 MOFs 个体的性能预测

本文分别采用前文所述的 BPNN 和 RBFNN 对 GA 搜索到的 763 个新型 MOFs 个体进行 CH_4 气体的吸附值预测, 取二者预测的对 CH_4 气体吸附值最高的前 10 位 MOFs 个体进行比较, 如表 5 所示.

表 5 BPNN 和 RBFNN 分别对新型 MOFs 个体的 CH_4 气体吸附值预测结果 TOP10 对比

序号	BPNN预测前10位新型MOFs			RBFNN预测前10位新型MOFs		
	基因编码	预测吸附值	RBFNN预测值	基因编码	预测吸附值	BPNN预测值
1	0-0-1-38-21-12	552.14	468.70	1-0-0-29-37-10	521.20	344.61
2	0-0-1-38-21-13	551.11	404.42	1-0-1-38-21-10	520.82	540.58
3	0-0-1-37-21-12	549.71	467.50	2-2-1-36-36-11	519.78	254.89
4	1-0-1-37-21-13	546.77	427.37	1-0-0-29-30-10	519.09	489.62
5	1-0-1-37-21-12	542.00	482.27	**1-0-1-37-21-10**	**518.94**	**532.45**
6	0-0-1-36-21-12	541.59	465.17	1-0-1-36-21-10	515.96	524.32
7	1-0-1-38-21-10	540.58	520.82	1-0-0-29-34-8	513.42	401.12
8	**1-0-1-38-21-9**	**535.81**	**513.21**	**1-0-1-38-21-9**	**513.21**	**535.81**
9	1-0-1-36-21-12	533.87	478.72	1-0-1-35-21-10	512.00	516.19
10	**1-0-1-37-21-10**	**532.45**	**518.94**	0-0-0-29-31-10	511.53	458.39
	平均值	542.60	474.71		516.60	459.80

从两者结果比较可以看出, BPNN 预测的前 10 种新型 MOFs 材料, 其 CH_4 气体吸附能力的均值为 542.60, 略高于 RBFNN 预测结果前 10 名的 516.60. 有趣的是, BPNN 预测的前 10 位 MOFs, 对 CH_4 气体的吸附性能均在 530 以上, 高于原始数据集内的最大值 528, 突破了训练集的限制, 具有更好的泛化能力. 而 RBFNN 预测的 CH_4 气体吸附值最大为 521.20, 未能突破训练集的范围. 从基因编码的结构上看, BPNN 对于结构相近的 MOFs 个体, 预测的 CH_4 气体吸附值也较为接近. 例如, 预测基因编码结构为 0-0-1-38-21-12 的 CH_4 气体吸附值为 552.14, 与其相近的基因编码结构为 0-0-1-38-21-13 的 CH_4 气体吸附值为 551.11. 这个结果具有一定的合理性. 另一方面, RBFNN 的预测结果具有更强的多样性, 得到的高 CH_4 气体吸附值的基因编码结构与 BPNN 预测得到的有很大不同.

具体地, BPNN 预测得到的高 CH_4 气体吸附值的 MOFs 个体, 其潜在互穿性仅限于 1 或 0, 而实际互穿性均保持在 0; 而 RBFNN 预测得到的高 CH_4 气体吸附值的 MOFs 个体, 潜在互穿性和实际互穿性均出现了 2 的取值. 对于第 3 个基因编码, BPNN 预测得到的

高 CH_4 气体吸附值的前 10 名均为 1, 而 RBFNN 的结果中还包含 0. 根据文献 [28] 的补充材料中的设定, 该位基因为 0 和 1 所表示的无机配体, 如图 6 所示.

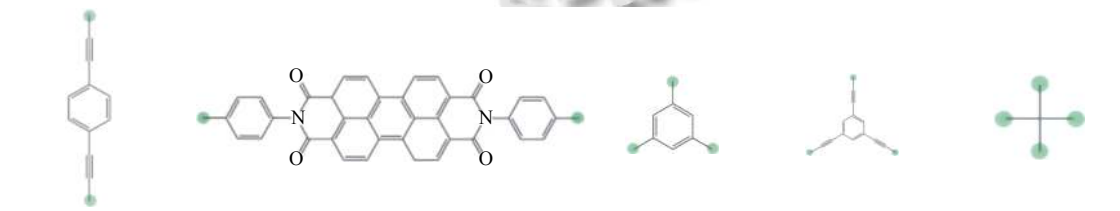
(a) 值为 0 表示的无机配体 (b) 值为 1 表示的无机配体

图 6 无机配体的表示[28]

BPNN 的结果中, 主要有机连接单元出现了 36、37、38, 而 RBFNN 的结果中还出现了 29 和 35. 第 5 位的次要有机连接单元, BPNN 预测得到的结果中均为 21, RBFNN 的结果中除了 21, 还出现了 30、31、34、36、37. 根据文献 [28] 的补充材料中的设定, 它们表示的结构如图 7 所示.

最后一位基因值表示的化学官能团, BPNN 预测的结果中出现了 9、10、12、13, 而 RBFNN 的结果中则为 8、9、10、11. 根据文献 [28] 的补充材料中的设定, 它们表示的结构如图 8 所示.

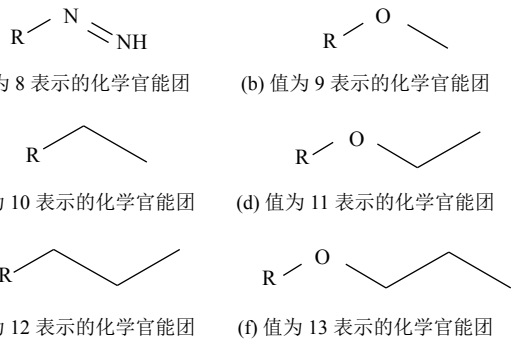
以上结果表明, 特定的几种结构将给 MOFs 带来较高的 CH_4 气体吸附值. 值得注意的是, BPNN 和 RBFNN 均预测出 1-0-1-38-21-9、1-0-1-37-21-10 结构的 MOF 具有相对较高的 CH_4 气体吸附值, 这值得后续的研究工作进一步关注.



(a) 值为 21 表示的有机连接单元 (b) 值为 29 表示的有机连接单元 (c) 值为 30 表示的有机连接单元 (d) 值为 31 表示的有机连接单元 (e) 值为 34 表示的有机连接单元

(f) 值为 35 表示的有机连接单元 (g) 值为 36 表示的有机连接单元 (h) 值为 37 表示的有机连接单元 (i) 值为 38 表示的有机连接单元

图 7 有机连接单元的表示[28]



(a) 值为 8 表示的化学官能团 (b) 值为 9 表示的化学官能团

(c) 值为 10 表示的化学官能团 (d) 值为 11 表示的化学官能团

(e) 值为 12 表示的化学官能团 (f) 值为 13 表示的化学官能团

图 8 化学官能团的表示[28]

4 结论与展望

本文首先根据我们提出的力场参数, 基于现有的 MOFs 数据集, 通过 GCMC 模拟计算构建了面向一定条件下 CH_4 气体吸附能力的 MOFs 数据集, 并通过上采样技术调整了数据集的分布. 其次, 以该数据集分别训练了 BPNN 和 RBFNN 模型, 使其具备较强的预测 CH_4 气体吸附性的能力. 然后, 通过 GA 基于 MOFs 数据库搜索新型的 MOFs 个体. 搜索时, 对于搜索到的数据集中已有的 MOFs, 直接查询数据集中其对应的 CH_4 气体吸附值; 对于搜索到的不在数据集中的新型 MOF, 则暂存它们. 最后搜索出 763 个新型 MOFs 个体, 并分别用训练好的 BPNN 与 RBFNN 对其 CH_4 气体吸附性进行预测, 得到了优于原始数据集的结果. 通过以上过程, 实现了通过 GA 搜索新型 MOFs, 并用 ANN 对其进行性能预测, 从而实现高性能 MOFs 的高效搜索与评估.

实验结果表明, BPNN 在模型的准确性, 泛化能力方面略优于 RBFNN, 而 RBFNN 预测结果则更具备多样性. 二者的预测结果均表现出特定的结构对 MOF 性能有一定的影响. 对于二者均预测出具有较高 CH_4 气体吸附值的两种 MOF 新型结构, 则需要进一步的研究

和验证.

未来可进一步拓展现有工作, 从而引入多种机器学习方法作为参照进行比较和相互佐证. 对 GA 参数更深入的优化研究也是一个有挑战性的课题方向. 同时, 可考虑结合实际化学材料领域中的自组装技术, 通过材料组装, 模拟, 优化出新材料的分子结构, 然后再利用 GCMC 手段, 添加材料的实际化学特征值数据, 完善成果.

参考文献

1 Silver D, Huang A, Maddison CJ, *et al*. Mastering the game of Go with deep neural networks and tree search. Nature, 2016, 529(7587): 484–489. [doi: 10.1038/nature16961]

2 Silver D, Schrittwieser J, Simonyan K, *et al*. Mastering the game of Go without human knowledge. Nature, 2017, 550(7676): 354–359. [doi: 10.1038/nature24270]

3 Shin HC, Roth HR, Gao MC, *et al*. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Transactions on Medical Imaging, 2016, 35(5): 1285–1298. [doi: 10.1109/TMI.2016.2528162]

4 Cambria E, White B. Jumping NLP curves: A review of natural language processing research. IEEE Computational Intelligence Magazine, 2014, 9(2): 48–57. [doi: 10.1109/MCI.2014.2307227]

5 Liu HT, Xu CS, Liang JY. Dependency distance: A new perspective on syntactic patterns in natural languages. Physics of Life Reviews, 2017, 21: 171–193. [doi: 10.1016/j.plrev.2017.03.002]

6 Tsai CW, Lai CF, Chiang MC, *et al*. Data mining for Internet of Things: A survey. IEEE Communications Surveys & Tutorials, 2014, 16(1): 77–97.

7 Feng N, Wang HJ, Li MQ. A security risk analysis model for information systems: Causal relationships of risk factors and vulnerability propagation analysis. Information Sciences, 2014, 256: 57–73. [doi: 10.1016/j.ins.2013.02.036]

8 Argall BD, Chernova S, Veloso M, *et al*. A survey of robot learning from demonstration. Robotics and Autonomous Systems, 2009, 57(5): 469–483. [doi: 10.1016/j.robot.2008.10.024]

9 Cully A, Clune J, Tarapore D, *et al*. Robots that can adapt like animals. Nature, 2015, 521(7553): 503–507. [doi: 10.1038/nature14422]

10 Kononenko I. Machine learning for medical diagnosis: History, state of the art and perspective. Artificial Intelligence in Medicine, 2001, 23(1): 89–109. [doi: 10.1016/S0933-3657(01)00077-X]

11 Kononenko I. Inductive and Bayesian learning in medical diagnosis. Applied Artificial Intelligence, 1993, 7(4): 317–337. [doi: 10.1080/08839519308949993]

12 Lu WC, Xiao RJ, Yang J, *et al*. Data mining-aided materials discovery and optimization. Journal of Materiomics, 2017, 3(3): 191–201. [doi: 10.1016/j.jmat.2017.08.003]

13 Ramprasad R, Batra R, Pilania G, *et al*. Machine learning in materials informatics: Recent applications and prospects. npj Computational Materials, 2017, 3: 54. [doi: 10.1038/s41524-017-0056-5]

14 Li JR, Sculley J, Zhou HC. Metal-organic frameworks for separations. Chemical Reviews, 2012, 112(2): 869–932. [doi: 10.1021/cr200190s]

15 Ockwig NW, Delgado-Friedrichs O, O'Keeffe M, *et al*. Reticular chemistry: Occurrence and taxonomy of nets and grammar for the design of frameworks. Accounts of Chemical Research, 2005, 38(3): 176–182. [doi: 10.1021/ar020022l]

16 Wilmer CE, Leaf M, Lee CY, *et al*. Large-scale screening of hypothetical metal-organic frameworks. Nature Chemistry, 2012, 4(2): 83–89. [doi: 10.1038/nchem.1192]

17 Simon CM, Kim J, Gomez-Gualdron DA, *et al*. The materials genome in action: Identifying the performance limits for methane storage. Energy & Environmental Science, 2015, 8(4): 1190–1199.

18 Simon CM, Kim J, Lin LC, *et al*. Optimizing nanoporous materials for gas storage. Physical Chemistry Chemical Physics, 2014, 16(12): 5499–5513. [doi: 10.1039/c3cp55039g]

19 Sikora BJ, Wilmer CE, Greenfield ML, *et al*. Thermodynamic analysis of Xe/Kr selectivity in over 137000 hypothetical metal-organic frameworks. Chemical Science, 2012, 3(7): 2217–2223. [doi: 10.1039/c2sc01097f]

20 Martin RL, Simon CM, Smit B, *et al*. *In silico* design of porous polymer networks: High-throughput screening for methane storage materials. Journal of the American Chemical Society, 2014, 136(13): 5006–5022. [doi: 10.1021/ja4123939]

21 Lin LC, Berger AH, Martin RL, *et al*. *In silico* screening of carbon-capture materials. Nature Materials, 2012, 11(7): 633–641. [doi: 10.1038/nmat3336]

22 Chung YG, Camp J, Haranczyk M, *et al*. Computation-ready, experimental metal-organic frameworks: A tool to enable high-throughput screening of nanoporous crystals. Chemistry of Materials, 2014, 26(21): 6185–6192. [doi: 10.1021/cm502

594j]

23 Le T, Epa VC, Burden FR, *et al*. Quantitative structure–property relationship modeling of diverse materials properties. Chemical Reviews, 2012, 112(5): 2889–2919. [doi: 10.1021/cr200066h]

24 Fernandez M, Trefiak NR, Woo TK. Atomic property weighted radial distribution functions descriptors of metal-organic frameworks for the prediction of gas uptake capacity. The Journal of Physical Chemistry C, 2013, 117(27): 14095–14105. [doi: 10.1021/jp404287t]

25 Fernandez M, Woo TK, Wilmer CE, *et al*. Large-scale Quantitative Structure-Property Relationship (QSPR) analysis of methane storage in metal-organic frameworks. The Journal of Physical Chemistry C, 2013, 117(15): 7681–7689. [doi: 10.1021/jp4006422]

26 Fernandez M, Boyd PG, Daff TD, *et al*. Rapid and accurate machine learning recognition of high performing metal organic frameworks for CO_2 Capture. The Journal of Physical Chemistry Letters, 2014, 5(17): 3056–3060. [doi: 10.1021/jz501331m]

27 Sezginel KB, Uzun A, Keskin S. Multivariable linear models of structural parameters to predict methane uptake in metal-organic frameworks. Chemical Engineering Science, 2015, 124: 125–134. [doi: 10.1016/j.ces.2014.10.034]

28 Chung YG, Gómez-Gualdrón D, Li P *et al*. *In silico* discovery of metal-organic frameworks for precombustion CO_2 capture using a genetic algorithm. Science Advances, 2016, 2(10): e1600909. [doi: 10.1126/sciadv.1600909]

29 Colón YJ, Gómez-Gualdrón DA, Snurr RQ. Topologically guided, automated construction of metal–organic frameworks and their evaluation for energy-related applications. Crystal Growth & Design, 2017, 17(11): 5801–5810.

30 Zhang C, Wang L, Maurin G, *et al*. In silico screening of MOFs with open copper sites for C_2H_2/CO_2 separation. AIChE Journal, 2018, 64(11): 4089–4096. [doi: 10.1002/aic.16376]

31 Ewald PP. Die berechnung optischer und elektrostatischer gitterpotentiale. Annalen der Physik, 1921, 369(3): 253–287. [doi: 10.1002/andp.19213690304]

32 Vlugt TJH, García-Pérez E, Dubbeldam D, *et al*. Computing the heat of adsorption using molecular simulations: The effect of strong coulombic interactions. Journal of Chemical Theory and Computation, 2008, 4(7): 1107–1118. [doi: 10.1021/ct700342k]

33 Jasuja H, Walton KS. Effect of catenation and basicity of pillared ligands on the water stability of MOFs. Dalton Transactions, 2013, 42(43): 15421–15426. [doi: 10.1039/c3dt51819a]

34 Burtch NC, Jasuja H, Walton KS. Water stability and adsorption in metal-organic frameworks. Chemical Reviews, 2014, 114(20): 10575–10612. [doi: 10.1021/cr5002589]

35 Jasuja H, Burtch NC, Huang YG, *et al*. Kinetic water stability of an isostructural family of zinc-based pillared metal−organic frameworks. Langmuir, 2013, 29(2): 633–642. [doi: 10.1021/la304204k]