工作年限：一年

# 杨文远

**求职意向：算法工程师/图像处理工程师**

生日：1993.12.08

籍贯：湖南 · 邵阳

## 教育背景

**2017.09-2020.06　　广州大学　　化学工程　　工学硕士**

研究主题：机器学习和大数据筛选多孔材料的气体分离性能

**2013.09-2017.06　　吉首大学　　化学工程与工艺　　工学学士**

## 科研经历

**2019.11-2020.03　　马克斯普朗克研究所　　德国•马格德堡**

研究助理

工作描述：

- 采用分子模拟，遗传算法和 LSTM 结合，实现多孔材料的定向优化，设计；
- 采用图像分分割 Unet 对多孔材料进行特征提取，并分割出“站立”材料，微调 coco 数据集完成的预训练权重，用数据增强增加样本量，计算晶体厚度；
- 结合粒子群多目标优化算法和套索回归对多孔材料结构-性能间的关系进行优化，改进。

**2017.09-2020.06　　广州大学　　广东•广州**

工作描述：　**（发表论文：4 篇，SCI）**

1. 对 21 种气体组分组成的 15 种性能指标，采用 PCA 降维，结合 DT,SVM, BPNN, RF 进行预测，并规划处一条性能最优的结构路径。
2. 采用 MS 软件对基因材料数据库自组装的材料进行合成和晶体图片导出，结合 CNN 提取材料特征，对提取的特征采用 SVM 做目标性能分类，实现原料到性能的筛选。
3. 大数据筛分多孔材料对硫醇的分离，利用 BPNN 和 RF 实现结构-性能的预测并探索材料结构特征的权重大小。

## 学校荣誉

- **荣誉类：**获得广州大学国家奖学金；第五届国际可持续化学产品和工艺工程最佳墙报奖；广州大学化学化工学院研究生年会论坛一等奖
- **活动类**：主持广东省攀登计划一般项目 1.5 万元资助，“基于高通量计算和人工智能的 MOF 吸附剂自动设计方法的研究。”

## 技能篇

**机器学习**：了解传统机器学习算法（DT, SVM， RF，聚类算法等）

**深度学习**：掌握 CNN 和 DCGANS，熟悉 Tensorflow，pytorch 等深度学习框架，了解多 GPU，分布式（一般 horovod），并行式计算（CUDA, cuDNN 加速器）

**图像识别**：掌握 RCNN 和 Faster RNN 以及盛行的 YOLO 系列（v1~v3）和 SSD，图像分割等。

**软件**：MATLAB, python, linux（操作系统），postman，Material Studio

**(对前段所使用的 html，css，JavaScript 有部分了解，能对照看懂)**

## 自我评价

- 做事踏实，认真，好学，抗压能力强，效率高；
- 勇于创新，敢于挑战；喜欢独立思考，喜欢探索新事物；
- 青睐有激情有活力的团队；
- 虚心，坦诚

**工作经历：2020.7.20 ——**

**公司名称： 深圳市创智链科技有限公司**

**工作岗位：图像算法工程师 （大数据算法组组长 1/4）**

**项目内容：**

一．智能称对农副产品的精准识别 （**2020.8——至今**）

终端产品：智能识别称，**SDK**

**主要工作：**

1. 负责图像数据库的采集标准制定；

2. 基于传统和前沿工作，负责算法模型的优化，压缩（量化，蒸馏和剪枝）；

3. 负责前期图像数据的预处理算法开发（降噪，去雾，白平衡）；

4. 负责对同行业公司算法模型应用的分析，评估；

5. 负责前沿文献的解读与 SOTA 模型的选取；

6. 负责算法模型的迁移，部署（嵌入式或云端，基于 docker）；

7. 参与公司的运维管理及目标定位；

8. 协助日常算法模型的推送，维护及市场反馈数据的统计分析。

**业绩及部分优化策略：**

1. 基于前沿开源网络（2019 年），得到基准 baseline；

2. 参考人脸相似性比对，研究中心损失，三元损失和边际损失（center loss，triplet loss）的识别方案；

3. 采用 CBAM 和 self-attention 等注意力机制优化，提升模型 1%的准

确性；

4. 基于空间金字塔优化模型和 SVD 分解权衡模型参数量和准确性；

5. 考虑图片因压缩而失真，采用 Facebook（2020）研究，调整模型输入输出分辨率，准确性提高 1.5%；

6. 基于最新前沿的优化方法，取代经典的提升模型鲁棒性的方法（例：softpool 取代 pool；autodropout 取代 dropout（均 2020 和 2021 年研究成果））；

6. 采用传统的降噪和去雾方法对图像做前期预处理。准确性提升 2%；

7. 基于 Noisestudent，添加标签噪音，提升整体 1%的准确性。

8. 撰写两篇相关发明专利（一种级联分类识别方法和设备；一种农副品动态智能识别方法）[均通过初审]

**二、物体检测 （自学自研）： （2021.4——2021.6）**

**对人，物体进行目标检测：**

●基于 coco 数据集的预训练权重，对已有数据集进行迁移学习（已有数据：网上爬取，已标注好的），通过整体的准确性，误差分析和测试图片中物体画框的位置，来调优算法模型的超参数，并对 YOLOV3 的基础模型 darknet53 添加注意力机制，更改激活函数（hswish 替换 leakyrelu），以及用 LN,GN 等来取代 BN 等，对比了 Faster RCNN，并利用 google 最新的算法模型 EfficientNet 来提取图片特征。

1. 熟悉经典的目标检测算法 yolo 和 Faster RCNN 等

2. 熟悉 pytorch 和 tensorflow

3. 熟悉 Linux 操作系统

4. 查阅和研究了部分最新目标检测相关的研究文献

人选情况：
人选认为和公司的业务、岗位还挺匹配的，地点可以接受的。
已经提了离职了，目前17k，期望薪资是20-25k。

下面有学历证书和一系列证明

## 硕士研究生学历证明：

硕士研究生

毕业证书

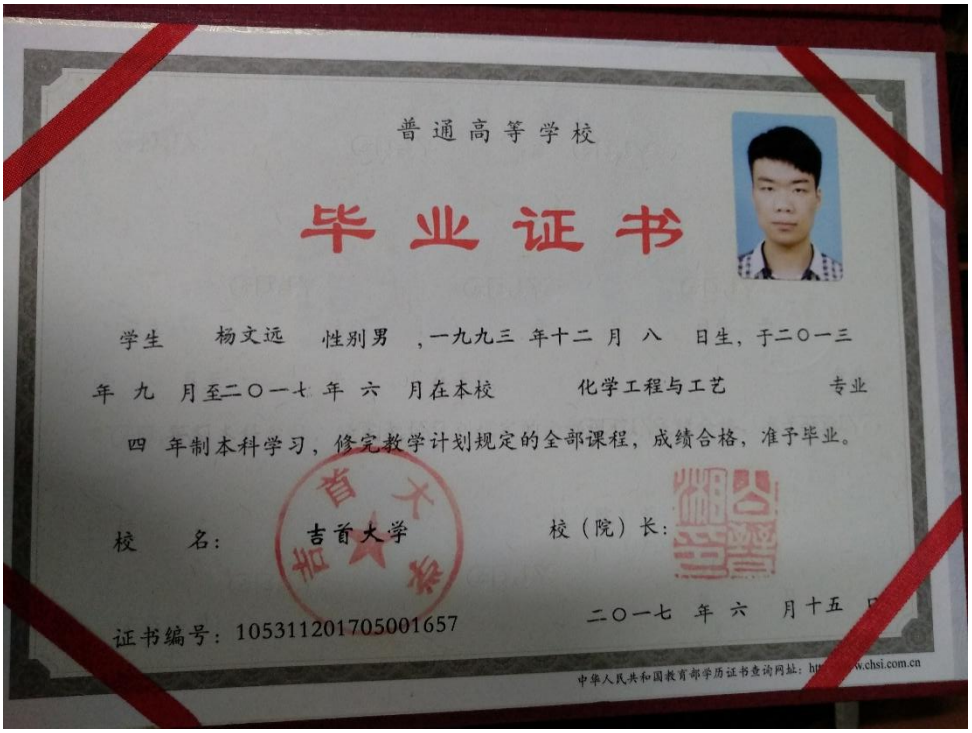
研究生 杨文远 性别 男，一九九三 年 十二 月 八 日生，于二〇一七 年 九 月至二〇二〇 年 六 月在 化学工程 专业 全日制 学习，学制 三 年，修完硕士研究生培养计划规定的全部课程，成绩合格，毕业论文答辩通过，准予毕业。

培养单位：广州大学 校 长：

证书编号：110781202002000263 二〇二〇 年 六 月二十九日

中华人民共和国教育部学历证书查询网址：http://www.chsi.com.cn

## 本科学历证明：



普通高等学校

毕业证书

学生 杨文远 性别男 ，一九九三 年十二 月 八 日生，于二〇一三 年 九 月至二〇一七 年 六 月在本校 化学工程与工艺 专业 四 年制本科学习，修完教学计划规定的全部课程，成绩合格，准予毕业。

校 名：吉首大学 校（院）长：

证书编号：105311201705001657 二〇一七 年 六 月十五 日

中华人民共和国教育部学历证书查询网址：http://www.chsi.com.cn

## 广州大学国家奖学金证明：


## 第五届国际可持续化学产品和工艺工程最佳墙报奖证明：

The 5th International Conference on Sustainable Chemical Product and Process Engineering

Best Poster Presentation Award

Presented to Wenyan Yang

2019 SCPPE

Prof. Xinbin Ma
Co-Chair, SCPPE 2019

Prof. Yinlun Huang

# 马克斯普朗克研究所_研究助理_证明：

OTTO VON GUERICKE
UNIVERSITÄT
MAGDEBURG

**Wenyuan**
**Yang**

**Matrikel-Nr.: 230751**

STUDIERENDENAUSWEIS - STUDENT IDENTITY CARD
Semesterticket / M
gueltig bis: 31.03.2020

Die Max-Planck-Gesellschaft zur Förderung der Wissenschaften e.V.,
vertreten durch den – Geschäftsführenden – Direktor des

Max-Planck-Instituts für Dynamik komplexer technischer Systeme
und

Herr Wenyuan Yang

schließen folgenden

**Arbeitsvertrag**

**§ 1**

Herr Wenyuan Yang, geboren am 08.12.1993, wird ab 01.12.2019 an dem Max-Planck-Institut für Dynamik komplexer technischer Systeme als wissenschaftliche Hilfskraft im Sinne der Richtlinien für die Förderung des wissenschaftlichen Nachwuchses und die wissenschaftliche Zusammenarbeit mit dem Ausland in ihrer jeweils geltenden Fassung befristet eingestellt.

Das Arbeitsverhältnis endet mit Ablauf des **29.02.2020**, ohne dass es einer Kündigung bedarf.

Die Befristung des Arbeitsverhältnisses beruht auf dem Gesetz über befristete Arbeitsverträge in der Wissenschaft (Wissenschaftszeitvertragsgesetz – WissZeitVG). Die Erklärungen im Ergänzungsblatt zum Personalfragebogen „Vorbeschäftigungen in befristeten Arbeitsverhältnissen" sind Grundlage dieses Arbeitsvertrages. Die Vertrag schließenden Parteien sind sich einig, dass unrichtige oder lückenhafte Angaben im Fragebogen einen wichtigen Grund zur außerordentlichen Kündigung darstellen. Sie sind verpflichtet, eine jeweils gültige Immatrikulationsbescheinigung vorzulegen.

**§ 2**

Herr Wenyuan Yang obliegen folgende Tätigkeiten:

Der regelmäßige Dienstort ist Max-Planck-Institut, Magdeburg.

Herr Wenyuan Yang ist verpflichtet, aus dienstlichen oder betrieblichen Gründen andere gleichwertige Tätigkeiten zu übernehmen.

**§ 3**

Die Arbeitszeit beträgt *während der Vorlesungszeit* ausschließlich der Pausen

☐ wöchentlich durchschnittlich [Wert] Stunden.
☒ **monatlich durchschnittlich 80 Stunden.**
☐ nach Arbeitsanfall höchstens [Wert] Stunden wöchentlich.
☐ nach Arbeitsanfall höchstens [Wert] Stunden monatlich.

Die Arbeitszeit beträgt *in der vorlesungsfreien Zeit* ausschließlich der Pausen

☐ wöchentlich durchschnittlich [Wert] Stunden.
☐ monatlich durchschnittlich [Wert] Stunden.
☐ nach Arbeitsanfall höchstens [Wert] Stunden wöchentlich.
☐ nach Arbeitsanfall höchstens [Wert] Stunden monatlich.

Die Urlaubsansprüche richten sich nach dem Bundesurlaubsgesetz vom 8. Januar 1963 in seiner jeweils geltenden Fassung.

**§ 4**

*[Bei Vereinbarung einer wöchentlichen Arbeitszeit bzw. bei Bezahlung nach Arbeitsanfall]*
☐ Die Vergütung beträgt je Stunde [Wert] Euro.

*[Bei Vereinbarung einer monatlichen Arbeitszeit]*
☒ **Die Vergütung beträgt monatlich 971,20 Euro.**

Die Vergütung wird nur für tatsächlich geleistete Arbeit gezahlt. Vergütungsfortzahlung im Krankheitsfall richtet sich nach den gesetzlichen Bestimmungen.

*Auf Erfindungen* und technische Verbesserungsvorschläge finden die gesetzlichen *Bestimmungen* (Arbeitnehmererfindungsgesetz) und die Erfinderregelungen[1] der Max-*Planck*-Gesellschaft in ihrer jeweils geltenden Fassung Anwendung. Vereinbarungen *mit Dritten*, die Diensterfindungen oder technische Verbesserungsvorschläge betreffen, *bedürfen* der vorherigen Zustimmung durch die Generalverwaltung der Max-Planck-*Gesellschaft*. Bereits getroffene Vereinbarungen sind der Generalverwaltung der Max-Planck-Gesellschaft mitzuteilen.

Die Regeln zur Sicherung guter wissenschaftlicher Praxis der Max-Planck-Gesellschaft vom 24. November 2000 in ihrer jeweils geltenden Fassung sind Bestandteil des Arbeitsvertrages.

**§ 8**

Änderungen und Ergänzungen dieses Vertrages sind nur wirksam, wenn sie schriftlich vereinbart werden. Einer stillschweigenden Verlängerung des Arbeitsverhältnisses wird gem. § 625 BGB schon jetzt widersprochen. Nebenabreden sind nur wirksam, wenn sie schriftlich vereinbart werden.

Herr Wenyuan Yang hat eine Ausfertigung dieses Vertrages erhalten.

Magdeburg, 12.11.2019

| Für die Max-Planck-Gesellschaft | Studentische Hilfskraft |
|---|---|
| Prof. Dr. Ing. U. Reichl | Wenyuan Yang |

[1] Derzeit in der Fassung des Verwaltungsratsbeschlusses vom 9. März 1967, veröffentlicht mit RS-GV Nr. 23/1967.

# 广州大学研究生学籍异动申请表

| 姓名 | 楠文选 | 性别 | 男 | 出生日期 | 1993.12.08 | 学号 | 2117605015 |
|---|---|---|---|---|---|---|---|
| 学科、专业名称 | 化学工程 | | | | | 入学年月 | 2017.09 |
| 培养类别 | ☑全日制 ○非全日制 ○非定向 ○定向 ○委培 | | | | | | |
| 个人联系电话 | 18624278787 | 现所在学院 | 化学化工学院 | 现导师姓名及职称 | 梁红 教授 | | |

| 申请异动情况 | | |
|---|---|---|
| | 1.转专业或导师 | 拟由 专业 导师转至 专业 导师 |
| | 2.延期毕业 | 拟延期 年，至 年 月 日毕业 |
| | 3.提前毕业 | 拟提前 年，至 年 月 日毕业 |
| | 4.休学 | 拟休学 年，至 年 月 日复学 |
| | 5.复学 | 拟于 年 月 日复学 |
| | 6.退学 | 拟于 年 月 日退学 |
| | 7.其它 | 保留学籍（2019.11.22～2020.2.28） |

申请学籍异动原因（可加附页说明）：

本人于2019年11月22日前往德国马克斯普朗克研究所访学交流三个月，于2020年2月28日返校

申请人签字：楠文选 2019年10月31日

| 原指导教师意见： | 现指导教师意见： |
|---|---|
| 同意<br>签字：梁红 2019年10月31日 | 签字： 年 月 日 |
| 原专业负责人意见： | 现专业负责人意见： |
| 签字： 年 月 日 | 签字： 年 月 日 |
| 学院（所）审核意见：<br>同意<br>主管领导签字：<br>（公章）2019年10月31日 | 研究生院审批意见：<br>主管院长签字：<br>（公章） 年 月 日 |
| 主管校长签批： | |
| 签字：<br>（公章） 年 月 日 | |

注：1、因病者需县级以上或学校指定医院医生证明；2、定向、委培研究生须征求工作单位意见；3、保留入学资格、休学、保留学籍期满，学生应向学院和研究生院提交复学申请，以便学校安排复学相关工作；未按期办理复学申请者，学校不再保留其学籍，作取消学籍处理；4、本表一式两份，学院、研究生院各一份。

# 科研成果证明：

化 学 学 报
ACTA CHIMICA SINICA

研究论文
Article

DOI: 10.6023/A18070293

## 高通量筛选金属-有机框架: 分离天然气中的硫化氢和二氧化碳

杨文远　梁红　乔智威*

(广州大学化学化工学院　能源与催化研究所　广州 510006)

**摘要**　采用分子模拟高通量筛选的方法研究了 6013 种实验已经合成的金属-有机框架(MOFs)对天然气五元混合物($CH_4$, $C_2H_6$, $C_3H_8$, $H_2S$ 和 $CO_2$)中 $H_2S$ 和 $CO_2$ 的吸附分离性能. 为了综合吸附量和选择性这两项指标, 我们首先比较了三种权衡方法[权衡 $\alpha$ 法(Tradeoff between $S_{H2S-CO2/C1-C3}$ and $N_{H2S+CO2}$, TSN), 标准值法(Standard normal method, SNM)和权衡 $\beta$ 法(Tradeoff between selectivity and capacity, TSC)]. 接着, 针对四种 MOF 描述符[最大孔径(LCD), 孔隙率($\phi$), 比表面积(VSA)和吸附热($Q_{st}^0$)], 通过 Pearson 相关系数分析了每种描述符分别对三种权衡变量的相关性. 结果显示 TSC 法与四种 MOF 描述符的相关性最佳. 然后, 使用多元线性回归方法定量地分析了四种 MOF 描述符分别对 TSC 的影响程度; 而决策树模型则被用于规划性能高效 MOFs 的设计路径. 最后, 20 种性能最优 MOFs 从数据库中脱颖而出, 它们将为净化天然气技术的发展提供坚实的理论指导.

**关键词**　分子模拟; 金属-有机框架; 吸附; $H_2S$; $CO_2$

## High-Throughput Screening of Metal-Organic Frameworks for the Separation of Hydrogen Sulfide and Carbon Dioxide from Natural Gas

Yang, Wenyuan　Liang, Hong　Qiao, Zhiwei*

*(Guangzhou Key Laboratory for New Energy and Green Catalysis, School of Chemistry and Chemical Engineering, Guangzhou University, Guangzhou 510006, China)*

**Abstract**　In this work, the adsorption performance of 6013 computation-ready, experimental metal-organic frameworks (CoRE-MOFs) for the capture of $H_2S$ and $CO_2$ from natural gas mixture ($CH_4$, $C_2H_6$, $C_3H_8$, $H_2S$ and $CO_2$) is calculated by high-throughput screening of grand canonical Monte Carlo (GCMC) simulation in 298 K and 10 bar. For the comprehensive consideration of both adsorption capacities and selectivities of $H_2S+CO_2$, first, we compare three different tradeoff methods (α tradeoff method (Tradeoff between $S_{H2S-CO2/C1-C3}$ and $N_{H2S+CO2}$, TSN), standard normal method (SNM), $\beta$ tradeoff method (Tradeoff between selectivity and capacity, TSC)). The effect of selectivity on the new tradeoff variables are appropriately reduced by these tradeoff methods, because some of selectivities are very high. Thus, the new tradeoff variables can comprehensively evaluate the adsorption performance of CoRE-MOFs. Moreover, the correlation of each MOF descriptor (including the largest cavity diameter (LCD), void fraction ($\phi$), surface area (VSA) and isosteric heat ($Q_{st}^0$)) with three tradeoff variables are analyzed by Pearson correlation coefficient, respectively. The LCDs are calculated by Zeo++ software, but the $\phi$ and VSA are simulated by RASPA using probes of He and $N_2$, respectively. The $Q_{st}^0$ of each adsorbate gas are calculated at infinite dilution condition using NVT-MC method. All GCMC simulations for the screening are carried out using RASPA software. The results show that TSC has the best correlation with four MOF descriptors and the linear model could sufficiently describe the relationship between TSC and four MOF descriptors. Pearson correlation coefficients of four descriptors were −0.613, −0.717, −0.673 and 0.536 on TSC, respectively. Multiple linear regression is applied to quantitatively determine the influencing degree of four descriptors on performance, respectively. Among the four descriptors, $Q_{st}^0$, $\phi$, and LCD have larger standardized regression coefficients compared with VSA. This indicates that $Q_{st}^0$, $\phi$, and LCD are more useful in describing the performances of the MOFs. Thus, these three descriptors are used in the decision tree modeling to define an effective path for screening high-performance MOFs. It is concluded that a maximum probability (77.6%) of finding the good MOFs can be obtained from the three descriptors. Finally, the 20 best MOFs stand out from the whole database, and find that the alkali or alkaline earth metals in MOFs could effectively enhance the separation performance of $H_2S$ and $CO_2$. The microscopic insights and guidelines by this computational study can provide significant theoretical guidance for the development of adsorbent for the purification of natural gas.

**Keywords**　molecular simulation; metal-organic frameworks; adsorption; $H_2S$; $CO_2$

* E-mail: zqiao@gzhu.edu.cn

Received July 22, 2018; published September 13, 2018.

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

Project was supported by the National Natural Science Foundation of China (Nos. 21676094 and 21576058).

项目受国家自然科学基金(Nos. 21676094 和 21576058)资助.

*nanomaterials*

MDPI

*Article*

# Computational Screening of Metal–Organic Framework Membranes for the Separation of 15 Gas Mixtures

**Wenyuan Yang [1], Hong Liang [1], Feng Peng [1,2], Zili Liu [1], Jie Liu [3] and Zhiwei Qiao [1,2,*]**

[1] Guangzhou Key Laboratory for New Energy and Green Catalysis, School of Chemistry and Chemical Engineering, Guangzhou University, Guangzhou 510006, China; 2111710559@e.gzhu.edu.cn (W.Y.); lhong@gzhu.edu.cn (H.L.); fpeng@gzhu.edu.cn (F.P.); gzdxlzl@gmail.com (Z.L.)

[2] School of Chemistry and Chemical Engineering, South China University of Technology, Guangzhou 510640, China

[3] School of Chemistry and Chemical Engineering, Wuhan University of Technology, Wuhan 430072, China; ljie@wit.edu.cn

* Correspondence: zqiao@gzhu.edu.cn

Received: 26 January 2019; Accepted: 17 March 2019; Published: 20 March 2019

**Abstract:** The Monte Carlo and molecular dynamics simulations are employed to screen the separation performance of 6013 computation-ready, experimental metal–organic framework membranes (CoRE-MOFMs) for 15 binary gas mixtures. After the univariate analysis, principal component analysis is used to reduce 44 performance metrics of 15 mixtures to a 10-dimension set. Then, four machine learning algorithms (decision tree, random forest, support vector machine, and back propagation neural network) are combined with $k$ times repeated $k$-fold cross-validation to predict and analyze the relationships between six structural feature descriptors and 10 principal components. Based on the linear correlation value $R$ and the root mean square error predicted by the machine learning algorithm, the random forest algorithm is the most suitable for the prediction of the separation performance of CoRE-MOFMs. One descriptor, pore limiting diameter, possesses the highest weight importance for each principal component index. Finally, the 30 best CoRE-MOFMs for each binary gas mixture are screened out. The high-throughput computational screening and the microanalysis of high-dimensional performance metrics can provide guidance for experimental research through the relationships between the multi-structure variables and multi-performance variables.

**Keywords:** metal–organic framework; gas separation; machine learning; molecular simulation; linear dimension reduction

## 1. Introduction

With the rapid development of the social economy, people increasingly depend on energy; however, energy is not inexhaustible. In recent years, the acceleration of the energy crisis has prompted people to think about how to use cleaner, more environmentally friendly, more efficient energy. Separation technology plays an indispensable part in the chemical industry, and is widely used in medicine, food, petroleum, chemical engineering, metallurgy, and other fields. However, separation also consumes energy, especially high-throughput gas separation; for example, deep cryogenic separation was used to separate $N_2/O_2$ in industry [1], where the energy consumption is much higher and the recovery rate is lower. In addition, chemical and physical absorption methods are used to remove acidic components ($H_2S$ and $CO_2$) from $CH_4$ in industry, including the low-temperature methanol, alcohol amine, and alkali methods. Energy consumption is extremely high during the

# Combining large-scale screening and machine learning to predict the metal-organic frameworks for organosulfurs removal from high-sour natural gas

Cite as: APL Mater. 7, 091101 (2019); doi: 10.1063/1.5100765
Submitted: 20 April 2019 • Accepted: 14 August 2019 •
Published Online: 3 September 2019

Hong Liang,[1] Wenyuan Yang,[1] Feng Peng,[1,2] Zili Liu,[1] Jie Liu,[3] and Zhiwei Qiao[1,2,a]

## AFFILIATIONS

[1]Guangzhou Key Laboratory for New Energy and Green Catalysis, School of Chemistry and Chemical Engineering, Guangzhou University, Guangzhou 510006, People's Republic of China
[2]School of Chemistry and Chemical Engineering, South China University of Technology, Guangzhou 510640, People's Republic of China
[3]Key Laboratory for Green Chemical Process of Ministry of Education, School of Chemical Engineering and Pharmacy, Wuhan Institute of Technology, Wuhan 430073, People's Republic of China

**Note:** This paper is part of the Special Topic on Open Framework Materials for Energy Applications.
[a]**Author to whom correspondence should be addressed:** zqiao@gzhu.edu.cn

## ABSTRACT

High-sour natural gas usually contains organosulfurs besides $H_2S$, the majority of which exist in the form of mercaptans. These impurities of organosulfurs are required to be removed efficiently and economically for commercial application and the environment. In this work, the adsorption performance of organic sulfur gases [methanethiol (MeSH) and ethanethiol (EtSH)] in 137 953 hypothetical metal-organic frameworks (hMOFs) and 4764 computation-ready experimental MOFs (CoRE-MOFs) were evaluated by a high throughput computational screening technique. The highest adsorption capacities are predicted to be approximately 700 and 980 mg/g for MeSH and EtSH, respectively, which is substantially higher than that in zeolites (~100 mg/g). Quantitative structure-performance relationships are established between adsorption capacities and MOF textural/energetic properties (including the largest cavity diameter, surface area, void fraction, and isosteric heat). Two machine learning techniques, the back propagation neural network (BPNN) and the partial least-square (PLS) methods, are applied to predict 4764 CoRE-MOFs after training all the data of hMOFs from the large-scale screening. Compared with PLS, BPNN shows better prediction accuracy for MeSH and EtSH, and finds that the isosteric heat among seven MOF features possesses the highest weight for the adsorption of organosulfurs. Finally, the best 8 MOFs are identified for the removal of gaseous organosulfurs from the high-sour natural gas in a variety of industrial situations.



## INTRODUCTION

In recent years, industrial development and population growth have led to a surge demand for energy in the global. The exploitation and utilization of a clean energy resource, such as natural gas, becomes particularly critical and urgent. However, a significant portion of the global natural gases still remains untapped due to the high sour concentrations in reservoirs. $H_2S$ and organosulfurs in natural gas both have great negative effect on the recovery and utilization of gas. Although there are many researchers paying attention to the separation of $H_2S$, it is still worth noting that the removal of gaseous organic sulfurs is rarely considered in the literature, especially for the high-sour natural gases. Even in high-sour gas fields, the concentration of gaseous organic sulfur contents has exceeded 600 mg/m$^3$ or 900 ppm.[1–] In sour gas fields, one of the main components of gaseous organic sulfurs is mercaptan. During combustion, they will

REVIEW

# Machine-learning-assisted High-throughput Computational Screening of High Performance Metal–Organic Frameworks

Received 00th January 20xx,
Accepted 00th January 20xx

DOI: 10.1039/x0xx00000x

Zenan Shi, Wenyuan Yang, Xiaomei Deng, Chengzhi Cai, Yaling Yan, Hong Liang, Zili Liu, and Zhiwei Qiao*

Over the past two decades, the number of works on metal–organic frameworks (MOFs) in the fields of gas adsorption and separation has experienced explosive growth due to their high void fraction and ultra-high specific surface area. With the rapid increase of MOF database, high-throughput computational screening (HTCS) has become the main method for selecting high-performance target materials from the large quantity of MOFs. Traditional HTCS methods, e.g. grand canonical Monte Carlo (GCMC) and density functional Theory (DFT), could accelerate the discovery of material, however, there are some shortcomings in these methods such as high computational cost and slow speed, considering the vast and almost infinite MOF database as well as different separation systems and diverse operating conditions. Machine learning (ML) is a potential screening method with the ability to accurately predict the high-performance materials through the training of data, which were obtained by HTCS, and the ML model that fits accurately the complex system can improve the screening speed by 2–3 orders of magnitude. In this work, in view of ML-assisted HTCS of MOFs in recent years, the relevant research progress including $CH_4$ storage, $H_2$ storage and $CO_2$ separation and etc. is summarized, aiming to clarify the potential problems and challenges about ML-assisted HTCS by categorizing the application and development of ML in this field. Then, a series of ML algorithms were designed and developed to adapt to different MOF systems, and to search key descriptors based on ML to reverse design new MOFs with excellent performance. Therefore, the ML-assisted HTCS method could accelerate the development of MOFs and promote their applications in various fields.

## 1. Introduction

Metal–organic frameworks (MOFs) are a new type of organic-inorganic hybrid porous polymer material[1]. Since their structure was determined by Ludi *et al.*[2] in 1977, MOFs have been extensively studied. Tens of thousands of new MOFs have been synthesized or properly designed by changing the composition of different metal ions and organic ligands as well as functional groups, and their gravimetric surface area (GSA) kept increasing, reaching more than 6000 $m^2/g$. Because most of MOFs have a three-dimensional network structure, the structural properties of material can be tuned by changing the length and functional group of organic ligands, and if a MOF possesses the open metal sites, it could have a strong interaction between the adsorbates and the open metals. Because of their superior characteristics, the MOFs could be the potential candidates in broader range of applications than traditional materials. So far, MOFs have been studied and applied in many fields, such as adsorption[3], storage[4], separation[5], catalysis[6], drug delivery[7] and sensing[8]. Traditionally, to develop new materials, which the trial and error method was the synthesis of new MOFs by changing different metal ions and organic ligands as well as functional groups. After the performance testing and characterization, this material may be not necessarily ideal, thus, it would spend very high cost, energy and time. As the synthetic technology is strengthened and experience is accumulated, the types of MOFs synthesized in the laboratory are constantly increasing, and the application level is continuously verified. The structural properties of synthesized materials in the experiments were collected and classified into a database, i.e. the Cambridge Crystallographic Data Centre (CCDC), to be continuously updated, maintained and improved. According to Moghadam *et al.*[9], the number of MOFs has increased to 69,666. Chung *et al.*[10] established computation-ready experimental metal–organic framework (CoRE-MOFs) database by removing solvent molecules and refining structural parameters. Furthermore, by combining computer science, statistics, probability, database theory and computational mathematics, some performance of MOFs in the CCDC were theoretically calculated and analyzed. For example, Monte Carlo (MC) and molecular dynamics (MD) were integrated to explore the adsorption, storage and separation of MOFs on gases with different components, and the experimental verification was thereby carried out. Through the analysis of CoRE-MOFs, different structural units and topologies could be automatically connected to form the MOF structure by the software packages, such as Tobacco[11], in different ways to automatically design new hypothetical MOFs (hMOFs), greatly expanding the database. Snurr *et al.*[12] used

*Guangzhou Key Laboratory for New Energy and Green Catalysis, School of Chemistry and Chemical Engineering, Guangzhou University, Guangzhou, 510006, P. R. China. E-mail: zqiao@gzhu.edu.cn*