DOI：10.19789/j.1004-9398.2026.03.005　　**Open Access**
文献引用：彭灵雅，窦小雅，刘乐洁，等．生成式AI与逆向设计：高通量预测新型荧光染料的分子结构[J]．首都师范大学学报（自然科学版），2026，47(3)：61-73. PENG L Y，DOU X Y，LIU L J，et al. Generative AI for inverse design：high-throughput discovery of fluorescent dye molecules [J]. Journal of Capital Normal University（Natural Science Edition），2026，47(3)：61-73.

生成式AI与逆向设计：高通量预测新型荧光染料的分子结构*

彭灵雅**，窦小雅，刘乐洁，江　艳

（陕西师范大学化学化工学院，应用表面与胶体化学教育部重点实验室，
陕西省新概念传感器与分子材料研究院，陕西 西安　710119）

摘要：荧光材料作为一种重要的发光材料，在照明显示、生物成像、化学传感等方面都有着重要应用，传统荧光材料的合成主要依赖于经验导向的“试错法”，使其研发周期大大延长，在一定程度上阻碍了该领域的快速发展。众所周知，近年来人工智能（AI）技术的发展为化学领域带来了新的机遇，将推动荧光分子设计由经验导向数据驱动的转型。那么，基于AI赋能新型荧光染料的设计与开发，可用生成模型和训练策略是什么？常用的高品质数据集是哪些？如何选择分子描述符？逆向设计策略的应用范例有哪些？为了寻找这些问题的答案，本文首先概述了生成式AI与逆向设计的基本原理，从复杂材料研发流程的关键环节——分子设计出发，汇总了目前比较常用的几种分子设计生成模型，并对比优缺点；其次从常用荧光染料数据库和描述符介绍了如何预测荧光染料的关键性质；最后，分别汇总了国内、国外生成式AI在荧光染料设计应用进展，并探讨了可能的发展方向。希望通过聚焦一些可实施的研究路径，为AI成为驱动荧光染料及功能分子设计与发现的可靠引擎提供参考。

关键词：生成式AI；荧光染料；逆向设计；机器学习；分子设计
中图分类号：O64　　**文献标志码**：A

Generative AI for inverse design：high-throughput discovery of fluorescent dye molecules*

PENG Lingya**，DOU Xiaoya，LIU Lejie，JIANG Yan

（Shaanxi Provincial Key Laboratory of New Concept Sensors and Molecular Materials，Key Laboratory of Applied Surface and Colloids Chemistry，Ministry of Education，School of Chemistry and Chemical Engineering，Shaanxi Normal University，Xi’an Shaanxi　710119）

Abstract：Fluorescent materials have been recognized as a series of important luminescent materials，which has important application in display，bioimaging，and chemical sensing. Conventional synthesis of fluorescent materials often relied on the trial-and-error method，suppressing the development of this realm and prolonging the development cycle. Generally，the recent advances in artificial intelligence（AI）technology have brought new opportunities to chemistry，furnishing the transformation of

收稿日期：2025-10-31
*国家自然科学基金项目（22403061）；陕西省科协青年人才计划托举计划项目（20250608）；中央高校基础科研业务费（GK202406028）
**通信作者：lingya.peng@snnu.edu.cn

fluorescent molecular design from experience-based to data-driven approach. In this context, several key questions arise regarding AI-empowered design and development of novel fluorescent dyes. What generative models and training strategies are applicable? What high-quality datasets are commonly used? How should molecular descriptors be selected? What are representative examples of inverse-design strategies? To address these questions, this article first outlines the fundamental principles of generative AI and inverse design. Starting from the key step of complex material development-molecular design, we summarize several commonly used generative models for molecular design and compare their strengths and limitations. Subsequently, we introduce the prediction of key properties of fluorescent dyes based on commonly employed fluorescent dye databases and molecular descriptors. Finally, we review domestic and international progress in applying generative AI to fluorescent dye design and discuss potential future directions. By focusing on actionable research pathways, we aim to provide a reference for establishing AI as a reliable engine driving the design and discovery of fluorescent dyes and functional molecules.

Keywords: generative AI; fluorescent dye; inverse design; machine learning; molecular design

CLC: O64 **DC**: A

0 引 言

有机荧光材料(organic fluorescent materials, OFMs)具有成本低廉、生物相容性高及合成便捷等特点,在生物成像、化学传感、信息显示及固态照明等领域应用广泛[1]。OFMs发展至今,已有一系列代表性的荧光团,如荧光素、罗丹明、香豆素以及萘酰亚胺等,铸成了OFMs领域的基石。针对不同应用场景,设计合理的OFMs分子结构是实验和理论研究的关键。合理设计的前提是理解OFMs的光物理过程:在光激发下,OFMs分子会吸收光能到达激发态;然而,激发态分子往往是不稳定的,会通过能量耗散的方式回到基态。具体来说:激发态的分子会通过无辐射的振动弛豫和内转换,回到第一激发单重态的最低振动能级;随后,第一激发单重态的分子将会通过弛豫的方式,回到基态或者通过系间窜跃到达三重态。对于前一个过程可以通过辐射跃迁即荧光发射回到基态,也可以通过无辐射跃迁回到基态,失去活性。刚性的有机分子通常以辐射跃迁的方式失活,而柔性分子容易以无辐射跃迁的方式失活[2]。分子发生系间窜跃过程时,激发态分子可以通过磷光失活或将能量转移给其他分子。可知荧光辐射、无辐射通道以及系间窜跃过程是相互竞争的。因此,为了获得实际情况所需的性质,理性设计荧光分子是非常重要的。如用于生物成像的荧光染料,为确保成像结果可靠,要求其斯托克斯位移通常较大且光稳定性较高[3-5]。通常OFMs的结构设计原则为[6-12]:(1)设计具有大共轭面的刚性平面结构,原因是共轭结构有助于红移荧光材料的吸收或发射波长,且刚性的分子骨架会限制分子的振动和转动以提高荧光的亮度;(2)在荧光基团中引入给电子基团以增加OFMs材料的亮度,形成供体-受体(donor-acceptor, D-A)或供体-π-受体(donor-π-acceptor, D-π-A)等结构;(3)使用纳米颗粒包封,以限制分子的转动或降低分子间相互作用,提高染料的亮度;(4)引入屏蔽基团,以削弱分子间的相互作用;(5)利用阴离子屏蔽效应,通过引入大空间位阻的阴离子作为屏蔽基团减弱分子间相互作用。因此,针对设计需求,传统荧光染料的开发方法通常是基于大量的实验去筛选合适的操作,这种基于实际结果筛选的方法也称为“试错法”。由于OFMs的潜在构效关系并不清晰,且材料发光过程会受外界环境影响等,导致OFMs研发周期较长、成本高昂等,限制了其有效应用。

OFMs的机理研究通常依赖于荧光光谱的研究和理论计算。常用的计算方法包括密度泛函理论(density function theory, DFT)、半经验方法和多参考方法等。目前应用最广泛的是DFT方法,其可精确预测发光材料的吸收和发射波长,但是预测的准确性仍然有待提高,当采用不同Hartree-Fock(HF)成分的交换相关泛函时,计算结果往往差距较大。另外,计算成本随着分子尺度的增大呈现指数增长,在计算较大体系时受限。如何发展一种高效、

可针对较大分子体系的研究方法是OFMs分子设计的前沿热点问题。

近年来,随着计算机的高速发展以及机器学习(machine learning,ML)技术的不断改进,人工智能(artificial intelligence,AI)技术为荧光材料的设计注入了新的活力[13-14]。如Alán课题组通过融合量子化学、化学信息学、ML、有机合成、分子表征、器件制造和光电子测试等手段(图1),探索了160万种分子的化学空间,使用DFT方法甄选出数千种有应用前景的有机发光分子,所合成分子的量子产率高达22%[13]。不同于传统的计算方法,AI可充分学习大量实验和计算数据,快速预测OFMs的几何和电子结构。主要研究过程为,将分子结构转化为图神经网络、分子指纹或简化分子线性输入规范(simplified molecular input line entry system,SMILES)字符串,建立OFMs数据库以及描述符,采用ML和深度学习(deep learning,DL)等模型预测发光性质。目前,荧光材料的ML模型已实现了吸收发射能量及量子产率的预测[15-18]。除了预测OFMs发光性质,有研究表明基于已知的化学空间,AI模型经分子构效关系训练后,也可设计全新的分子结构[19-20]。这促使了传统分子设计方法将由从"经验驱动"进化为"计算驱动",本文拟概述荧光染料开发过程中,生成式AI与逆向设计的应用,尝试寻找以下关键问题的答案:基于荧光材料设计的生成模型和训练策略是什么?当前常用的高品质数据集是哪些?如

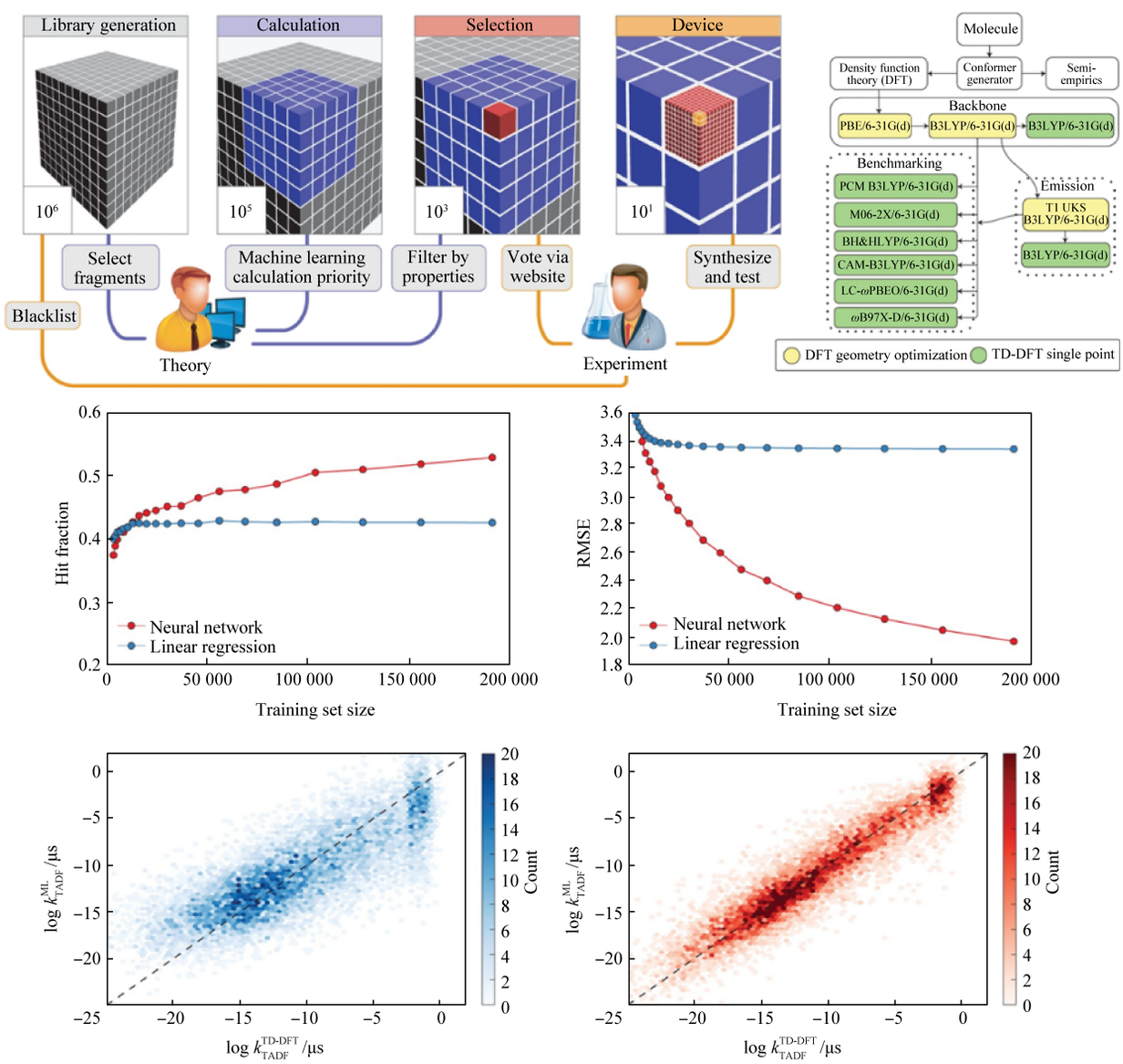


图1　机器学习(ML)在有机发光材料领域的应用[13]

Figure 1　the application of machine learning (ML) in organic luminescent materials[13]

何选择分子描述符？逆向设计策略设计的应用范例有哪些？未来该领域的机遇和挑战是什么？希望在寻找上述问题答案的过程中，为荧光染料的开发提供新的思考和建议。

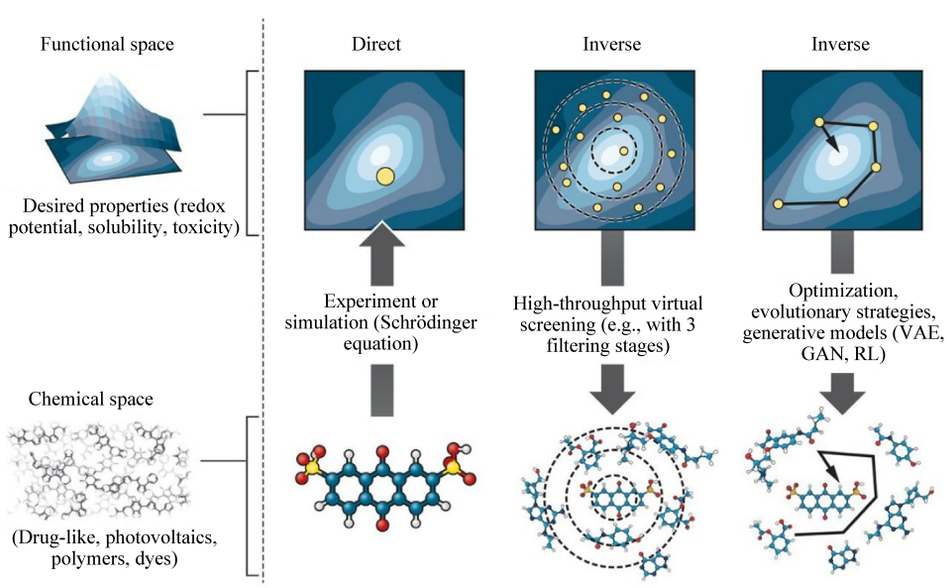
1　逆向设计与生成式AI的基本原理

1.1　逆向设计

分子结构设计是复杂材料研发流程中的关键环节。考虑到荧光体系的应用，OFMs的分子设计是生物学研究和材料科学研究的核心。实现有效设计的关键在于建立和明确结构与功能之间的联系。传统量子化学方法通过设定原子核参数与三维坐标，揭示分子特性，但是往往难以平衡计算速度与准确性，因此仅在部分体系有应用。统计模型的成功，为逆向设计和定制化搜索化合物带来了可能。与传统方法不同，逆向设计以期望功能为出发点，反向设计分子结构(图2)[21]。在此框架下，输入目标功能，输出几何结构。值得注意的是，一种功能未必对应唯一结构，而可能映射至一组概率性的结构分布。逆向设计的实现依赖于分子描述符的构建，以将分子结构转为机器可读的格式。在荧光材料领域，需先明确荧光材料性质，关键的发光性质包括吸收波长、发射波长、荧光量子产率和摩尔吸光系数等。当分子结构、溶剂环境及取代基效应改变时，发光性质也会相应改变。如当设计具有溶剂效应的荧光分子，通常需要设计具有D-A片段结合的分子结构。

典型的逆向设计策略主要有高通量虚拟筛选(high-throughput virtual screening，HTVS)、优化方法(optimization)、进化策略(evolution strategies，ES)、通过序列蒙特卡洛改进贝叶斯采样、构建可逆哈密顿量、推导分子体系性质的分析梯度、优化化学势能面、利用数据挖掘技术发现规律和源于ML的生成模型等。上述典型逆向设计策略的参数对比列于表1。HTVS是其中一个典型范例[22-26]：首先，基于经验获得初始分子数据库，再根据合成可行性和分子稳定性筛选分子，然后由研究人员判断以确定能否合成，最终将构筑单元纳入再迭代优化。虽然HTVS方法看似集合了直接设计方法，但是其核心理念却截然不同。数据驱动发现是HTVS的重中之重，通过与自动化流程、时效性与计算筛选机制相结合，以筛选目标分子，并通过高精度方法验证。因此HTVS实现了理论模拟与实验验证之间的闭环反馈。相反，Optimization具有访问构型少的优势，主要通过功能流形的几何信息，根据总体趋势、方向和曲率进行索引[27]。ES可优化全局，在化学空间映射中有所应用[28]。ES通过自然进化的启发式搜索、多轮迭代，对群体中的参数向量引入扰动，随后评估目标函数值。在参数变量和扰动的反复迭代下，ES可解决复杂的优化问题。

因此，针对不同体系选择适宜的逆向设计策略



注：VAE为变分且编码器；GAN为生成对抗网络；RL为强化学习。

图2　分子设计策略[21]

Figure 2　strategies toward molecular design[21]

十分重要,虽然以上策略在一定程度上可实现分子的筛选和设计,但是常受限于一些实际问题:如HTVS方法虽然可进行大规模的计算和实验,但筛选结果依赖数据库且搜索资源消耗大;Optimization算法在优化过程中容易陷入局部极小点,依赖于空间的连续性和可微性,当分子结构出现较大变化时,往往难以处理;ES方法则容易收敛到局部极小点,且收敛速度较慢。生成模型可在一定程度上解决这些问题,如可探索巨大的化学空间,并基于合理的化学经验筛选,以下简要介绍生成模型的特点及主要算法分类。

1.2　生成模型

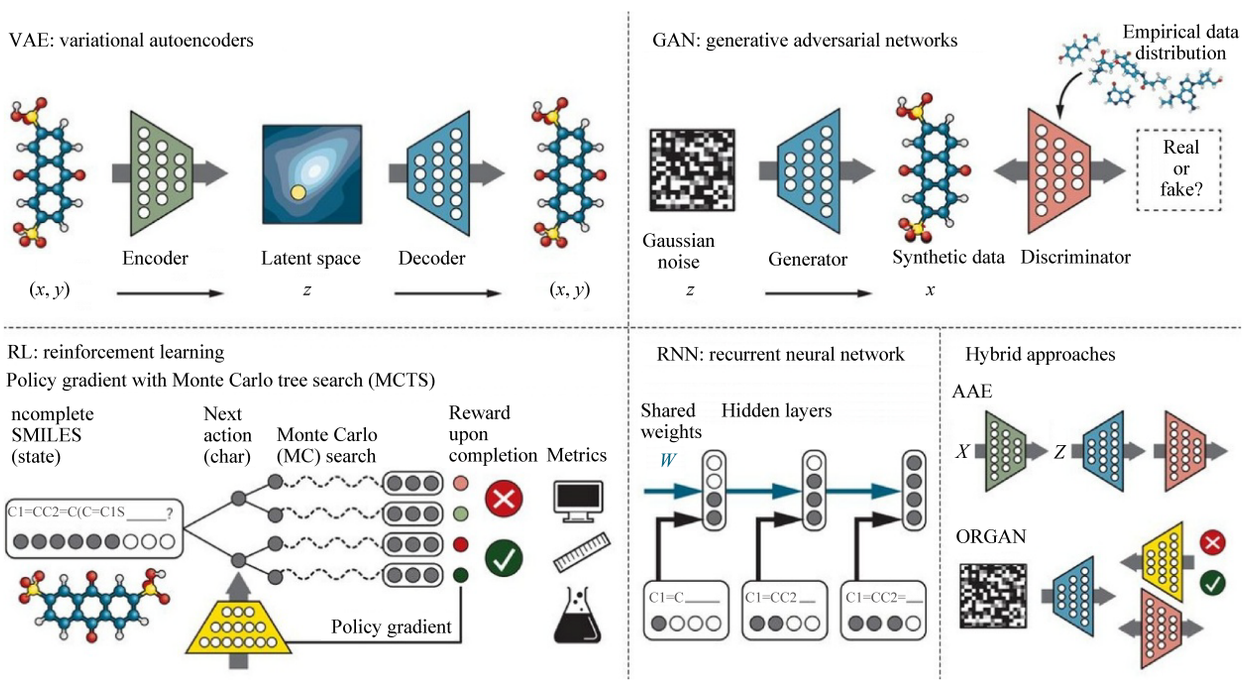
生成模型是逆向设计的核心策略之一,是DL的特殊分支,旨在构建结构与性质的联合概率分布模型,并建立非线性关联[21,29]。深度生成模型比直接的ML方法更困难,但是随着DL算法和计算策略的发展,生成模型已在分子设计领域获得了一系列进展,实现了从数据中学习化学结构规则并映射到潜空间,能在潜空间中通过插值和梯度寻找新分子,最终输出候选结构并完成逆向设计循环(图3)[21]。现有可供研究者使用的生成模型算法主要包括变分自编码器(variational autoencoders,VAE),强化学习(reinforcement learning,RL),生成对抗网络(generative adversarial network,GAN)以及循环神经网络(recurrent neural network,RNN)等。

(1)VAE算法[30-31],在潜空间中引入概率分布(通常为多元高斯分布)。此时,分子在潜空间不再为固定点而是概率分布,这既保证了空间的连续性,又准确反映了邻近区域分子间的结构相似性。在潜空间中,分子被表示为连续可微的向量。以给

表1　经典逆向设计策略的对比

Table 1　comparative of classic inverse design strategies

策略	特点	应用场景	性能对比
高通量虚拟筛选	大规模筛选荧光数据库	用于早期潜在荧光骨架的发现、多目标筛选及作为实验高通量筛选的预筛选或后分析工具	覆盖广度大,依赖于数据库和筛选标准,资源消耗大
优化方法	梯度或启发式算法,在连续或离散空间寻找最优解	用于在已知荧光团上微调、连续参数调优	局部搜索效率高、精度高,易陷入局部最优、易过拟合,依赖目标函数性质
进化策略	模拟自然进化,通过种群迭代进行启发式搜索	用于荧光染料的全局优化和骨架迁移设计,无梯度优化,可以处理"黑箱"问题	具有一定的全局搜索能力,计算成本高,收敛速度慢



注:AAE为对抗自编码器;ORGAN为目标强化生成对抗网络。

图3　生成式模型常见架构[21]

Figure 3　architectures used in generative models[21]

定分子为参照,在邻近区域解码可得到相似结构的分子,当距离增大时会得到差异较大的分子。此外,通过在潜在空间对2个分子的潜在向量进行插值,可生成新的分子。如VAE算法可用于优化发射能量和光致发光量子产率等实验测量的关键光物理性质,通过连续潜空间实现从目标性质到分子结构的梯度优化,以保持有限数据集的多样性和精准度[30]。然而,该算法仍然受限于实验数据集的规模与品质,在复杂光物理性质预测时受限。为解决这一问题,通过数据融合扩充化学空间,以及引入参数确保精准度和生成多样性,最终辅以实验验证确保策略可行性。

(2)RL算法,指智能体可与环境进行试错式交互以学习序列决策策略[32]。在这一过程中,预设的输入-输出样本并不起主要作用,而是由反馈奖励信号起引导作用。RL智能体可以首先探索不同行为,在行为产生目标结果时得到奖励。比如可以主动探索新的分子结构,当预测新的分子结构具有较高的量子产率时获得奖励,故在荧光材料的逆向设计中应用广泛。具体操作为:每次在智能体上添加1个化学片段,当达到目标分子的要求时则获得奖励,通过分步组装的方式获得目标分子。此外,当得到违反稳定性规则的结构时,RL还能设置奖励函数进行惩罚,从而实现多步决策过程的自然处理。因此,RL在实现荧光染料领域目标导向的基础上,可提供主动发现和优化工具,充分搜索化学空间,从而与被动预测模型互补。采用RL算法,基于荧光分子的垂直激发能、振子强度与荧光结构相似性,研究者能在化学子空间中定向搜索,从而筛选有潜力的候选分子。然而,该策略目前受限于对更加复杂的激发态光物理过程的研究,后续可通过整合更丰富的实验数据集、引入激发态动力学描述符等解决局限[32]。

(3)GAN算法[33-35],由生成器与判别器2个相互竞争的神经网络组成。生成器负责生成随机SMILES字符串,判别器则区分这些随机分子与训练集中的目标分子。通过迭代训练,生成器逐步优化生成能力,创造出与目标分子高度相似的分子结构。当判别器无法再区分生成器产生的分子与目标分子时,训练终止。为使生成器能够采样具有特定性质的分子,GAN训练常结合辅助任务——生成器产出具有目标性质的SMILES字符串,同时强化学习在整个训练过程中优化生成器,在生成高有效性、独特性与新颖性的分子方面表现卓越。为实现荧光染料的设计,可采用以下计算驱动流程:以初始分子结构为输入,经编码器映射到潜空间,训练GAN,并通过解码器转回新的分子结构;最后结合高通量的理论计算评估分子性能[35]。需要指出的是,GAN算法并不直接用于染料设计,因此在使用过程中,需要与性质预测模型相结合以实现逆向设计。

(4)RNN算法[36],由循环连接存储以保留用于后续处理的复杂数据,其预测能力大幅增强。当使用SMILES表示化学结构后,RNN可从最初的信号处理和自然语言处理开发到预测分子性质并生成分子结构。主要工作以递归方式处理,流程为:每次迭代中,接受数值向量以及前一步的隐藏状态矩阵,并将得到结果输出作为下一步的输入,直至完成数据处理。上述流程也会带来一些问题,如梯度爆炸或消失问题导致长期依赖关系捕捉失败。可通过长短期记忆(long short-term memory,LSTM)和门控循环单元(gated recurrent units,GRU)选择性记忆输入片段解决。RNN通过符号序列学习条件概率,这一点与马尔科夫链类似。但是,RNN可编码长期依赖关系,因此可用于掌握复杂语法规则。此外,RNN还可用于处理分子图结构,但也会带来输出空间庞大、表示方式不唯一及依赖关系复杂等挑战。在OFMs设计中,RNN本身不直接建模或预测OFMs的具体性质,而是专注于分子结构的生成,最终得到分子SMILES。因此需要将RNN与量子化学计算相结合,并通过强化学习优化,以实现从0开始设计出具有目标荧光性质的新型分子。

随着ML理论的发展,目前的生成式模型已经取得了很多突破,但仍存在一些挑战:如对于分子结构图像同构象问题的算法尚未实用化,如何正确地描述分子结构尚未解决;VAE方法常存在生成分子不够新颖,以及生成的分子结构无效或不太合理,另外训练目标与实际期望的结构存在偏差较大等问题;GAN常存在训练不稳定以及生成分子多样性不足的问题;RL采样效率低且训练不稳定;RNN缺乏显示的逆向设计机制且容易生成无效分子。因此,建议采用多种方法融合的策略可在一定程度上解决这些问题。另外,采用不同方法进行基准测试也非常重要。

2 荧光染料关键性质预测

2.1 荧光染料数据库

高品质的OFMs数据库是逆向设计的基础,决

定了上述生成模型的预测精度与泛化能力,提升了模型的效果和准确性。目前数据库主要来源于已发表的论文和开源平台的实验数据,这些数据为模型的训练提供了充足和必要的实验资源。当前可供研究者使用的OFMs数据库包括Reaxys、ASBase和Deep4Chem等。

Reaxys数据库是由荷兰爱思唯尔公司开发的旨在优化合成路线设计的数据库[37],其包括超过4 300万条反应数据和5亿条物化参数,可生成20步以内的合成路径并提供文献依据。ASBase数据主要集中在荧光材料领域[38],其中有4 535种化合物,涉及聚集诱导发光、聚集猝灭效应、室温磷光及热激活延迟荧光等多种发光机制。该数据库包括了化合物等光物理参数和理化特性,形成多维度数据矩阵。此外,ASBase可提供API接口与计算服务,这显著提升了数据利用率和模型开发效率。Deep4Chem数据库针对光学性质研究需求[39],整合了来自1 358篇权威期刊的7 016种有机发色团在365种环境体系中的光学数据,以标准化(comma-separated values,CSV)格式提供包括SMILES字符串、光谱参数在内的20 236条精准测量记录,为光谱预测模型建立了坚实的数据基准。通过DFT、半经验方法等计算的理论数据库为OFMs的设计可提供支撑。量子化学计算可以获得分子的几何结构、Mulliken电荷分布、前线轨道能级和静电势等几何和电子结构参数[40]。如基于QM7b和QM9等量子化学数据集,研究人员从电子结构层面理解荧光分子的激发态行为与发光机制,从而实现计算化学与数据科学的方法融合[41]。

因此,实验和理论数据库的发展对OFMs的研究起了重要作用[1]。实验数据库为模型验证提供基准数据,理论数据库揭示构效关系以增强模型的可解释性。Reaxys数据库包括了许多分子合成的相关信息,ASBase和Deep4Chem可以检索得到分子的光物理性质。尽管当前数据库已经包括了众多荧光分子的信息,但是高品质数据库依然有限,且分子的数量也十分有限,有关开源数据库的建设仍然是未来需要考虑的问题。相信随着数据库的不断标准化与丰富,OFMs的智能设计会持续发展。

2.2 荧光染料描述符

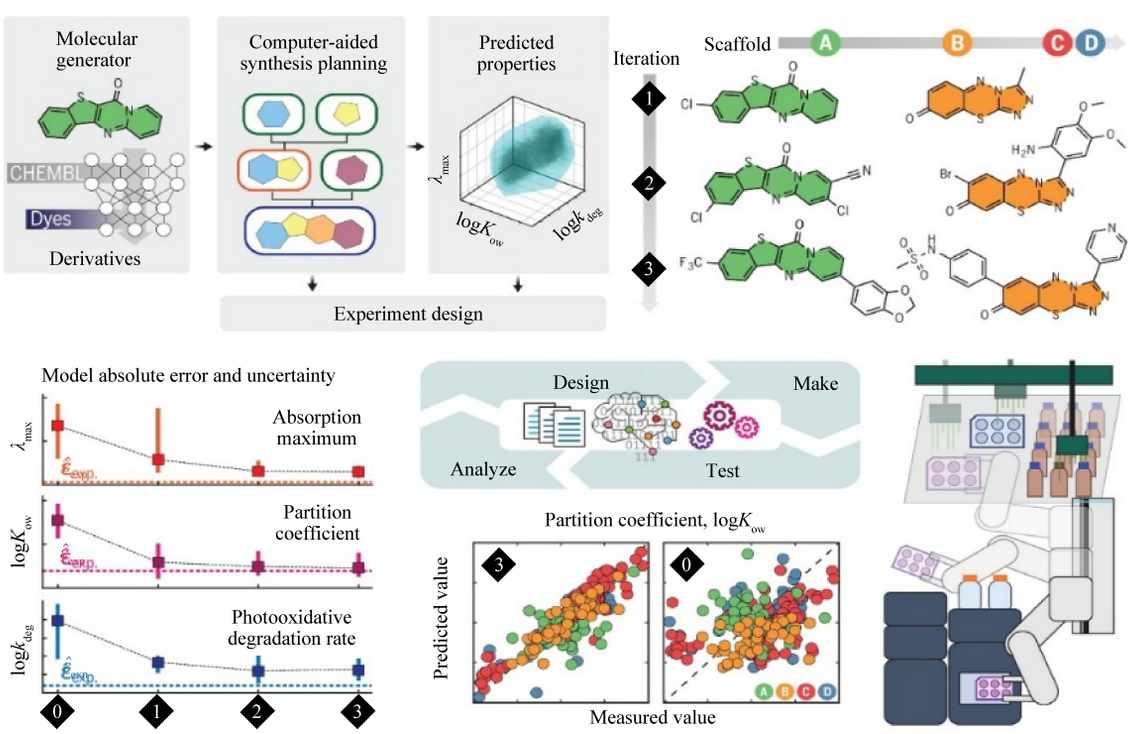
在计算化学与化学信息学中,分子描述符是将复杂的化学信息(分子的电子特性、拓扑结构、几何参数及子结构模式等)转化为可量化数据的数值表示,从而实现分子性质的数学处理。因此,实现定量结构与性质关系模型,需要先构建描述符,既要捕捉分子的物理化学规律,又不能冗余或者具备线性关系,否则可能引入噪声,导致有效性较低。现有可供选择的OFMs描述符包括电子结构信息、拓扑特征、几何结构参数以及结构基元等[27,40-43]。其中几何结构(如分子的键长、键角、二面角和分子表面积等)及结构基元(主要的功能基团)可分别通过测量和结构识别确认,有关电子结构信息可以通过理论计算,获得最高占据分子轨道(highest occupied molecular orbital,HOMO)、最低未占据分子轨道(lowest unoccupied molecular orbital,LUMO)能级,电荷分布、静电势和偶极矩等。另外也可以通过Python脚本生成拓扑描述符,如Wiener指数和连通性指数等。目前:SMILES描述符的应用最为广泛,将其与数值描述符结合,可增强特征表征能力[44];基于片段的表示指纹(finger print,FR)能够通过径向分布函数以增强立体化学区分能力[45];加权原子中心对称函数(atom centered symmetry function,ACSF)可以引入环境依赖性权重以提升可扩展性[46];原子位置平滑重叠(smooth overlap of atomic positions,SOAP)描述符可模拟周期性体系中的电子密度分布,但受到简并性限制[47];结合三维构象动力学和溶剂环境相互作用方法,也为解释构效关系提供了新的解决方案[48]。

为描述荧光染料的光物理过程,选择合适的描述符非常关键,如几何结构、电子结构、拓扑描述符,以及SMILES、指纹、ACSF和SOAP等。电子结构描述符直接与荧光分子激发能和溶剂化效应相关,决定了发射波长与Stokes位移;几何与拓扑描述符反映了分子刚性与共轭程度,直接影响无辐射跃迁通道以及最终的发光量子产率;而三维原子环境描述符(ACSF和SOAP)可细致描述激发态结构弛豫与局部相互作用,为预测光谱的位移及环境提供关键信息。尽管分子指纹与SMILES字符串属于隐式编码,但是通过数据驱动学习能够识别与荧光性能强相关的化学结构式。因此,以上描述符系统覆盖了荧光产生与调控的关键机制,为构建荧光性质预测模型,以及建立生成模型奠定了基础。但是,选择适当的描述符进行模型训练依然面临多重共线性和维度问题的持续挑战,需进一步检测验证。

3　生成式AI在荧光染料设计中的应用

3.1　国际研究进展

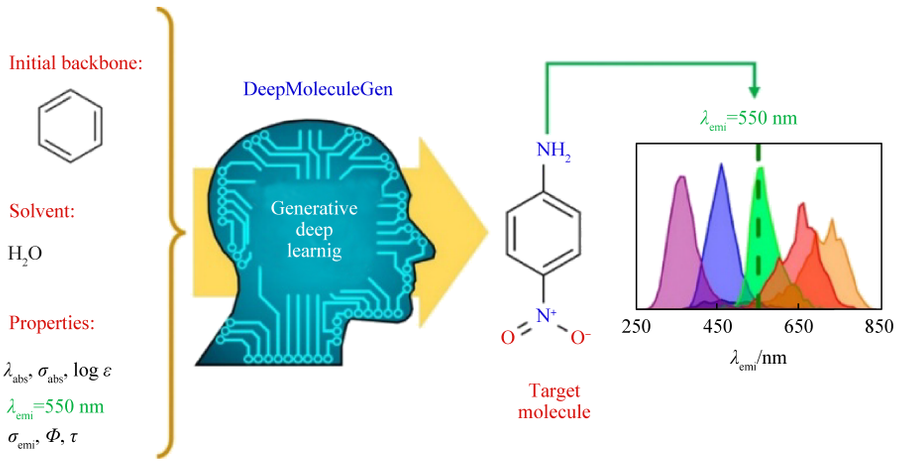
麻省理工学院Jensen课题组报道了一个无需人工实验即可探索化学空间并优化已知化学结构的design-make-test-analyze(DMTA)循环(图4)[49],该平台提出并合成表征了303种未见文献报道的类染料分子,这些分子分布在4个探索性骨架和1个优化性骨架中,可用于搜索具有目标性质的结构来探索未知的结构-性质空间,也能利用已表征结构优化有前景候选物的性质;实现了最大吸收波长、辛醇-水分配系数和光氧化降解速率共3个目标性质的自动表征,并优化模型预测,指导后续实验选择,该闭环集成平台的进一步发展有望显著加速分子设计与制备进程。韩国大学Park课题组构建了包含71 424个分子和溶剂对实验数据库的生成式DL模型[50],可生成具有特定光学性质的分子(图5),如具有特定的电子吸收和发射峰位置与带宽、消光系数、光致发光量子产率以及光致发光寿命等;该模型在设计具有目标光学性质的分子时,能有效运用溶剂效应、共轭效应等关键设计原理,有望用于有机光伏、有机电致发光二极管、有机光电二极管和生物成像染料等领域材料的制备。此外,该课题组还通过结合DL模型与实验数据库[51],实现了对有机化合物吸收峰位置与带宽、消光系数、发射峰位置与带宽、光致发光量子产率以及发射寿命的预测;通过引入发色团-溶剂相互作用机制,训练了30 094种发色团与溶剂的组合,以反映环境对发光的影响,训练的模型可以快速预测荧光分子在不同状态(溶液、气相、薄膜和粉末态)的性质,实现了蓝光材料的设计和开发。日本RIKEN高级智能项目中心的Tsuda课题组通过结合量子化学计算与从头分子生成器(de novo molecule generator, DNMG)[36],生成了3 643种候选分子,选取了1种新分子及7种已知分子验证合成,表明该生成系统设计的荧光分子准确率达到75%,并成功合成了1种备受跨学科领域关注的具有明亮荧光的新分子。芝加哥大学Ju课题组以实验导向的逆向设计原则[30],提出以自适应β-VAE耦合潜向量预测模型的策略,在潜空间中梯度优化荧光发射能量与量子产率,动态调控Kullback-Leibler(KL)散度因子,得到了稳定性高且多样性丰富的分子,并实验合成验证。哈佛大学Alán课题组报道了一种集成的有机功能材料设计流程[13],融合了理论洞察、量子化学、化学信息学、



注:$\hat{\varepsilon}_{exp}$为实验误差;λ_{max}为最大吸收波长;$\log K_{ow}$为辛醇-水分配系数;$\log k_{deg}$为光氧化降解速率。

图4　通过迭代实验探索分子染料骨架所决定的结构-性质关系[49]

Figure 4　iterative experiments to explore the structure-property space defined by molecular dye scaffolds[49]



注:λ_{abs}为吸收波长;σ_{abs}为吸收带宽;ε为消光系数;λ_{emi}为发射波长;σ_{emi}为发射带宽;Φ为量子产率;τ为发光寿命。

图5　基于生成深度学习的有机分子定制化高效设计[50]

Figure 5　generative deep learning-based efficient design of organic molecules[50]

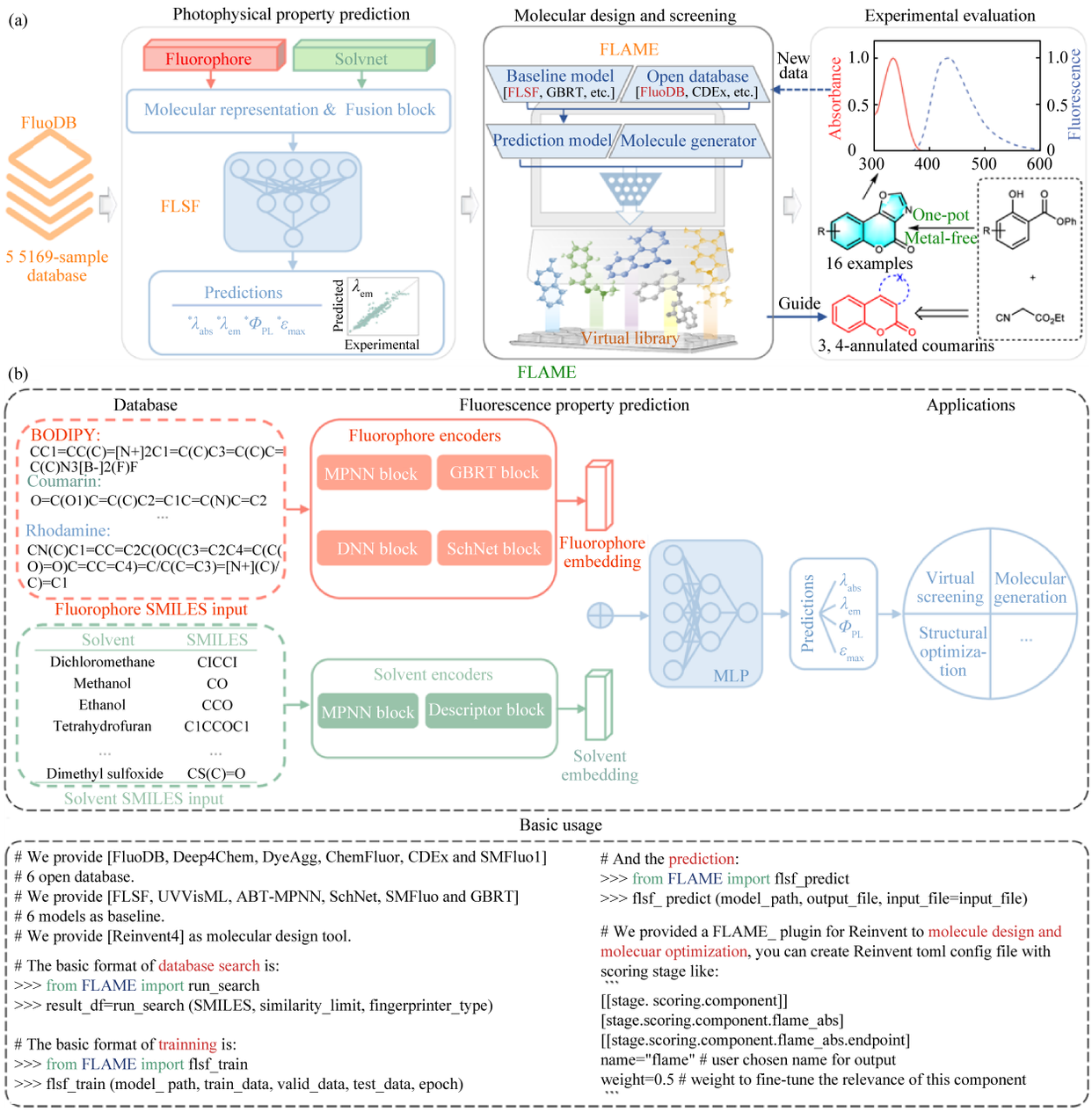
ML、行业经验、有机合成、分子表征、器件制备及光电测试等环节,探索了160万种分子的化学空间,并利用DFT对超过40万种分子进行筛选,成功鉴定了数千种覆盖整个可见光谱的新型有机发光二极管候选分子,并遴选出最优候选分子,且经实验测定其量子效率最高可达22%。

3.2　国内研究进展

近年来,国内科研人员也在生成式AI赋能OFMs领域取得了一系列成果,逐步形成了从数据库搭建、算法模型创新和实验闭环验证的完整性布局。台湾师范大学Tsai课题组通过引入溶液环境并采用2种特征化方案改进了SchNET模型[52],从而显著提高了对包括吸收能量、发射能量和光致发光量子产率等实验光物理数据的预测精度。浙江大学钱玲慧课题组报道了荧光团设计加速(fluorophore design acceleration ModulE, FLAME)的AI框架(图6(a))[53],构建了迄今规模最大的开源荧光染料数据库FluoDB(包含55 169个荧光染料和溶剂数据),并采用荧光骨架驱动预测模型(Fluorescence prediction with fluoroScaFfold-driven model, FLSF)在内的多种模型,实现了最大吸收波长、最大发射波长、发光量子产率和摩尔吸光系数等光学性质的快速精准预测;通过集成分子生成器提供具有目标荧光性质的新化合物,并且通过对代表性3,4-噁唑并香豆素衍生物的合成与性能验证,成功合成了一类具有明亮荧光的新型化合物;FLAME框架整合了最新的数据库与预测模型,用于虚拟筛选、分子生成及结构优化等多种应用场景(图6(b)),包括数据库检索、光物理性质预测、不同荧光团数据库整合、预测模型与分子生成器及未报道分子生成等。电子科技大学张万里课题组基于门控循环单元自编码器构建模型,以SMILES字符串作为输入,通过编码与解码过程处理分子结构信息[35],该模型生成分子的激发态性质与原始样本性质相同,且易于合成,得到了振子强度高且发光效率高的分子。电子科技大学谭等课题组搭建了一个集成框架[29],通过融合算法合成、深度预测、对抗生成和精细筛选等模块,设计了热活化延迟荧光分子,基于经验给定的D-A单元,采用逆合成规则算法生成D-A复合物,随后利用深度神经网络进行高通量标记与预测;通过生成模型得到的D-A分子激发态性质分布与原始样本完全匹配,并基于高精度的激发态电子结构计算,在化学空间中选择了具有显著电荷转移特征、极小绝热单三重态能级差以及适宜自旋轨道耦合的热活化延迟荧光候选分子。北京交通大学宋丹丹课题组建立了集成框架以描述圆偏振热活化延迟荧光分子的不对称性因子并预测了新的荧光分子[54],首先对于已报道的实验值及其理论预测结果进行了全面的统计分析,揭示了实验与计算结果之间的差异;其次以摩根指纹为特征的机器学习模型预测不对称因子,进行详细的结构分析;最后基于改变变分自编码器的生成模型,以不对称因子为目标函数,实现了圆偏振热活化延迟荧光分子的逆向设计。

3.3　国内外发展差异

国内在生成式AI辅助荧光染料设计已形成较为完整的研发布局,并取得了一系列成果。与国际



注:FluoDB为开源荧光数据库;FLSF为荧光骨架驱动的预测模型;FLAME为荧光团设计加速;MLP为多层感知器;λ_{abs}为最大吸收波长;λ_{em}为最大发射波长;Φ_{PL}为发光量子产率;ε_{max}为摩尔吸光系数。

图6　FLAME荧光团智能设计:模块化人工智能(AI)驱动框架[53]

(a)加速荧光团设计;(b)设计示意

Figure 6　a modular artificial intelligence (AI) framework to facilitate fluorophore design[53]

(a) facilitating fluorophore design; (b) schematic overview

研究相比,主要有以下差异:国际团队搭建的实验数据库规模更大、开发性更强,而国内对于数据整合和标准化共享机制仍存在一定提升空间;另外,相比较于国外的全自动、高通量实验闭环,国内对于全自动化集成平台的研究仍处于上升阶段,对于多项性能指标的优化策略仍然需要加强。

目前,国内研究在算法原创性、应用场景拓展及平台化建设方面发展迅速,并以实际需求为导向,注重模型可解释性与实验验证的发展趋势。随着数据资源的不断积累,学科合作的继续推进以及自动化实验平台的完善,生成式AI有望加快设计进程,为新分子的设计提供了理论参考和指导意见。

4 结束语

生成式AI与逆向设计的结合为高通量预测和设计新型荧光染料提供了强大工具。通过集成数据库、预测模型和分子生成器,研究人员可实现从目标性能到分子结构的全流程设计,试错成本显著降低,研发周期明显缩短。然而,该领域仍面临诸多挑战,比如数据稀缺、复杂性质表征困难等问题。为应对这些挑战,推动该领域进一步发展,未来的研究工作应着力于:构建标准化、可扩展的高品质数据库,覆盖合成路径、纯度、光谱测试条件等原始数据;融合领域知识,建立引入经典的光物理约束的生成模型等,如Kasha和能隙定律等规则,提升生成分子的合理性和准确性;建立复杂光物理性质的预测框架,结合量子化学计算和分子动力学模拟,预测关键微观参数;创建开放、集成的跨领域的多效用数据库,体现分子种类与溶剂环境、稳定性与电化学等数据;建设闭环平台,从理论计算、AI设计、自动化合成、实验反馈的模型调整与应用,实现闭环建设。

©The editorial department of Journal of Capital Normal University (Natural Science Edition) 2025. This is an open access article under the CC BY-NC-ND 4.0 license (https://creativecommons.org/licenses/by-nc-nd/4.0/).

参 考 文 献

[1] ZHONG J M, ZHU W, SHEN S T, et al. Machine learning for organic fluorescent materials [J]. Aggregate, 2025,6(9):e70089. DOI:10.1002/agt2.70089.

[2] LIU S J, LI Y Y, KWOK R T K, et al. Structural and process controls of AIEgens for NIR-II theranostics [J]. Chemical Science, 2021, 12 (10): 3427-3436. DOI: 10.1039/d0sc02911d.

[3] AHMAD N, EID G, EL-TOONY M M, et al. Harnessing machine learning for the rational design of high-performance fluorescent dyes [J]. Spectrochimica Acta Part A (Molecular and Biomolecular Spectroscopy), 2025,334:125918. DOI:10.1016/j.saa.2025.125918.

[4] YANG Z H, WANG X F. Machine learning-based prediction of fluorescence lifetimes of zero-dimensional antimony halides [J]. Journal of Materials Chemistry C, 2025,13(24):12271-12286. DOI:10.1039/d5tc00499c.

[5] ZHAO Y H, CHEN K, ZHU L, et al. Data-driven machine learning models for quick prediction of the Stokes shift of organic fluorescent materials [J]. Dyes and Pigments, 2023, 220: 111670. DOI: 10.1016/j.dyepig.2023.111670.

[6] WU J X, SHI Z X, ZHU L L, et al. The design and bioimaging applications of NIR fluorescent organic dyes with high brightness [J]. Advanced Optical Materials, 2022,10(8):2102514. DOI:10.1002/adom.202102514.

[7] SHENG W L, WU Y Y, YU C J, et al. Synthesis, crystal structure, and the deep near-infrared absorption/emission of bright azaBODIPY-based organic fluorophores [J]. Organic Letters, 2018, 20(9): 2620-2623. DOI: 10.1021/acs.orglett.8b00820.

[8] WU X Y, BELL T D M, YEOW E K L. Electron transport in the long - range charge - recombination dynamics of single encapsulated dye molecules on TiO_2 nanoparticle films [J]. Angewandte Chemie International Edition, 2009, 48(40): 7379-7382. DOI: 10.1002/anie.200902596.

[9] SILVEIRA G DE Q, CHEN Z W, BARRY E F, et al. Energy transfer induced by dye encapsulation in a hybrid nanoparticle-purple membrane reversible assembly [J]. Advanced Functional Materials, 2019, 29(43): 1904899. DOI: 10.1002/adfm.201904899.

[10] YANG Q L, HU Z B, ZHU S J, et al. Donor engineering for NIR-II molecular fluorophores with enhanced fluorescent performance [J]. Journal of the American Chemical Society, 2018, 140 (5): 1715-1724. DOI: 10.1021/jacs.7b10334.

[11] MELNYCHUK N, KLYMCHENKO A S. DNA-functionalized dye-loaded polymeric nanoparticles: ultrabright FRET platform for amplified detection of nucleic acids [J]. Journal of the American Chemical Society, 2018, 140 (34): 10856-10865. DOI: 10.1021/jacs.8b05840.

[12] ANDREIUK B, REISCH A, LINDECKER M, et al. Fluorescent polymer nanoparticles for cell barcoding in vitro and in vivo [J]. Small, 2017, 13 (38): 1701582. DOI: 10.1002/smll.201701582.

[13] GÓMEZ-BOMBARELLI R, AGUILERA-IPARRAGUIRRE J, HIRZEL T D, et al. Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach [J]. Nature Materials, 2016, 15 (10): 1120-1127. DOI: 10.1038/nmat4717.

[14] SZYMKUĆ S, GAJEWSKA E P, KLUCZNIK T, et al. Computer - assisted synthetic planning: the end of the beginning [J]. Angewandte Chemie International Edition, 2016,55(20):5904-5937. DOI:10.1002/anie.201506101.

[15] BIAN Q H, WANG X F. Machine learning-driven design

of fluorescent materials: principles, methodologies, and future directions [J]. Nanomaterials, 2025, 15(19): 1495. DOI:10.3390/nano15191495.

[16] FURUKORI M, NAGAMUNE Y, NAKAYAMA Y, et al. High-throughput transient photoluminescence spectrometer for deep learning of thermally activated delayed fluorescence materials [J]. Journal of Materials Chemistry C, 2023, 11(13): 4357-4364. DOI: 10.1039/d3tc00482a.

[17] HE Z M, BI H, LIANG B Y, et al. Frontier molecular orbital weighted model based networks for revealing organic delayed fluorescence efficiency [J]. Light (Science & Applications), 2025, 14(1): 75. DOI: 10.1038/s41377-024-01713-w.

[18] CHEBOTAEV P P, BUGLAK A A, SHEEHAN A, et al. Predicting fluorescence to singlet oxygen generation quantum yield ratio for BODIPY dyes using QSPR and machine learning [J]. Physical Chemistry Chemical Physics, 2024, 26(38): 25131-25142. DOI: 10.1039/d4cp02471k.

[19] JU C W, BAI H Z, LI B, et al. Machine learning enables highly accurate predictions of photophysical properties of organic fluorescent materials: emission wavelengths and quantum yields [J]. Journal of Chemical Information and Modeling, 2021, 61(3): 1053-1065. DOI:10.1021/acs.jcim.0c01203.

[20] NIKHITHA R, MONDAL A. Enhancing the prediction of TADF emitter properties using Δ-machine learning: a hybrid semi-empirical and deep tensor neural network approach [J]. The Journal of Chemical Physics, 2025, 162(14): 144103. DOI:10.1063/5.0263384.

[21] SANCHEZ-LENGELING B, ASPURU-GUZIK A. Inverse molecular design using machine learning: generative models for matter engineering [J]. Science, 2018,361(6400):360-365. DOI:10.1126/science.aat2663.

[22] CURTAROLO S, HART G L W, NARDELLI M B, et al. The high-throughput highway to computational materials design [J]. Nature Materials, 2013, 12(3): 191-201. DOI:10.1038/nmat3568.

[23] DRAL P O, BARBATTI M. Molecular excited states through a machine learning lens [J]. Nature Reviews Chemistry, 2021, 5(6): 388-405. DOI: 10.1038/s41570-021-00278-1.

[24] MENG Q Y, WANG R, SHAO H Y, et al. Precise regulation on the bond dissociation energy of exocyclic C—N bonds in various N-heterocycle electron donors via machine learning [J]. Journal of Physical Chemistry Letters, 2024, 15(16): 4422-4429. DOI: 10.1021/acs.jpclett.4c00705.

[25] PENG J H, SHEN S T, ZHU W, et al. Ground-state descriptor enables machine learning-assisted virtual screening of AIE-active mechanofluorochromic molecules with high contrast [J]. Journal of Chemical Information and Modeling, 2025, 65(12): 5912-5920. DOI:10.1021/acs.jcim.5c00658.

[26] THAPA K, JONES J I, RATCLIFF L E. A high throughput virtual screening approach for identifying new thermally activated delayed fluorescence-based emitters [J]. Journal of Materials Chemistry C, 2025, 13(37): 19212-19225. DOI:10.1039/d5tc01938a.

[27] MAHATO K D, KUMAR U. Optimized machine learning techniques enable prediction of organic dyes photophysical properties: absorption wavelengths, emission wavelengths, and quantum yields [J]. Spectrochimica Acta Part A (Molecular and Biomolecular Spectroscopy), 2024, 308: 123768. DOI: 10.1016/j.saa.2023.123768.

[28] YOSHIKAWA N, TERAYAMA K, SUMITA M, et al. Population-based de novo molecule generation, using grammatical evolution [J]. Chemistry Letters, 2018, 47(11): 1431-1434. DOI:10.1246/cl.180665.

[29] TAN Z, LI Y, ZHANG Z Y, et al. Efficient adversarial generation of thermally activated delayed fluorescence molecules [J]. ACS Omega, 2022, 7(21): 18179-18188. DOI:10.1021/acsomega.2c02253.

[30] XU Y Z, LUO Y R, LI B, et al. Exploring optimized organic fluorophore search through experimental data-driven adaptive β-VAE [J]. JACS Au, 2025, 5(7): 3082-3091. DOI:10.1021/jacsau.5c00052.

[31] SHI Y M, SHI H C, WANG H C, et al. Machine learning-driven precise design of stable OLED materials: predicting and enhancing multi-state C—N bond dissociation energies [J]. Chemical Engineering Journal, 2024, 500: 157082. DOI:10.1016/j.cej. 2024.157082.

[32] KIM B, ASHYRMAMATOV I, UCAK U V, et al. Fluorophore design via generative modeling in a curated subspace [J]. ACS Omega, 2025, 10(51): 62693-62703. DOI:10.1021/acsomega.5c07333.

[33] KADURIN A, NIKOLENKO S, KHRABROV K, et al. druGAN: an advanced generative adversarial autoencoder model for de novo generation of new molecules with desired molecular properties in silico [J]. Molecular Pharmaceutics, 2017, 14(9): 3098-3104. DOI:10.1021/acs.molpharmaceut.7b00346.

[34] HONG Y, HOU B, JIANG H L, et al. Machine learning and artificial neural network accelerated computational

discoveries in materials science [J]. WIREs Computational Molecular Science, 2019, 10(3): e1450. DOI: 10.1002/wcms.1450.

[35] TAN Z, LI Y, WU X, et al. De novo creation of fluorescent molecules via adversarial generative modeling [J]. RSC Advances, 2023, 13(2): 1031-1040. DOI: 10.1039/d2ra07008a.

[36] SUMITA M, TERAYAMA K, SUZUKI N, et al. De novo creation of a naked eye-detectable fluorescent molecule based on quantum chemical computation and machine learning [J]. Science Advances, 2022, 8: eabj3906. DOI: 10.1126/sciadv.abj3906.

[37] GOODMAN J. Computer software review: reaxys [J]. Journal of Chemical Information and Modeling, 2009, 49: 2897-2898. DOI: 10.1021/ci900437n.

[38] GONG J Y, GONG W J, WU B, et al. ASBase: the universal database for aggregate science [J]. Aggregate, 2022, 4(1): e263. DOI: 10.1002/agt2.263.

[39] JOUNG J F, HAN M, JEONG M, et al. Experimental database of optical properties of organic compounds [J]. Scientific Data, 2020, 7(1): 295. DOI: 10.1038/s41597-020-00634-8.

[40] MAO J L, WEI H L, LUO Z H. Machine learning-assisted study on structure-property relationships of dyes: a review [J]. Dyes and Pigments, 2026, 245: 113273. DOI: 10.1016/j.dyepig.2025.113273.

[41] HUANG B, VON LILIENFELD O A. Communication: understanding molecular representations in machine learning: the role of uniqueness and target similarity [J]. The Journal of Chemical Physics, 2016, 145(16): 161102. DOI: 10.1063/1.4964627.

[42] NEMATOV D, HOJAMBERDIEV M. Machine learning-driven materials discovery: unlocking next-generation functional materials-a review [J]. Computational Condensed Matter, 2025, 45: e01139. DOI: 10.1016/j.cocom.2025.e01139.

[43] PAROLIN G, GARAIN B C, MUKHERJEE S, et al. Conformational dynamics of the pyrene excimer [J]. Physical Chemistry Chemical Physics, 2024, 26(47): 29351-29363. DOI: 10.1039/d4cp03947e.

[44] TU C, HUANG W J, LIANG S, et al. Combining machine learning and quantum chemical calculations for high-throughput virtual screening of thermally activated delayed fluorescence molecular materials: the impact of selection strategy and structural mutations [J]. RSC Advances, 2022, 12(48): 30962-30975. DOI: 10.1039/d2ra05643g.

[45] DONG J, QIAN J, YU K Q, et al. Rational design of organelle-targeted fluorescent probes: insights from artificial intelligence [J]. Research, 2023, 6: 0075. DOI: 10.34133/research.0075.

[46] BEHLER J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials [J]. The Journal of Chemical Physics, 2011, 134(7): 074106. DOI: 10.1063/1.3553717.

[47] POZDNYAKOV S N, WILLATT M J, BARTÓK A P, et al. Comment on "manifolds of quasi-constant SOAP and ACSF fingerprints and the resulting failure to machine learn four-body interactions" [J]. The Journal of Chemical Physics, 2022, 157(17): 34302. DOI: 10.1063/5.0088404.

[48] OCHIAI H, KANEKO H. Construction of machine learning models to predict the maximum absorption wavelength considering the solute and solvent and inverse analysis of the models [J]. ACS Omega, 2025, 10(1): 665-672. DOI: 10.1021/acsomega.4c07490.

[49] KOSCHER B A, CANTY R B, MCDONALD M A, et al. Autonomous, multiproperty-driven molecular discovery: from predictions to measurements and back [J]. Science, 2023, 382(6677): eadi1407. DOI: 10.1126/science.adi1407.

[50] HAN M, JOUNG J F, JEONG M, et al. Generative deep learning-based efficient design of organic molecules with tailored properties [J]. ACS Central Science, 2024, 11(2): 219-227. DOI: 10.1021/acscentsci.4c00656.

[51] JOUNG J F, HAN M, HWANG J, et al. Deep learning optical spectroscopy based on experimental database: potential applications to molecular design [J]. JACS Au, 2021, 1(4): 427-438. DOI: 10.1021/jacsau.1c00035.

[52] HUNG S H, YE Z R, CHENG C F, et al. Enhanced predictions for the experimental photophysical data using the featurized schnet-bondstep approach [J]. Journal of Chemical Theory and Computation, 2023, 19(14): 4559-4567. DOI: 10.1021/acs.jctc.3c00054.

[53] ZHU Y C, FANG J B, AHMED S A H, et al. A modular artificial intelligence framework to facilitate fluorophore design [J]. Nature Communications, 2025, 16: 3598. DOI: 10.1038/s41467-025-58881-5.

[54] SHI H C, SHI Y M, JIANG P C, et al. Integrating machine learning and generative models for the intelligent design of TADF materials with circularly polarized luminescence [J]. Journal of Materials Chemistry C, 2025, 13(42): 21408-21417. DOI: 10.1039/d5tc02451j.

(责任编辑:王 媛)