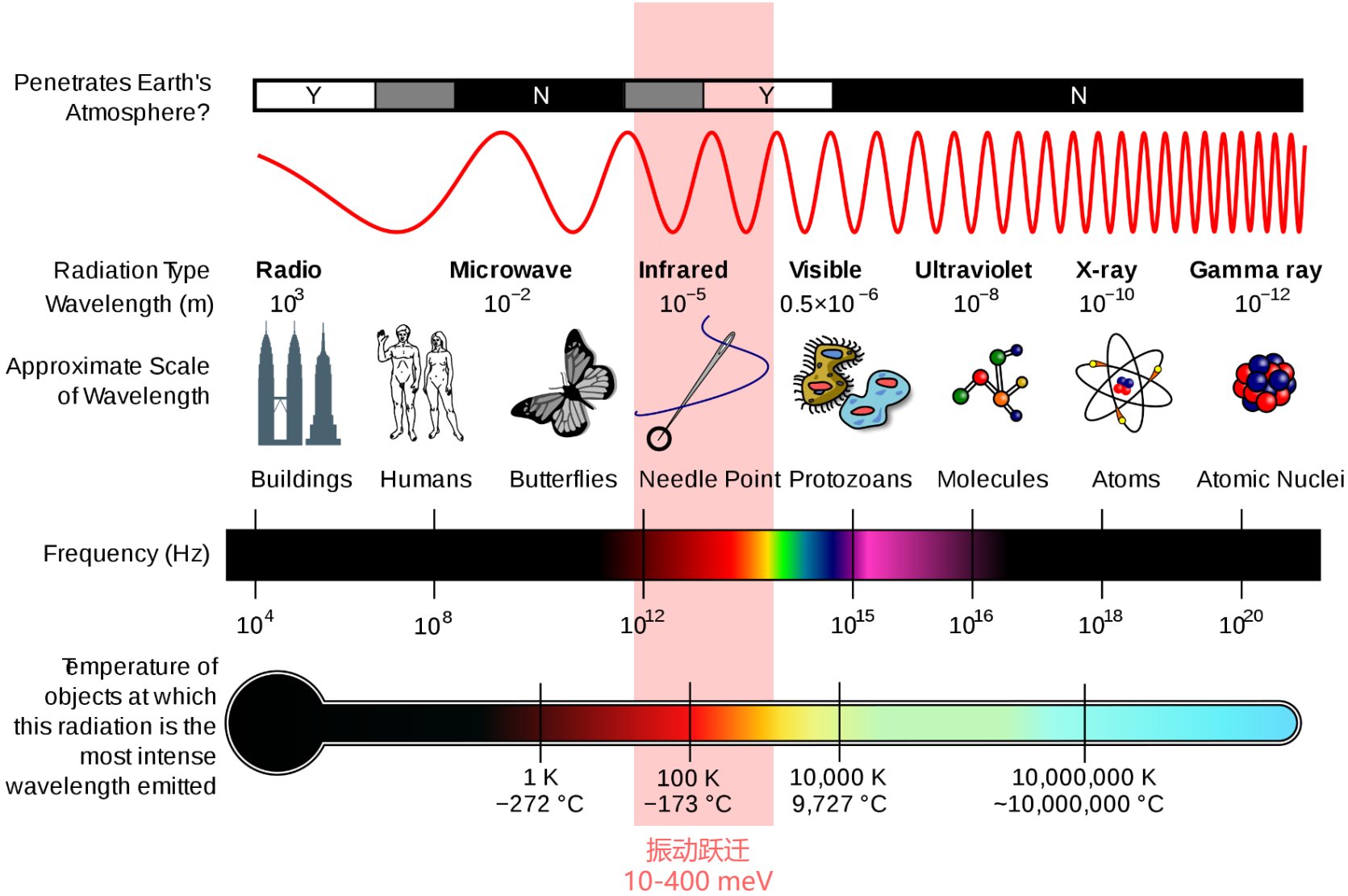
基于机器学习的
振动谱快速预测和结构智能识别

任浩
中国石油大学（华东）
renh@upc.edu.cn

主 要 内 容

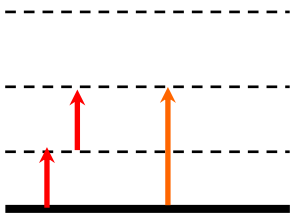
- 简介：光谱与表征、振动光谱
- 振动光谱快速预测
- 基于振动光谱的结构识别
- 总结和讨论

简介：分子振动和振动光谱

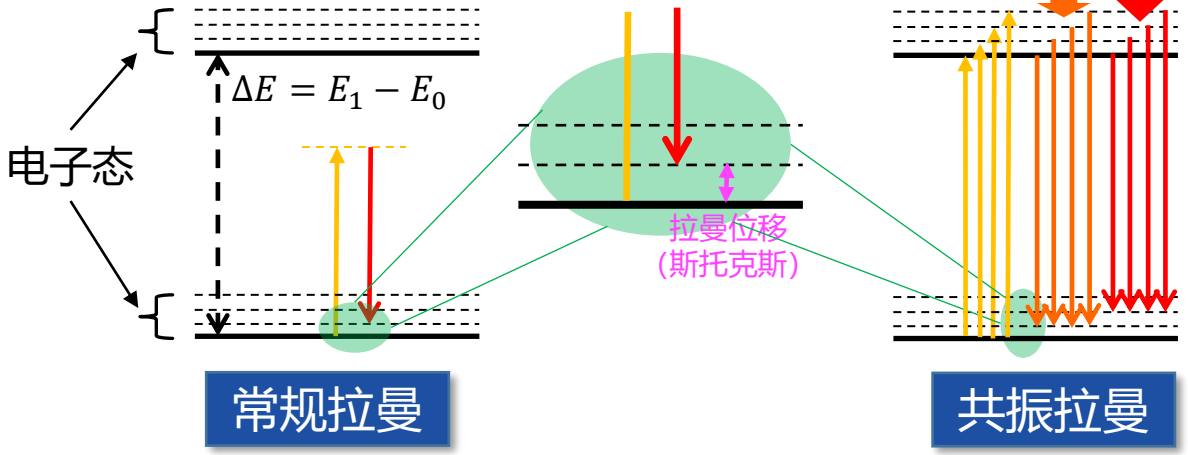


简介：振动的表征方法

传统振动光谱



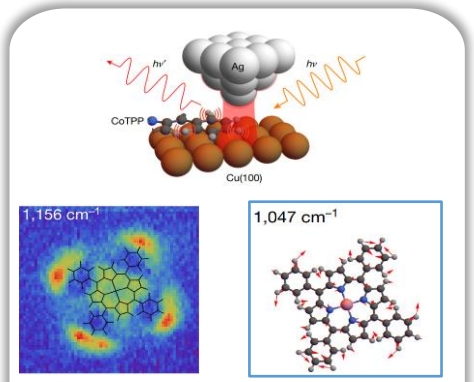
红外吸收



常规拉曼　　共振拉曼

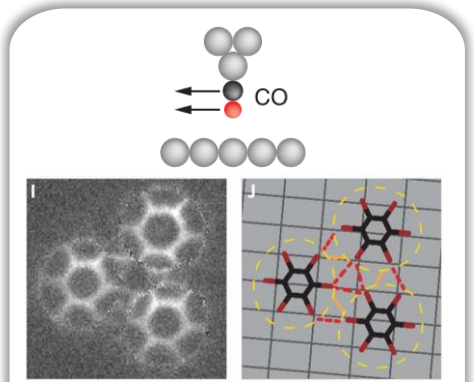
其他新型非线性光谱: 二维红外、合频产生、相干拉曼……

其它非光学和光-实物粒子结合的振动谱



Nature 2019, 568, 78

隧穿诱导拉曼/发射



Science 2017, 358, 206

非弹性电子隧穿

- 低能电子衍射
- 中子非弹性散射
- ……

谱学信号解释和结构识别的现状

正向预测：由结构计算（预测）光谱

- 原则上可行：第一性原理、半经验方法、响应理论
- 存在的困难：高阶非线性响应、凝聚态复杂体系
- 优化空间：算得更快、更准、更大

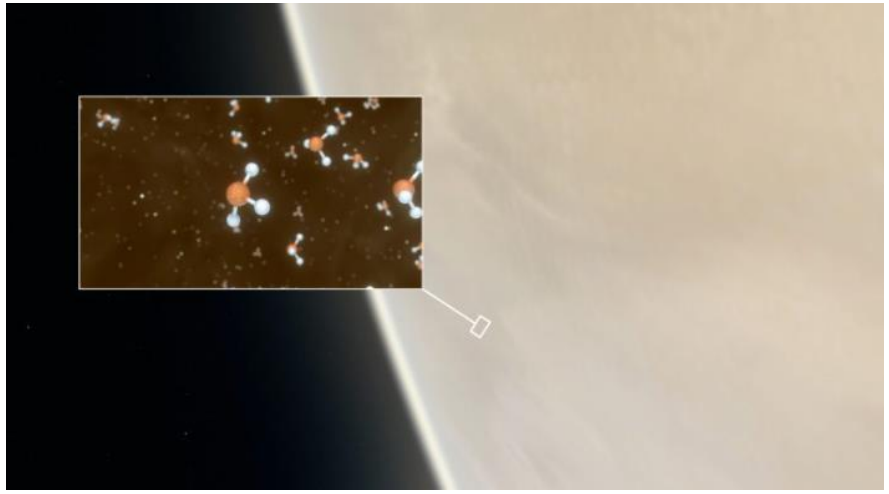
反演结构：由光谱识别结构（或化学环境）

- 人工搜索和比对
- 自动模式识别（已有NMR、XAS等相关工作出现）

挑战1：未知物质的结构多样性带来反演困难

□ 星际物质谱图

紫外线环境下显示出弓形激波的狮子座

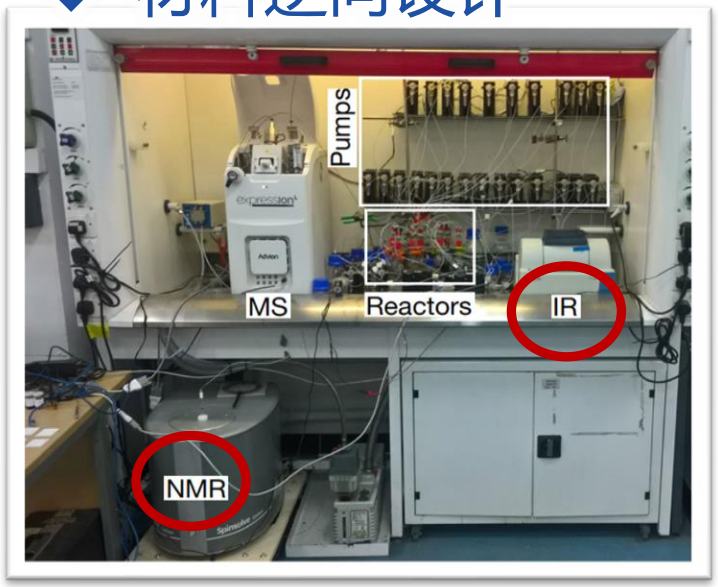


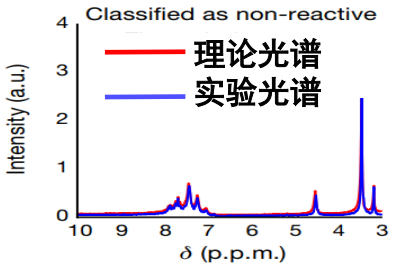
金星大气层存在PH_3？ *Nat Astron* (cautioned)

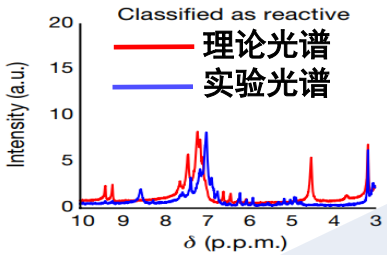
- 2014年5月，美国NASA建立宇宙光谱数据库
- 宇宙中>20%的碳和多环芳香烃相关，或是生命形成的起始物质
- 需要发展新的工具，从海量谱图中自动反演未知化学结构

挑战2：不清晰的内在关系制约光谱解读

- 分子合成机器人
- 材料逆向设计







英国Glasgow大学L. Cornin 院士

Nature 2018, 559, 377

Controlling an organic synthesis robot with machine learning to search for new reactivity

Jaroslaw M. Granda[1], Liva Donina[1], Vincenza Dragone[1], De-Liang Long[1] & Leroy Cronin[1]*

The discovery of chemical reactions is an inherently unpredictable and time-consuming process[1]. An attractive alternative is to predict reactivity, although relevant approaches, such as computer-aided reaction design, are still in their infancy[2]. Reaction prediction based on high-level quantum chemical methods is complex[3], even for simple molecules. Although machine learning is powerful for data analysis[4,5], its applications in chemistry are still being developed[6]. Inspired by strategies based on chemists' intuition[7], we propose that a reaction system controlled by a machine learning algorithm may be able to explore the space of chemical reactions quickly, especially if trained by an expert[8]. Here we present an organic synthesis robot that can perform chemical reactions and analysis faster than they can be performed manually, as well as predict the reactivity of possible reagent combinations after conducting a small number of experiments, thus effectively navigating chemical reaction space. By using machine learning for decision making, enabled by binary encoding of the chemical inputs, the reactions can be assessed in real time using nuclear magnetic resonance and infrared spectroscopy. The machine learning system was able to predict the reactivity of about 1,000 reaction combinations with accuracy greater than 80 per cent after considering the outcomes of slightly over 10 per cent of the dataset. This approach was also used to calculate the reactivity of published datasets. Further, by using real-time data from our robot, these predictions were followed up manually by a chemist, leading to the discovery of four reactions.

pool of reagents and with bits representing reagents that were present in a given reaction mixture to one, similarly to one-hot encoding. Figure 1d shows example vector representations for the model substrate pool consisting aniline, benzaldehyde, acetyl chloride, phenylhydrazine and furan.

This approach to representing chemical space renders it structure-independent and allows the robotic platform to operate without prior knowledge about reactivity and chemical structure (Fig. 2). Initially, the chemical space was sampled by performing reactions with random combinations of starting materials, evaluating their reactivity as reactive or non-reactive using the SVM model (to determine expected values of reactivity, Y) and encoding them in vector form (to obtain a training set, X). The process of random selection is important because the system avoids making prior assumptions about the possible reactivity of the reagents, ensuring that the initial run results are unknown. Even if the reaction mixture decomposes or is non-reactive, this information is still useful for the navigation of the chemical space, allowing real-time assessment of the reactivity of the starting materials. After the reaction database has been built, a linear discriminant analysis (LDA)[18] model is trained on the data obtained to construct a model of the chemical space. The remaining reactions are then rated by predicting the probability of reactivity using the LDA model. This allows for autonomous decision making, and the reaction with the highest score is performed and analysed by the robotic system, thus avoiding many non-reactive combinations and speeding up the search. The loop

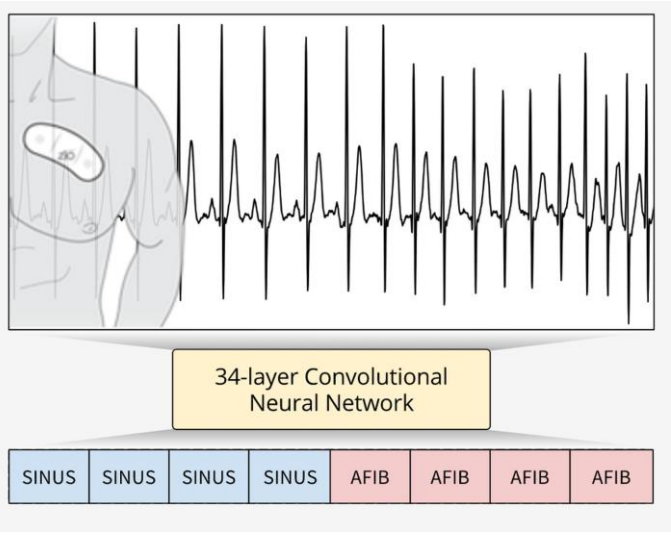
通过NMR和IR的理论和实验光谱的自动比对，为分子合成机器人分辨产物的成分和比例，提供重要的反应判据

基于已知目标的化学结构做光谱计算：
意料之外的中间体或产物？前所未有的新物质？

探索新路径：突破光谱的预测瓶颈和反演困难

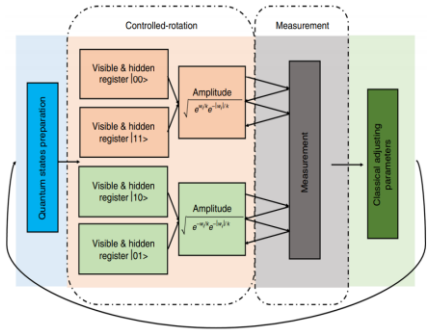
2018年8月
Facebook & NYU
ML & NMR interpretation

2020年3月
布鲁克公司
将机器学习算法集成于NMR固件



2019年1月
吴恩达等
自心电图识别14种心脏病，达到医学专家诊断水平

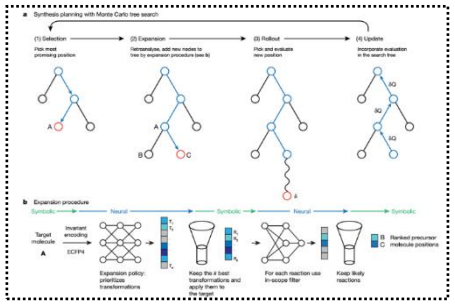
人工智能结合计算化学的科学机遇



采用人工智能加速分子结构优化和势能面构建

Nature Commun. 2018, 9, 4195

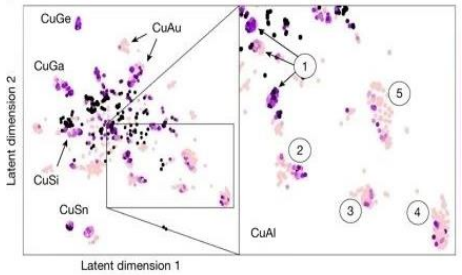
突破计算瓶颈



人工智能结合化学模板发现新的合成路线

Nature 2018, 555, 604

发现内在关系

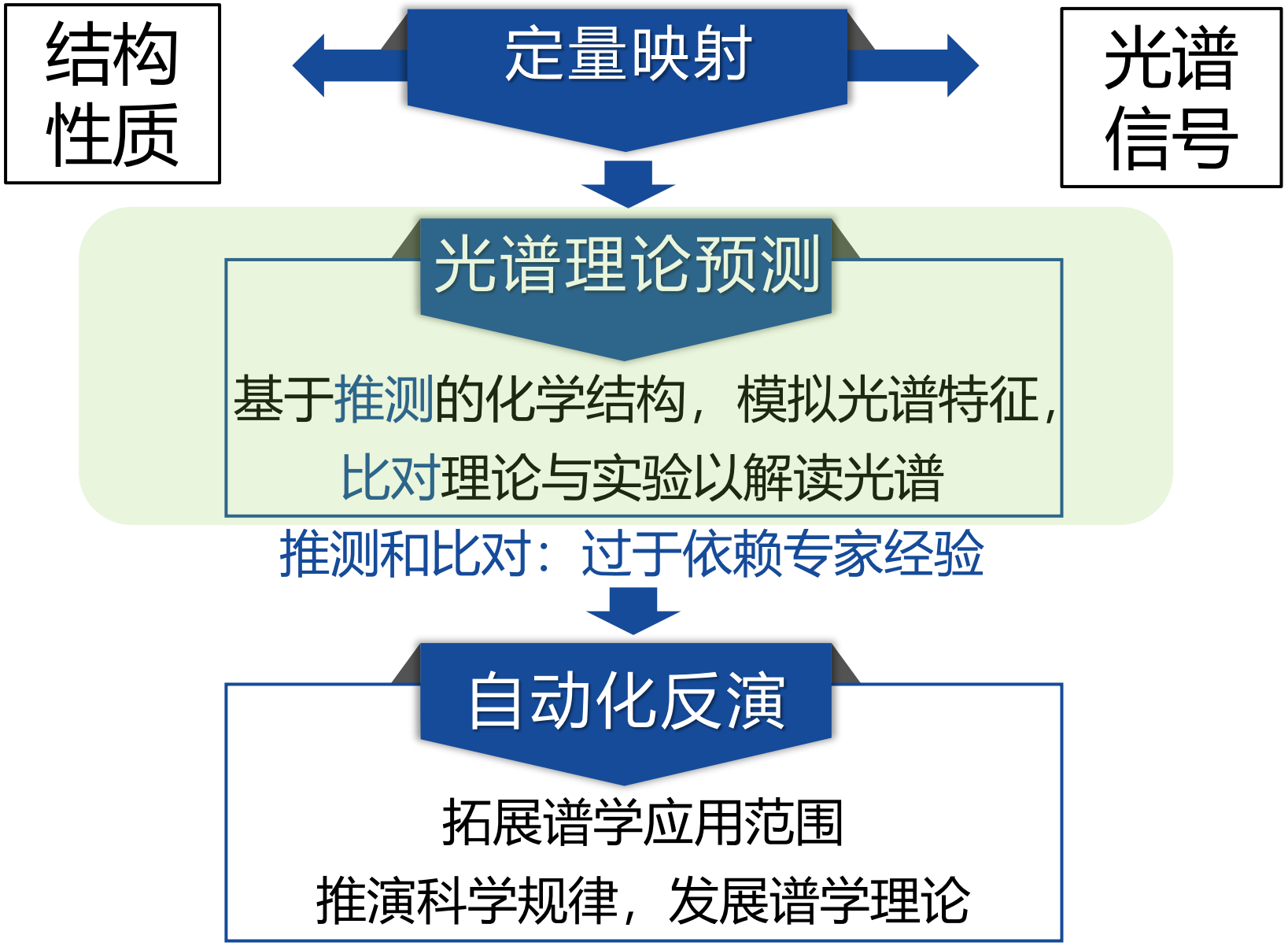


人工智能结合量子化学计算预测高效电催化剂

Nature 2020, 581, 178

预测新的物质

光谱 *vs.* 结构、性质


结构
性质
定量映射
光谱
信号
光谱理论预测
基于推测的化学结构，模拟光谱特征，
比对理论与实验以解读光谱
推测和比对：过于依赖专家经验
自动化反演
拓展谱学应用范围
推演科学规律，发展谱学理论

探索人工智能与光谱研究的结合

两个关键点

1、机器学习的基础，需要**大量精确的化学数据**

2、需要将化学语言转化成机器能懂的数字语言，也就是提供化学信息的**描述符（descriptor）**

基于机器学习的振动光谱快速预测

数据集：QM8全部（22k 分子）、QM9（3k）、QM10（3k）

振动光谱信号的局域性：

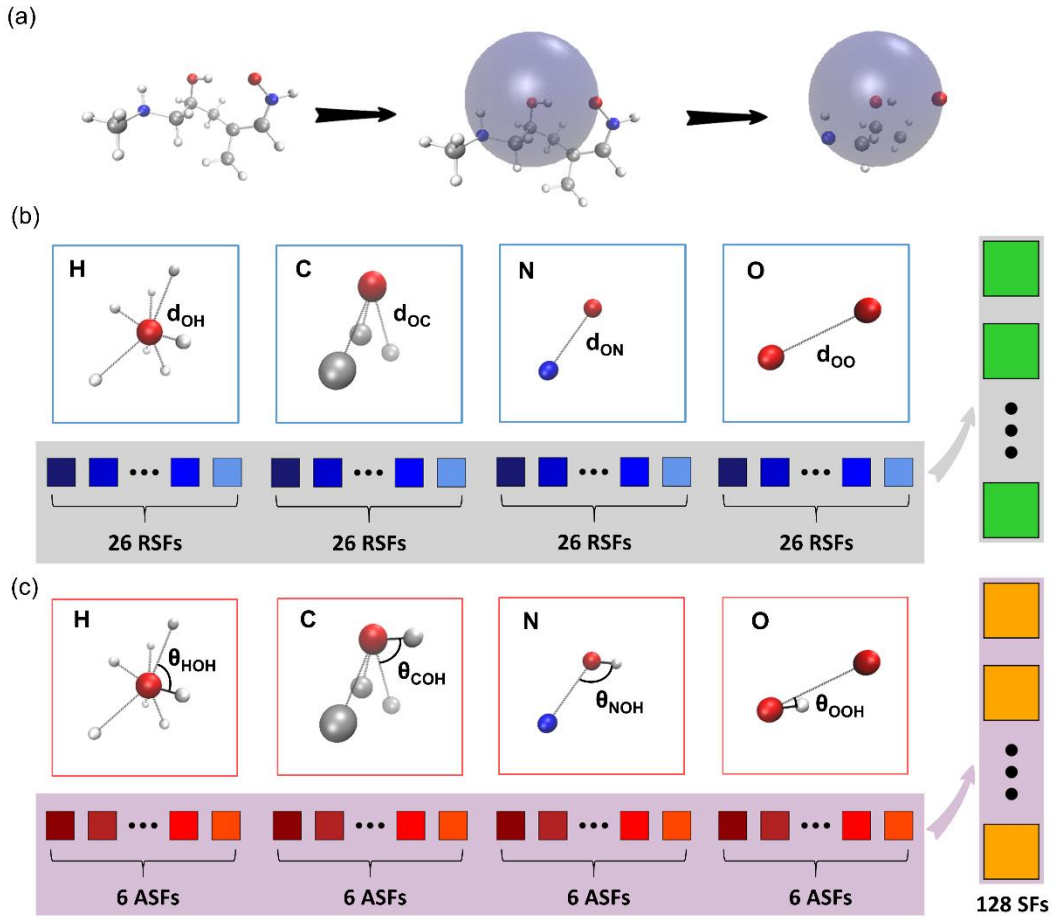
- 频率：能量二阶梯度
- 红外强度：偶极矩梯度
- 拉曼强度：极化率梯度

对称函数描述符：

$$f_c(R_{ij}) = \begin{cases} 0.5 \cdot \left[\cos\left(\frac{\pi R_{ij}}{R_c}\right) + 1\right] & \text{for } R_{ij} \le R_c \\ 0 & \text{for } R_{ij} > R_c. \end{cases}$$

$$G_i^{r,X} = \sum_{j \neq i} e^{-\eta(R_{ij} - R_s)^2} f_c(R_{ij})$$

$$G_i^{a,X} = 2^{1-\zeta}\left[(1 + \lambda \cos\theta_{ijk})^{\zeta} \cdot e^{-\eta(R_{ij}^2 + R_{ik}^2)} \cdot f_c(R_{ij}) \cdot f_c(R_{ik})\right]$$



Gastegger, et al. *J. Chem. Phys.* **2018**, *148*, 241709.
Behler & Parrinario, *Phys. Rev. Lett.* **2007**, *98*, 146401.

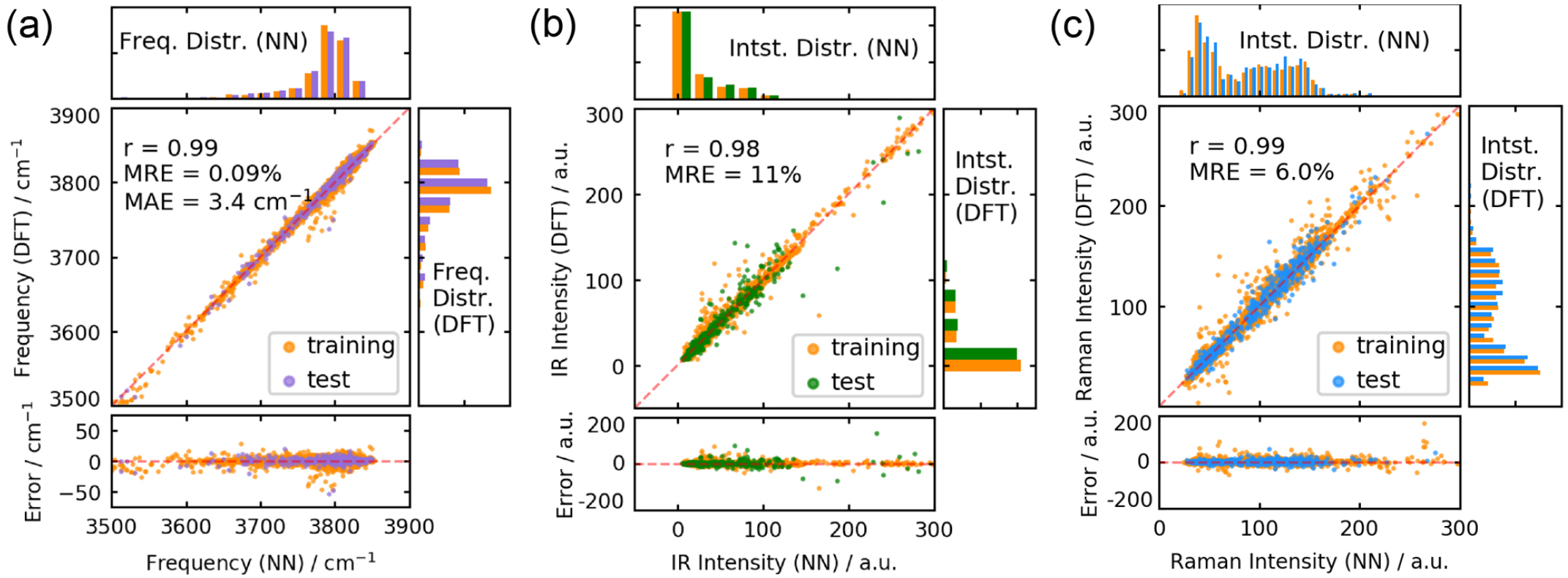
Ren, H.; et al. Luo, Y.;* and Jiang, J.* *Fundm. Res. 2021, 1,* 1880.

回归模型

振动频率、红外和拉曼强度使用各自的神经网络：

- 全链接前馈神经网络
- 三个隐藏层（256、128、64节点）
- 网格搜索优化对称函数截断
- 训练集/测试集 9:1划分，10重交叉验证
- 零均值归一化回归对象

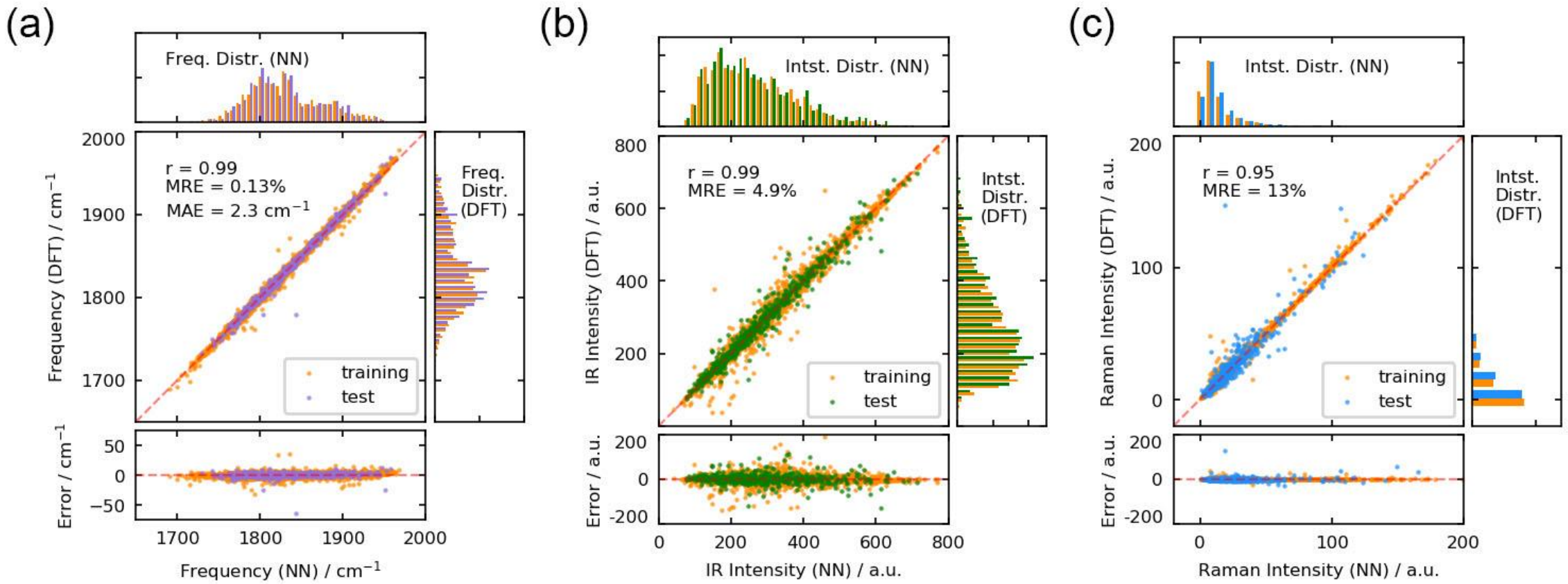
羟基伸缩振动信号预测结果



频率精度：3.4 cm^{-1}； 红外强度：11%；拉曼强度：6%

Ren, H.; et al. Luo, Y.;* and Jiang, J.* *Fundm. Res. 2021, 1, 488.*.

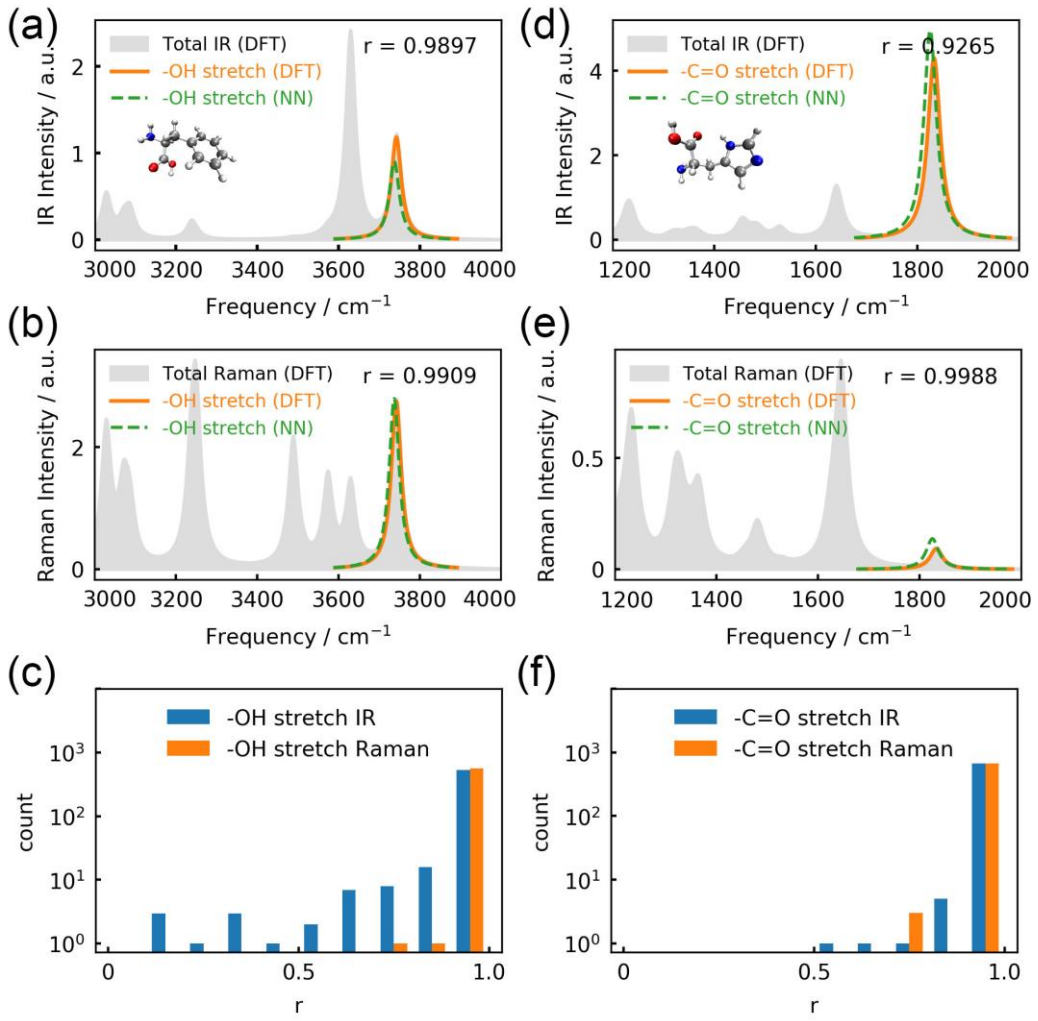
羰基伸缩振动信号预测结果



频率精度：2.3 cm^{-1}； 红外强度：4.9 %；拉曼强度：13 %

Ren, H.; et al. Luo, Y.;* and Jiang, J.* *Fundm. Res. 2021, 1, 488.*.

预测振动光谱信号的精度和效率



Ren, H.; et al. Luo, Y.;* and Jiang, J.* *Fundm. Res. 2021, 1, 488.*.

与DFT计算结果相比：

- 频率-强度数据展宽为频域序列
- 皮尔森相关系数表征相似度
- 超过98%的分子谱线相似度>0.9

计算速度：

- DFT
  - B3LYP/6-31G(2df,p), @Xeon E5v4
  - 5×10^5 CPU minutes
- 神经网络模型
  - 训练：80 min @1080Ti
  - 预测：~60 ms （7000 分子）

人工智能从大数据中推演内在规律

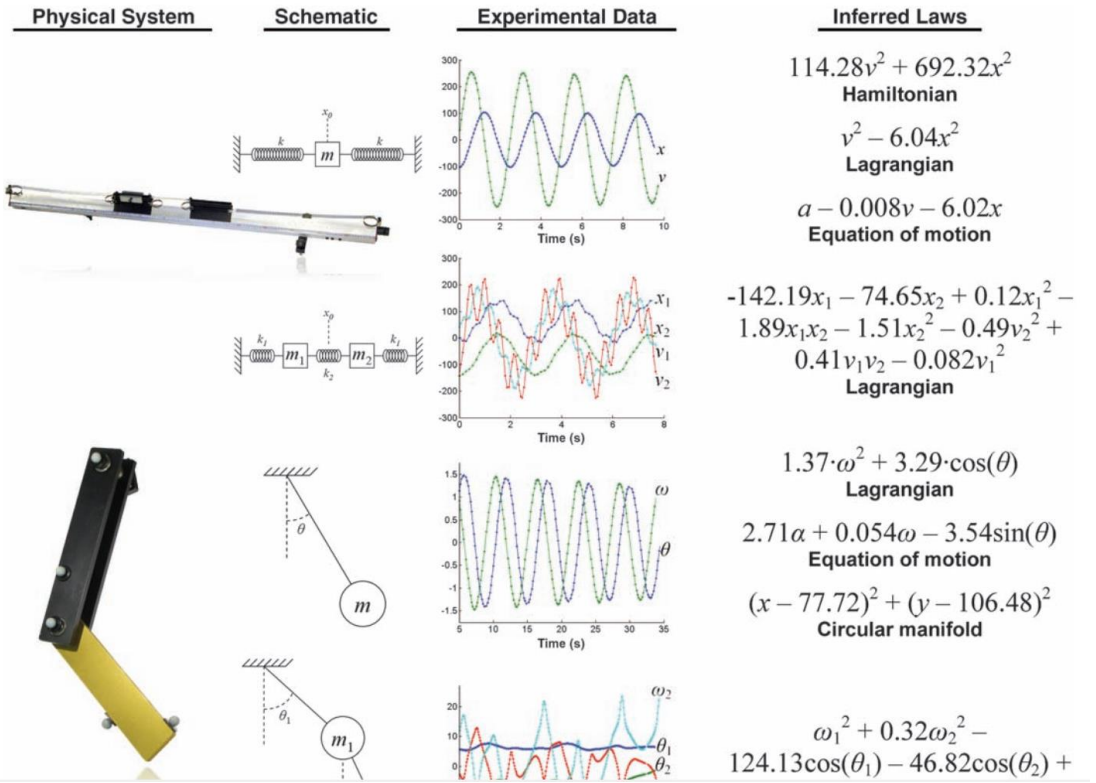
AAAS Become a Member

Science Contents News Careers Journals

REPORT

Distilling Free-Form Natural Laws from Experimental Data

Science 03 Apr 2009:
Vol. 324, Issue 5923, pp. 81-85
DOI: 10.1126/science.1165893



不依赖先验知识，由实验数据出发，推导出哈密顿力学、拉格朗日力学动力学规律

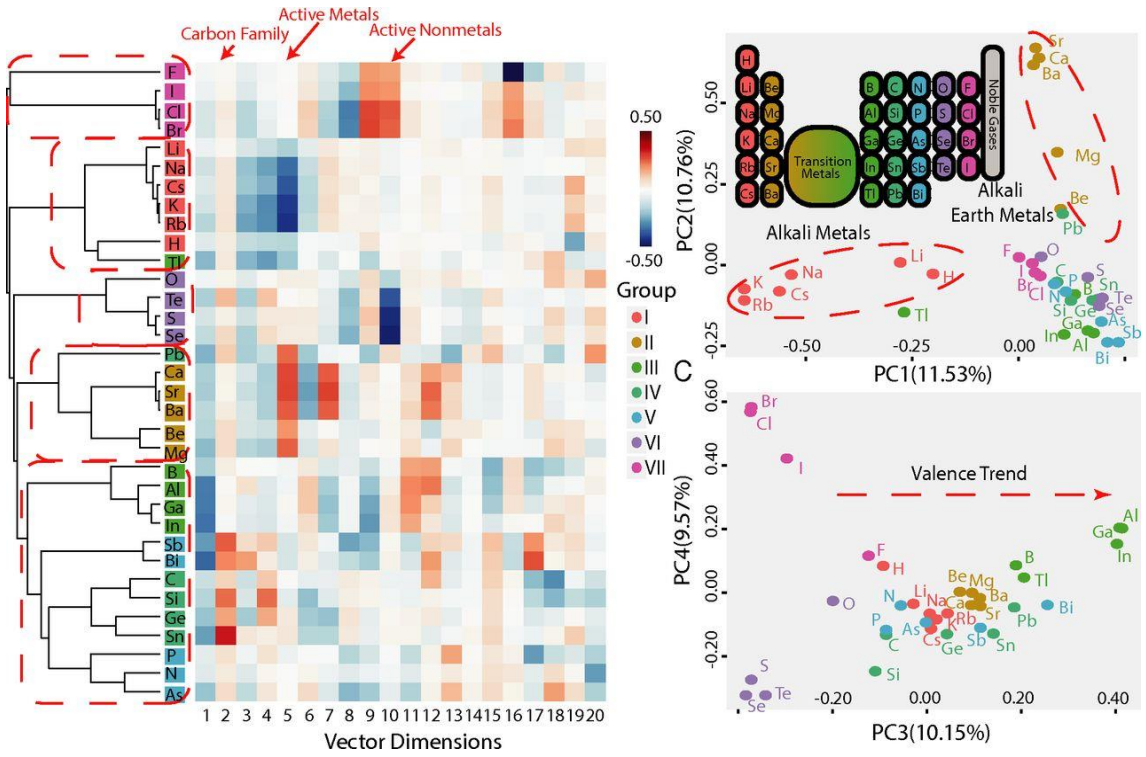
人工智能从大数据中推演内在规律

PNAS Proceedings of the National Academy of Sciences of the United States of America

Learning atoms for materials discovery

Quan Zhou, Peizhe Tang, Shenxiu Liu, Jinbo Pan, Qimin Yan, and Shou-Cheng Zhang

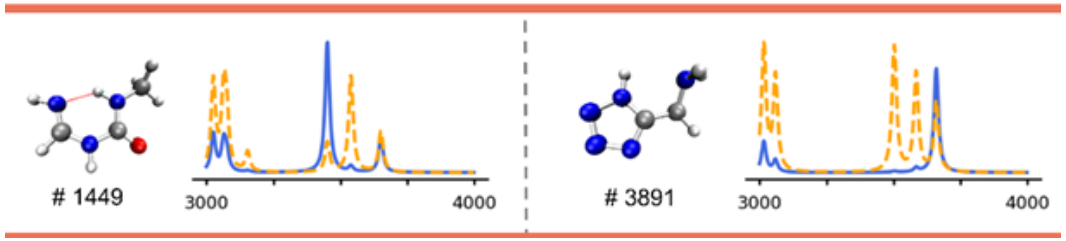
PNAS July 10, 2018 115 (28) E6411-E6417; first published June 26, 2018; https://doi.org/10.1073/pnas.1801181115

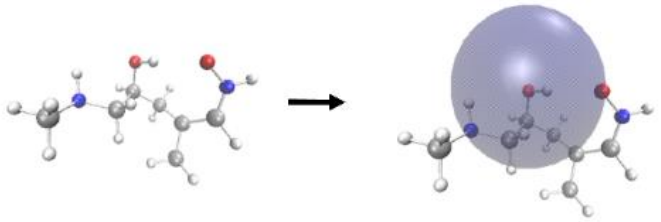


从大量材料数据中，“重新发现”元素周期表

描述符截断半径隐含物理量的定域性描述

基于对称函数描述符(symmetry functions)
预测红外、拉曼光谱





$$f_c(R_{ij}) = \begin{cases} 0.5 \cdot \left[\cos\left(\frac{\pi R_{ij}}{R_c}\right) + 1\right] & \text{for } R_{ij} \le R_c \\ 0 & \text{for } R_{ij} > R_c. \end{cases}$$

RSF cutoff Radii (Å)	Frequency	IR intensity	Raman intensity
–OH	4.2	5.4	6.2
–C=O	4.0	5.0	6.0

对称函数描述符的截断半径
反映了激发态跃迁偶极矩的尺寸

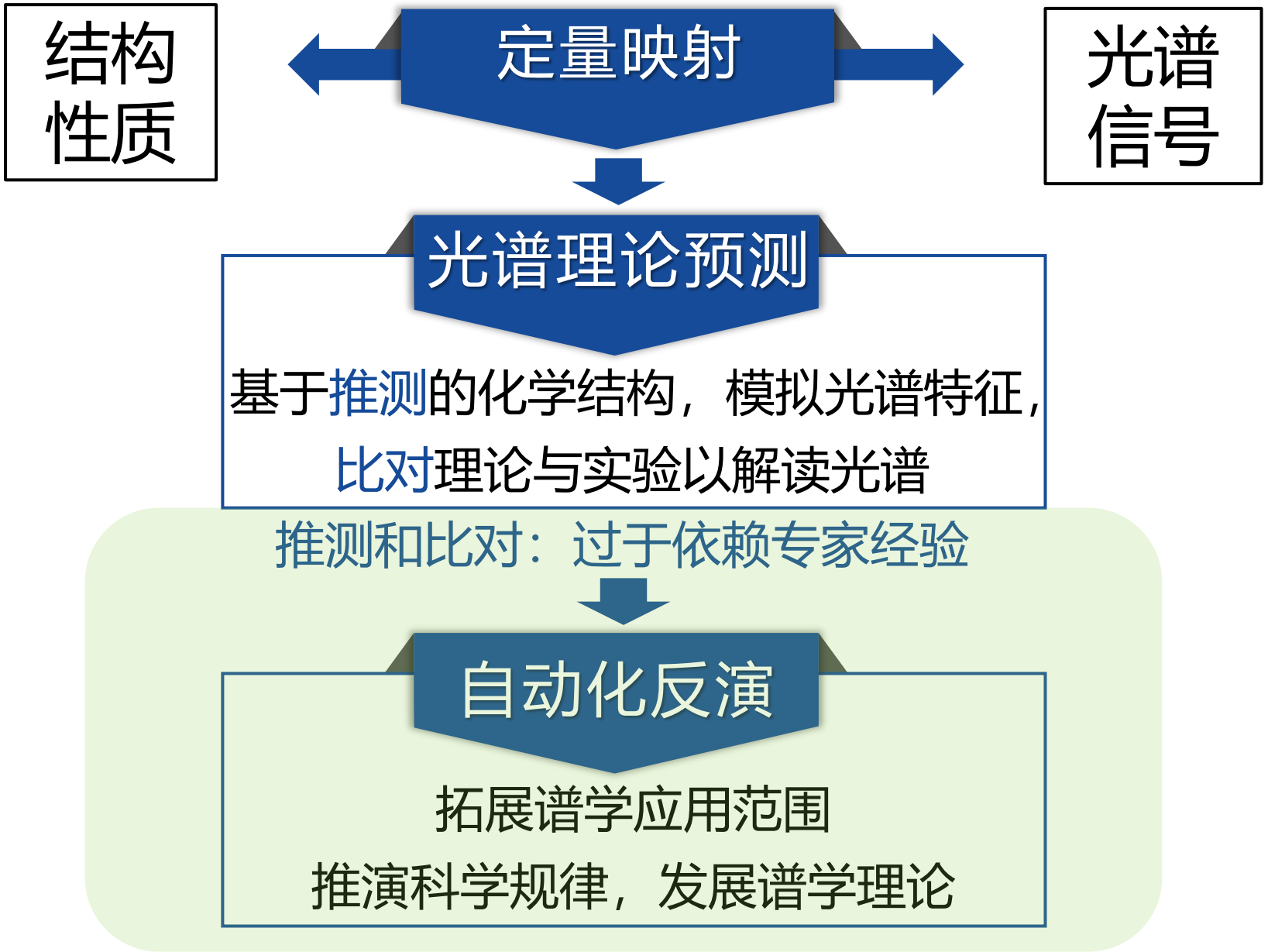
$$I^{IR} = \frac{8\pi^3 N_A}{3hc}\nu \sum_k p_g f(\nu_{gk},\nu) \left|\langle\psi_g|\boldsymbol{\mu}|\psi_k\rangle\right|$$

$$I^{Raman} = Ik^4 \left| \sum_{r\neq 0} \frac{\langle m|\boldsymbol{\mu}\cdot\mathbf{e}|r\rangle\langle r|\boldsymbol{\mu}\cdot\mathbf{e}'|0\rangle}{\hbar(\omega_{r0}-\omega_{in})} + \frac{\langle m|\boldsymbol{\mu}\cdot\mathbf{e}'|r\rangle\langle r|\boldsymbol{\mu}\cdot\mathbf{e}|0\rangle}{\hbar(\omega_{r0}+\omega_{sc})} \right|^2$$

- 最优截断半径：振动频率 < 红外强度 < 拉曼强度
- 能否用于确定描述特定物理性质的截断半径？如QM/MM?

Ren, H.; et al. Luo, Y.;* and Jiang, J.* *Fundm. Res. 2021, 1, 488..*

光谱 *vs.* 结构、性质


结构
性质
定量映射
光谱
信号
光谱理论预测
基于推测的化学结构，模拟光谱特征，
比对理论与实验以解读光谱
推测和比对：过于依赖专家经验
自动化反演
拓展谱学应用范围
推演科学规律，发展谱学理论

基于振动谱的结构自动反演

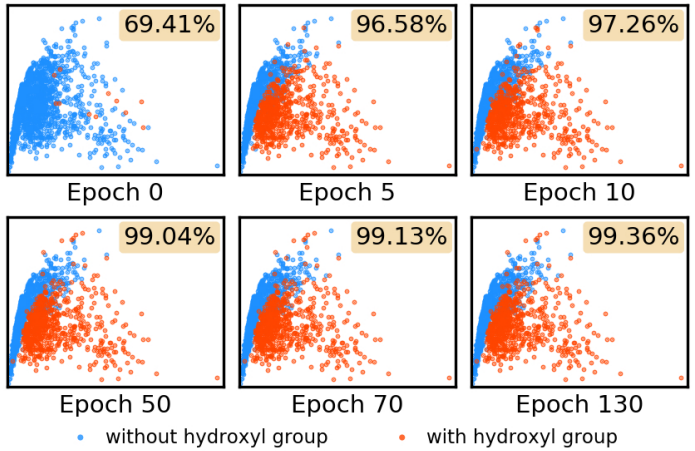
定义问题

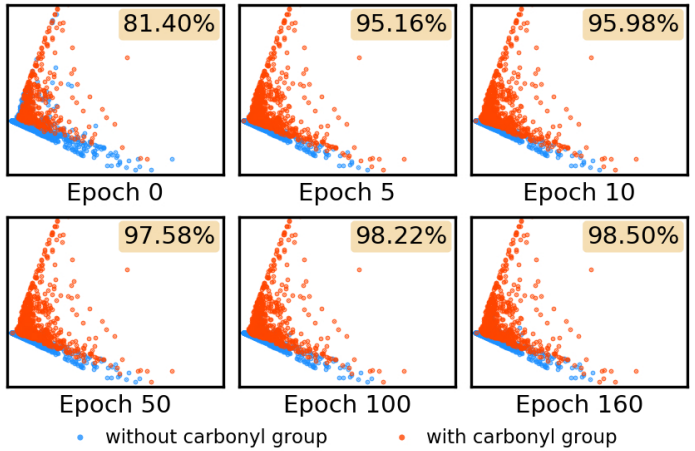
- 给定特定频段范围的振动谱
- 识别分子中是否存在特定官能团（羟基和羰基）
- 利用多种光谱选择定则差异，从多个维度重构化学信息

机器学习模型

- 长短时记忆神经网络（LSTM）
- 光谱数据作为频域序列输入
- 可依次读入不同光谱学数据
- 从不同谱学数据读取（学习）化学信息

官能团识别过程及模型性能





-OH Acc. 99.36%		Predicted	
		No	Yes
Actual	No	1510	2
	Yes	12	670

-C=O Acc. 98.50%		Predicted	
		No	Yes
Actual	No	1407	22
	Yes	11	754

识别准确率：羟基 99.36%； 羰基 98.50%

Ren, H.; et al. Luo, Y.;* and Jiang, J.* *Fundm. Res. 2021, 1, 488.*

迁移至更大分子

(a)

-OH Acc. 98.97%		Predicted	
		No	Yes
Actual	No	1973	9
	Yes	22	996

(b)

-OH Acc. 96.07%		Predicted	
		No	Yes
Actual	No	2017	25
	Yes	93	865

（a-b）羟基识别模型迁移至QM9和QM10分子集

(c)

-C=O Acc. 97.47%		Predicted	
		No	Yes
Actual	No	1876	36
	Yes	40	1048

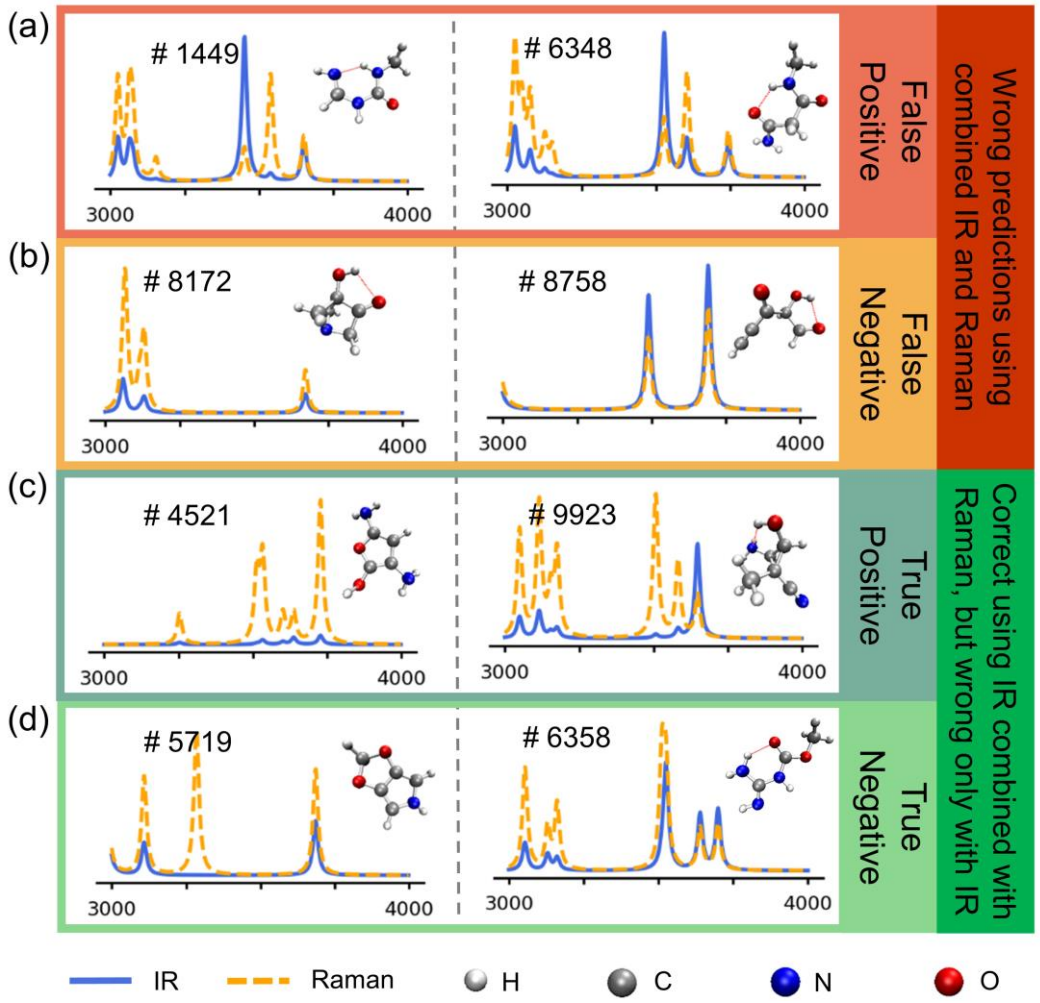
(d)

-C=O Acc. 93.27%		Predicted	
		No	Yes
Actual	No	1980	85
	Yes	117	818

（c-d）羰基识别模型迁移至QM9和QM10分子集

Ren, H.; et al. Luo, Y.;* and Jiang, J.* *Fundm. Res. 2021, 1, 488.*

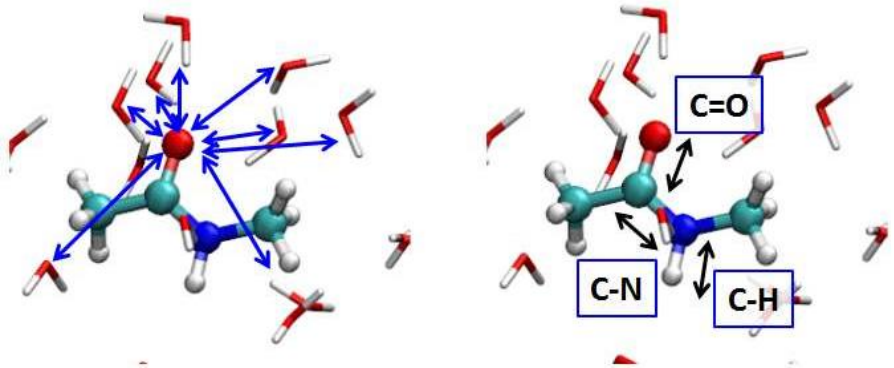
官能团识别模型错误分析

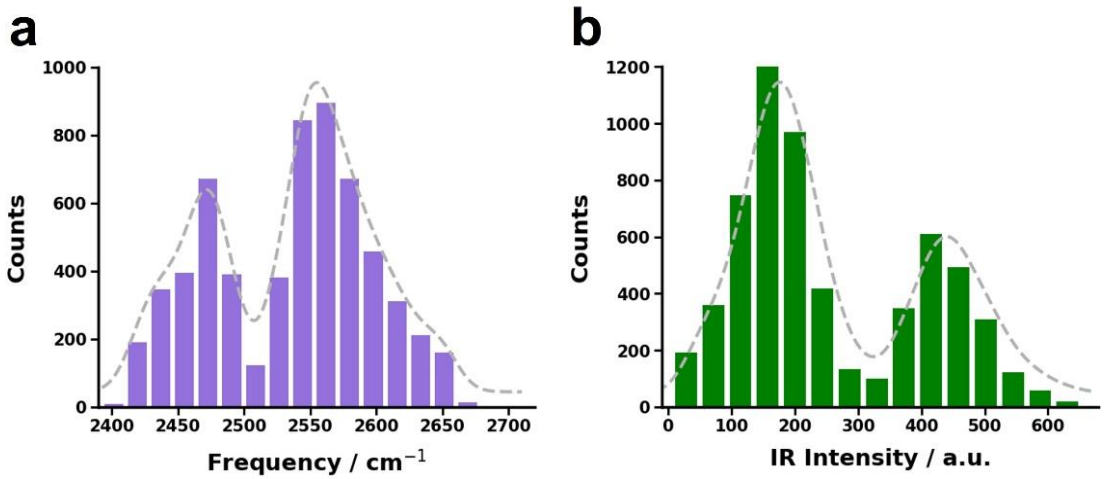


- 分子内氢键可能导致混淆N-H和O-H伸缩振动，但有例外
- 若分子红外和拉曼信号差别不够明显，则综合考虑两种光谱信号不会提高准确率
- 扩展至其他谱学或非谱学特征，应能提高识别准确率

Ren, H.; et al. Luo, Y.;* and Jiang, J.* *Fundm. Res. 2021, 1, 488.*

凝聚相中氢键的影响





- 氢键个数对振动谱有显著影响
- 可通过振动信号差别反演局域化学结构（氢键网络）
- 自然扩展至凝聚相体系

暗示了新的氢键探测手段

Zhang, Q; et al. Jiang, J.*; **Ren, H.*** *in manuscript.*

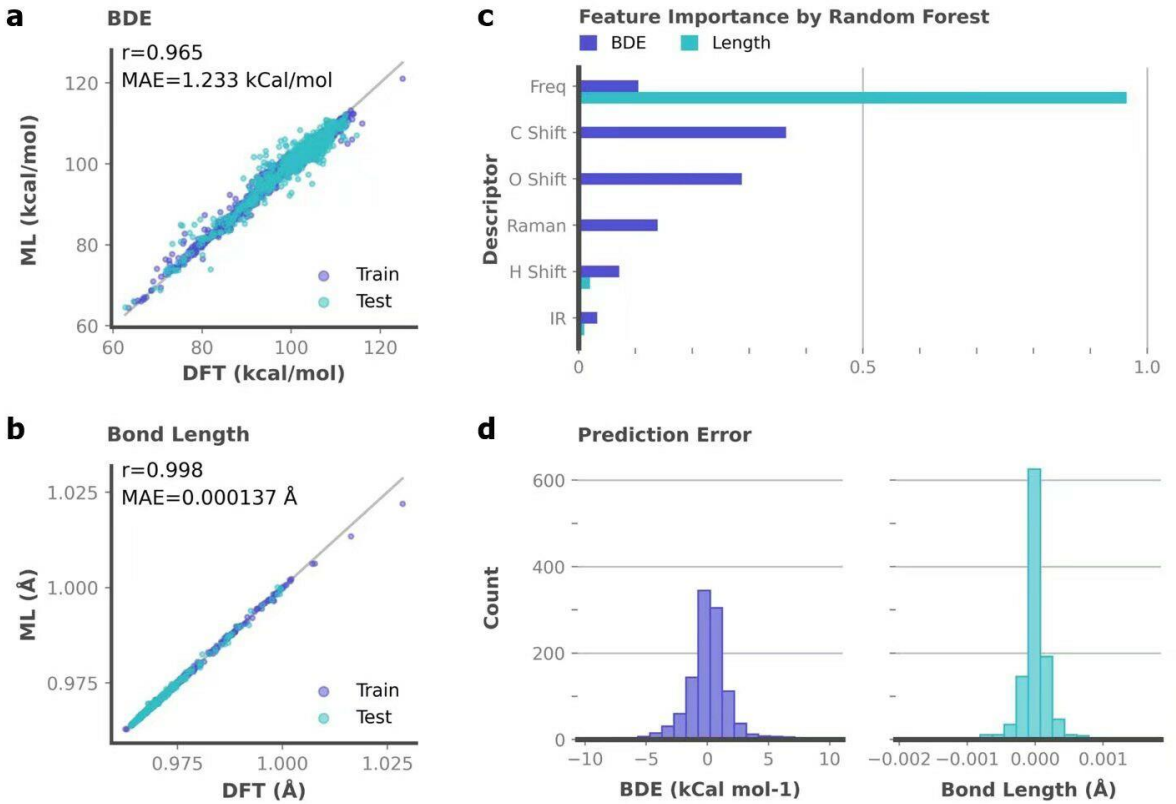
从多个维度重构化学信息

Table 2: Overall −OH and C=O recognition accuracies and error rates (in parentheses) using IR, Raman, and combined (IR+Raman) spectra.

	IR	Raman	IR+Raman
−OH	98.50% (1.50%)	98.58% (1.42%)	99.36% (0.64%)
C=O	98.04% (1.96%)	95.49% (4.51%)	98.50% (1.50%)

- 结合红外和拉曼，降低识别错误率（30~60%）
- 振动谱信息不足以完全判定，应提供其他信息（NMR？XAS？）
- LSTM对序列读入顺序敏感，试用其它模型？（Transformer，注意力机制）
- QM9/GDB17 构形空间远大于常见分子，改善现有模型进行（迁移学习和主动学习）

基于振动谱反演键能信息



- 键能：远离平衡态，O电负性相关
- 键长：平衡态性质
- 振动谱：平衡态附近
- 化学位移：电子分布密度及屏蔽

- 振动谱与键长相关性强
- 化学位移与键能相关性强

Guo, S.; et al. **Ren, H.**; and Jiang, J. *in preparation.*

线上部署

A Demo of LSTM Carbonyl Group Recognition

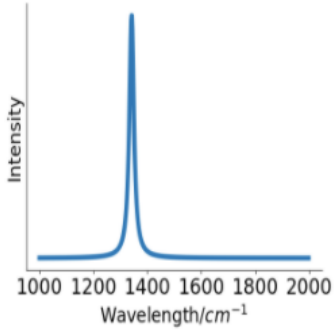
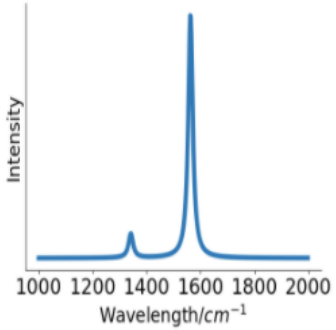
We can predict the presence of carbonyl groups in chemical structures

Please upload a IR/Raman spectral file, support ".npy".
Download example data

Select Data File

Select Upload

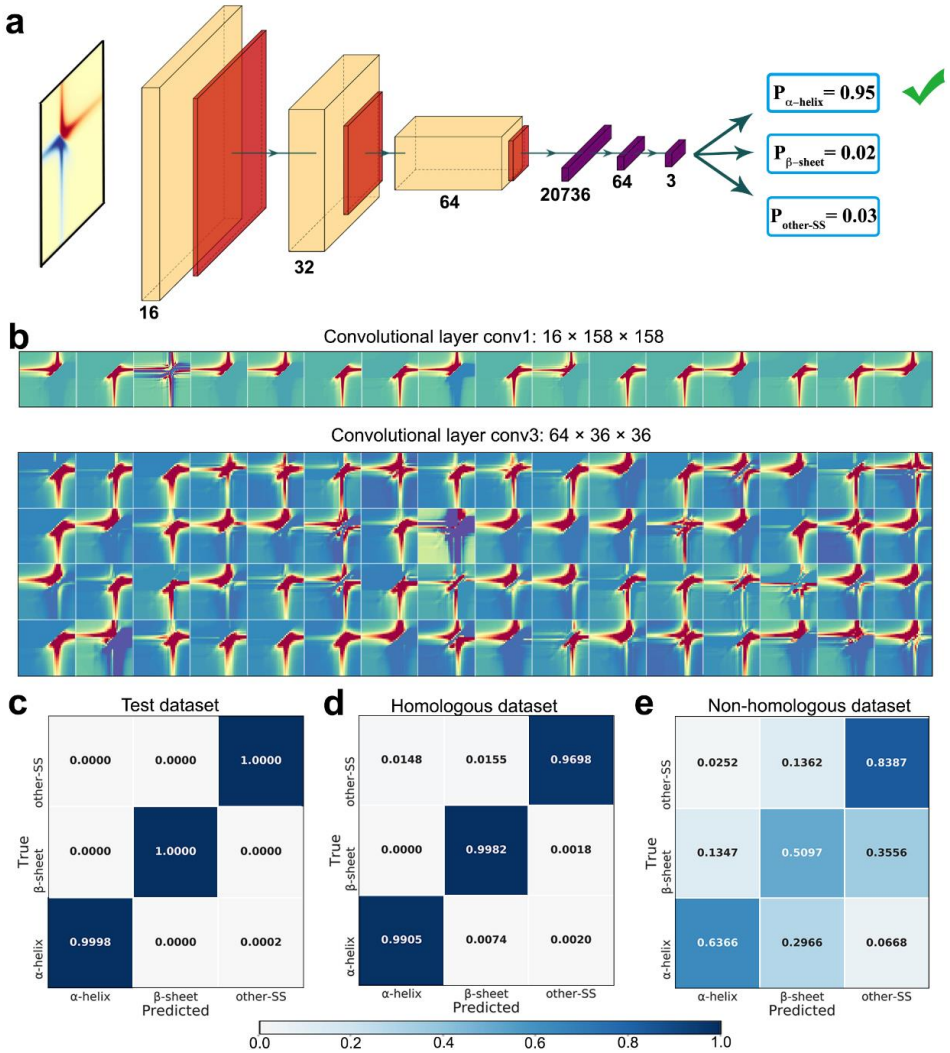
Results

IR Spectrum	Raman Spectrum	Prediction
		The model predicts no carbonyl group in the structure.

总结和讨论

- 实现了振动谱快速预测机器学习模型
- 实现了基于振动谱的结构（官能团）识别模型，可结合不同实验数据提取的化学信息进行判断
- 识别模型允许读入其他特征数据，利用更丰富的化学信息提高识别精度
- 可自然扩展至凝聚相体系（下一步工作重点）

迁移学习提高模型迁移精度

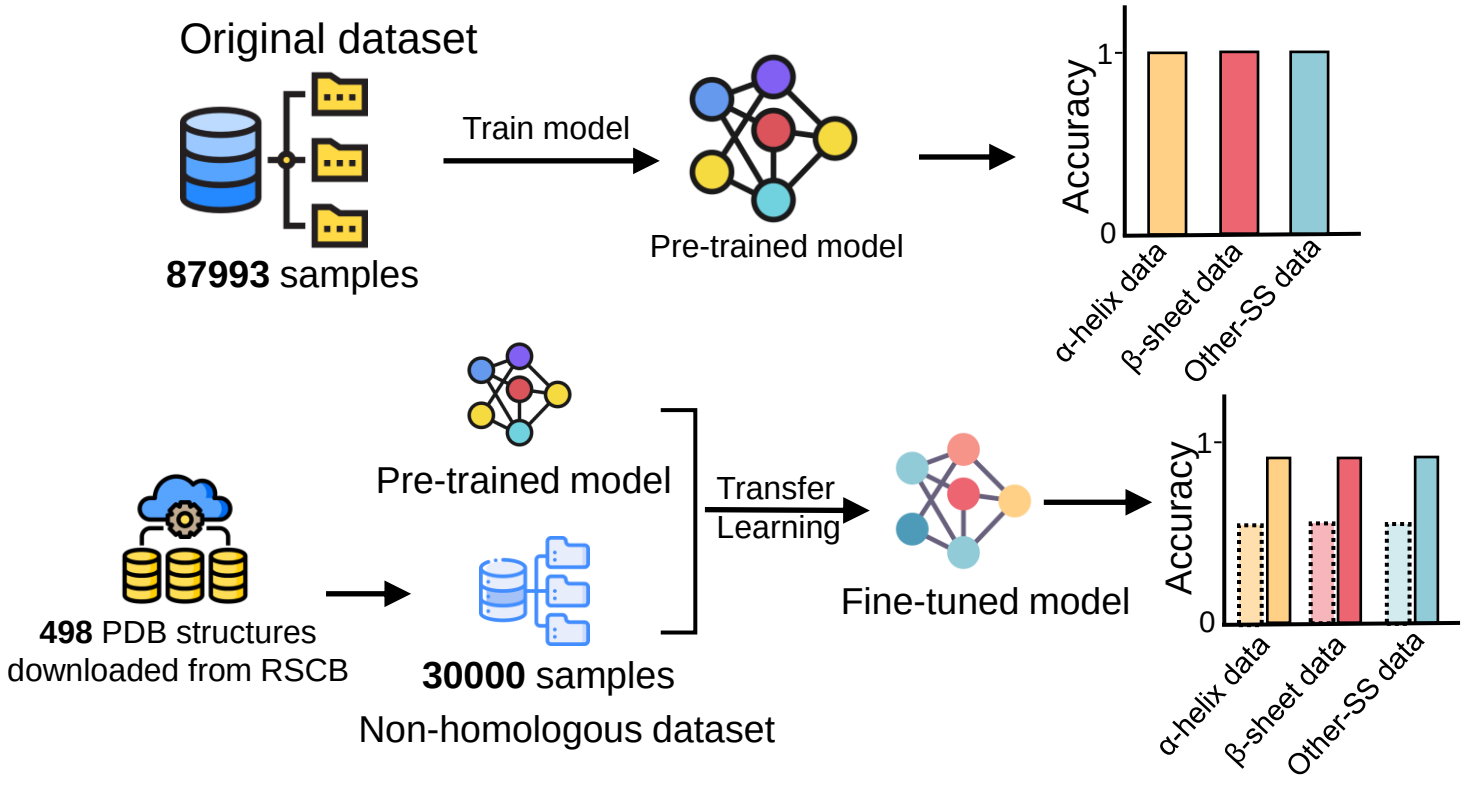


迁移学习训练集大小	100	200	400	1000	2000	4000	6000	8000
总的acc	0.76	0.82	0.87	0.9	0.91	0.93	0.94	0.95
helix	0.86	0.85	0.83	0.85	0.91	0.91	0.93	0.94
sheet	0.58	0.7	9.84	0.91	0.91	0.94	0.94	0.94
other	0.82	0.9	0.93	0.92	0.92	0.93	0.95	0.97

- 大批量典型数据（~100k）提取特征
- 少量扩展数据（~5k）提高迁移性

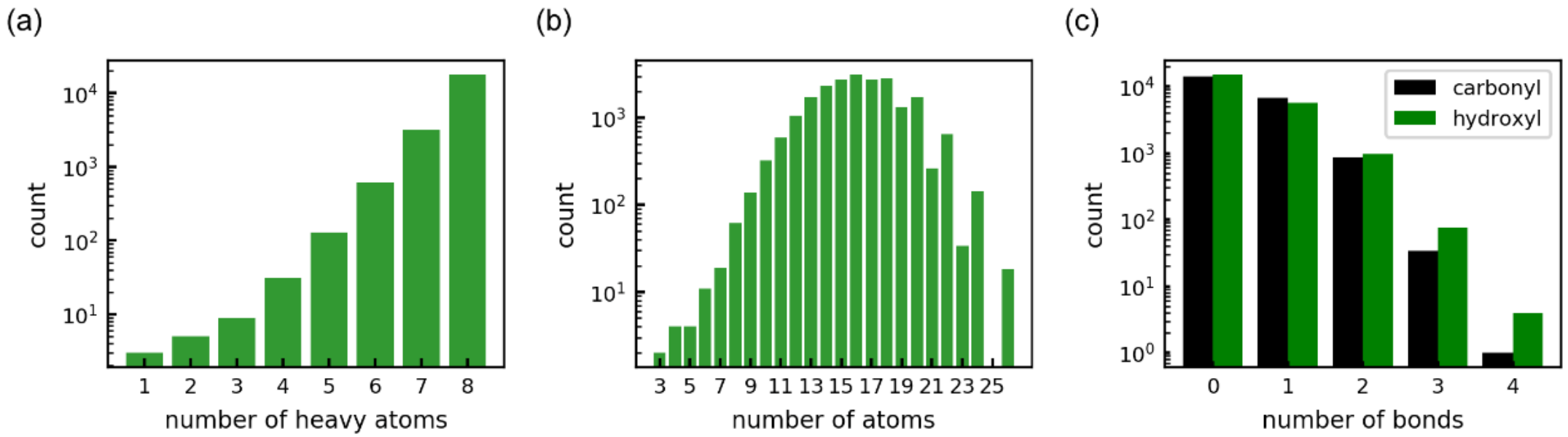
Ren, H.; et al. Mukamel, S.*; and Jiang, J.* *in manuscript.*

基于卷积神经网络的2DUV光谱-结构识别研究：迁移学习

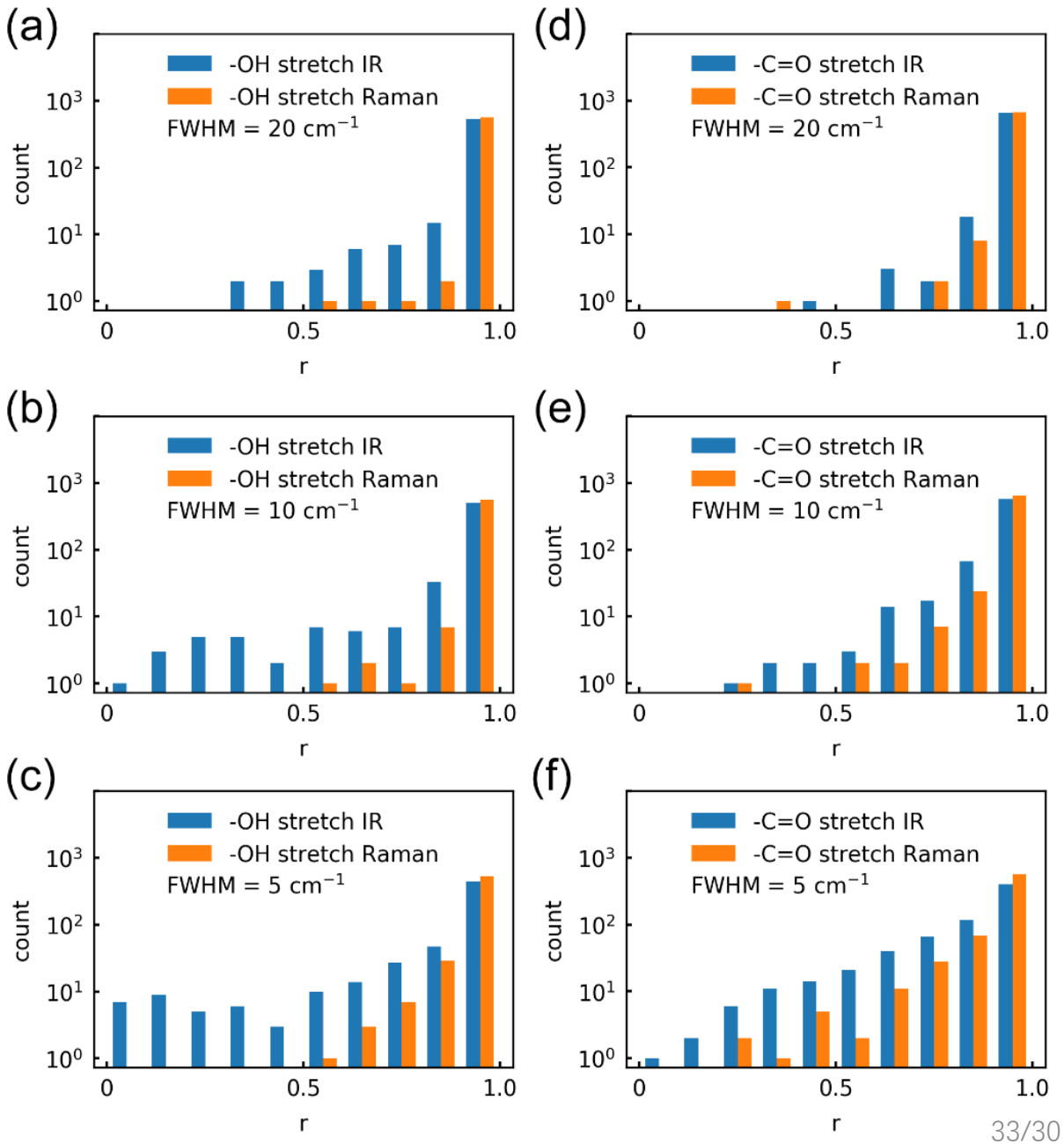


Ren, H.; et al. Mukamel, S.*; and Jiang, J.* *in manuscript.*

数据集


(a)
count
number of heavy atoms
(b)
count
number of atoms
(c)
count
number of bonds
carbonyl
hydroxyl

展宽参数


(a)
-OH stretch IR
-OH stretch Raman
FWHM = 20 cm^{-1}
(d)
-C=O stretch IR
-C=O stretch Raman
FWHM = 20 cm^{-1}
(b)
-OH stretch IR
-OH stretch Raman
FWHM = 10 cm^{-1}
(e)
-C=O stretch IR
-C=O stretch Raman
FWHM = 10 cm^{-1}
(c)
-OH stretch IR
-OH stretch Raman
FWHM = 5 cm^{-1}
(f)
-C=O stretch IR
-C=O stretch Raman
FWHM = 5 cm^{-1}
count
r

谱学信号解释和结构识别中的问题

试错模式：实验谱比对

- 数据库或手册中的标准谱（化学多样性受限）
- 猜测结构的高精度计算谱

成本高、效率低、易出错：

- 人力成本：具备化学和谱学专业知识和经验的专家
- 计算资源：猜测大量结构、高通量、高精度计算
- 时间成本：熟悉体系、查询资料、建模计算、比对分析
- 人为误差：人工操作难以避免

简介：光谱与表征

光谱是一种重要表征手段

- 光谱的解释：量子理论发展历史
- 本质是物质对光子的散射
- 著名的早期实验：黑体辐射、光电效应、波尔模型……

现代光谱学

- 基于电磁场和物质相互作用模型：量子电动力学（计算复杂）
- 电磁场驱动样品内带电粒子运动，体系响应即为光谱
- 化学上常用半经典模型处理（经典电磁场）