基于机器学习的红外光谱特征峰位移规律的探究及教学实践

摘要：红外光谱法是分子结构表征的重要方法，红外谱图解析为“仪器分析实验”课程的重要教学内容。其中，影响官能团发生位移的因素在国内外教学中多为单一因素的定性分析鲜有官能团位移规律的定量化描述。基于机器学习的数字化技术能有效提取已知数据中的隐藏知识。结合机器学习技术，探索红外特征峰的位移规律并量化，是对理论和实践教学的有益补充。为此，本项目选取红外光谱中典型的羰基峰，通过实验测试和标准谱图提供大量的数据，利用不同的机器学习方法对羰基峰位移进行拟合，得到最优模型最后，利用该模型预测了4组未知标准谱图化合物的羰基峰位移并成功量化了羰基峰的位移规律，将传统教学中的定性位移因素转化为了可视化的定量数据。本项目的教学设计与实施，通过结合机器学习技术，探索红外特征峰的位移规律。有助于学生深刻理解共轭效应、诱导效应等对羰基峰位移的影响提升学生利用数字化手段解决实际问题的能力，培养创新精神和团队合作意识。

关键词：机器学习；红外光谱；特征峰位移；教学实践

Investigation and teaching practice of infrared spectral feature peak displacement law based on machine learning

Abstract: Infrared spectroscopy is one of the important methods for molecular structure characterization, and the analysis of infrared spectra is an important teaching content in the course of “Instrumental Analysis Experiment”. Among them, the factors affecting the displacement of functional groups are mostly analyzed qualitatively in domestic and international teaching, and there are few quantitative descriptions of the displacement law of functional groups. Digitalization techniques based on machine learning can effectively extract the hidden knowledge in known data. Combining machine learning techniques to explore the displacement laws of infrared characteristic peaks and quantify them is a useful supplement to theoretical and practical teaching. To this end, this project selects typical carbonyl peaks in infrared spectra, provides a large amount of data through experimental tests and standard spectra, and utilizes different machine learning methods to fit the displacement law of carbonyl peaks, and explores the effects of different groups on carbonyl characteristic peaks. Finally, the carbonyl peak displacements of four groups of unknown standard spectrogram compounds were predicted and interpreted using the best machine learning model to successfully quantify the displacement law of carbonyl peaks. The transformation of qualitative displacement factors in traditional teaching into visualized quantitative data helps students to further deeply understand the influence of conjugation effect, induction effect, etc. on the displacement of carbonyl peaks. The pedagogical design and implementation of this project explores the displacement law of infrared characteristic peaks by combining machine learning techniques, and successfully realizes the quantification of the factors affecting the displacement of infrared characteristic peaks of functional groups in infrared spectroscopy. It helps to enhance students' ability to solve practical problems by digital means, deepen their understanding of related theoretical knowledge, and cultivate their innovative spirit and sense of teamwork.

Key Words: Machine learning; Infrared spectroscopy; Characteristic peak displacement; Teaching practice

1 引言

红外光谱法因其独特的官能团特征吸收而成为分子结构表征的重要方法之一。高效准确地解析红外谱图是化学科研工作者必备的关键技能。因此，在《仪器分析实验》课程实践教学中，红外谱图解析及影响官能团红外特征峰位移的因素通常作为教学的重点和难点内容。目前，国内外教材在描述“影响官能团红外特征峰位移”因素时，多为单个因素的定性分析，鲜有对官能团位移的影响因素进行定量描述[1-5]。若能将影响官能团位移的因素量化，将更有利于学生理解不同的作用机制，对理论和实践教学都将是大有裨益的。

大数据和人工智能的结合被称为“科学的第四次范式”和“第四次工业革命”，并在化学领域的应用迅速增长[6]。机器学习作为人工智能的一个分支，能够从大量数据中获取隐藏的、有效的、

可理解的知识，目前被广泛应用于生物信息学、经济学、工程学、医学、生物学和化学等多个领域。其中，机器学习在化学中的应用包括：提高有机反应的产率[7]、发现新的候选药物[8]、开发具有预定性质的新材料[9]、预测化合物的化学性质[10,11]等。通过结合机器学习技术，探索红外特征峰的位移规律，有望实现对影响红外光谱中官能团影响红外特征峰位移因素的量化。

羰基（C=O）是有机化学中常见的官能团，存在于醛、酮、羧酸、酯、酰胺等多种化合物中。羰基在红外光谱中表现为强烈的吸收峰，其位置受相邻基团的影响显著，在不同化学环境中，羰基峰位移范围为 1600~1800cm^{-1}。因此，本论文选取羰基峰为研究对象，通过实验测试和检索标准谱图获取大量数据，再利用机器学习方法分析数据，构建羰基峰位移和官能团之间的耦合映射关系，成功量化了红外光谱中不同官能团对羰基特征峰的位移规律，并预测了未知标准谱图化合物的羰基峰位移。其中，我们利用最佳机器学习模型预测了 4 组未知标准谱图化合物的羰基峰位移，表明该机器学习模型有望经过优化后在特定范围内具有较高的准确性。本项目的教学设计与实施，有助于提升学生利用数字化手段解决实际问题的能力，加深学生对红外课程理论知识的理解，培养学生勇于探索的创新精神和分工合作的团队意识。

2 数字化设计方案

2.1 原型实验

红外光谱法因其独特的官能团特征吸收而成为分子结构表征的重要方法之一。高效准确地解析红外谱图是化学科研工作者必备的关键技能。目前，实验教学中对于已知化合物，一般以测试为主，查阅资料进行官能团的归属，再对照标准谱图最终确定化合物的结构。

对于目前的原型实验教学内容，通用流程主要为“样品制备-仪器测试与解析-数据处理”3 个步骤。实验目的主要是让学生“了解傅里叶变换红外光谱仪的基本结构、工作原理；熟练掌握的样品制备技术和仪器规范操作流程；掌握由红外光谱鉴定未知物的一般规律，了解红外谱图解析及标准谱图检索的一般过程。”学生在红外光谱教学实验课后的最大收货往往就是学会了红外光谱的使用以及谱图的解析，很难通过实验加深对红外光谱理论知识的理解。

2.2 教学目标

影响官能团红外特征峰位移的因素是红外谱图解析的重点和难点，红外光谱实验教学中极少涉及到位移因素的阐述，国内外的理论教材很少定量描述这些因素。通过结合机器学习技术，探索红外特征峰的位移规律，有望实现对影响红外光谱中影响官能团红外特征峰位移因素的量化。在此过程中学生不仅可以掌握基本的红外光谱测试方法，还能学习到前沿的机器学习手段，通过量化红外光谱中不同官能团对特征峰的位移规律，让学生深刻理解不同基团的共轭效应、诱导效应以及这些效应对红外光谱产生的影响（**图 1**）。

原型实验教学目标	数字化设计后 实验教学目标
了解红外光谱仪的基本结构及工作原理	了解红外光谱仪的基本结构及工作原理
掌握样品制备技术和仪器规范操作	掌握样品制备技术和仪器规范操作
掌握红外谱图解析及标准谱图检索	熟练掌握红外谱图解析及标准谱图检索
	掌握基础的机器学习方法
	量化红外光谱中不同官能团对特征峰的位移规律
	助力学生深刻理解不同基团的共轭效应、诱导效应

图 1 原型实验教学目标与数字化实验教学目标对比

2.3 数字化设计思路

羰基（C=O）作为有机化合物中非常典型的官能团，广泛存在于醛、酮、羧酸、酯、酰胺等多种化合物中，是红外光谱中吸收较强的特征谱带，非常适用于探究官能团特征峰位移因素的影响。本文选取羰基峰为研究对象，通过改变羰基两侧的基团（分别记为 R_1、R_2）来改变羰基所处的化学环境。其中 R_1 基团共包含了 33 种常见基团，分别营造了电子效应、空间效应、氢键等多种化学环境；

而 R_2 有氢原子、甲基、羟基、氯原子、氨基，分别代表了醛、酮、羧酸、酰氯、酰胺这五种羰基化合物。确定化合物结构之后，我们利用实验测试和标准谱图收集到大量的羰基峰数据。

基于这些数据，采用 5 种不同的机器学习方法（SVR、随机森林、线性回归、KNN、KernelRidge）拟合不同结构的羰基峰位移，构建羰基峰位移和官能团之间的耦合映射关系。在优化参数和横向对比后，最终选择效果最佳的 SVR 方法成功量化了红外光谱中不同官能团对羰基特征峰的位移规律，并预测了未知标准谱图化合物的羰基峰位移（**图 2**）。

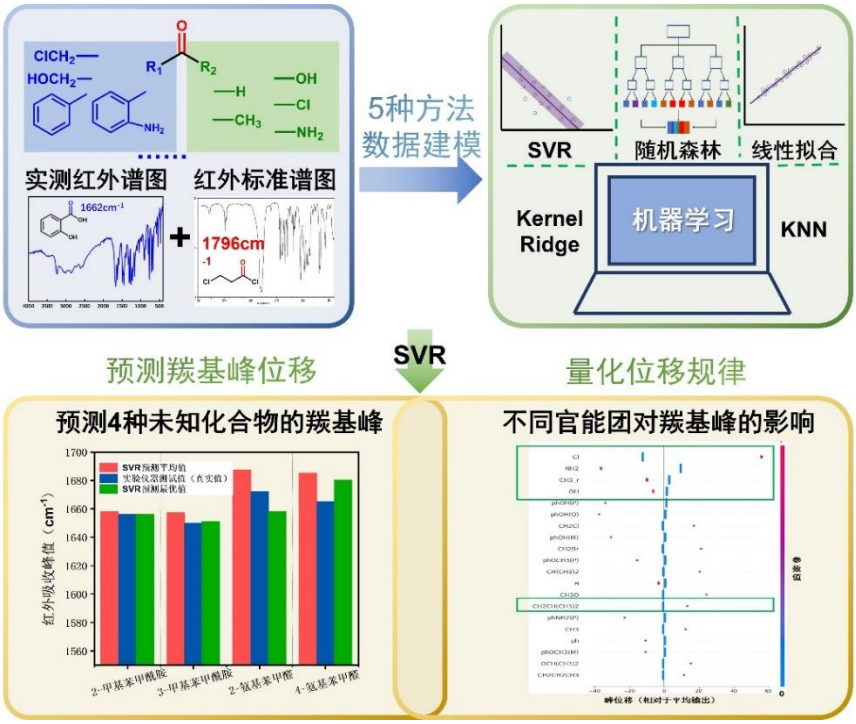


图 2 基于机器学习的红外光谱特征峰位移规律探究的整体设计思路

2.3 数字化内容：数据库及算法

数据库表格

以一个碳氧双键为中心，左右两侧各连一个基团。将所有出现的不同的基团分别作为一个变量进行处理，如果某基团出现在该物质中，其数据为 1；如果不出现，其数据则为 0。由于每个物质只会相连两个基团，每行数据只有两个变量赋值为 1，其余皆为 0。以不同基团为横列，以羰基特征峰为纵列，可以得到由 0、1 组成的数据表格，用于之后的机器学习过程。

数据导入与预处理

读取 EXCEL 文件并删除其中的缺失红外光谱特征峰值的行

```
data = pd.read_excel('Data.xlsx', header = 0, sheet_name = 0)
data = data.dropna().reset_index(drop=True)
```

X 选择了从“CH2CH3”到“H”的特征列，y 选择了目标变量红外光谱特征峰值

```
X = data.loc[:, 'CH2CH3':'H']
y = data['Actual']
```

通过不同的随机数划分数据中的 80%为训练集，20%为测试集

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=1)
```

利用 transform 方法标准化训练集和测试集的特征，提高模型的性能

```
S = StandardScaler().fit(X_train.values)
X_train = S.transform(X_train.values)
X_test = S.transform(X_test.values)
```

模型选择

该机器学习中一共选取了 5 种模型进行训练，分别是线性回归、随机森林、KNN、SVR、KernelRidge。在寻找最优模型时，为了避免因划分训练集和测试集的随机数引起的误差，我们对于 5 种模型均选取随机数 1 到 99 去进行训练。每一个模型的训练集均得到 99 组 MAE（平均绝对误差）和 R^2（用来评价训练集预测值和真实值的线性拟合度）。最后从 99 个随机数中随机抽取了 8 个，分别在这 8 种随机数环境下，对 5 个模型的平均绝对误差从小到大进行排序，最后发现 SVR 模型的平均排名最小，说明 SVR 模型用于预测羰基特征峰值更为可靠。

SVR

SVR 是一种基于支持向量机（SVM）原理的监督学习算法，基本原理是找到一个超平面来对数据进行分割，使得两个不同类别的数据在该超平面两侧尽量分布的均匀，该方法可以处理更为复杂的数据集，特别是当数据不是线性可分时非常有效。

在使用 SVR 模型时，我们使用网格搜索来确定最优超参数。由于要训练 99 次不同随机数的模型，为了避免训练时间过于长，我们没有把所有的超参数进行网格搜索，而是通过尝试后确定了一个最大可能出现最优超参数的范围来进行搜索。下面是我们超参数搜索的范围：

```
 param_grid_4 = {
'kernel': ['linear', 'rbf', 'sigmoid'],
'gamma': ['scale', 'auto'],
'coef0': [0],
'C': list(np.linspace(11,30,20, dtype=int)),
    'epsilon': [0.5, 1,2,3,4,5]
}
```

上面每个超参数的意思如下：

（1）“kernel”：核函数，用于将输入数据映射到高维空间，便于在该空间进行线性回归。我们选择了三种核函数，分别是“linear”--线性核函数，适用于线性问题；“rbf”--径向基函数核，适用于非线性问题；“sigmoid”--称为双曲正切核函数或 S 型核函数，也适用于非线性问题；

（2）“gamma”：控制核函数的影响范围，决定了一个单一训练样本对决策边界的影响。“scale”--根据特征的数量和方差来计算得到一个 gamma 值；“auto”--计算得到的 gamma 值为特征数量的倒数；

（3）“coef0”：用于核函数的常数项，主要影响多项式核和 sigmoid 核的灵活性。在该参数选择中，由于测试时所有情况均选择了“linear”这个核函数，因此我们将此超参数定为 0，并不影响训练效果且减少计算成本；

（4）“C”：惩罚参数，控制对误差的容忍度。一般来说，C 值越大，模型越关注训练集的准确性，可能导致过拟合。该超参数我们选择了 11-30 的整数；

（5）“epsilon”：在直线附近定义了一个边界，在这个边界内的误差不会被惩罚，也就是说，当预测值和真实值的差小于 epsilon 时，这部分不会影响模型的训练。该超参数我们选择了 0.5，1，2，3，4，5。

然后通过网格搜索确定最优超参数，并将最优超参数传递给 SVR 模型进行训练，得到 SVR 训练集模型

```
grid_search_4 = GridSearchCV(estimator = svr, param_grid = param_grid_4, scoring='neg_mean_squared_error', cv = 5, n_jobs = -1)
grid_search_4.fit(X_train, y_train)
m4 = SVR(**grid_search_4.best_params_).fit(X_train, y_train)
```

用训练好的模型去预测测试集数据的红外光谱特征峰值，与真实值对比，输出测试集得到的 MAE 和 R^2 来表示这次模型的好坏。

```
m4_test = m4.predict(X_test)
print('Mean Absolute Error:',mean_absolute_error(y_test, m4_test))
```

print('R^2 score:',r2_score(y_test, m4_test))

为了避免随机数引起的误差，在上面程序之前采用循环来对 SVR 多次训练，得到 random1-99 的共 99 组数据。

for i in range(1,100):

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=i)

具体运行程序见**图 3**：

SVR

```
param_grid_4 = {
  'kernel': ['linear', 'rbf', 'sigmoid'],
  'gamma': ['scale', 'auto'],
  'coef0': [0],
  'C': list(np.linspace(11,30,20, dtype=int)),
  'epsilon': [0.5, 1,2,3,4,5,]
}
svr = SVR()
grid_search_4 = GridSearchCV(estimator = svr, param_grid = param_grid_4, scoring='neg_mean_squared_error', cv = 5, n_jobs = -1)
grid_search_4.fit(X_train, y_train)

m4=SVR(**grid_search_4.best_params_).fit(X_train,y_train)
m4_test=m4.predict(X_test)
print('Mean Absolute Error:',mean_absolute_error(y_test,m4_test))
print('R^2 score:',r2_score(y_test,m4_test))
```

Python

图 3 SVR 算法的运行程序

机器学习可解释性--SHAP 方法

模型的可解释性一直是机器学习中关注的焦点，新兴的 SHAP（SHapley Additive exPlanations）算法提供了一种全新的视角来理解复杂的机器学习模型。SHAP 算法是基于博弈论中的 Shapley 值的一个解释模型的方法，是一种事后解释框架。该算法将模型的输出分解为各个特征的贡献和，通过计算每个特征对模型输出的边际贡献，计算出其重要性值（Shapley 值），进而帮助理解模型是如何做出决策的。

我们首先选取了 SVR 模型中表现相对较好的随机数——66 进行超参数搜索，用最优超参数训练 SVR 模型，然后利用 SHAP 中的 Explainer 来解释训练得到的模型结果，最后使用 SHAP 中的 summary_plot 输出宏观特征密度散点图（**图 8**）。该图将每一个变量的贡献可视化，可以显示特征的重要性和特征值的影响方向。

model = SVR(**grid_search_4.best_params_)

model_prediction = model.fit(X, y)

explainer=shap.Explainer(model.predict,X)

shap_values=explainer.shap_values(X)

shap.summary_plot(shap_values,X)

3 实验部分

3.1 实验原理

红外光谱部分原理：

物质分子中的各种不同基团，受到频率连续变化的红外光照射时，分子吸收了某些频率的辐射，并由其振动或转动运动引起偶极矩的净变化，产生分子振动和转动能级从基态到激发态的跃迁，形成的分子光谱称为红外光谱。由于化合物的分子结构不同，所产生的红外光谱即不同，据此实现化合物的定性与定量分析[12]。本实验选取羰基的红外吸收峰位移规律进行定量地探究。

数字化部分原理：

采用线性回归、随机森林、KNN、SVR、KernelRidge 五种不同的机器学习算法，既运用传统的单变量方法，又使用基于多变量特征的非线性分析方法，并从中找到最适用于提取红外峰潜在位移规律的算法。

在找到适合的 SVR 方法后，我们进一步使用 SHAP 模型对 SVR 模型进行事后解释，将不同基团对羰基峰峰值的影响进行定量描述，最后得到一个宏观特征密度散点图。图中基团对应的红

点与蓝点横坐标的平均差值，即为该基团对羰基峰位移的定量影响。具体的算法内容详见“**2.4 数字化内容：数据库及算法**”及正文后“**附件材料-1**”。

3.2 试剂或材料

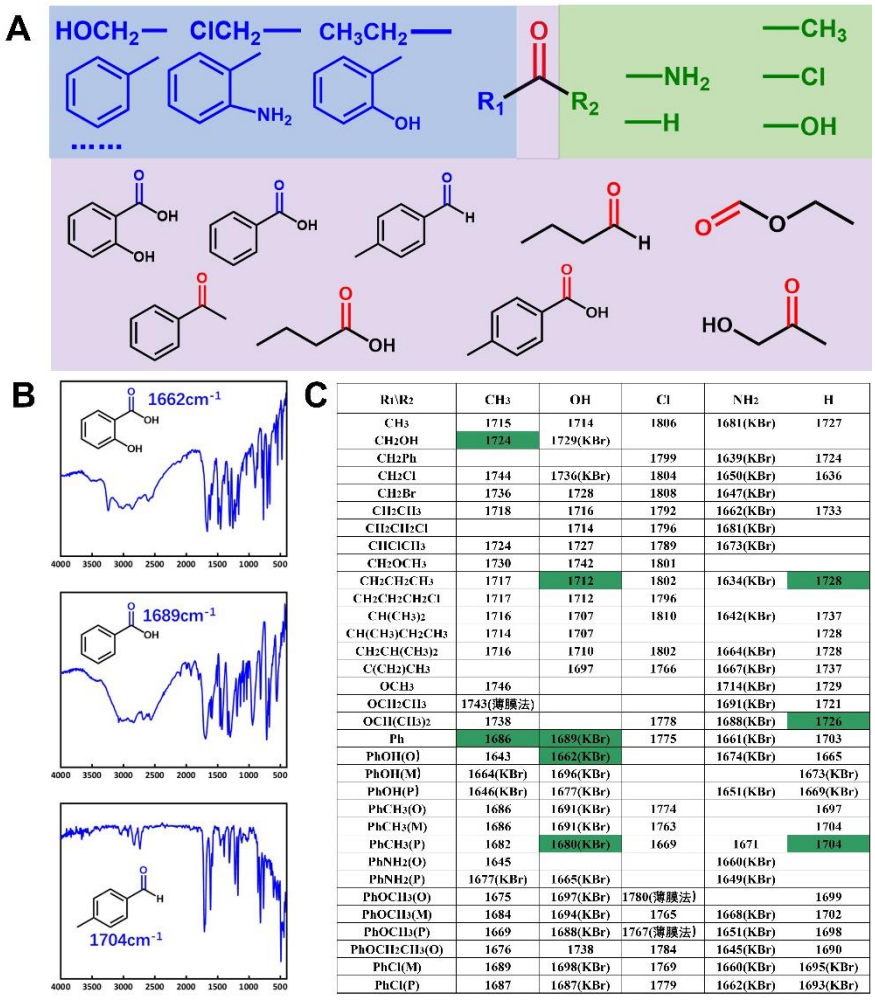
苯甲酸、水杨酸、丙酮醇、对甲基苯甲酸、对甲基苯甲醛、2-甲基苯甲酰胺、3-甲基苯甲酰胺、4-氨基苯甲醛、2-氨基苯甲醛、苯乙酮、正丁酸、正丁醛、甲酸乙酯。以上实验试剂均来自阿拉丁化学试剂有限公司，所用试剂未经进一步处理。

3.3 仪器和表征方法

实验仪器：傅里叶变换红外光谱仪（德国 Bruker 公司，TENSOR II 型）。

计算软件：Visual Studio Code User 1.93.1(Python 版本 3.12.4)；电脑：Lenovo XiaoXinPro 16AH 2021。

3.4 实验过程及结论



R1\R2	CH3	OH	Cl	NH2	H
CH3	1715	1714	1806	1681(KBr)	1727
CH2OH	1724	1729(KBr)			
CH2Ph			1799	1639(KBr)	1724
CH2Cl	1744	1736(KBr)	1804	1650(KBr)	1636
CH2Br	1736	1728	1808	1647(KBr)	
CH2CH3	1718	1716	1792	1662(KBr)	1733
CH2CH2Cl		1714	1796	1681(KBr)	
CHClCH3	1724	1727	1789	1673(KBr)	
CH2OCH3	1730	1742	1801		
CH2CH2CH3	1717	1712	1802	1634(KBr)	1728
CH2CH2CH2Cl	1717	1712	1796		
CH(CH3)2	1716	1707	1810	1642(KBr)	1737
CH(CH3)CH2CH3	1714	1707			1728
CH2CH(CH3)2	1716	1710	1802	1664(KBr)	1728
C(CH2)CH3		1697	1766	1667(KBr)	1737
OCH3	1746			1714(KBr)	1729
OCH2CH3	1743(薄膜法)			1691(KBr)	1721
OCH(CH3)2	1738		1778	1688(KBr)	1726
Ph	1686	1689(KBr)	1775	1661(KBr)	1703
PhOH(O)	1643	1662(KBr)		1674(KBr)	1665
PhOH(M)	1664(KBr)	1696(KBr)			1673(KBr)
PhOH(P)	1646(KBr)	1677(KBr)		1651(KBr)	1669(KBr)
PhCH3(O)	1686	1691(KBr)	1774		1697
PhCH3(M)	1686	1691(KBr)	1763		1704
PhCH3(P)	1682	1680(KBr)	1669	1671	1704
PhNH2(O)	1645			1660(KBr)	
PhNH2(P)	1677(KBr)	1665(KBr)		1649(KBr)	
PhOCH3(O)	1675	1697(KBr)	1780(薄膜法)		1699
PhOCH3(M)	1684	1694(KBr)	1765	1668(KBr)	1702
PhOCH3(P)	1669	1688(KBr)	1767(薄膜法)	1651(KBr)	1698
PhOCH2CH3(O)	1676	1738	1784	1645(KBr)	1690
PhCl(M)	1689	1698(KBr)	1769	1660(KBr)	1695(KBr)
PhCl(P)	1687	1687(KBr)	1779	1662(KBr)	1693(KBr)

图 4（A） 常见的含羰基化合物，以及羰基两端基团的选择（B） 水杨酸、苯甲酸、对甲基苯甲醛三种物质的红外光谱实测图及羰基峰位移（C）不同化合物的羰基峰位移（绿色部分为实验测试的羰基峰位移数据，实验测试和标准谱图中的羰基峰位移几乎没有差别）

首先，对羰基两端的基团进行分类组合，选择不同类型的含羰基化合物，从而获得了不同的化合物结构。对易于购买且安全性较高的常见试剂进行了实验测试，比如苯甲酸、水杨酸、丙酮醇、对甲基苯甲酸、对甲基苯甲醛、苯乙酮、正丁酸、正丁醛、甲酸乙酯（**图 4A**）。最后得到了不同的实验测试谱图（**图 4B、图 S5**），并提取出其中羰基峰的位移值。然而，由于实验能测试的数据有限，且大部分含卤素化合物和部分酰胺类化合物具有很强的毒性，为满足实验的安全性、

数据的广泛性及绿色化的需求，我们从 ChemicalBook 平台上检索了部分化合物的红外谱图，提取了其中的羰基峰位移（**图 4C**）。通过比对**图 4A** 中化合物实际测量峰值和标准谱图峰值，发现两者的羰基峰位移几乎没有差异，因此利用标准谱图对化合物的羰基峰位移进行补充是合理的。

在得到 132 组数据的前提下，我们选用了 5 种不同的机器学习方法（随机森林、KernelRidge、线性回归、KNN、SVR。SVR 具体程序请见“**2.4 数字化内容：数据库及算法**”，余下四种算法具体程序请见正文后“**附件材料-1**”）对以上的数据进行了处理，通过“test size”和不同的随机数划分数据中的 80%为训练集，20%为测试集。第一次运行程序时，我们的随机数（从给定的数组里随机抽取值，随机数不同代表抽取的数据不同）选为 1，此时可以得到五种不同的拟合模型（**图 5A-E**）。通过对这 5 个模型的 MAE（平均绝对误差）和 R^2（用来评价训练集预测值和真实值的线性拟合度）进行比较（**图 5F**），我们发现 KernelRidge 这种方法在随机数为 1 的前提下，其平均绝对误差最小（12%左右），R^2 最大，说明其拟合的效果最好。然而由于羰基峰的数据量整体较小，且进一步扩充的难度较大，通过随机数划分训练集和测试集的时候可能会引起不小的误差，只训练某一个随机数的话，可能会出现偶然的情况。

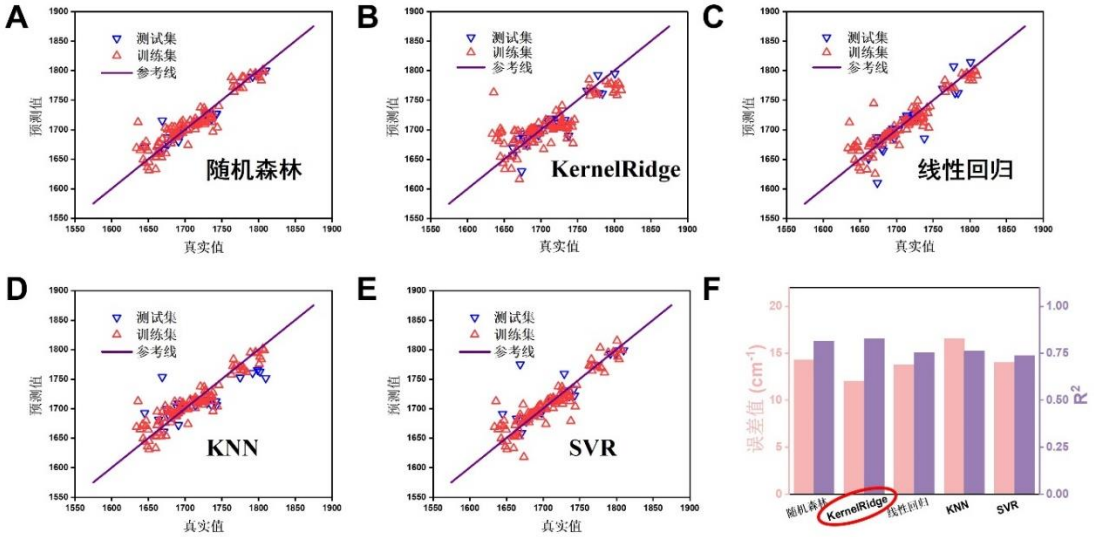
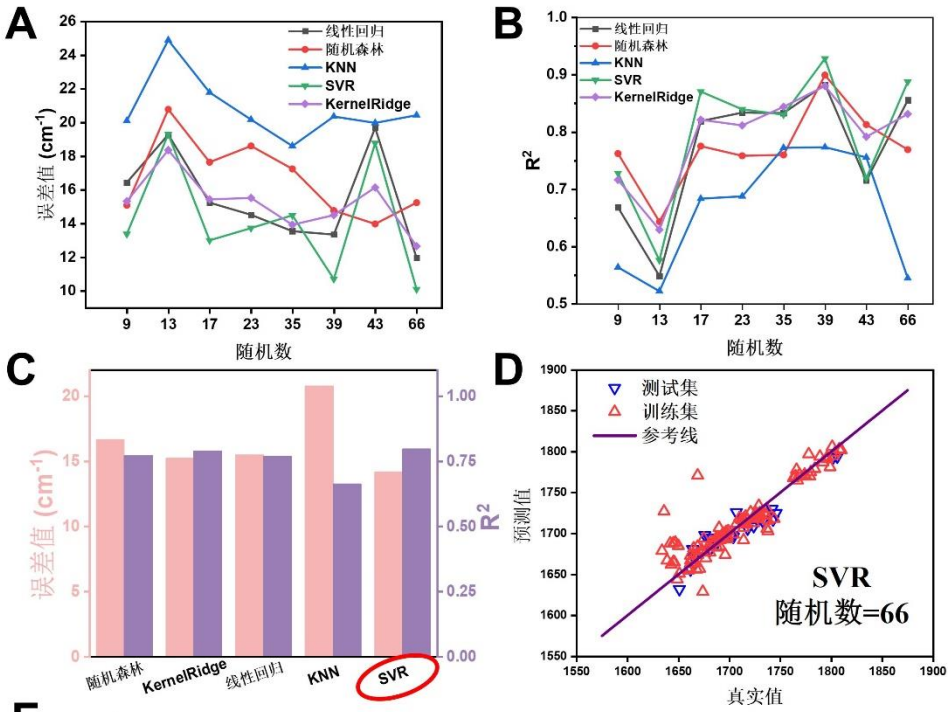


图 5 机器学习模型训练：分别利用（A）随机森林、（B）KernelRidge、（C）线性回归、（D）KNN、（E）SVR 对羰基峰位移规律进行拟合（F）五种方法的误差值及 R^2 对比（随机数=1）

因此，为了降低偶然性，寻找最合理的预测模型，避免因划分训练集和测试集的随机数引起的误差，我们对于每个模型均选取随机数 1 到 99 去进行训练。每一个模型的训练集都得到 99 组的 MAE（平均绝对误差）和 R^2（线性拟合程度）最后求这 99 组的平均 MAE 和平均 R^2，发现 SVR 比其他的模型表现得更为优异（**图 6C**）。最后我们从 99 个随机数中随机抽取了 8 个随机数（9、13、17、23、35、39、43、66），并将其平均 MAE（**图 6A**）和平均 R^2（**图 6B**）进行了展示。同时我们对这 8 种随机数环境下 5 个模型的平均绝对误差从小到大进行排序，最后发现 SVR 模型的平均排名最小。通过大量的随机数测试、比较，我们判断 SVR 模型可能是最适合用于该场景的拟合模型。SVR 模型中拟合效果最好的随机数为 66，其平均绝对误差达到 10%以下（**图 6D**）。之后我们利用 SVR 模型（随机数=66）对**图 3C** 中未找到标准谱图化合物的羰基峰位移进行预测，得到的结果见**图 6E**（标黄部分为 SVR 模型预测的羰基峰位移）。



E

R1\R2	CH3	OH	Cl	NH2	H
CH3	1715	1714	1806	1681(KBr)	1727
CH2OH	1724	1729(KBr)	1806.053	1686.947	1735.153
CH2Ph	1704.04	1706.94	1799	1639(KBr)	1724
CH2Cl	1744	1736(KBr)	1804	1650(KBr)	1636
CH2Br	1736	1728	1808	1647(KBr)	1727.871
CH2CH3	1718	1716	1792	1662(KBr)	1733
CH2CH2Cl	1715.758	1714	1796	1681(KBr)	1724.81
CHClCH3	1724	1727	1789	1673(KBr)	1724.11
CH2OCH3	1730	1742	1801	1695.424	1743.603
CH2CH2CH3	1717	1712	1802	1634(KBr)	1728
CH2CH2CH2Cl	1717	1712	1796	1674.922	1723.102
CH(CH3)2	1716	1707	1810	1642(KBr)	1737
CH(CH3)CH2CH3	1714	1707	1792.303	1673.223	1728
CH2CH(CH3)2	1716	1710	1802	1664(KBr)	1728
C(CH2)CH3	1703.555	1697	1766	1667（KBr)	1737
OCH3	1746	1742.592	1819.645	1714(KBr)	1729
OCH2CH3	1743(薄膜法)	1731.258	1808.311	1691(KBr)	1721
OCH(CH3)2	1738	1730.258	1778	1688(KBr)	1726
Ph	1686	1689(KBr)	1775	1661(KBr)	1703
PhOH(O)	1643	1662(KBr)	1729.427	1674(KBr)	1665
PhOH(M)	1664(KBr)	1696(KBr)	1758.503	1639.424	1673(KBr)
PhOH(P)	1646(KBr)	1677(KBr)	1750.028	1651(KBr)	1669(KBr)
PhCH3(O)	1686	1691(KBr)	1774	1651.947	1697
PhCH3(M)	1686	1691(KBr)	1763	1650.556	1704
PhCH3(P)	1682	1680(KBr)	1669	1671	1704
PhNH2(O)	1645	1674.964	1752.017	1660(KBr)	1681.116
PhNH2(P)	1677(KBr)	1665(KBr)	1755.066	1649(KBr)	1684.166
PhOCH3(O)	1675	1697(KBr)	1780(薄膜法)	1643.347	1699
PhOCH3(M)	1684	1694(KBr)	1765	1668(KBr)	1702
PhOCH3(P)	1669	1688(KBr)	1767(薄膜法)	1651(KBr)	1698
PhOCH2CH3(O)	1676	1738	1784	1645(KBr)	1690
PhCl(M)	1689	1698(KBr)	1769	1660(KBr)	1695(KBr)
PhCl(P)	1687	1687(KBr)	1779	1662(KBr)	1693(KBr)

图 6 机器学习模型优化及预测：不同随机数（9、13、17、23、35、39、43、66）下五种机器学习方法的（A）平均绝对误差和（B）R^2（C）五种方法的误差值及方差 R^2 对比（随机数=1）（D）SVR 模型（随机数=66）的拟合模型图（E）SVR 模型预测的羰基峰位移，其中标黄部分为预测值，标绿部分为实验测试值，其它为标准谱图中的数据

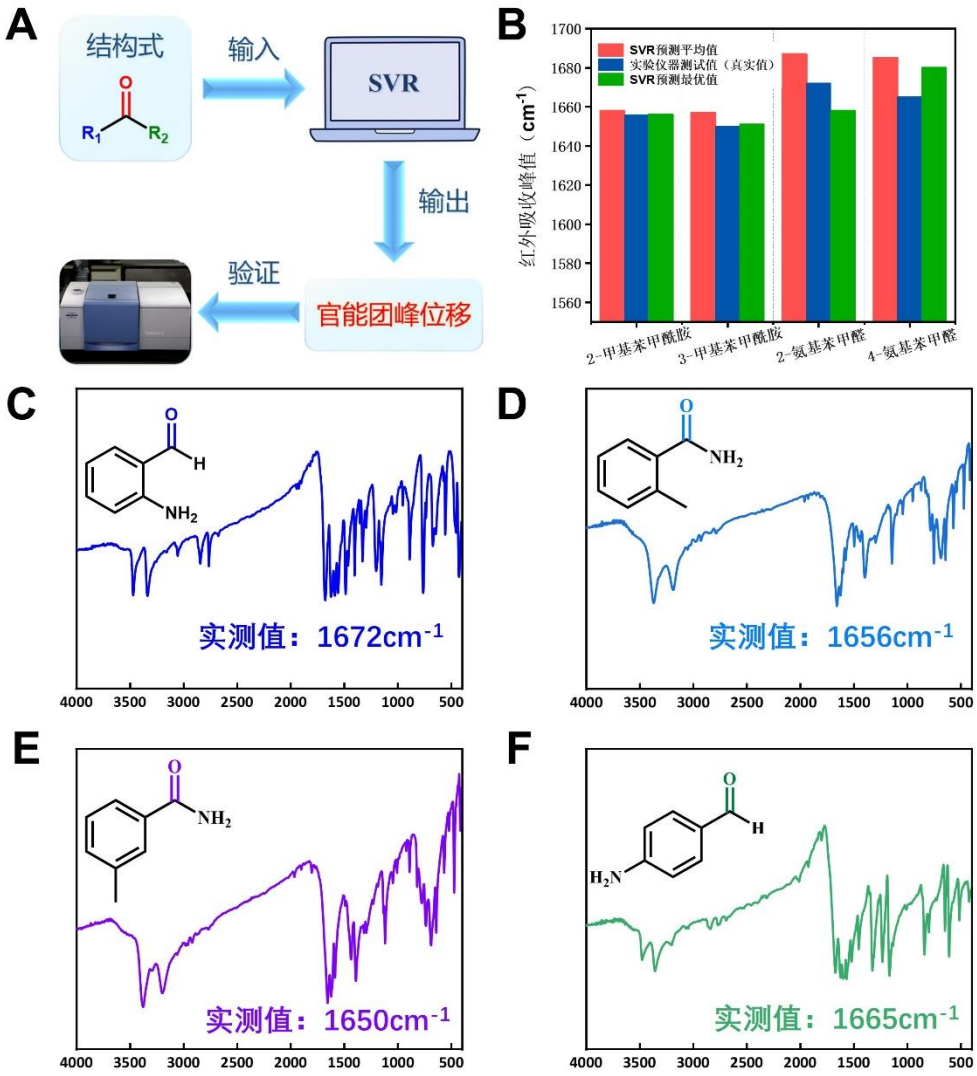


图 7 机器学习模型预测及验证：（A）利用机器学习模型（SVR 方法）建模、预测及验证的思路。（B）2-甲基苯甲酰胺、3-甲基苯甲酰胺、2-氨基苯甲醛、4-氨基苯甲醛的真实测试值，8 种随机数（9、13、17、23、35、39、43、66）条件下的 SVR 预测值（C）2-氨基苯甲醛、（D）2-甲基苯甲酰胺、（E）3-甲基苯甲酰胺、（F）4-氨基苯甲醛的红外测试谱图及羰基峰位移

通过大量的数据训练和横向对比后，我们得到一个较为可靠的机器学习模型（SVR 方法）。通过该模型预测未知化合物的羰基峰位，并利用实验测试对其进行验证，有助于进一步理解机器学习等数字化技术对于传统实验科学的帮助与意义（**图 7A**）。我们通过选取了未能搜集到标准谱图的四种化合物：2-甲基苯甲酰胺、3-甲基苯甲酰胺、2-氨基苯甲醛、4-氨基苯甲醛，并利用红外光谱仪对其进行测试，得到了实际测试谱（**图 7C-F**）。然后，用 SVR 模型分别在上述 8 种随机数条件下，对羰基特征峰进行预测，由此得到 SVR 预测平均值和 SVR 预测最优值，并与真实测试值进行比较（**图 7B**）。这 4 组未知标准谱图化合物的羰基峰位移预测值中，2-甲基苯甲酰胺、3-甲基苯甲酰胺的预测值与实验测试值的平均误差不超过 5 cm^{-1}，表明该机器学习模型在特定范围内具有较高的准确性。2-氨基苯甲醛、4-氨基苯甲醛的预测误差较大，平均误差在 15 cm^{-1} 左右。

在找到适合的 SVR 方法后，我们进一步使用 SHAP 模型对 SVR 模型进行事后解释，将不同基团对羰基峰峰值的影响进行定量描述，最后得到一个宏观特征密度散点图（**图 8**），成功得到了不同官能团对羰基峰位移的影响权重，量化了羰基峰的位移规律。**图 8** 中的横坐标为 SHAP 值，其大于 0，为正面影响；小于 0，为负面影响,且平均绝对值越大，代表对应的基团对羰基峰值的影响越明显。纵列每一行代表一个基团，按照 SHAP 值的平均绝对值由高到低进行排序。图中一个点代表

一个样本，红点对应数据表格中的“1”，蓝点对应“0”。宽的地方表示有大量的样本聚集。**图 8** 中，基团对应的红点与蓝点横坐标的平均差值，即为该基团对羰基峰位移的定量影响。

我们选择 R_2 包含的五种基团（具有代表性且对羰基峰值影响明显的四种基团，“Cl”、“NH_2”、“CH_3”、“OH”）以“H”为基准，其中“Cl”对羰基峰值产生近 75 波数的正面影响，“NH_2”、“CH_3”、“OH”对羰基峰值依次产生近 40 波数、15 波数、10 波数的负面影响，与标准谱图反映的情况相符，成功量化羰基特征峰位移规律。通过本谱图我们也可以比较不同基团更换时，羰基峰会发生的位移。比如将“Cl”换成 “NH_2”时，羰基峰的峰值会降低大约 115 个波数，这与我们观察到的数据比较接近（**图 6E**），比如 CH_3CH_2COCl（1792cm^{-1}）和 $CH_3CH_2CONH_2$（1662cm^{-1}）之间的波数差距为 130 个波数，$ClCH_2CH_2COCl$（1792cm^{-1}）和 $ClCH_2CH_2CONH_2$（1662cm^{-1}）之间的波数差距为 130 个波数。通过大量数据的总结归纳与可视化，能够让学生更加深刻理解共轭效应、诱导效应对于红外光谱中不同基团的影响。

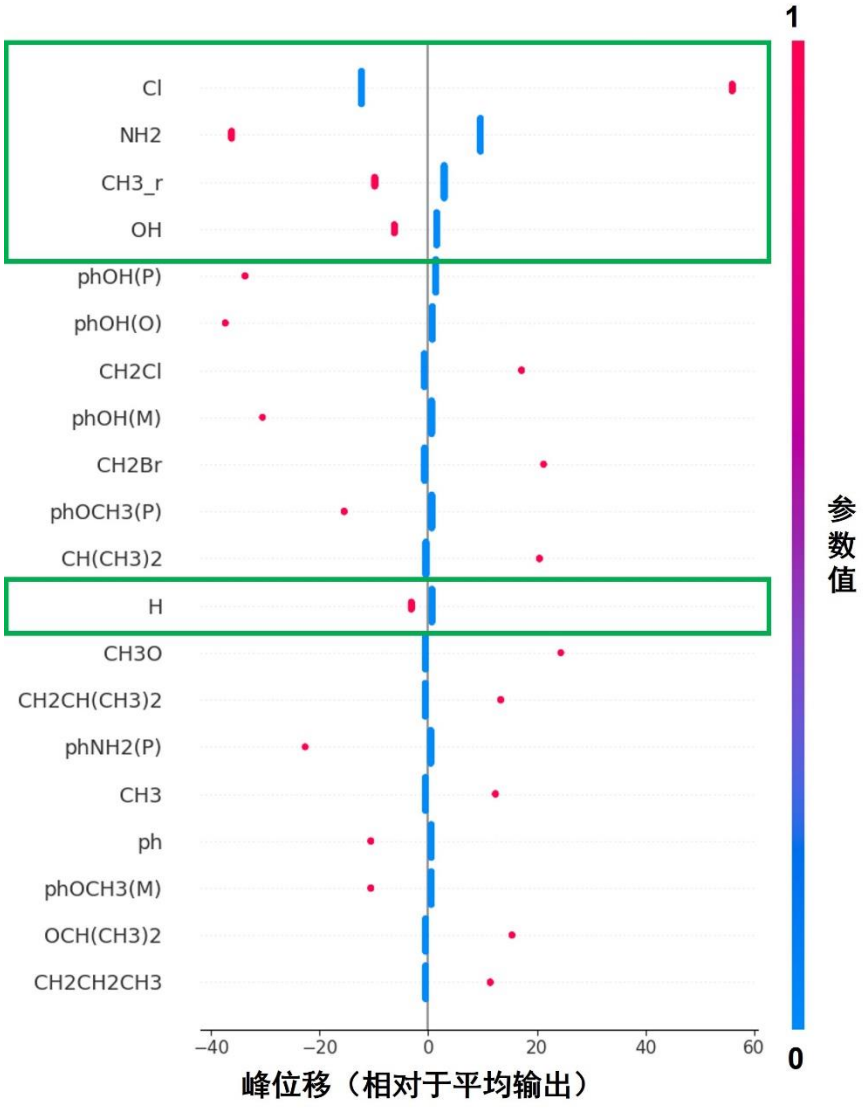


图 8 机器学习模型对不同基团影响权重的统计（绿色方框内是 R_2 基团，数据量相对较多）

4 教学实践

课前，学生对实验内容进行预习，包括红外光谱实验原理、机器学习背景及 SVR 模型和 SHAP 模型的相关知识。课中，教师指导学生测试部分化合物的红外光谱，并结合标准谱图建立数据库，利用 Python 编写 SVR 模型预测未知化合物的红外特征峰，编写 SHAP 程序对其进行解释。课后，学生对数据进行对比分析，量化官能团红外光谱特征峰的位移规律，完成实验报告。为合理安排时

间并提高学生的协作能力，红外测定并建立数据库、编写程序，均以小组为单位合作完成。实验项目总学时约为 5 学时。

5 结语

本项目把经典的红外光谱理论和实验实践教学与人工智能领域相联系，探索了红外特征峰的位移规律并量化，改变了以测试为主的传统红外光谱实验教学内容。该机器学习的方法不仅可以量化羰基的特征峰位移规律，预测羰基特征峰值，同样对于其他官能团，如碳碳双键等，也具有适用性。本项目的实施有助于提升学生利用数字化手段解决实际问题的能力，加深学生对红外光谱理论知识的理解，培养学生创新精神和团队合作意识。

6 特色与创新

（1）教学内容创新：经典实验与机器学习结合，实现了羰基特征峰位的预测并可视化其位移规律，改变了单一依赖测定和谱图检索确定峰位的传统方法，加深对理论知识的理解；

（2）教学方法可推广：该机器学习方法通用且易于理解，程序较为简单，在原有的实验教学基础上，学生仅用个人电脑就能尝试多种机器学习方法；

（3）增加课程挑战度：本实验涉及数字化建模（SVR）和模型解释（SHAP）两种梯度化的机器学习方法，在挑战中丰富学生的理论知识，培养学科交叉处理问题的能力。

参考文献：

[1] 陈怀侠, 王升富，叶勇. 仪器分析. 北京: 科学出版社, 2022:93-109

[2] 孙东平，李羽让，纪明中，唐卫华. 现代仪器分析实验技术(上册). 北京: 科学出版社, 2021:334-360

[3] 郁桂云, 钱晓荣.仪器分析实验教程(第二版).上海: 华东理工大学出版社,2015:103-12

[4] Robinson, J. W.; Skelly Frame, E. M.; Frame II, G. M. *Undergraduate instrumental analysis*, Seventh edition.; CRC Press, Boca Raton, FL,USA .2014:250-268

[5] Robinson, K. A..; Robinson, J. F. *Contemporary instrumental analysis* Upper Saddle River, NJ : Prentice Hall, c2000:439-466

[6] Butler, K. T.; Davies ,D.W.; Cartwright, H.; Isayev, O.; Walsh ,A. *Nature*, **2018**, *559*,547.

[7] Blanco, D. E.; Lee, B.; Modestino, M.A. *PNAS*, **2019**, *116* (36), 17683.

[8] Chan, S. H.C.; Shan, H.B.; Dahoun , T.; Vogel, H.; Yuan, S.G. *Trends in Pharmacological Sciences*, **2019**, *40*(8),592.

[9] Janet, J. P.; Ramesh, S.; Duan, C.; *ACS Cent. Sci.* **2020**, *6*, 513.

[10] He, H.; Xu, M, x.; Zong, C.; Zheng, P.; Luo, L. L.; Wang, L.; Ren, B. *Anal. Chem.*, **2019,** *91*, 7070.

[11] Lussier, F.; Missirlis, D.; Spatz, J. P.; Masson, J. F. *Nano*, **2019,** *13*, 1403.

[12] 北京大学化学系仪器分析教学组. 仪器分析教程. 北京：北京大学出版社, 1997: 53.

附件材料

附件材料-1
余下 4 种机器学习模型的解释：

线性回归

与 SVR 一样的处理，我们选取了同样的 99 个随机数对模型进行训练，并输出模型在测试集预测结果的 MAE 和 R^2

```
m1 = LinearRegression().fit(X_train, y_train)
m1_test = m1.predict(X_test)
print('Mean Absolute Error:',mean_absolute_error(y_test, m1_test))
print('R^2 score:',r2_score(y_test, m1_test))
```

线性回归模型没有超参数，并不需要进行超参数的处理。

线性回归

```
m1 = LinearRegression().fit(X_train, y_train)
m1_test = m1.predict(X_test)
print('Mean Absolute Error:',mean_absolute_error(y_test, m1_test))
print('R^2 score:',r2_score(y_test, m1_test))
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=66)

S = StandardScaler().fit(X_train.values)
X_train = S.transform(X_train.values)
X_test = S.transform(X_test.values)
```

Python

图 S1 线性回归算法的运行程序

随机森林

随机森林由多个决策树组成，决策树通过一系列问题来决策出各个属性对预测结果的影响，随机森林则可以整合多个决策树的输出为一个统一的结果。

随机森林与 SVR 训练模型的思路与程序基本一致，程序上只需要把训练时的模型调用修改成随机森林即可，并同样的输出训练集的 MAE 和 R^2。

```
m2=RandomForestRegressor(bootstrap=True,random_state=0,**grid_search_2.best_params_).fit(X_train, y_train)
```

随机森林的超参数选择我们同样采取了与 SVR 一致的方法，通过测试找到了一个最大可能出现超参数的范围，在这个范围内进行超参数的网格搜索。（后面的两种模型将不再说明这个问题，只展示我们所确定的超参数范围）下面是我们随机森林的超参数搜索范围：

```
param_grid_2 = {
'max_depth': list(np.linspace(10,100,10, dtype=int)),
'max_features': list(np.linspace(11,30,20, dtype=int)),
'n_estimators': list(np.linspace(10,100,10, dtype=int))
}
```

上面每个超参数的意思如下：

（1）“max_depth”:树的最大深度，超过树的最大深度后的数据会被剪掉。限制树的深度可以在一定程度上防止过拟合现象。我们采用了从 10-100 中的 10 个等间距整数，即 10，20，30……；

（2）“max_features”:在每一个树的分裂节点时需要考虑的特征数量。我们采用了 11-30 之间的所有共 20 个整数。

（3）“n_estimators”:森林中树的数量。一般更多的树会提高模型的性能，但是会增加计算成本导致训练时间变长。我们平衡了计算成本和模型性能，采用了 10-100 中的 10 个等间距整数，即 10，20，30……；

图 S2 随机森林算法的运行程序

KNN

KNN 算法是一种非常特别的机器学习算法，因为它没有一般意义上的学习过程。当预测一个新样本的类别时，根据它距离最近的 K 个样本点是什么类别来判断该新样本属于哪个类别。

KNN 与 SVR 训练模型的思路与程序基本一致，程序上只需要把训练时的模型调用修改成 KNN 即可，并同样的输出训练集的 MAE 和 R^2。

m3=KNeighborsRegressor(**grid_search_3.best_params_).fit(X_train, y_train)

下面是我们 KNN 的超参数搜索范围：

param_grid_3 = {

'n_neighbors': list(np.linspace(1,10,10, dtype=int)),

'leaf_size': list(np.linspace(1,100,100, dtype=int))

}

上面每个超参数的意思如下：

（1）“n_neighbors”：KNN 中的 K 值，决定要选取几个距离目标点最近的样本点，一般默认为 5，我们这里用网格搜索了 1-10 中的所有整数；

（2）“leaf_size”：这个值控制了使用 kd 树或者球树时， 停止建子树的叶子节点数量的阈值。这个值越小，则生成的 kc 树或者球树就越大，层数越深，建树时间越长，反之，则生成的 kd 树或者球树会小，层数较浅，建树时间较短。默认是 30。这个值一般依赖于样本的数量，随着样本数量的增加，这个值必须要增加，否则不光建树预测的时间长，还容易过拟合。可以通过交叉验证来选择一个适中的值。我们这里用网格搜索了 1-100 中的所有整数。

图 S3 KNN 算法的运行程序

KernelRidge

核岭回归就是为线性回归加上“核”和“岭”，“核”也就是在计算预测值时分别给各个属性值加上权重，但是只用“核”的话会导致程序有过拟合的风险，所以还要引入“岭”来作为惩

罚项，即在之前的基础上加上正则化，可以避免某个单个属性值因被赋予很大的权重而导致误差过大。

KernelRidge 与 SVR 训练模型的思路与程序基本一致，程序上只需要把训练时的模型调用修改成 KernelRidge 即可，并同样的输出训练集的 MAE 和 R^2。

m5 = KernelRidge(**grid_search_5.best_params_).fit(X_train, y_train)

下面是我们 KernelRidge 的超参数搜索范围：

```
param_grid_5 = {
    'alpha': [0.1, 0.5, 1, 2, 5, 10],
    'kernel': ['linear', 'rbf', 'sigmoid'],
    'gamma': [None, 0.1, 0.5, 1, 2],
    'coef0': [0.1, 0.5, 1, 2]
}
```

上面每个超参数的意思如下：

（1）“alpha”：正则化强度，表示岭回归中的正则化程度，越大表示正则化越强；

（2）“kernel”：核函数的类型，有“linear”、“poly”、“rbf”、“sigmoid”，我们选择了三种核函数，分别是“linear”--线性核函数，适用于线性问题；“rdf”--径向基函数核，适用于非线性问题；“sigmoid”--称为双曲正切核函数或 S 型核函数，也适用于非线性问题；

（3）“gamma”:rbf、poly、sigmoid 核函数的系数，决定了一个单一训练样本对决策边界的影响。我们采用了 None（使用默认的选择机制，具体取决于所使用的库），0.1，0.5，1，2。

（4）“coef0”:用于核函数的常数项，主要影响 poly 核和 sigmoid 核的灵活性。我们采用了 0.1，0.5，1，2。

KernelRidge

```python
param_grid_5 = {
  'alpha': [0.1, 0.5, 1, 2, 5, 10],
  'kernel': ['linear', 'rbf', 'sigmoid'],
  'gamma': [None, 0.1, 0.5, 1, 2],
  'coef0': [0.1, 0.5, 1, 2]
  }
krr = KernelRidge()
grid_search_5 = GridSearchCV(estimator = krr, param_grid = param_grid_5, scoring='neg_mean_squared_error', cv = 5, n_jobs = -1)
grid_search_5.fit(X_train, y_train)

m5 = KernelRidge(**grid_search_5.best_params_).fit(X_train, y_train)
m5_test = m5.predict(X_test)
print('Mean Absolute Error:',mean_absolute_error(y_test, m5_test))
print('R^2 score:',r2_score(y_test, m5_test))
```

Python

图 S4 KernelRidge 算法的运行程序

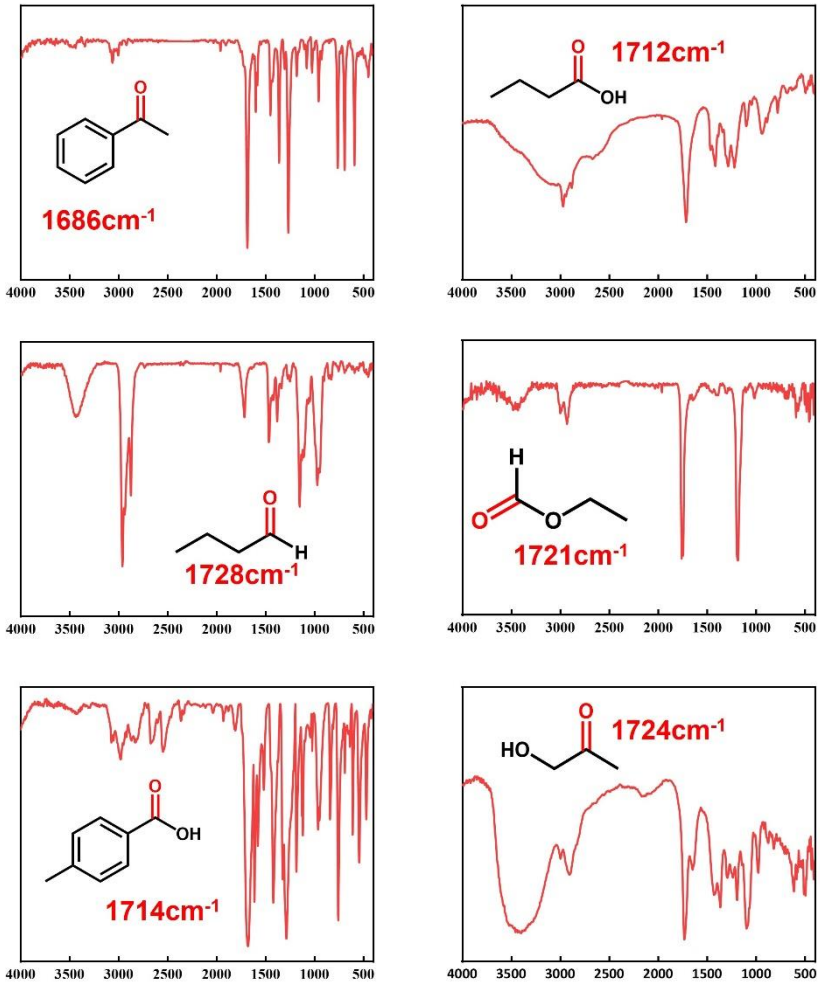


图 S5 丙酮醇、对甲苯甲酸、苯乙酮、正丁酸、正丁醛、甲酸乙酯 6 种物质的红外光谱实测图及羰基峰位移