科学通报
中国天文化学发展专题 评 述
《中国科学》杂志社
SCIENCE CHINA PRESS

机器学习在天体化学中的应用

李广平[1], 王均智[1], 王昭[1,2*]

1. 广西大学物理科学与工程技术学院, 广西相对论天体物理重点实验室, 南宁 530004
2. 广西应用数学中心(广西大学), 南宁 530004
* 联系人, E-mail: zw@gxu.edu.cn

2024-10-30 收稿, 2025-03-21 修回, 2025-03-24 接受, 2025-03-25 网络版发表
国家自然科学基金(12463005)资助

摘要 天体化学作为天文学与化学的交叉学科, 旨在研究宇宙中分子的特性与分布. 近年来, 机器学习技术在天体化学领域的应用取得了显著进展, 尤其在提高光谱分析的精度和效率、识别与表征星际物质中的分子方面表现突出. 深度学习方法能够从复杂的观测数据中提取关键信息, 预测天体物理环境中的化学参数与反应路径, 从而为研究分子在不同星际条件下的形成与演化过程提供重要工具. 本文回顾了机器学习在天体化学研究中的重要应用, 重点介绍了几项创新进展, 包括基于机器学习算法的新方法提出和传统模型的优化. 这些研究深化了对宇宙分子组成的理解, 为探索天体化学中的未解之谜提供了新的视角.

关键词 机器学习, 天体化学, 星际介质, 人工智能

天体化学主要研究星际空间中分子的丰度和演化, 探讨星际分子、尘埃与辐射之间的相互作用[1]. 该学科的一个核心问题是探索分子气体云的形成、演化及其在不同天体环境中的化学反应[2]. 尽管研究涉及从分子尺度到星际尺度的复杂现象, 尤其集中于恒星形成区、行星状星云、星际分子云以及星际物质的化学特性[3,4], 其目标仍是解答一些根本性问题, 如生命的起源、恒星形成过程中的化学演化机制, 以及这些过程如何与星际物质发生相互作用[5,6].

这一学科的发展可追溯至1937年, 当时首次在星际空间中检测到分子的存在. 1940年, McKellar等人[7]通过光谱观测发现了CH和CN的星际吸收线, 这不仅确认了星际分子的存在, 还标志着天体化学的正式起步. 随后, 随着1963年羟基自由基(OH)和1969年甲醛(CH_2O)分子的发现, 科学家逐渐认识到有机分子广泛存在于星际空间中, 进一步证明了宇宙中复杂化学过程的广泛性[8,9]. 1970年, Wilson等人[10]发现了一氧化碳(CO)分子; 同年, Snyder等人[11,12]在星际中探测到了甲酰正离子(HCO^+), 为分子离子的研究奠定了基础. 进入21世纪, 天体化学的研究取得了显著进展, 特别是在复杂有机分子的研究方面. 诸如甲胺(CH_3NH_2)[13]、氨基乙腈(NH_2CH_2CN)[14]、甲酰胺(CH_3NO)[15]、羟基乙腈($HOCH_2CN$)[16]等四种甘氨酸前体分子相继被探测到. 此外, 氰基甲亚胺(HNCHCN)[17]的发现, 作为HCN二聚体的组成部分, 亦被认为是腺嘌呤(一种RNA碱基)的潜在前体. 这些发现促使科学界开始探讨生命的基本构件是否可能源自星际介质, 并推动了对宇宙化学演化的深入研究. 2018年, Suzuki等人[18,19]在一颗大质量恒星的形成区域成功探测到丙烯腈(H_2CCHCN). 这一发现表明, 腈类化合物可能在星际介质中发挥重要作用, 尤其在恒星形成的热核区域, 它们可能是复杂有机分子合成的关键前驱体. 2021年, McGuire等人[20]通过堆叠和匹配滤波技术, 在金牛座分子云(Taurus molecular cloud 1, TMC-1)中识别出氰萘分子的两种异构体, 这些是首次在星际空间中认证的多环芳香烃(polycyclic aromatic hydrocarbons, PAHs)分子.

引用格式: 李广平, 王均智, 王昭. 机器学习在天体化学中的应用. 科学通报
Li G, Wang J, Wang Z. Applications of machine learning in astrochemistry (in Chinese). Chin Sci Bull, doi: 10.1360/TB-2024-1139

这些天体化学领域的进展主要得益于观测技术的进步. 这些技术不仅提高了分子探测的分辨率和灵敏度, 还扩大了研究的时空尺度, 使科学家能够以前所未有的深度和广度来研究星际物质和分子云的化学特性. 西班牙IRAM 30米望远镜通过其强大的毫米波高分辨率光谱观测能力, 为研究银河系内及邻近星系的分子云化学、恒星形成过程, 乃至早期宇宙中最遥远星系的分子发射提供了重要支持[21]. 詹姆斯·韦伯空间望远镜则进一步拓展了我们的视野, 它在红外波段的观测能力极大地促进了分子云和尘埃形成区的研究, 揭示了恒星诞生时伴随的复杂化学反应[22,23]. 与此同时, 地基设施如阿塔卡马大型毫米/亚毫米阵列则通过对冷星际介质中分子发射的观测, 补充了空间望远镜的研究, 为我们提供了前所未有的分子云和原行星盘的高分辨率图像[24,25]. 得益于这些望远镜技术的迅速发展, 人们已经探测到超过300种星际分子[26].

这些技术的进步提高了天体化学的研究能力, 也带来了新的挑战, 特别是在如何处理海量观测数据方面. 与以往的小样本研究不同, 现代天体化学面临着数据爆炸的问题. 每一次观测任务都会产生数太字节计的数据, 这些数据的处理、存储和分析需要全新的方法和工具. 在这种背景下, 天体化学家逐渐将目光转向机器学习等现代计算工具.

机器学习, 特别是在有监督学习中的分类与回归任务, 已成为天文学中的重要工具. 分类任务通过构建模型来学习输入特征与类别标签之间的关系, 通常通过寻找决策边界最大化不同类别的区分度. 机器学习分类算法广泛应用于天文数据分类, 如Dorn等人[27]利用一种支持向量机(support vector machine, SVM)分类算法将恒星分类. 回归任务则用于预测连续变量, 通过优化损失函数拟合输入特征与目标变量之间的关系. 在天文学中, 回归算法被用来定量预测天体化学组成或演化历史, 揭示星际物质的复杂构成[28]. 例如, 通过回归模型可预测恒星形成区内分子的丰度随时间变化, 从而深入理解星际物质的化学动态过程[29].

机器学习在天体化学中的应用始于21世纪10年代左右, 最初被用于研究分子云的性质. 2014年, Makrymallis等人[30]应用贝叶斯推断方法估算暗分子云中的气体密度和宇宙射线电离率, 为推断星际环境参数提供了新思路. 随着研究的深入, Grassi等人[31]通过自动化简化化学网络模型模拟分子云中的化学演化. Zaverkin等人[32]采用一种神经网络提升了计算效率并保证结果一致性. 传统量子化学方法的高计算成本也促使了基于机器学习的低成本计算方法的提出, 如Villadsen等人[33]通过高斯过程回归(Gaussian process regression, GPR)预测星际物种的结合能. 之后, 随着深度学习算法的引入, 进一步拓宽了天体化学数据分析的应用范围, 提升了观测数据与模型的结合能力, 为研究星际介质中的化学过程提供了强有力的支持.

尽管人工智能在天体化学中的应用大大提升了自动化分析的能力, 但其有效性和可靠性仍然存在争议. 其结果往往依赖于训练数据的质量和算法假设, 因此, 确保算法的稳健性和解释其物理意义仍是挑战. 例如, Meng等人[34]利用随机森林(random forest, RF)算法分析了14124种PAHs分子结构与红外光谱的关系, 推动了PAHs结构-光谱关系的研究. 然而, 尽管RF通过特征重要性指标揭示分子光谱与结构之间的关系, 其结果依然基于不确定的统计关联, 这凸显了机器学习在自动化分析中的局限性. 同样, Lira-Barria等人[35]提出的基于数据驱动的外行星大气化学网络简化方法, 虽然提供了有效的简化工具, 但过度依赖数据驱动方法, 忽视了化学反应的物理机制, 特别是复杂光化学过程. 因此, 如何平衡数据驱动分析与理论建模, 成为未来研究中的关键方向. 在这一方面, 物理信息神经网络(physics-informed neural networks, PINNs)作为一种新兴方法展现了很大的潜力.

总体而言, 天体化学的快速发展与技术进步息息相关. 观测设备的提升使我们能够更精确地描绘星际分子的分布与化学过程, 而机器学习等新兴技术为处理和分析海量数据提供了强大支持. 尽管这些技术为天体化学研究带来了机遇, 也带来了新的挑战. 本文将回顾该领域的一些最新进展, 并展望机器学习技术在星际化学研究中的潜力与挑战, 期望这些技术能够推动对星际分子的进一步理解.

1 机器学习基础

机器学习的主要目标是通过数据处理, 使机器能够自主学习和做出决策, 从而模拟人类的认知过程. 在天文学中, 机器学习与传统方法的区别在于处理任务和决策的方式不同. 传统方法依赖预先设定明确天体物理规则, 而机器学习则通过大量数据自主学习这些规则. 通过模式识别和性能优化, 机器学习能够提供更加灵活、自适应的预测方法.

根据数据的标记情况, 机器学习可分为监督学习

和无监督学习. 监督学习利用带标签的数据集进行训练, 识别特定模式, 常用于恒星分类等任务; 无监督学习则处理无标签的数据, 帮助揭示潜在的结构规律, 通常用于观测数据的预处理.

各种机器学习算法为天文学中的不同任务提供了基础工具. 在无监督学习中, K-均值聚类(K-means)是一种常见的算法, 通过分析数据点之间的相似性将数据划分为多个簇, 有助于模式识别和探索性数据分析. 在监督学习中, 极度梯度回升(extreme gradient boosting, XGBoost)是一种高效的分类和回归算法, 因其强大的集成学习能力和鲁棒性而广受欢迎[36]. Shang等人[37]采用XGBoost算法, 针对郭守敬望远镜第8次数据发布中的一类和二类化学异常星(chemical peculiars, CP1和CP2)进行了深入研究. 通过这一算法, 他们成功地识别出6917颗CP1星和1652颗CP2星. GPR是一种用于回归任务的概率非参数方法, 它能够提供关于预测结果的不确定性和方差信息[38]. 决策树是一种简单且易解释的监督学习方法, 通过层次结构对数据进行决策; 而基于决策树的RF算法则通过构建多个多样化的决策树来增强模型的稳定性和准确性, 减少方差并防止过拟合[39]. 此外, RF还具备内置的特征重要性度量功能, 在处理复杂数据集和需要解释数据时尤为有效. SVM和K-近邻(K-nearest neighbors, KNN)算法在天文学研究中也得到了广泛应用. SVM是一种广泛应用的机器学习算法, 旨在通过找到最优超平面以最大限度地分离不同类别的数据[40]. Maravelias等人[41]使用SVM分类器将大质量恒星分类为热星、冷星和发射线星. KNN是一种基于实例的学习算法, 通过分析特征空间中k个最近邻的数据点来进行分类或预测[42]. KNN的非参数性质使其能够适应不同类型的数据分布, 特别适合处理噪声较大或结构较复杂的数据集. Li等人[43]采用KNN算法对多波长天文对象进行了自动分类, 获得了97.73%的高分类准确率. 该方法可以有效地区分活跃天体、恒星和普通星系, 从而为天文学及虚拟天文台面临的自动分类问题提供了有力的解决方案.

神经网络在天文学中的应用同样日益广泛. 神经网络通过模拟生物大脑的结构和信息处理过程, 实现复杂函数的近似计算. 一个典型的神经网络由输入层、隐藏层和输出层组成, 每一层包含多个神经元, 这些神经元通过加权连接相互作用. 神经网络的学习过程包括使用反向传播算法调整权重和偏差, 以最小化模型输出与实际值之间的误差. 根据不同的架构设计, 神经网络可以实现多种模型, 例如多层感知器(multi-layer perceptron, MLP)、卷积神经网络(convolutional neural network, CNN)、循环神经网络(recurrent neural network, RNN)和概率神经网络(probabilistic neural network, PNN). MLP通过多层前馈结构适用于多种广泛的任务; CNN特别擅长处理图像分类任务, 如图1所示, CNN的架构从输入图像开始, 通过一系列卷积层和池化层逐层提取特征. 每个卷积层提取越来越复杂的特征, 而池化层则通过减少空间维度来提高计算效率. 最终的特征图被展平并通过全连接层, 最后在输出层使用softmax函数将图像分类到预定义的类别中.

Škoda等人[44]利用CNN和主动学习技术, 从郭守敬望远镜的410万条光谱中自动识别出具有复杂发射线形状的天体. 通过多次迭代训练, 模型成功发现了948个新的发射线天体候选者. RNN能够处理序列数据, 适用于天文学中的时间序列分析任务. Tsang等人[45]提出了一种周期性光变曲线分类器, 使用RNN进行无监督特征提取, 对变星的光变特征进行分类, 实现了约99%的交叉验证分类准确率, 而且在加入光度特征后, 能够以0.90的精确率、0.96的召回率和0.93的F1分数检测到之前未见过的变星类型. PNN引入了概率特性, 常用于处理天文观测数据中的不确定性. Kheirdastan等人[46]探讨了PNN、SVM和K-means在自动化分类海量恒星光谱中的应用, 使用了来自斯隆数字巡天计划的40万条光谱数据, 研究显示PNN在光谱型分类上较为准确, 分类效果优于K-means和SVM.

随着星际化学研究的逐步深入, 传统的物理模型和经验公式在处理复杂天体物理过程时逐渐暴露出局限性. 近年来, 深度学习技术的快速发展为解决这些问题提供了新的视角和方法, 特别是将物理知识嵌入神经网络模型的创新技术, 如PINNs. 这种方法通过在神经网络训练过程中融入已知的物理规律, 有效地提高了模型的精度和泛化能力, 尤其适用于多尺度、非线性和高维度问题. 如Liang等人[47]提出了一种新的PINNs方法, 通过将PINNs与傅里叶变换相结合, 在频域中解决偏微分方程, 从而减轻了神经网络在模拟多频率函数时的频谱偏差问题. 该方法被用于结构动力学中的正向模拟和参数反演识别问题, 特别是在移动载荷下的结构响应. He等人[48]提出了一种多层级PINNs框架, 通过结合多个神经网络来解决结构力学中的高阶偏微分方程, 显著提高了计算的准确性和效率, 为数字孪生系统的智能计算提供了新范式.

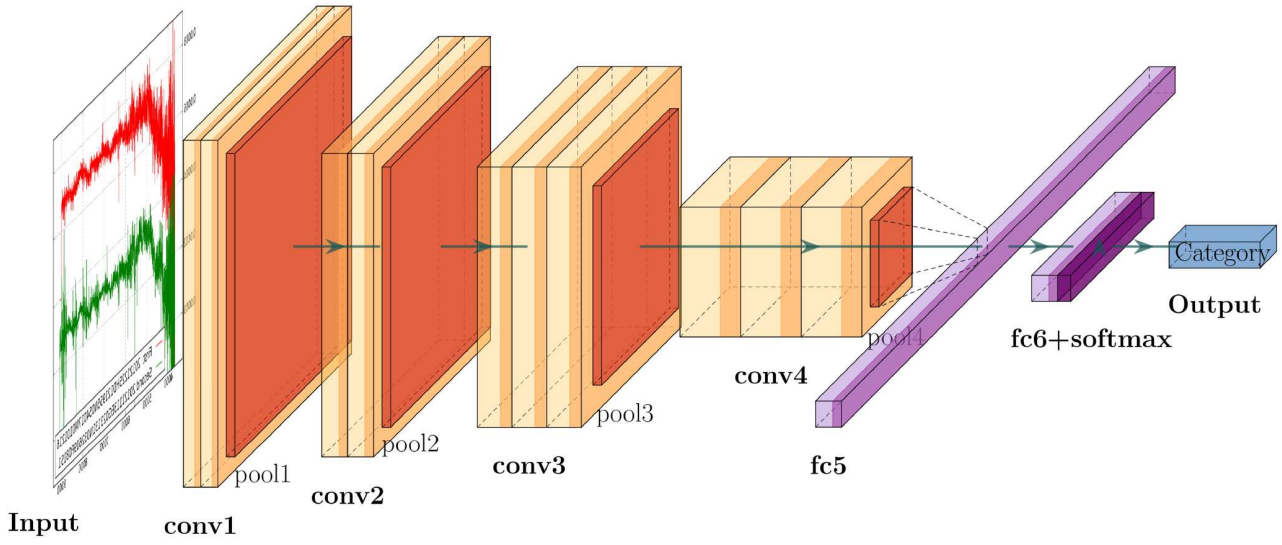

图 1 (网络版彩色)CNN的模型结构图. 输入图像首先通过多个卷积层(convolutional layer, conv), 逐层提取从简单到复杂的特征. 池化层(pooling layer, pool)随后减少空间维度, 提高计算效率. 最终的特征图被展平并通过全连接层(fully connected layer, fc), 最后由softmax函数在输出层(output)将图像分类到预定义的类别中

Figure 1 (Color online) The model architecture of a CNN. The input image first passes through multiple convolutional layers (conv), extracting features from simple to complex in a hierarchical manner. A pooling layer (pool) then reduces the spatial dimensions, improving computational efficiency. The final feature map is flattened and passes through fully connected layers (fc), and ultimately, the softmax function at the output layer classifies the image into predefined categories

在天文领域中, Ni等人[49]提出了结合物理信息的无监督学习方法, 用于天文图像的反卷积. 通过引入望远镜的点扩散函数作为先验知识, 优化了高分辨率图像的处理和重建. 这些研究表明, PINNs不仅在工程和物理学领域取得了重要进展, 也为处理复杂的多尺度问题提供了有效工具. 尽管目前星际化学中的应用还相对较少, 但PINNs在模拟星际介质中的分子光谱、化学反应动力学等复杂系统方面具有巨大潜力, 值得进一步探索.

在天文学的机器学习应用中, 数据来源对于模型的训练和评估至关重要. 数据的质量和预处理过程显著影响模型的性能, 是决定机器学习算法成功与否的关键因素[50]. 高质量的数据意味着观测数据具有高的信噪比、准确的测量值以及完整的覆盖范围, 这些都能有效提升模型的泛化能力. 然而, 天文数据往往包含噪声、缺失值、不均匀分布等问题, 因此需要精心设计的预处理步骤来清洗和准备数据. 例如, 可以通过去除异常点、填补缺失值、标准化特征尺度、选择相关特征等方法来改善数据质量[51]. 此外, 数据增强技术也被用于扩充有限的数据集, 提高模型的鲁棒性和准确性.

2 天体化学中的机器学习

机器学习技术在天体化学中的应用包括以下几个方面.

2.1 星际分子预测

在天体化学领域, 研究星际介质中的化学成分可以为我们提供关于天体源中化学和物理过程的宝贵见解. 随着天体化学中可能存在的分子种类数量呈指数增长, 即使只考虑热力学上最稳定的分子, 发现新星际分子也变得越来越困难. Lee等人[52]结合了化学信息学和机器学习的方法, 试图理解和模拟星际化学成分. 该研究运用了无监督学习算法Mol2Vec, 这是一种基于分子图的嵌入方法, 可以将分子转化为多维向量表示, 捕捉分子结构和化学性质特征的相似性. 通过计算分子在特征向量空间中的距离, 研究人员能够识别与星际介质中已知分子相似的候选分子. 此外, 研究还结合了SVM回归和GPR等监督学习算法, 利用这些回归模型再现已知化学成分的丰度, 并预测尚未观测到的分子的丰度. 研究所使用的数据主要来自观测数据和小分子化学数据库, 这些数据模拟了星际介质中分子的形成和演化过程, 帮助研究人员探索和预测未知分子的

存在及其丰度. 作为验证, 他们研究了TMC-1的化学成分, 这一星际区域已知为化学成分最丰富的空间区域之一. 他们最终搜寻了1510个分子作为TMC-1的候选分子, 并使用GPR模型预测了这些分子的柱密度.

在探索星际化学复杂性的过程中, Scolati等人[53]通过机器学习方法深入分析了猎户座(Orion Kleinmann-Low, Orion KL)星云的化学成分. 与之前在TMC-1上的应用相比, 这项研究扩展到了化学环境更为复杂的区域. Orion KL星云包含多个具有不同结构和化学环境的区域. 这些环境之间的分子柱密度呈现出非线性相关性, 可能导致某些预期物种在不同环境中意外出现或缺失. 他们成功应用了类似于Lee等人[52]使用的回归模型, 基于观测数据, 准确再现了通过谱线拟合程序得到的柱密度数据. 这项工作不仅展示了机器学习在天体化学中的应用潜力, 还为理解复杂天体源的化学特性提供了新的视角. 图2展示了该团队

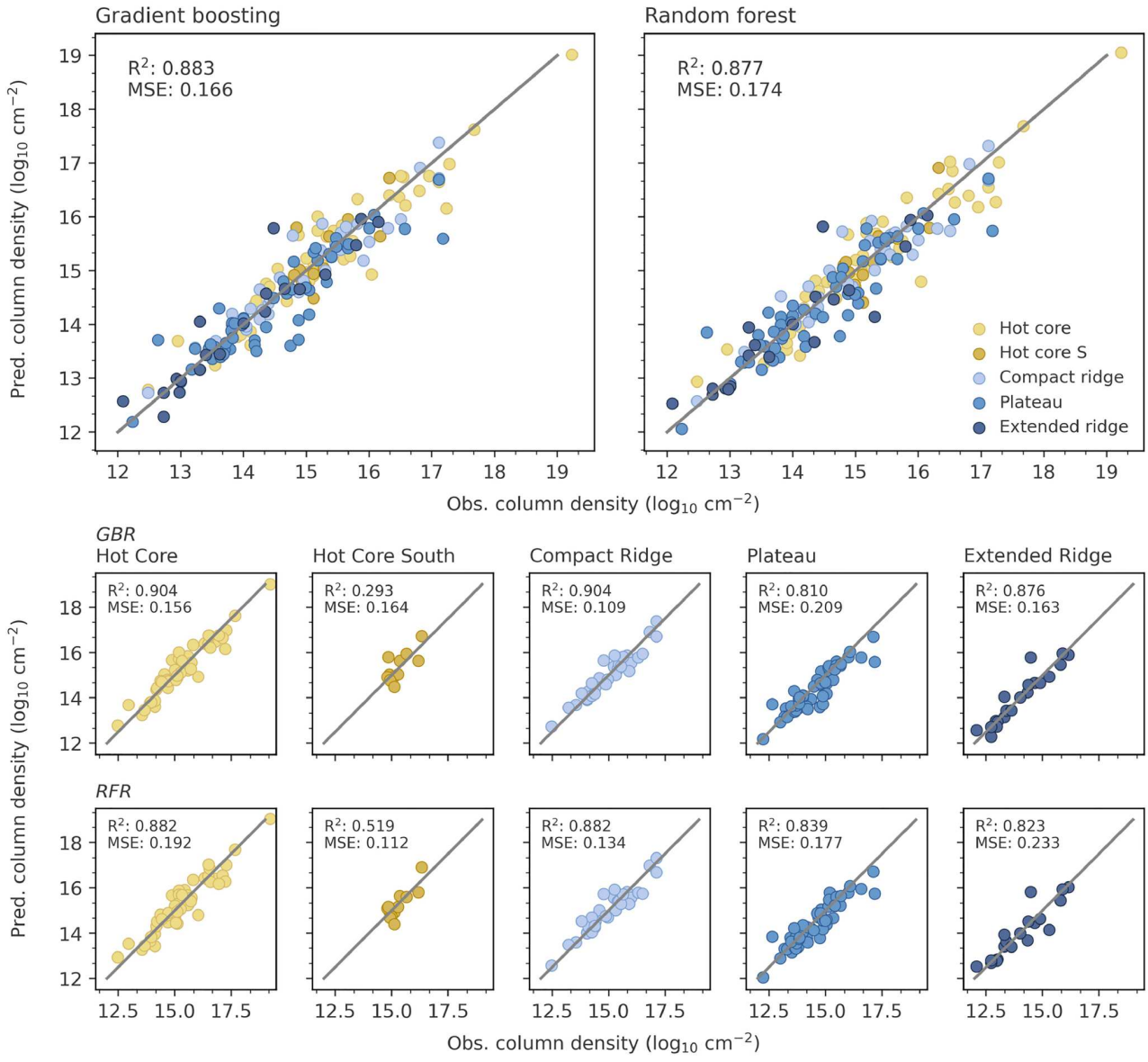

图 2 (网络版彩色)图顶部展示了在Orion KL源内不同环境中观测到的柱密度与各模型预测值的对比, 底部的图展示了各个独立区域的结果, 使用相同的斜率为1的直线来指示模型性能的线性度[53]

Figure 2 (Color online) The top of the figure shows a comparison between observed column densities and different model predictions in various environments within the Orion KL source. The bottom plot presents results for individual regions, using the same slope-1 line to indicate the linearity of the model's performance[53]

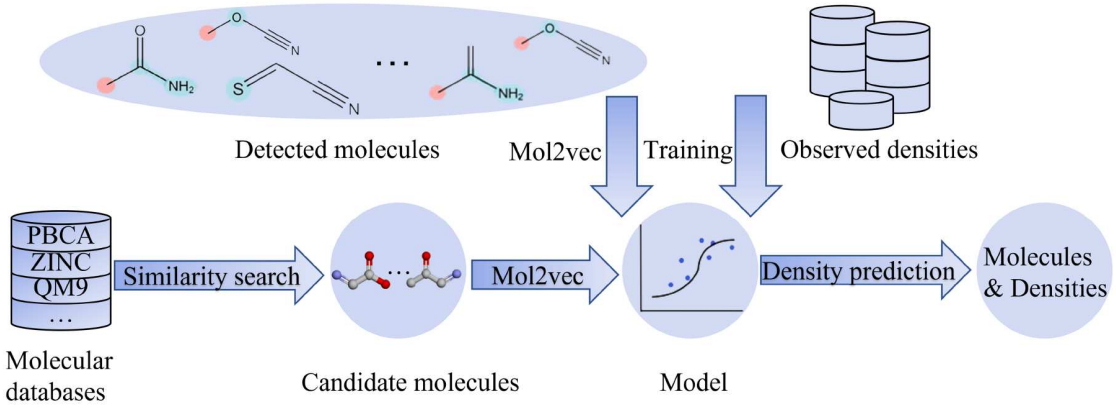

图 3 (网络版彩色)IRC+10216星周包层分子预测流程图. 在候选分子搜寻阶段, 收集已探测的分子结构, 构建包含超过200万个分子的数据库, 并通过相似性搜索识别出与已探测分子相似的候选分子. 在柱密度预测阶段, 使用预训练的Mol2vec模型将分子结构转换为机器学习可读的向量, 结合已知丰度信息训练回归模型, 最终预测候选分子的柱密度

Figure 3 (Color online) Flowchart of molecular predictions in the circumstellar envelope of IRC+10216. In the candidate molecule search phase, the structures of detected molecules were collected to construct a database containing over 2 million molecules and similarity searches were performed to identify candidate molecules similar to the detected ones. In the column density prediction phase, a pre-trained Mol2vec model was used to convert the molecular structures into machine-readable vectors, and a regression model was trained by combining these vectors with known abundance information, ultimately predicting the column densities of the candidate molecules

对Orion KL星云环境中分子柱密度预测的表现, 评估了分子柱密度与其环境之间的关联性. 这种关联性表明, 不同的化学环境对分子柱密度产生显著的影响, 可能与局部物理条件、化学反应途径的差异有关. 与传统回归模型相比, 机器学习方法能够有效捕捉这些复杂的关系, 为天体化学研究提供了更深刻的视角. 评估这些关联性对于理解复杂天体源的化学特性至关重要, 有助于进一步优化天体化学模型.

在一项正在进行的研究中, 我们运用化学信息学手段在小分子化学数据库中对狮子座CW星(IRC+10216)星周包层中的分子进行了系统性探索, 旨在发现与已知物种结构相近的新候选分子. 在此基础上, 我们借助机器学习算法, 基于已探测到的分子数据, 进一步预测了这些候选分子的相对丰度, 为理解星际介质中的化学动力学提供了数据支持. 为了深入剖析候选分子的物理化学特性, 采用了量子化学计算方法, 系统地预测了它们的几何构型、能量、偶极矩、零点振动能、转动常数以及红外光谱特征. 这些计算结果不仅丰富了对富碳星环境的化学成分的认知, 还为后续的天体化学观测和实验研究提供了基础. 图3展示了这项研究的流程.

2.2 分子云研究中的机器学习应用进展

分子云是星际介质中的关键组成部分, 对恒星和行星的形成具有重要影响. 随着天文观测数据的增加, 传统数据分析方法面临巨大挑战. 机器学习为解决这一问题提供了有效的工具, 通过深度学习、SVM等技术, 能够更精确地分析和识别分子云中的化学成分. 近年来, 机器学习在预测分子云的物理与化学状态、揭示分子组成以及研究其动态演化等方面取得了显著进展. 该领域的不断创新使机器学习成为分子云研究中的一项重要技术.

Luo等人[54]引入了一个用于验证分子团块的半监督CNN模型. 该模型在包含三个不同密度区域的数据集上进行训练, 在测试集上实现了高准确度和高召回率, 这些结果与手动验证的结果非常接近. Feng等人[55]提出了一种基于高斯分解和图论的方法, 用于识别分子云结构并解决过度链接问题. 他们的方法有效地区分了密集区域中的气体结构, 保留了大部分流量, 而不需要进行全局数据剪辑或对结构几何形状做出假设, 并且能够适应复杂线轮廓的多个高斯分量.

Xu等人[56]采用不同磁场强度和大尺度动力学的磁流体动力学模拟数据, 使用去噪扩散概率模型从投影质量表面密度图中推断巨分子云的体积密度或数量密度, 并将其结果与传统幂律拟合方法和CASI-2D (convolutional approach to structure identification-2D)神经网络进行了比较, 结果显示去噪扩散概率模型在预测准确性上具有显著优势. 基于这一工作, 该团队进

一步探讨了去噪扩散概率模型在星际化学领域的潜力, 强调了该模型在数据转换中的维度保持、信息丢失的减少以及可追溯性和可解释性方面的独特优势[56]. 此外, Xu等人[57]结合磁流体动力学模拟和机器学习技术(包括CASI-2D和扩散模型), 系统研究了碰撞诱导磁重连机制所形成的分子云细丝. 在测试数据集上, 他们的方法展示了超过80%的检测率, 并在赫歇尔尘埃观测数据中成功识别出高置信度的磁重连细丝候选体. 与此同时, 他们通过深度学习模型CASI-3D, 全面分析了Ophiuchus、Taurus、Perseus和Orion分子云中的原恒星外流, 揭示了外流质量与年轻恒星物体数量之间显著的线性关系, 并估算这些外流所注入的能量足以抵消湍流耗散[58]. 最新研究中, 团队利用去噪扩散概率模型对分子云的星际辐射场强度进行了预测. 他们基于气态环境中的恒星形成项目的磁流体动力学模拟生成尘埃发射图, 并训练模型以估计辐射场强度. 结果显示, 预测值与真实值之间的差异小于0.1倍, 且模型有效地将相对强度误差约束在两倍以内[59].

2.3 优化天体化学模型

在天体化学建模中, 大型化学网络通常需要大量的计算资源. 然而, 最近的机器学习技术显著提高了计算效率. 例如, Grassi等人[31,60]提出了一种方法, 他们通过使用机器学习减少了需要考虑的物种和反应数量, 将化学网络的计算速度提高了65倍.

在探索原行星盘化学成分复杂性方面, Smirnov-Pinchukov等人[61]采用了一种创新的方法, 运用机器学习技术加速化学模型的预测. 他们使用KNN回归算法, 通过一个小子集的物理条件, 快速预测其他盘模型的化学组成. 图4展示了该机器学习模型对CO分子预测的性能, 包括预测的相对丰度(图4(a))、预测值与测试集数据之间的偏差(图4(b))、标准差(图4(c)), 以及数据点的相对密度直方图(图4(d)). 这些结果不仅证明了机器学习在天体化学建模中的潜力, 还展示了如何以毫秒级速度实现对化学成分的快速预测, 这对从观测数据中提取盘物理参数具有重要意义. 这项研究不仅展示了机器学习在准确复现化学成分方面的能力, 还讨论了这种方法的不确定性和局限性, 为原行星盘化学建模领域带来了新的视角, 并为未来的天文观测和数据分析提供了强有力的工具.

在构建天体化学动态模型时, 化学模型不仅对理解微观物理过程至关重要, 还为观测提供了可识别的

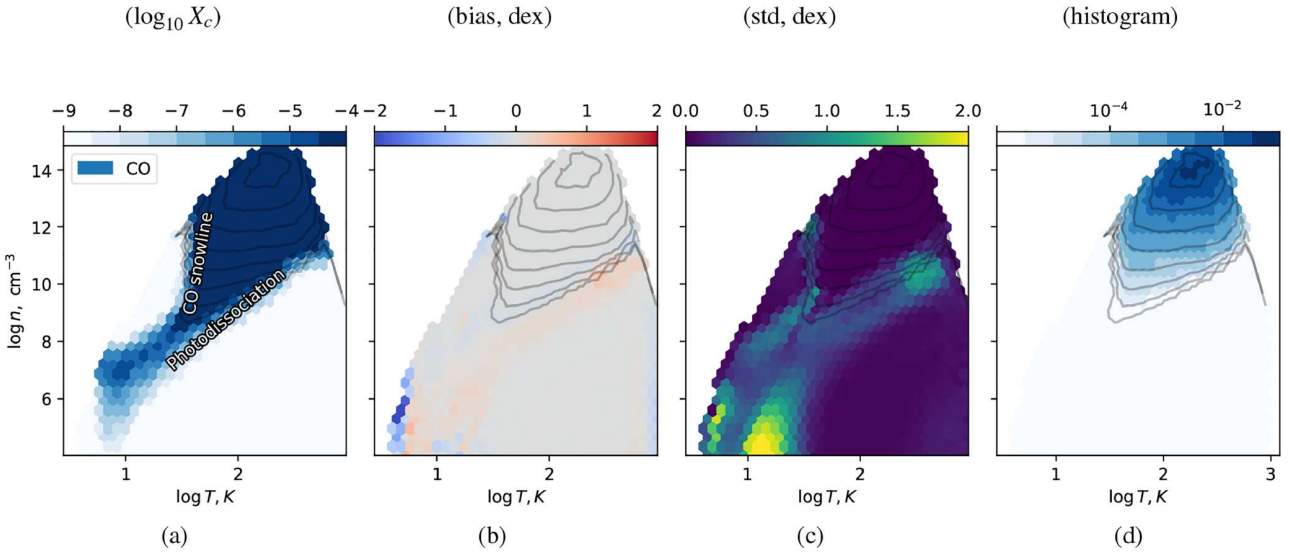

图 4 (网络版彩色)ML加速化学预测CO的性能[61]. (a) 预测的相对丰度(以10为底的对数), 作为局部温度、气体密度和电离率的函数. 颜色较深的区域表示相对丰度(相对于氢原子)较高. (b) 预测值与测试集数据之间的差异中位数, 作为温度和密度的函数, 灰色区域表示无偏拟合. (c) 预测值与测试集数据之间的差异标准偏差. (d) 相对密度直方图

Figure 4 (Color online) ML-accelerated chemical prediction performance for CO[61]. (a) Predicted relative abundance (log base 10) as a function of local temperature, gas density, and ionization rate. Darker regions indicate higher relative abundance (relative to hydrogen atoms). (b) The median difference between predictions and test set data as a function of temperature and density, with gray areas representing unbiased fits. (c) The standard deviation of the difference between predictions and test set data. (d) Histogram of the relative density of species in the data points

特征. 然而, 将这些模型直接整合到流体动力学模拟中, 由于计算成本过高, 常常难以实现. Holdship等人[62]提出了一种基于神经网络的化学演化模拟器(Chemulator), 用于预测气体在恒星形成区中的温度和化学成分随时间的变化. 该模拟器采用了简化模型, 包括一个时间依赖的气体-颗粒化学反应方程组和一个单点模型, 以减少计算成本并提高效率. 同时, 为了克服非局部效应的影响, 模拟器还使用了线性化近似法来处理分子冷却和加热过程, 并通过一个自编码器将输入参数转换为低维空间, 以便更好地采样物理和化学参数空间.

Chemulator相对于传统的化学演化模拟器具有更高的效率和准确性. 它利用了简化的模型和自编码器技术, 显著提升了模拟器在计算气体的温度和化学成分变化方面的速度和精度. 自编码器是一种深度学习方法, 通过将输入数据压缩到一个低维的潜在空间, 再通过解码器将其重建, 从而实现对高维数据的高效表示. 在化学反应建模中, 自编码器能够捕捉化学过程中的非线性关系, 将复杂的化学反应和物理过程转化为便于模拟和分析的低维表示. 这种方法在提高计算效率的同时, 避免了传统方法中所面临的高计算成本和不均匀采样问题. 此外, 自编码器技术还能够更好地采样物理和化学参数空间, 使得模拟器在多次运行中实现更大的时间步长, 从而进一步提高计算效率. 通过这一创新方法, Chemulator能够在较短时间内准确计算气体的温度和化学成分变化, 提供对恒星形成区气体演化过程的深入理解, 并为天文观测提供更加可靠的数据支持.

Molpeceres等人[63]探讨了在无定形固体水表面上P+H → PH反应的动态过程和能量分配情况, 旨在揭示磷氢化物生成时化学能的微观散布. 该研究利用基于神经网络的分子间势, 模拟该反应在不同结合位点上的迁移和能量消散过程. 结果显示, 生成的PH分子在反应初期经历了快速迁移, 趋向于高结合能位点, 其平动能量占反应总能量的1%~5%. 这意味着PH分子在表面上具有一定的非热扩散和解吸附倾向.

图5展示了P + H → PH反应初期的动力学过程, 具体来说, 它展示了大约50 fs内的动态变化[63]. 这项研究揭示了P + H → PH反应中非热扩散和解吸的重要机制, 这些机制有助于解释为什么在冷分子云中检测到的磷化氢和其他磷化物的丰度较低. 非热过程导致PH分子在冰晶表面快速迁移并与其他分子发生反应, 从而减少了其在冷星际介质中的积累. 这些发现为星际化学

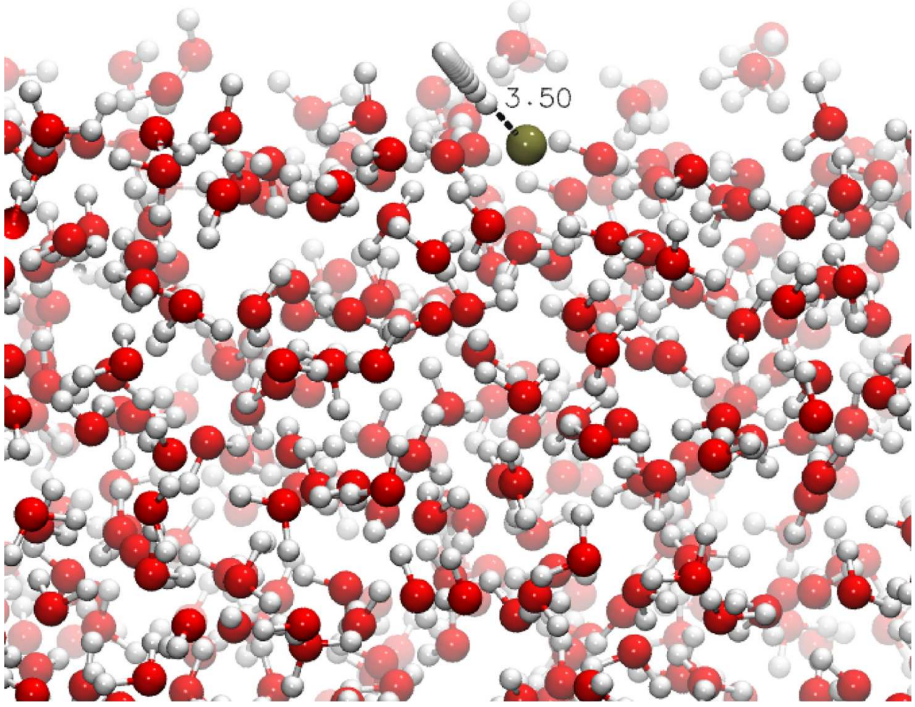

图 5 (网络版彩色)反应的起始阶段[63]. 图中显示了总动态过程的约50 fs(总计50 ps). 3.5 Å为P–H的初始距离

Figure 5 (Color online) Onset for the title reaction in our simulations[63]. The figure represents around 50 fs of the total dynamics (50 ps). The distance of 3.5 Å represents the P–H initial distance

模型提供了重要的参数，有助于更好地理解磷化物在宇宙中的化学演化.

2.4 反应动力学建模

机器学习有助于理解星际分子的演化，并克服在不同星际环境中模拟化学反应的挑战. 通过从大量量子化学数据中学习，机器学习力场模型能够对复杂星际分子之间的相互作用和反应进行建模. 这为传统的昂贵量子化学方法提供了一种经济高效的替代方案[64~69]. 如Villadsen[33]设计了一种基于GPR算法的机器学习模型, 用于预测天体化学相关分子的结合能. 研究表明, 该模型能够快速且准确地预测分子能量, 从而有效补充实验室数据和量子化学计算数据. Meng等人[66]通过反应分子动力学模拟和量子化学计算，研究了富碳渐近巨星分支恒星周围的星周包层中富勒烯的结构演化及其对红外发射特性的影响，揭示了不同氢浓度下形成的洋葱状碳纳米结构或单层富勒烯，发现随着富勒烯的增长，其红外光谱中会出现新的发射特征, 且双层富勒烯表现出显著的蓝移或减弱效应. Qi等人[70]使用分子动力学模拟来表征有机分子与碳纳米颗粒之间的相互作用，采用自适应原子间反应经验键序势函数模拟键旋转和扭转，揭示了碳纳米颗粒作为选择性催化剂在星际化学演化中的重要作用. 研究发现, 碳纳米颗粒在星际介质中选择性吸附芳香族有机分子, 促进PAHs和类似富勒烯结构的逐层形成.

2.5 天体分子光谱分析

在天体分子研究中, PAHs尤其引人注目, 因为它们在星际介质的演化中起着关键作用. 对这些分子结构在红外范围内的识别研究，得益于广泛检测到的芳香红外带. 然而, 传统的量子化学计算在处理这些分子结构的巨大变化时遇到了困难. 大数据时代的到来为解决这一挑战提供了新的方法和工具. 如今, 包含大量有机化合物光谱的数据库已经建立，收录了超过4000种PAHs光谱[71]. 这些数据库促进了机器学习方法的应用, 使人们对复杂的光谱-结构关系有了更深入的理解[72,73].

PAHs是星际介质中广泛存在的有机分子，理解其光谱对揭示星际介质的化学组成至关重要. 传统的计算方法如密度泛函理论虽然能够准确预测光谱，但计算成本非常高. 为了解决这一问题, Kovacs等人[74]使用美国航天局艾姆斯研究中心PAHs红外光谱数据库中的数据, 开发了一个多层神经网络(neural network, NN)来预测PAHs的红外光谱. 模型不仅利用分子的拓扑结构进行预测，还引入了RF方法来分析不同化学特征在预测中的作用. 研究表明，NN对光谱的预测精度高，而RF虽然预测精度稍低，但可以揭示哪些分子特征对光谱有显著影响.

图6展示了不同模型的预测精度比较，尤其是低频段的光谱部分[74]. 图6的顶部展示了NN预测的EMD分布，大多数EMD值较低，表明NN模型能够很好地匹配数据库中的光谱数据. 中间部分显示了RF模型的EMD分布, 预测精度稍逊于NN, 但仍具有一定的解释力. 图6底部则为数据库中随机分子对之间的EMD分布, 用作参考. 通过对比可以看出，NN模型的预测结果明显优于RF模型，尤其是在准确预测低频光谱特征时表现突出.

同样，Meng等人[34]利用机器学习模型来识别星际PAHs分子中的结构片段，这些片段与红外发射特征相关. 通过构建扩展连通性指纹作为分子结构描述符, 研究者使用RF模型来分析14124种PAHs的光谱数据，从而评估10632个分子片段在特定红外波段中的重要性. 这一方法揭示了PAHs的化学结构与中红外和远红外发射光谱的联系，帮助进一步理解未识别红外发射谱带的可能载体.

图7展示了5个红外发射波段(3.3、7.7、8.6、11.2和12.0 μm)中最重要的分子片段[34]. 这些波段与PAHs的C-H和C-C键的特征振动相关. 图中的每个波段都列出了前五个最重要的分子片段，并依次排列它们的重要性. 该图揭示了某些分子片段对红外发射特征的主导作用. 例如，甲基对3.3 μm波段的贡献特别显著，而C-H和C-C键的不同振动模式主导了其他波段发射. 这一分析有助于理解这些红外谱带的分子起源，并为未来的星际PAHs观测提供了指导.

Gastegger等人[75]开发了一种结合环境依赖的神经网络电荷模型与神经网络势模型的方法，通过机器学习加速从头计算分子动力学模拟，高效预测红外光谱. 他们成功模拟了甲醇、含多达200个原子的烷烃及质子化丙氨酸三肽的红外光谱，结果与理论和实验光谱高度一致. 与此同时, Trujillo等人[76]利用量子化学计算方法生成了958种含磷分子的近似红外光谱数据库，选择了谐振*w*B97X-D/def2-SVPD化学模型，并对250个小分子测试了更复杂的非谐振化学模型. 尽管存在局限性，但他们的计算显著改善了快速近似光谱计算数据的缺陷. 这些改进包括定量强度的提升、在关键指纹

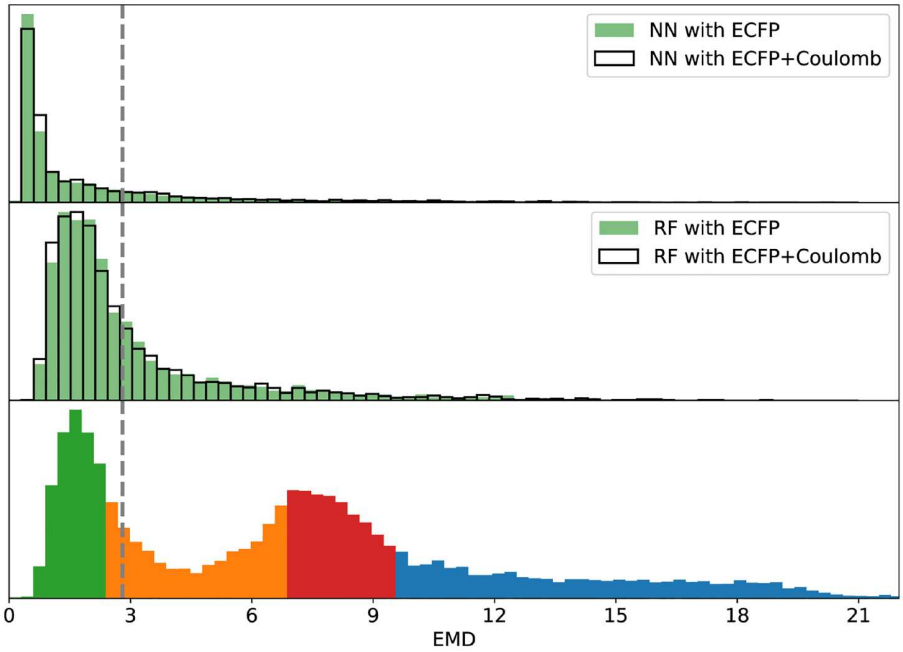

图 6 (网络版彩色)神经网络和RF模型预测红外光谱低频部分的性能比较[74]. 绿色条形表示仅使用拓扑指纹训练的模型结果, 空白条形表示同时使用拓扑指纹和库仑矩阵前10个特征值的模型结果. 图片顶部: 展示了NN模型预测的红外光谱低频部分与数据库光谱数据之间的地球移动距离(Earth mover's distance, EMD)分布. 图片中部: 显示了RF模型预测的红外光谱低频部分与数据库光谱之间的EMD分布. 图片底部: 展示了数据库中随机分子对之间的EMD分布, 并根据25%、50%和75%的四分位数变化用不同颜色进行了编码

Figure 6 (Color online) Comparison of the neural network and random forest models in predicting the low-frequency part of infrared spectra[74]. The green bars represent the results of models trained using only topological fingerprints, while the empty bars indicate the results of models that also utilize the top 10 eigenvalues of the Coulomb matrix and ECFP. The top panel displays the EMD distribution between the predicted low-frequency part of the infrared spectra from the neural network model and the database records. The middle panel shows the EMD distribution between the predicted low-frequency part of the infrared spectra from the random forest model and the database records. The bottom panel illustrates the EMD distribution between randomly selected pairs of molecules from the database, color-coded according to the 25%, 50%, and 75% quartiles

区域和高频区域的光谱覆盖的扩展, 以及根据具体化学环境改进了基频跃迁的准确性, 这些改进有助于识别金星大气中的磷化氢生物标志物及其他气态含磷分子. 此外, Stienstra等人[77]通过迁移学习调整图神经网络变换器模型, 以预测气态离子的红外光谱. 他们使用10336个计算光谱和312个实验光谱进行模型微调, 加入了描述分子电荷状态的编码, 使模型性能提升了21%, 并捕捉到了钠化引起的光谱红移现象. 这种方法可以快速预测未知小分子的红外离子光谱, 并确定其结构. McGill等人[78]开发了一种基于机器学习的化学性质预测软件包, 用于通过机器学习预测分子红外光谱. 他们通过双向信息传递神经网络对分子进行编码, 并结合光谱度量和归一化技术改进模型性能, 同时通过预训练和子模型集成方法提升模型的泛化能力. 他们的研究成果展示了在化学空间中处理复杂峰结构的强大能力, 表明该方法不仅可高效预测光谱, 还能捕捉复杂分子的多样性及光谱特性, 为分子光谱分析提供新的技术支撑.

这些研究通过先进的机器学习方法, 为天文学家远程探测和分析复杂分子提供了高效工具. 高精度红外光谱预测可以识别星际介质和系外行星大气中的分子成分, 如磷化氢等潜在生物标志物. 这些研究的成果不仅加速了对复杂分子体系的解析, 还为探索外星生命和理解遥远天体环境的化学过程提供了数据支持和理论工具, 推动了天文观测与生命科学研究的交叉发展.

3 总结

本文回顾了机器学习在天体化学研究中的应用现状, 涵盖了星际分子预测、化学模型优化、分子光谱分析和反应动力学建模等多个领域. 机器学习有效地预测了星际分子的丰度, 并推动了对复杂星际环境中化学成分的探索, 优化了化学模型的计算效率, 进而加速了恒星形成区气体温度和化学成分变化的模拟. 在

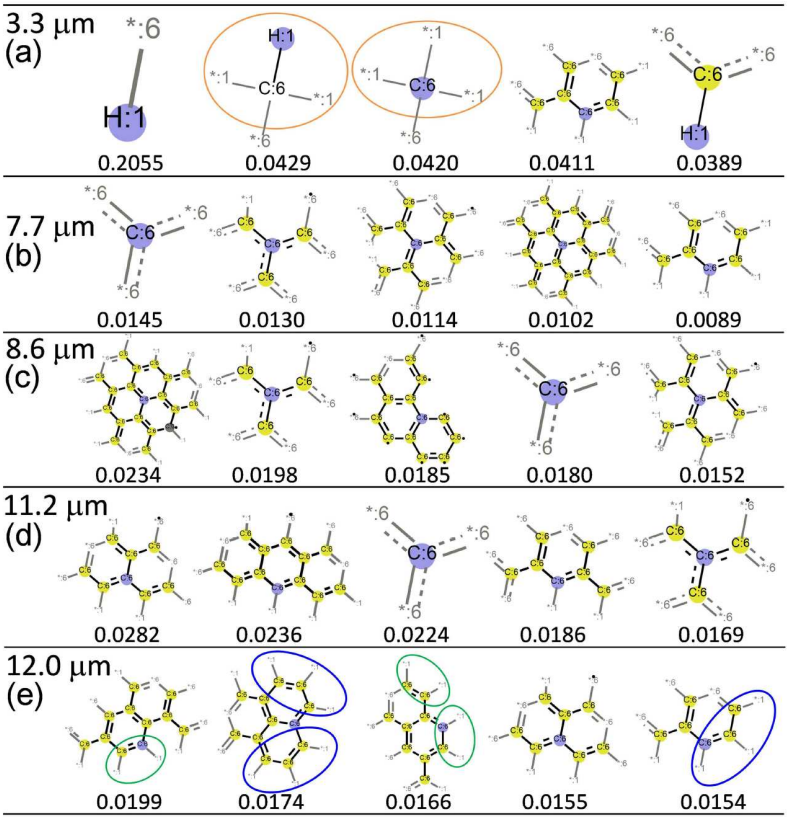

图 7 (网络版彩色)不同红外波段的主要分子片段及其重要性排序[34]. 展示了3.3、7.7、8.6、11.2和12.0 μm红外波段的5个最重要分子片段, 按其特征重要性值从高到低排序. 这些分子片段是通过机器学习分析确定的, 反映了它们在各自波段的贡献程度

Figure 7 (Color online) Main molecular fragments and their importance ranking across different infrared bands[34]. The figure presents the five most influential molecular fragments for the infrared bands at 3.3, 7.7, 8.6, 11.2, and 12.0 μm, arranged in descending order of their feature importance values. These fragments were determined through machine learning analysis, indicating their significant contribution to the respective bands

PAHs的光谱分析中, 机器学习增强了对光谱-结构关系的理解, 此外, 还为星际分子反应的模拟提供了高效的替代方案, 展现了巨大的应用潜力. 总体而言, 天体化学与机器学习的结合为星际分子形成与演化的研究开辟了新路径, 随着机器学习技术的进步, 它将能够处理更复杂的天文数据, 并揭示新的天体化学现象, 从而为理解宇宙中的化学过程带来重要突破.

尽管如此, 机器学习在天体化学中的应用仍面临诸多挑战, 尤其是数据质量问题. 天文数据来源复杂且庞大, 存在噪声、不完整性和多观测尺度等问题, 确保数据的一致性和高质量具有较大难度. 为应对这些挑战, 数据增强技术(如随机旋转、平移变换和噪声添加)可用于扩充样本量并提高模型的泛化能力. 此外, 异常点检测方法(如基于孤立森林、局部异常因子和自动编码器的技术)能够有效识别和剔除观测数据中的异常值, 从而提升数据可靠性. 另一方面, 模型复杂性带来的高计算成本和可解释性问题, 尤其在处理维度高度复杂的天体化学数据时尤为突出. 未来的研究应聚焦于开发适应天体化学独特需求的机器学习模型, 特别是增强算法的可解释性. 例如, 集成可视化技术可以深入理解模型在分子识别和特征提取过程中的决策依据. 跨学科合作, 结合天文学、化学与机器学习领域的先进方法, 将进一步推动机器学习在天体化学中的应用, 助力解决复杂的星际化学问题.

参考文献

1 Öberg K I. Photochemistry and astrochemistry: photochemical pathways to interstellar complex organic molecules. Chem Rev, 2016, 116: 9631–9663

2 Sun J, Du F. Chemical evolution during the formation of molecular clouds. Res Astron Astrophys, 2022, 22: 065022
3 Schinnerer E, Leroy A K. Molecular gas and the star-formation process on cloud scales in nearby galaxies. Annu Rev Astron Astrophys, 2024, 62: 369–436
4 Aller L H, Czyzak S J. Chemical compositions of planetary nebulae. Astrophys Space Sci, 1983, 51: 211–247
5 Ehrenfreund P, Irvine W, Becker L, et al. Astrophysical and astrochemical insights into the origin of life. Rep Prog Phys, 2002, 65: 1427–1487
6 van Dishoeck E F, Blake G A. Chemical evolution of star-forming regions. Annu Rev Astron Astrophys, 1998, 36: 317–368
7 McKellar A. Wave lengths of the CH band lines. Publ Astron Soc Pac, 1940, 52: 309–312
8 Barrett A H, Henry J C, Meeks M L, et al. Radio observations of OH in the interstellar medium. Nature, 1963, 200: 829–831
9 Snyder L E, Buhl D, Zuckerman B, et al. Microwave detection of interstellar formaldehyde. Phys Rev Lett, 1969, 22: 679–681
10 Wilson R W, Jefferts K B, Penzias A A. Carbon monoxide in the Orion Nebula. Astrophys J, 1970, 161: L43
11 Buhl D, Snyder L E. Unidentified interstellar microwave line. Nature, 1970, 228: 267–269
12 Wahlgren U, Liu B, Pearson P K, et al. Theoretical support for the assignment of X-ogen to the HCO+ molecular ion. Nat Phys Sci, 1973, 246: 4–5
13 Holtom P D, Bennett C J, Osamura Y, et al. A combined experimental and theoretical study on the formation of the amino acid glycine (NH_2CH_2COOH) and its isomer ($CH_3NHCOOH$) in extraterrestrial ices. Astrophys J, 2005, 626: 940–952
14 Belloche A, Menten K M, Comito C, et al. Detection of amino acetonitrile in Sgr B2(N). Astron Astrophys, 2008, 492: 769–773
15 Ferus M, Laitl V, Knizek A, et al. HNCO-based synthesis of formamide in planetary atmospheres. Astron Astrophys, 2018, 616: A150
16 Jiménez-Serra I, Martín-Pintado J, Rivilla V M, et al. Toward the RNA-world in the interstellar medium—Detection of urea and search of 2-amino-oxazole and simple sugars. Astrobiology, 2020, 20: 1048–1066
17 Rivilla V M, Martín-Pintado J, Jiménez-Serra I, et al. Abundant Z-cyanomethanimine in the interstellar medium: paving the way to the synthesis of adenine. Mon Not R Astron Soc-Lett, 2019, 483: L114–L119
18 Suzuki T, Ohishi M, Saito M, et al. The difference in abundance between N-bearing and O-bearing species in high-mass star-forming regions. Astrophys J Suppl Ser, 2018, 237: 3
19 Liao Q, Wang J, Xie P, et al. Density functional theory calculations on the interstellar formation of biomolecules. Res Astron Astrophys, 2023, 23: 122001
20 McGuire B A, Loomis R A, Burkhardt A M, et al. Detection of two interstellar polycyclic aromatic hydrocarbons via spectral matched filtering. Science, 2021, 371: 1265–1269
21 Jacob W M, Hooghoudt B G, Mezger P G, et al. The IRAM 30-m millimeter radio telescope on Pico Veleta, Spain. Astron Astrophys, 1987, 175: 319–326
22 Gardner J P, Mather J C, Clampin M, et al. The James Webb Space Telescope. Space Sci Rev, 2006, 123: 485–606
23 Pontoppidan K M, Barrientes J, Blome C, et al. The JWST early release observations. ApJL, 2022, 936: L14
24 Brogan C L, Pérez L M, Hunter T R, et al. The 2014 ALMA long baseline campaign: first results from high angular resolution observations toward the HL tau region. Phys Rev Lett, 2015, 808: L3
25 Long F, Pinilla P, Herczeg G J, et al. Gaps and rings in an ALMA survey of disks in the Taurus star-forming region. Astrophys J, 2018, 869: 17–24
26 Endres C P, Schlemmer S, Schilke P, et al. The cologne database for molecular spectroscopy, CDMS, in the virtual atomic and molecular data centre, VAMDC. J Mol Spectr, 2016, 327: 95–104
27 Dorn-Wallenstein T Z, Davenport J R A, Huppenkothen D, et al. Photometric classifications of evolved massive stars: preparing for the era of Webb and Roman with machine learning. Astrophys J, 2021, 913: 32–45
28 Wang R, Luo A L, Chen J J, et al. SPCANet: stellar parameters and chemical abundances network for LAMOST-II medium resolution survey. Astrophys J, 2020, 891: 23–36
29 Liu H, Ji K F, Jin Z Y, et al. Machine learning in solar physics (in Chinese). Sci Sin Phys Mech Astron, 2019, 49: 105–117 [刘辉, 季凯帆, 金振宇. 机器学习在太阳物理中的应用. 中国科学: 物理学 力学 天文学, 2019, 49:105–117]
30 Makrymallis A, Viti S. Understanding the formation and evolution of interstellar ices: a Bayesian approach. Astrophys J, 2014, 794: 45–55
31 Grassi T, Nauman F, Ramsey J P, et al. Reducing the complexity of chemical networks via interpretable autoencoders. Astron Astrophys, 2022, 668: A139
32 Zaverkin V, Kästner J. Gaussian moments as physically inspired molecular descriptors for accurate and scalable machine learning potentials. J Chem Theor Comput, 2020, 16: 5410–5421
33 Villadsen T, Ligterink N F W, Andersen M. Predicting binding energies of astrochemically relevant molecules via machine learning. Astron Astrophys, 2022, 666: A45
34 Meng Z, Zhang Y, Liang E, et al. Machine learning identified molecular fragments responsible for infrared emission features of polycyclic aromatic hydrocarbons. Mon Not R Astron Soc-Lett, 2023, 525: L29–L35
35 Lira-Barria A, Harvey J N, Konings T, et al. DARWEN: data-driven algorithm for reduction of wide exoplanetary networks. Astron Astrophys,

2024, 692: A158
36 Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In: Balaji K, Mohak S, eds. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, 2016. 785–794
37 Shang L H, Luo A L, Wang L, et al. Objective separation between CP1 and CP2 based on feature extraction with machine learning. Astrophys J Suppl Ser, 2022, 259: 63
38 Wang J. An intuitive tutorial to Gaussian process regression. Comput Sci Eng, 2023, 25: 4–11
39 de Ville B. Decision trees. WIREs Comput Stats, 2013, 5: 448–455
40 Huang P S, Damarla T, Hasegawa-Johnson M. Multi-sensory features for personnel detection at border crossings. IEEE Intell Transp Syst Mag, 2011, 236: 1–8
41 Maravelias G, Bonanos A Z, Tramper F, et al. A machine-learning photometric classifier for massive stars in nearby galaxies I. The method. Astron Astrophys, 2022, 666: 26
42 Cover T, Hart P. Nearest neighbor pattern classification. IEEE Intell Transp Syst Mag, 1967, 13: 21–27
43 Li L L, Zhang Y X, Zhao Y H. K-nearest neighbors for automated classification of celestial objects. Astron Astrophys, 2008, 51: 916–922
44 Skoda P, Podsztavek O, Tvrdik P. Active deep learning method for the discovery of objects of interest in large spectroscopic surveys. Astron Astrophys, 2020, 643: 14–24
45 Tsang B T H, Schultz W C. Deep neural network classifier for variable stars with novelty detection capability. ApJL, 2019, 877: L14
46 Kheirdastan S, Bazarghan M. SDSS-DR12 bulk stellar spectral classification: artificial neural networks approach. Astrophys Space Sci, 2016, 361: 304–315
47 Liang R, Liu W, Fu Y, et al. Physics-informed deep learning for structural dynamics under moving load. Int J Mech Sci, 2024, 284: 109766
48 He W, Li J, Kong X, et al. Multi-level physics informed deep learning for solving partial differential equations in computational structural mechanics. Commun Eng, 2024, 151: 3
49 Ni S, Qiu Y S, Chen Y C, et al. A physics-informed neural networks framework for model parameter identification of beam-like structures. Mech Syst Signal Process, 2025, 224: 112189
50 Alshdaifat E, Alshdaifat D, Alsarhan A, et al. The effect of preprocessing techniques, applied to numeric features, on classification algorithms' performance. Data, 2021, 6: 11
51 Maharana K, Mondal S, Nemade B. A review: data pre-processing and data augmentation techniques. Glob Transs Proc, 2022, 3: 91–99
52 Lee K L K, Patterson J, Burkhardt A M, et al. Machine learning of interstellar chemical inventories. ApJL, 2021, 917: L6
53 Scolati H N, Remijan A J, Herbst E, et al. Explaining the chemical inventory of Orion KL through machine learning. Astrophys J, 2023, 959: 108
54 Luo X, Zheng S, Jiang Z, et al. Semi-supervised deep learning for molecular clump verification. Astron Astrophys, 2024, 683: A104
55 Feng H, Chen Z, Jiang Z, et al. ISMGCC: finding gas structures in molecular interstellar medium using Gaussian decomposition and graph theory. Res Astron Astrophys, 2024, 24: 115005
56 Xu D, Tan J C, Hsu C J, et al. Denoising diffusion probabilistic models to predict the density of molecular clouds. Astrophys J, 2023, 950: 146
57 Xu D, Kong S, Kaul A, et al. CMR exploration. II. Filament identification with machine learning. Astrophys J, 2023, 955: 113
58 Yadav R K, Samal M R, Semenko E, et al. A comprehensive study of the young cluster IRAS 05100+3723: properties, surrounding interstellar matter, and associated star formation. Astrophys J, 2022, 926: 16
59 Xu D, Offner S S R, Gutermuth R, et al. Predicting the radiation field of molecular clouds using denoising diffusion probabilistic models. Astrophys J, 2023, 958: 97
60 Ichimura R, Nomura H, Furuya K. Carbon isotope fractionation of complex organic molecules in star-forming cores. Astrophys J, 2024, 970: 55
61 Smirnov-Pinchukov G V, Molyarova T, Semenov D A, et al. Machine learning-accelerated chemistry modeling of protoplanetary disks. Astron Astrophys, 2022, 666: 10
62 Holdship J, Viti1 S, Haworth T J. Chemulator: fast, accurate thermochemistry for dynamical models through emulation. Astron Astrophys, 2021, 653: 15
63 Molpeceres G, Zaverkin V, Furuya K, et al. Reaction dynamics on amorphous solid water surfaces using interatomic machine-learned potentials. Astron Astrophys, 2023, 673: A51
64 Pan L, Carrete J, Wang Z, et al. Machine learning boosted *ab initio* study of the thermal conductivity of Janus PtSTe van der Waals heterostructures. Phys Rev B, 2024, 109: 035417
65 Liao Q, Xie P, Wang Z. Enantiodetermining processes in the synthesis of alanine, serine, and isovaline. Phys Chem Chem Phys, 2023, 25: 28829–28834
66 Meng Z, Wang Z. Evolution of fullerenes in circumstellar envelopes by carbon condensation: insights from reactive molecular dynamics simulations. Mon Not R Astron Soc, 2023, 526: 3335–3341
67 Yang S, Xie P, Liang E, et al. Catalytic role of Hi in the interstellar synthesis of complex organic molecule. Res Astron Astrophys, 2023, 23:

055019
68 Lu S, Meng Z, Xie P, et al. Gas-phase formation of interstellar nucleobases from dehydrogenated formamide and vinyl cyanide. Astron Astrophys, 2021, 656: A84
69 Hanine M, Meng Z, Lu S, et al. Formation of interstellar complex polycyclic aromatic hydrocarbons: insights from molecular dynamics simulations of dehydrogenated benzene. Astrophys J, 2020, 900: 188
70 Qi H, Picaud S, Devel M, et al. Adsorption of organic molecules on onion-like carbons: insights on the formation of interstellar hydrocarbons. Astrophys J, 2018, 867: 133
71 Mattioda A L, Hudgins D M, Boersma C, et al. The NASA Ames PAH IR Spectroscopic Database: the laboratory spectra. Astrophys J Suppl Ser, 2020, 251: 22
72 Calvo F, Simon A, Parneix P, et al. Infrared spectroscopy of chemically diverse carbon clusters: a data-driven approach. J Phys Chem A, 2021, 125: 5509–5518
73 Laurens G, Rabary M, Lam J, et al. Infrared spectra of neutral polycyclic aromatic hydrocarbons based on machine learning potential energy surface and dipole mapping. Theor Chem Acc, 2021, 140: 66
74 Kovács P, Zhu X, Carrete J, et al. Machine-learning prediction of infrared spectra of interstellar polycyclic aromatic hydrocarbons. Astrophys J, 2020, 902: 100
75 Gastegger M, Behler J, Marquetand P. Machine learning molecular dynamics for the simulation of infrared spectra. Chem Sci, 2017, 8: 6924–6935
76 Zapata Trujillo J C, Syme A M, Rowell K N, et al. Computational infrared spectroscopy of 958 phosphorus-bearing molecules. Front Astron Space Sci, 2021, 8: 639068
77 Stienstra C M, van Wieringen T, Hebert L, et al. A machine-learned “chemical intuition” to overcome spectroscopic data scarcity. Front Astron Space Sci, 2024, 21: 256–266
78 McGill C, Forsuelo M, Guan Y, et al. Predicting infrared spectra with message passing neural networks. J Chem Inf Model, 2021, 61: 2594–2609

Summary for “机器学习在天体化学中的应用”

Applications of machine learning in astrochemistry

Guangping Li[1], Junzhi Wang[1] & Zhao Wang[1,2*]

[1] *Guangxi Key Laboratory for Relativity Astrophysics, School of Physical Science and Technology, Guangxi University, Nanning 530004, China*
[2] *Center for Applied Mathematics of Guangxi (Guangxi University), Nanning 530004, China*
* Corresponding author, E-mail: zw@gxu.edu.cn

Astrochemistry, a multidisciplinary field that bridges astronomy and chemistry, seeks to understand the properties and distribution of molecules throughout the universe. Recent advances in machine learning (ML) have played a crucial role in accelerating progress in this field, particularly by enhancing the precision and efficiency of spectroscopic analyses and aiding in the identification and characterization of molecules in interstellar space. In particular, deep learning techniques have proven effective at extracting critical information from complex observational data, allowing for the prediction of chemical parameters and reaction pathways in astrophysical environments. These tools are invaluable for studying the formation and evolution of molecules under varying interstellar conditions.

This progress reports on key applications of ML in astrochemistry, emphasizing several innovative developments, including the introduction of ML algorithms and the improvement of existing models. These advancements have significantly improved our understanding of the molecular composition of the universe and provided fresh perspectives on long-standing questions in astrochemistry. A central focus is on how ML is being applied to address core issues, such as the interactions between interstellar molecules, dust, and radiation, as well as the chemical evolution occurring in star-forming regions and molecular clouds.

The progress of ML in astrochemistry is closely tied to advancements in observational instruments, such as the IRAM 30-meter telescope and the Atacama large millimeter/submillimeter array. These instruments have greatly enhanced the resolution and sensitivity of spectroscopic measurements, leading to the discovery of over 300 interstellar molecules. However, the sheer volume of data generated by modern observations presents challenges in data processing and analysis. ML has become an essential tool in addressing the so-called “data explosion”, particularly through supervised learning methods like classification and regression. These techniques enable astronomers to classify celestial bodies and predict the chemical composition or evolutionary history of interstellar matter, offering valuable insights into the chemical processes involved in star formation.

ML applications in astrochemistry began to gain traction in the 2010s. Initial efforts focused on studying molecular cloud properties, using techniques such as Bayesian inference and neural networks to estimate parameters like gas density and cosmic ray ionization rates. More recent work has employed deep learning algorithms to model chemical evolution and predict molecular abundances in complex interstellar environments. However, the reliability of these models remains a subject of debate, as the quality of training data and the assumptions underlying the algorithms can influence their results. This highlights the need for robust validation methods. A key challenge moving forward will be balancing data-driven approaches with traditional theoretical modeling, ensuring that ML complements rather than replaces conventional methods. Achieving this balance is essential for overcoming the limitations of purely data-driven models and ensuring that they accurately reflect the underlying physical processes.

Despite the promise of ML in astrochemistry, several challenges persist. A major concern is the dependence of ML models on the quality of the training data. Poorly curated or biased datasets can lead to inaccurate or misleading outcomes, particularly when models are applied outside the scope of their training. Additionally, deep learning models are often criticized for their “black-box” nature, which makes it difficult to interpret the underlying physics driving their predictions. This lack of interpretability can hinder scientific progress. As a result, there is growing interest in developing interpretable ML models for astrophysical applications. Furthermore, integrating ML with physical principles, such as through physics-informed neural networks, offers a promising direction for future research. With continued advancements, ML is poised to play a transformative role in enhancing our understanding of the molecular universe.

machine learning, astrochemistry, interstellar medium, artificial intelligence

doi: 10.1360/TB-2024-1139