arXiv:2611.32133v1 [physics.chem-ph] 24 Nov 2026

# OMNI-P3x 面向大规模计算光化学工作流程的通用神经网络势

Miko-laj Martyka，† Joanna Jankowska，*、† 以及 Pavlo O. Dral*、‡、¶、§、‖

†华沙大学，化学学院，*Krakowskie Przedmie´scie 26/28*，*00-927*，华沙，波兰

‡固体表面物理化学国家重点实验室、化学与化工学院、福建省理论与计算化学重点实验室，厦门大学，厦门，福建 *361005*，中国

¶物理学研究所、物理、天文学与信息学学院，尼古拉·哥白尼大学（托伦），*ul. Grudziadzka 5*，*87-100 Toru´n*，波兰§高等研究院，尼古拉·哥白尼大学（托伦），*ul. Wile´nska 4*，*87-100 Toru´n*，波兰‖*Aitomistic*，深圳 *518000*，中国

电子邮箱：joanna.jankowska@uw.edu.pl；dral@xmu.edu.cn

**摘要**

计算光化学有助于新材料的设计与理解。然而，激发态量子力学（QM）计算成本高昂，使得大规模分子库筛选的费用高到难以承受。为解决这一限制，我们推出了 OMNI-P3x，这是一种通用的机器学习原子间势（MLIP），用于中性和带电有机分子的单重态基态和激发态模拟。该模型在基态和激发态构象上进行训练，能够完成完整的计算光化学工作流程，精度接近含时密度泛函理论（TD-DFT）。我们通过筛选一个包含 234,375 种黄烷鎓衍生物的组合分子库，展示了该模型的实际应用能力：预训练模型无需额外的 QM 计算，即可进行基态几何优化并预测吸收性质。经过微调后，我们进一步筛选了 162,960 种基于呫吨的染料，并发现了可用于红光、近红外和短波红外发射的新候选项。此外，在未使用参考原子电荷的情况下学习得到的态依赖神经网络电荷，为激发时的分子内电荷重排提供了可解释的图景。

## 引言

计算光化学和光物理学通过计算机模拟 预测吸收和发射光谱、[1–3] 评估分子光反应性和光稳

定性，[4,5] 并分析光诱导反应机制。[6,7] 这些能力有助于开发染料、[8,9] 用于生物成像的荧光探针、[10] 分子光开关、[11,12] 有机发光二极管、[13–15] 太阳能电池[16,17] 和分子太阳能热储能系统，[18–20] 这些系统的性能在很大程度上取决于其光学性质和光化学特性。通过将分子结构与光物理和光化学行为联系起来，计算方法有助于解释实验观察结果，并指导新候选体系的设计，从而缩小化学空间，减少昂贵的实验验证。[13,14]

然而在实际应用中，激发态计算所需的量子力学（QM）方法计算成本高得令人望而却步，阻碍了此类研究。即使采用广泛使用的近似方法，例如含时密度泛函理论（TD-DFT），[21,22]评估激发态性质的成本也会随分子规模迅速增长；而计算成本更高的基于波函数的方法，例如 ADC(2)[23] 或运动方程耦合簇（EOM-CC），[24]则更加昂贵。对单个发色团而言可行的计算，若扩展到数十万个候选分子，便会变得难以实施。因此，理论研究往往仅限于少量分子，甚至只研究某一类分子中单个具有代表性的分子；要筛选庞大的分子体系库仍然十分困难。[25,26]

机器学习原子间势（MLIPs）为加速此类工作负载提供了自然途径，即以机器学习（ML）近似取代量子力学计算。通过在海量分子性质数据集上进行预训练，这些模型无论是否经过领域特定的微调，都能以极低的误差、仅需一小部分成本取代昂贵的量子力学计算。[27–30] 对于基态化学而言，此类代理模型的应用正日益普及，使研究人员既能模拟更大的体系，也能更广泛地覆盖所研究的化学空间。[31] 然而，与此同时，机器学习辅助的光化学模拟和工作流程方面的进展则要有限得多。尽管已有许多主要基于神经网络（NNs）的激发态性质机器学习架构问世，[32–34] 但大多数应用仍依赖于针对特定单个分子体系训练的模型。[35–38]

近期的发展正推动激发态机器学习方法的前沿，使其能够在广泛的化学空间中实现更强的可迁移性。Axelrod 等人开发了一种在偶氮苯光开关领域具有可迁移性的神经网络势能，使得能够对其光开关量子产率和热半衰期等性质进行筛选。[39,40] 在一项相关研究中，Tiefenbacher 等人为芳基偶氮吡唑光开关训练了一种机器学习原子间势，并将其整合到自动化工作流中，用于定位极小值、过渡态和单重态–三重态交叉点。[41] 这使得对热弛豫速率进行筛选成为可能，并能够识别决定异构化路径之间竞争关系的结构特征。X-MACE[42,43] 实现了对荧光蛋白发色团衍生物的类似筛选：在针对相关系统进行微调之前，先对更广泛的发色团集合进行预训练。研究表明，预训练能够减少获得准确模型所需的数据点数量。[42,43] 与此同时，基于学习电子哈密顿量的方法[44,45] 以及基于极值池化的尺寸强度架构[46] 也正朝着提高可迁移性的方向发展。

多状态 ANI（multi-state ANI）模型架构进一步表明，单个神经网络能够以足够高的精度学习势能面，从而对不同分子开展非绝热分子动力学模拟。[47] 在这一多状态模型的基础上，OMNI-

P2x 是首个用于激发态性质预测的通用神经网络势，[48]能够直接预测各种有机分子的激发能和吸收光谱。然而，该模型受到训练数据的限制：激发态标签仅适用于基态平衡几何结构，因此在预测非平衡结构时可靠性有限，而这类结构对于几何优化或发射光谱预测等任务至关重要；这些任务也是光化学工作流程的关键环节。

本文介绍了 OMNI-P3x，这是一种用于基态和激发态模拟的通用机器学习原子间势，旨在支持大规模光化学工作流程；其设计与能力概述见图 1。该模型在前代 OMNI-P2x 的基础上进行了扩展，扩大了元素覆盖范围，能够描述带电体系，并可预测态依赖的永久偶极矩。模型基于新构建的训练集进行训练，该训练集通过采样基态和激发态势能面（PESs）生成，使模型能够对两种状态进行几何优化，并进一步预测吸收和发射性质。模型将成熟的 AIMNet2架构与多状态[47]和一体化[49]（多保真度）学习相结合，并纳入多个理论层级的额外基态参考数据，从而在预测基态性质时保持准确性。我们通过在多个测试集上将结果与参考方法及理论最佳估计进行比较，评估模型在基态和激发态预测方面的性能。此外，我们还表明，模型在未使用参考原子电荷的情况下学习到的态依赖神经网络电荷，可以直观地呈现激发时分子单元之间的电荷重分布。这种分析或有助于设计对电荷转移性质有关键要求的材料，例如染料敏化太阳能电池[50]或用于热激活延迟荧光的给体–受体发射体，[51]不过迄今为止，这仅作为神经网络势的一项概念验证能力得到展示。[52]

重要的是，我们展示了OMNI-P3x能够执行预测光化学和光物理性质所需的完整工作流程：从优化基态或激发态几何结构，到计算频率并生成用于获得振动分辨光谱的构象体，再到预测跃迁能量和强度。通过对包含234,375种黄烷鎓衍生物的组合库进行高通量筛选，我们展示了其实际应用能力。预训练模型能够执行包括基态几何优化和吸收光谱预测在内的完整任务，其准确度接近含时密度泛函理论（TD-DFT）。此外，我们将微调后的OMNI-P3x应用于包含162,960种基于氧杂蒽的候选染料的库进行发射筛选，从而能够提出新型发射体体系。

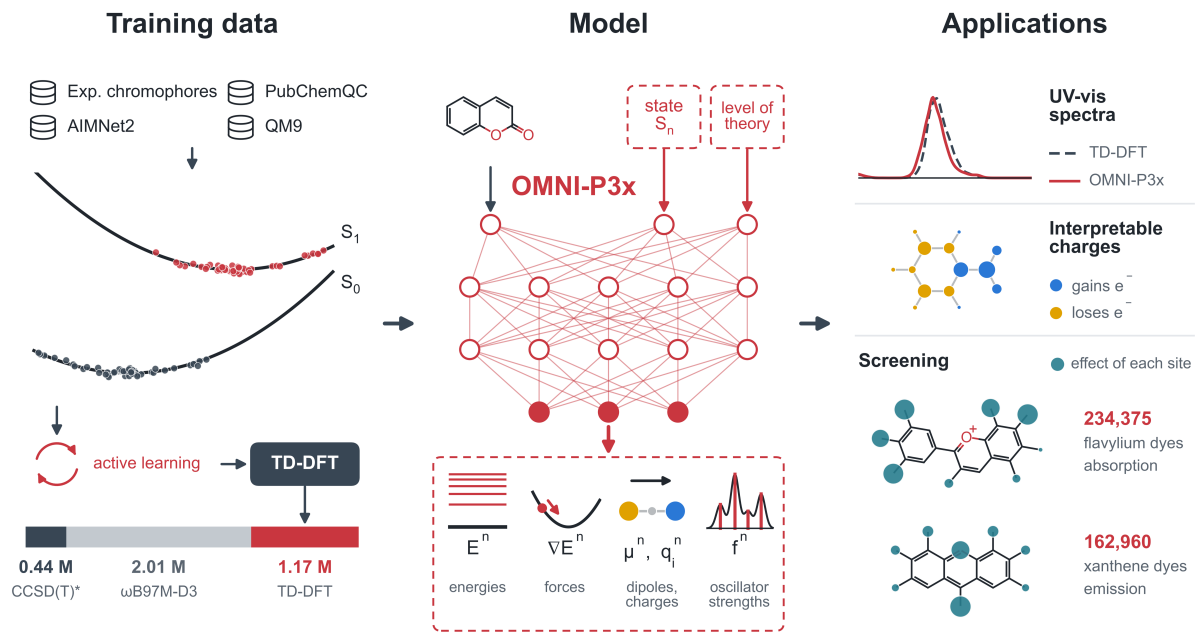



图 1：OMNI-P3x 模型开发与应用概览。我们从多个数据库中采样分子的基态和激发态构象，通过主动学习进行筛选，并使用 TDA-$\omega$B97X-D/def2-SVP 标注，从而构建了一个包含 117 万个结构的新数据集。该数据集与另外两个理论层级的基态数据结合后，用于训练 OMNI-P3x。该模型可预测紫外-可见光谱模拟、激发态电荷分析和高通量筛选所需的激发态性质。

# 结果与讨论

## 模型架构与训练

OMNI-P3x架构基于AIMNet2模型及其预训练权重（基础模型），[28] 通过参考文献47中开发的多态学习方案扩展至激发态 。通过在共享神经网络中高效学习任意数量的电子态及其性质，并在输入层对电子态进行编码，使整个网络以所关注的状态为条件。与其前身一样，OMNI-P3x采用一体化（多保真度）训练范式，[49] 使其能够跨不同量子力学理论水平高效学习。这是在神经网络的描述符层面实现的，如图2a所示：在为每个原子构建的原子特征向量中追加两个输入特征：电子态 $S_n$的标量索引 $n$ ，例如 $n = 0$对应 $S_0$ 、$n = 1$ 对应 $S_1$，以及独热编码理论水平向量 $\mathbf{d}$。在用于细化描述符的每次消息传递迭代中都会执行此操作，从而形成给定电子态的最终AIM（分子内原子）向量。根据$\mathbf{d}$的选择，对于基态性质，模型可在CCSD(T)*/CBS和 $\omega$B97M-D3(BJ)/def2-TZVPP理论水平下预测性质；对于基态和激发态性质，则可在TDA-$\omega$B97X-D/def2-SVP理论水平下进行预测。在本文中，我们将这一选择记为OMNI-P3x@CCSD(T)*,

OMNI-P3x@$\omega$B97M-D3(BJ) 和 OMNI-P3x@TDA-$\omega$B97X-D。AIMNet2 还会执行神经电荷均衡过程，分配原子部分电荷，使其总和等于分子的总电荷 $Q$，从而得到依赖于状态的学习型部分电荷分布。

OMNI-P3x 具有两个作用于 AIM（分子中的原子）向量的输出头，它们是包含 2 个隐藏层、每层 128 个神经元的多层感知器。第一个是能量头，与基础模型类似，通过原子贡献求和得到最终能量，其中 $i$ 是给定原子的索引，并加上长程静电项；该项表示神经网络部分电荷之间的库仑相互作用。第二个是振子强度头，用于预测 $S_0 \to S_n$ 跃迁的振子强度。力 $\mathbf{F^n} = -\nabla E^n$ 通过对能量输出进行自动微分获得。部分电荷求和为 $\sum r_i q_i^n$，其中 $r_i$ 是原子 $i$的位置， $q_i^n$是其处于状态 $n$时的神经网络电荷，从而形成电子状态 $n$ $\boldsymbol{\mu}^n$的永久偶极矩。这四项指标在训练过程中使用以下损失函数进行联合最小化：

$$\mathcal{L} = w_E \frac{\left\| E^{\mathrm{ML}} - E^{\mathrm{ref}} \right\|^2}{\sqrt{N_{\mathrm{at}}}} + w_F \left\| \mathbf{F}^{\mathrm{ML}} - \mathbf{F}^{\mathrm{ref}} \right\|^2 + w_\mu \left\| \sum_i \mathbf{r}_i q_i^{\mathrm{ML}} - \boldsymbol{\mu}^{\mathrm{ref}} \right\|^2 + w_f \left\| f^{\mathrm{ML}} - f^{\mathrm{ref}} \right\|^2, \tag{1}$$

其中，上标 ML 和 ref 分别表示模型预测值和参考值， $\|\cdot\|^2$为均方误差， $N_{\mathrm{at}}$ 为原子数。能量、力、偶极矩和振子强度的损失权重分别设为 1、0.1、0.1 和 0.5，这些数值由超参数搜索和经验确定。

## 训练集生成

构建成功的多层感知机，首先需要稳健且高质量的训练数据。然而，由于计算光化学的复杂性更高，[53] 目前可用于分子电子激发态的数据仍十分有限。具体而言，目前还没有包含激发态能量梯度，以及化学空间内非平衡结构性质的数据集。纳入后者尤为关键，因为几何优化或荧光光谱预测等任务要求多层感知机在远离 $S_0$ 极小值的区域也能保持准确。因此，在本研究中，我们通过对现有大型数据集进行抽样，并为新体系及其构象体生成数据，整理构建了一个新的激发态数据集。

该流程以 Joung 等人提供的、由实验确认的发色团体系 SMILES字符串数据库为起点。[54] 使用 RDKit 将分子转换为三维结构，[55] 再通过 GFN2-xTB 进行优化，[56] 并在基态和第一激发态下采用元动力学进行构象采样。采样过程针对总电荷介于 $-2$ 和 $+2$之间、且仅包含 12 种元素（氢、B、C、N、O、F、Si、P、S、Cl、Br 和 I）的体系进行，如图2所示。后续采样阶段采用基于不确定性的选择，该不确定性通过主模型与辅助模型之间的预测差异进行量化：主模型基于能量和梯度训练，辅助模型则仅基于能量训练，方法如参考文献 47、57 所述。从 PubChemQC、[58] AIMNet2,[28] 和 QM9 数据集中采样分子，[59] 并生成基态和

激发态构象，采用元动力学和正则模采样生成。该流程仅对分子构象进行采样，并针对所有构象在 TDA-ωB97X-D/def2-SVP 理论水平下生成参考标签，包括基态和最低的十个单重激发态的能量，以及 $S_0 \rightarrow S_n$ 跃迁的跃迁偶极矩和振子强度。计算 $S_0$ 和 $S_1$的能量梯度及永久偶极矩；对于更高激发态，则对每个结构的计算设定固定预算。图2b概述了激发态数据集的采样策略，详情见方法部分。新生成的训练集包含 1,168,869 个带有基态和激发态标签的含时密度泛函理论数据点，其激发能分布如图2c所示。

与其前身一样，OMNI-P3x 采用一体化学习[49]，在多个理论层级上同时进行训练。另使用了两个数据集：ANI-1ccx 数据集[60]包含约 500,000 个达到 CCSD(T)/CBS 精度的基态数据点；AIMNet2 数据集[28] 则仅取随机抽取的 10% 构象，额外提供 2,000,000 个在 $\omega$B97M-D3/def2-TZVPP 理论层级下标注的基态数据点，并在训练过程中充当回放集。添加这些额外基态数据，主要是为了提升模型在基态下的性能，并保留预训练 AIMNet2 的性能，这对吸收光谱模拟等光化学任务也至关重要。纳入基态数据后，训练集共包含 3,618,343 个构象。数据集的最终组成见图2b，图2d展示了按分子大小划分的分布，涵盖 2 至 78 个原子的体系（中位数为 24 个原子）。

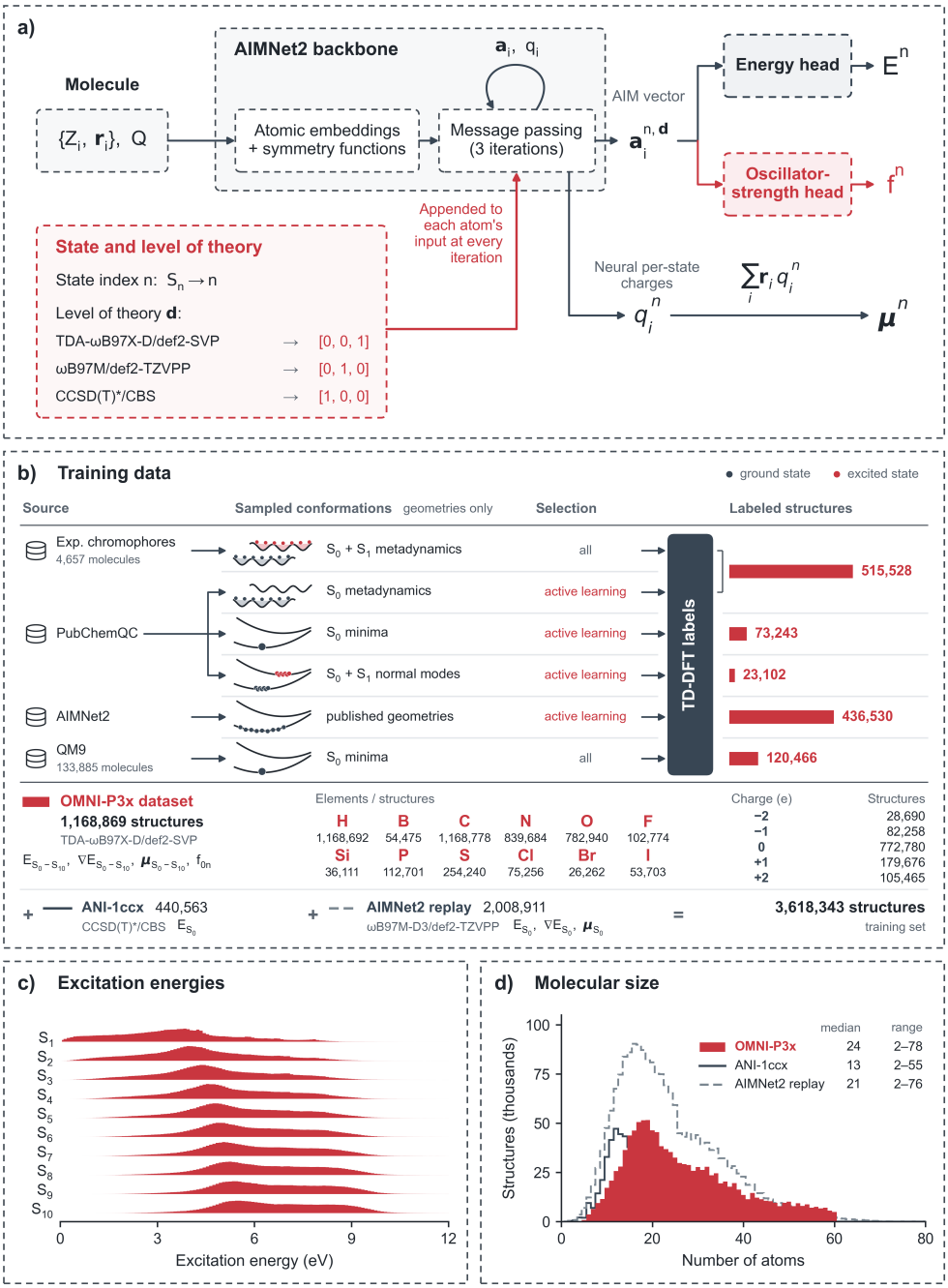



图2：OMNI-P3x 的模型架构和训练数据。a) OMNI-P3x 的神经网络架构，以 AIMNet2 骨干网络为基础，并扩展至激发态。(b) 用于构建新激发态数据集的采样过程及其组成概览。(c) 10 个标注的激发态的激发能分布，$E(S_n) - E(S_0)$。(d) 激发态数据集以及辅助 ANI-1ccx 和 AIMNet2 回放基态数据集中分子大小的分布。所报告的均值和中位数均指激发态数据集。所有计数均以分子结构为单位，包括同一分子的不同构象。

## OMNI-P3x 的性能

我们首先在 PubChemQC 中由 39,862 个此前未见过的、化学结构多样的分子组成的测试集上评估 OMNI-P3x 的激发态性质（图3），以量化训练误差和模型外推误差。对于最低激发，$S_0 \rightarrow S_1$，OMNI- P3x@TDA-$\omega$B97X-D 的平均绝对误差（MAE）为 0.078 eV，不到 OMNI-P2x 在同一组分子上的 0.170 eV 的一半（不包括 OMNI-P2x 不支持的带电体系以及含 B/Si/P/Br 的体系）。两种模型均与各自的参考值进行比较。在十个激发态中，激发能平均绝对误差从 S1态的 0.078 eV 上升至 $S_4$ 处的最高值 0.109 eV，随后逐渐降至 S10 的 0.089 eV （所有状态合并计算为 0.096 eV ）。绝对态能量也呈现相似趋势：平均绝对误差从基态的 0.0298 eV 上升至 $S_4$处的 0.107 eV，随后降至 $S_{10}$处的 0.087 eV。

对激发态而言，能量梯度和永久偶极矩的预测更具挑战性。基态的平均绝对误差分别为每个梯度分量 0.0424 eV/$^{\circ}$A 和 0.145 D，而高能态的误差更大。一个特别的难点在于，能量接近的状态可能具有截然不同的电子特征和性质。例如，具有较大偶极矩的电荷转移态，其能量可能几乎与偶极矩小得多的局域激发态简并。因此，即使两个状态的能量都预测准确，若错误地判定其排序或电子特征，也可能导致较大的偶极矩误差。值得注意的是，这种比较误差并非机器学习模型所固有；在比较任意两种方法的绝热排序性质时也可能出现这种误差，因为尽管能量相近，两种方法对应的状态对仍可能具有不同的电子特征。

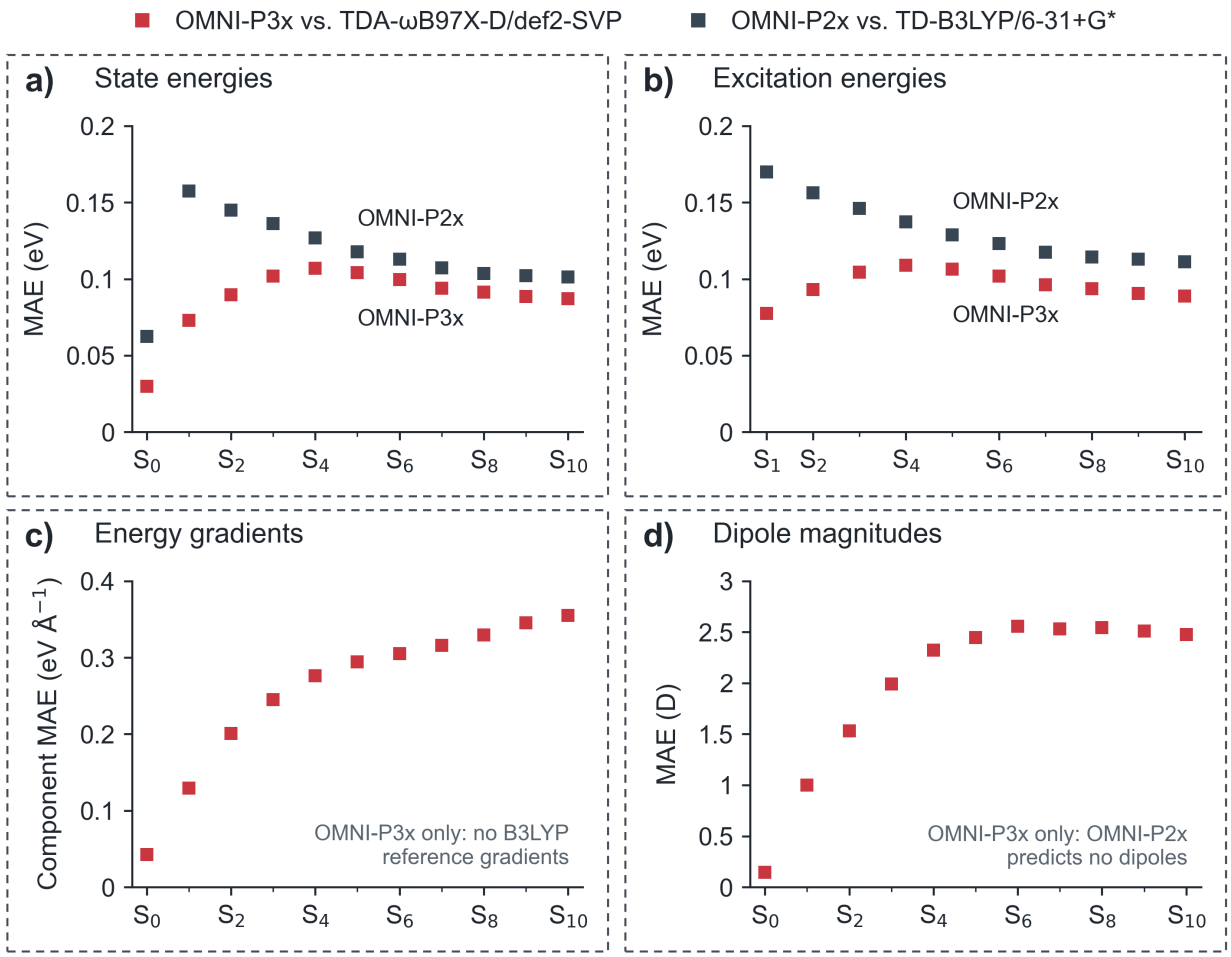



图3：OMNI-P3x 和 OMNI-P2x 在 PubChemQC 数据上的逐状态性能。两种模型均与各自的参考理论水平进行比较。a) 电子态能量的平均绝对误差（MAE），b) $S_0 \rightarrow S_n$ 电子跃迁能量的 MAE，c) 电子态能量梯度的 MAE，d) 电子态永久偶极矩大小的 MAE。

接下来，我们评估 OMNI-P3x 预测值相对于理论最佳估计（TBE）的准确度，而非相对于参考方法的准确度，以检验其预测与已知激发态方法相比有多准确。为此，我们采用 QUEST-3[61]和 QUEST-5[62] 数据集，其中包含小型和中型有机体系的 CC3[63] 优化几何结构，并在理论最佳估计水平上计算激发能。亚硝基分子（HNO）被排除在外，因为它是模型训练域之外的无机体系。所有方法均在已发表的几何结构上进行单点计算；对于所有方法都涵盖的 37 个中性分子，将预测的最低单重态激发能与各分子所列最低单重态的 TBE 进行比较，因为 OMNI-P2x 不支持电荷输入。图 4 展示了 TBE 预测的 $S_0 \rightarrow S_1$ 激发与 OMNI-P3x@TDA-$\omega$B97X-D 预测结果之间的比较。总体而言，对于所有价层激发，其平均绝对误差为 0.21 eV；对于 $n \rightarrow \pi^*$ 激发（0.09 eV），表现明显优于 $\pi \rightarrow \pi^*$（0.46 eV）。其表现与训练数据非常接近：TDA-$\omega$B97X-D/def2-SVP 在价层激发上的总体平均绝对误差为 0.17 eV

激发方面，与 $\pi \to \pi^*$ 相比，$n \to \pi^*$ 的表现显著较差。增大基组并未显著改善价层激发的误差，因为 TDA-$\omega$B97X-D/aug-cc-pVTZ 的总体平均绝对误差仅降低了 0.02 eV。可将这些数值与我们上一代模型 OMNI-P2x 进行比较：其价层平均绝对误差为 0.38 eV，而其训练数据所用的 TD-B3LYP/6-31+G* 的平均绝对误差为 0.28 eV。因此，OMNI-P3x 的改进既源于理论水平更准确，也得益于模型更出色，能更贴近地复现其参考方法。OMNI-P3x 在里德伯型激发方面的表现明显较差，平均绝对误差为 1.32 eV，如图 4d 所示。这一特性源自参考方法所用的小基组（平均绝对误差为 1.37 eV）；换用 aug-cc-pVTZ 基组后，误差几乎完全消除（为 0.10 eV）。OMNI-P2x 使用 TD-B3LYP/6-31+G* 进行训练，由于数据生成过程中加入了弥散函数，其表现更好，也与参考方法的表现一致。作为比较，精度更高的基于波函数的方法 ADC(2)/aug-cc-pVTZ 在价层激发上的平均绝对误差更低，为 0.12 eV；相较于密度泛函理论，其在 $\pi \to \pi^*$ 跃迁上的改进最为显著（平均绝对误差为 0.17 eV）。此外，该方法对里德伯激发也达到了近乎完美的精度（平均绝对误差为 0.04 eV）。

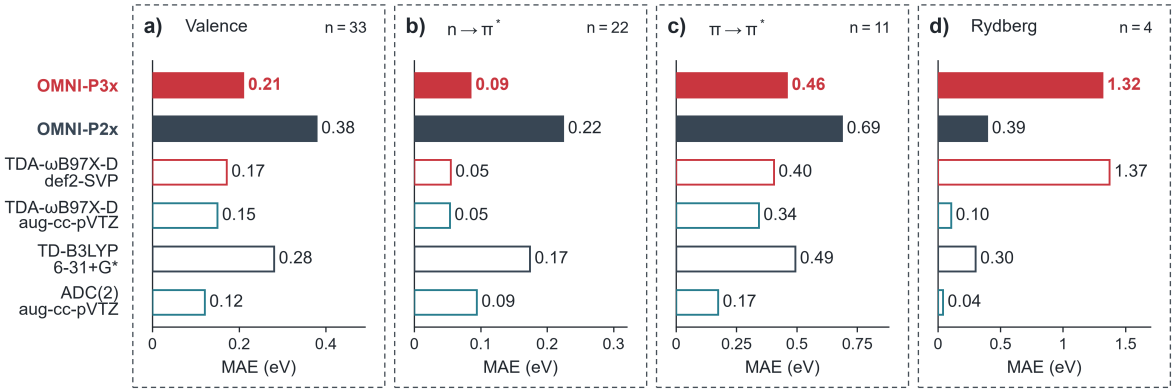



图 4：OMNI-P3x 与其他激发态方法在 QUEST-3 和 QUEST-5 数据集上的预测误差，以 $S_0 \to S_1$ 跃迁相对于理论最佳估计的平均绝对误差表示。（a）所有价层激发；（b）$n \to \pi^*$ 激发；（c）$\pi \to \pi^*$ 激发；（d）里德伯型激发。

接下来，我们评估 OMNI-P3x 在端到端光化学工作流程中的表现，即优化分子的基态或激发态，并计算其在所得极小值处的电子性质。这在实践中具有重要意义，例如可用于预测发射能量，因为这需要定位激发态极小值；也可用于预测吸收光谱，因为这需要在基态极小值处及其附近进行计算。现有的通用机器学习原子间势均不具备此类能力，而前一代 OMNI-P2x 模型的精度仅适用于基态极小值处的单点计算，且这些极小值必须通过单独的方法进行定位。

我们从 QUEST-3,[61]QUEST-5[62] 和 Thiel[64] 数据集中选取了用作光化学基准体系的分子（合并并去重后共 62 个分子），其重原子数介于 2 到 14 之间。对于给定体系，分别采用 OMNI-P3x@TDA-$\omega$B97X-D 和 TDA-$\omega$B97X-D，在 $S_0$ 和 $S_1$中进行优化，并比较两种方法在优化极小值处的性质。工作流程示意图如图5所示。同样，亚硝酰（HNO）作为训练域之外的无机体系被移除，吡咯和噻吩也被移除

自 $S_1$ 评估中移除，因为在 TDA-ωB97X-D/def2-SVP 理论水平下，它们的激发态势能面呈解离性。

图5b展示了此类工作流程的示例，其中比较了 $S_0$ 和$S_1$的机器学习优化几何结构（实心球棍模型）与参考量子力学优化结果（青绿色线框），以及相应的垂直激发能。测试的误差统计如图5c所示。总体而言，OMNI-P3x在基态优化中实现了0.005 °Å的平均重原子均方根偏差，在第一激发态中则为0.029 °Å。优化态最终能量的平均绝对误差对于 $S_0$为0.020 eV，对于 $S_1$为0.198 eV。优化态偶极矩向量的误差，即结构对齐后量子力学值与机器学习值之差向量的长度： $\|\Delta\vec{\mu}\| = \|\vec{\mu}_{\mathrm{ML}} - \vec{\mu}_{\mathrm{QM}}\|$，也从基态下的0.116 D上升至 $S_1$ 优化后的0.463 D。最后，在整个 $S_0 \to S_n$ 光谱范围内，从 $n = 1$ 到 $n = 10$ 平均得到的激发能平均绝对误差，对于 $S_0$ 和 $S_1$ 工作流程分别为0.200 eV和0.378 eV。

OMNI-P3x 的较小误差表明，它能够再现搜索到的极小值的结构，并准确预测这些极小值处的能量和偶极矩等性质。 $S_1$ 优化后出现的较大误差表明，激发态弛豫是更具挑战性的情况，这一点也在预期之中。尽管如此，OMNI-P3x 仍能通过单一模型同时提供弛豫结构及其电子性质，将通用激发态势能的实际应用范围从单点预测扩展到完整的基态和激发态优化及性质评估工作流程；我们将在本文中进一步探讨相关示例。

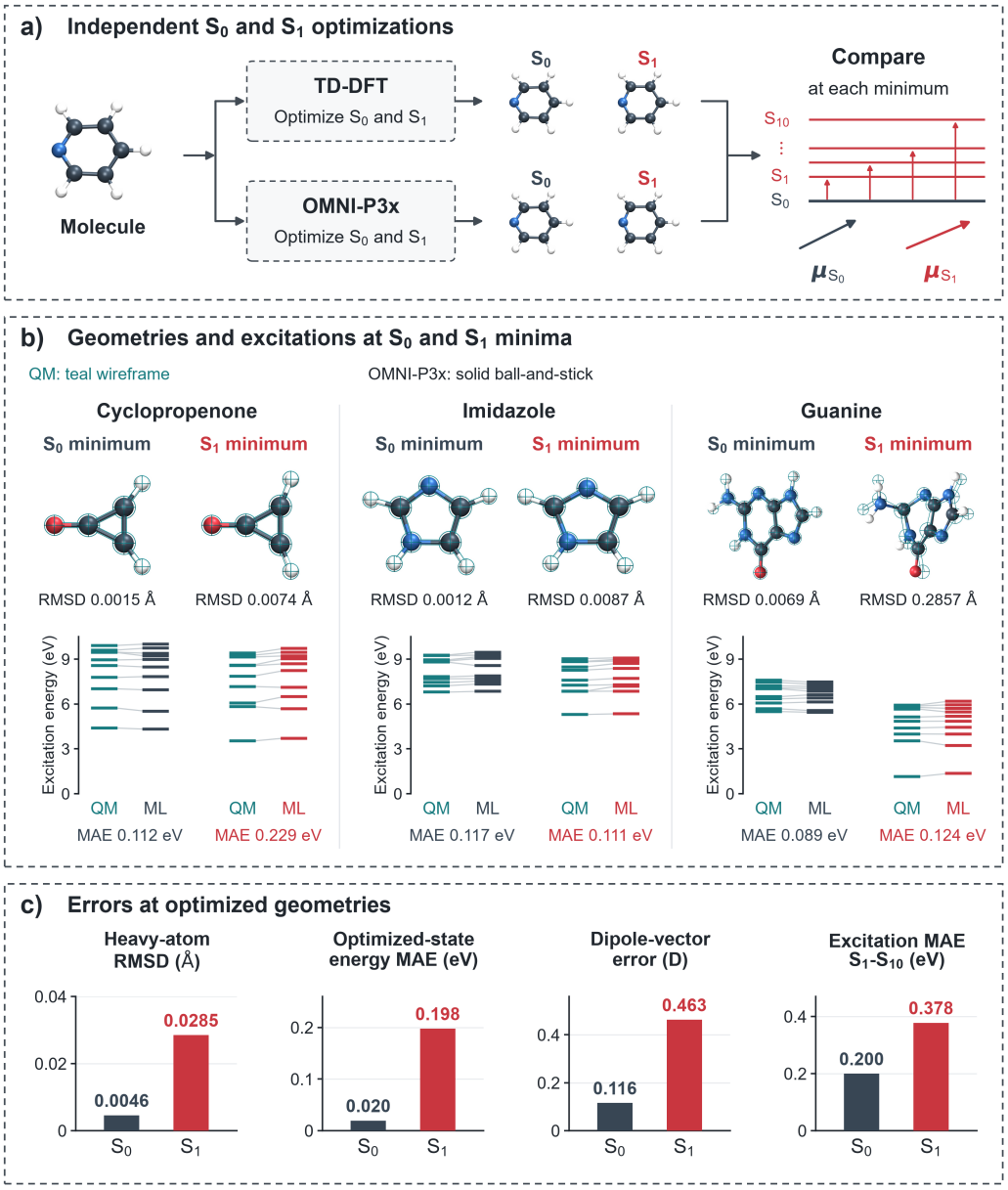



图5：OMNI-P3x在QUEST和Thiel基准测试中的几何优化与性质预测性能。a）基准测试示意图，包括使用TDA-$\omega$B97X-D和OMNI-P3x进行的$S_0$ 和 $S_1$ 优化，并比较所得极小值处的性质。b）基准测试中的示例分子，展示其机器学习优化几何（实心球棍模型），并叠加量子力学结果（青绿色线框），以及它们的垂直吸收光谱：OMNI-P3x（ $S_0$ 极小值为黑色， $S_1$极小值为红色）和TDA-DFT（黑色）。c）保留在各状态下的基准分子的优化极小值的平均重原子均方根偏差，以及优化态总能量的平均绝对误差、优化态偶极向量误差和前十个单重激发态的激发能平均绝对误差。黑色柱表示 $S_0$ 极小值处的统计数据，红色柱表示 $S_1$ 极小值处的统计数据。

尽管模型主要聚焦于激发态性质，但在基态能量学预测方面也保持了均衡且实用的精度。我们使用 GMTKN55 评估这一能力。GMTKN55 是一个广泛使用的基准数据库，包含 55 个子集，涵盖主族热化学、反应势垒高度和非共价相互作用，是评估基态方法精度的标准工具。[65] 测试集缩减为 35 个子集，仅包含 OMNI-P3x 完全支持的 12 种元素所涉及的闭壳层样例。我们采用文献 65 中定义的加权总平均绝对偏差（类型2）（WTMAD-2）指标量化性能。总体而言，OMNI-P3x@ωB97M-D3(BJ) 和 OMNI-P3x@TDA-ωB97X-D 的 WTMAD-2 值分别为 15.03 和 20.62 kcal·mol$^{-1}$。对于 679 个受支持的 CHNO反应，OMNI-P3x@CCSD(T)* 的 WTMAD-2 为 10.43 kcal·mol$^{-1}$ 。分类误差分解见图6a。基本性质和反应能的误差最低，而模型在非共价相互作用和反应势垒高度方面的表现明显较差。

图6b展示了 OMNI-P3x 与其他机器学习原子间势及基态量子化学方法的基态性能比较，为上述结果提供了背景。对于各自支持的反应集，OMNI-P3x@ωB97M-D3(BJ) 的 WTMAD-2 介于 AIMNet2 模型[28] 和 MACE-OFF23(L)[27] 之间，二者的数值分别为 11.66 和 19.53 kcal·mol$^{-1}$。它的表现也优于半经验 GFN2-xTB 方法（WTMAD-2 为 18.70 kcal/mol）[56] 以及 ωB97X-D 在较小的 def2-SVP 基组下的表现（WTMAD-2 为 16.92 kcal/mol）。采用更大基组的密度泛函理论参考结果仍显著更加准确。图6c展示了相同的比较，但仅限于涉及 CHNO元素的反应；这是 OMNI-P3x@CCSD(T)* 支持的子集。图6c列出了各方法在其支持的 CHNO反应上的表现。小基组参考结果 ωB97X-D/def2-SVP 在包含 35 个子集的 CHNO数据池上的 WTMAD-2 为 16.92 kcal·mol$^{-1}$ ，在 CHNO数据池上的 WTMAD-2 为 18.31 kcal·mol$^{-1}$ 。

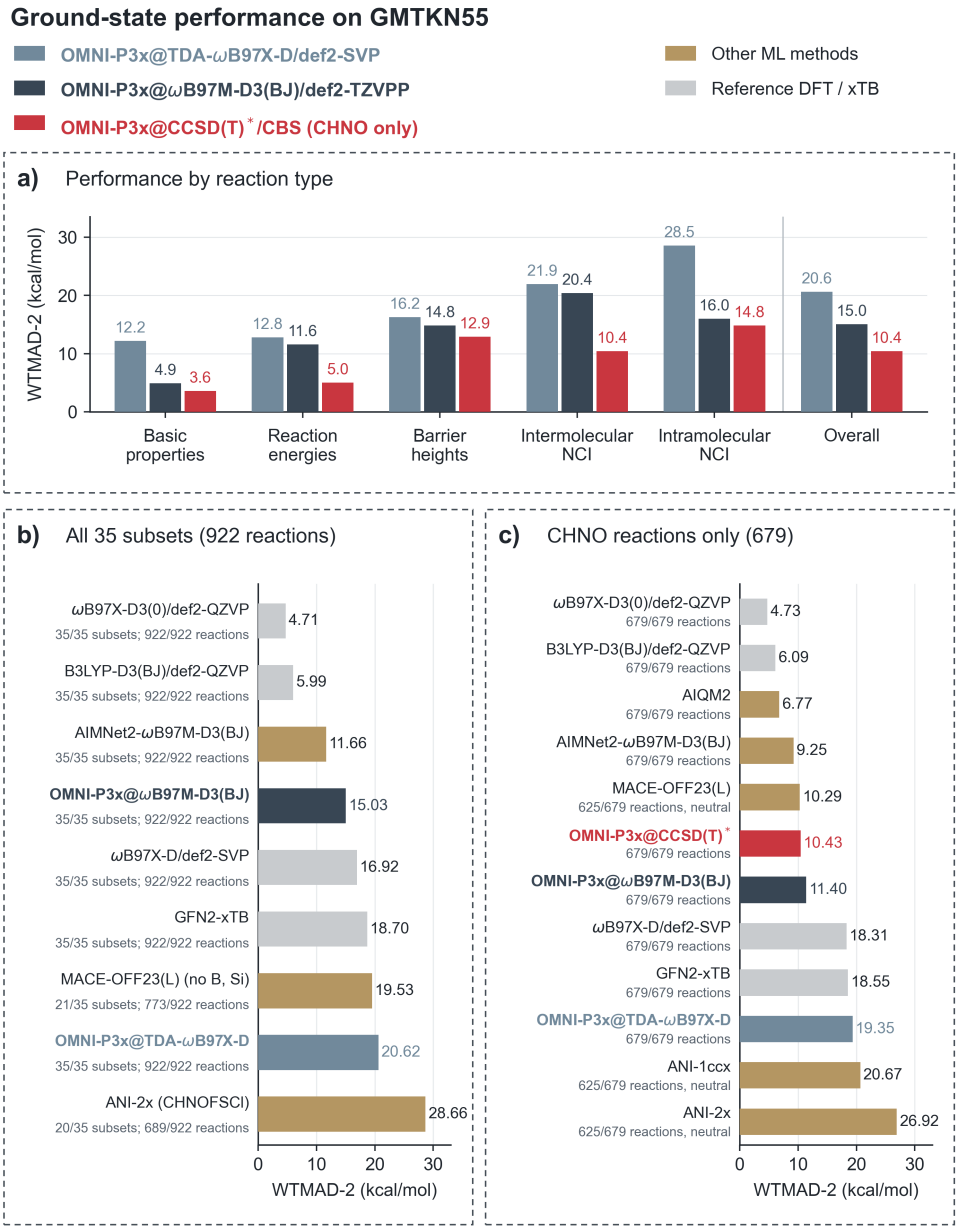



图6：OMNI-P3x 在其闭壳层 12元素训练数据集完全涵盖的 35 个 GMTKN55 子集上的基态性能。(a) 三种 OMNI-P3x 理论层级按类别划分的 WTMAD-2；每种密度泛函理论层级使用 922 个差异值，CCSD(T) 使用 679 个仅CHNO差异值。(b) 对同一 35 个子集范围内所有支持的反应进行比较。每种方法完全支持的子集和反应数量标注在其旁边。ωB97X-D3(0)/def2-QZVP 和 B3LYP-D3(BJ)/def2-QZVP 的数值取自 GMTKN55官方结果，[65] 其他所有结果均为本研究计算所得。ANI[29,66]和 MACE 仅在中性体系上进行评估。AIQM2 作为基线纳入比较。[67] (c) 对支持的 CHNO反应进行比较，并展示各方法的覆盖率；该反应池包含 679 个反应。

## 紫外-可见吸收光谱

接下来，我们测试 OMNI-P3x 能否很好地复现采用参考含时密度泛函理论计算进行的紫外-可见模拟。为此，我们从 Pub- ChemQC 测试集中随机选取 1,000 个未见过的结构，并使用 TDA-$\omega$B97X-D/def2-SVP 和 OMNI-P3x@TDA-$\omega$B97X-D 进行基态几何优化和单点光谱预测，如图7a所示（SPC 分支）。我们使用高斯函数对光谱进行展宽，展宽参数设为 0.3 eV （更多详情请参阅方法部分）。每种方法都会生成各自的极小值，因此该比较评估的是完整工作流程。图7b对比了机器学习和量子力学工作流程生成的光谱，并根据参考光谱与机器学习光谱之间的光谱重叠（见方法部分）从上到下按前 5% 排序。平均光谱重叠为 0.62；光谱重叠是一项严格的指标：仅 0.2 eV 的蓝移就会使得分降至 0.58，而强度减半或加倍则会使得分降至 0.50。OMNI-P3x 对跃迁能量的复现效果很好（平均绝对误差为 0.10 eV），优化结构也与参考结构吻合良好（重原子均方根偏差中位数为 0.09 Å）。光谱的整体形状通常都能得到良好复现，只有 5% 的分子光谱重叠低于 0.40，其余差异主要来自谱带强度。使用单个 GH200 GPU 运行 OMNI-P3x 完成一次模拟（包括优化）平均耗时 1.5 秒，而使用 16 个 CPU 核心运行参考方法则需 4.6 分钟，加速比达到 180 倍。

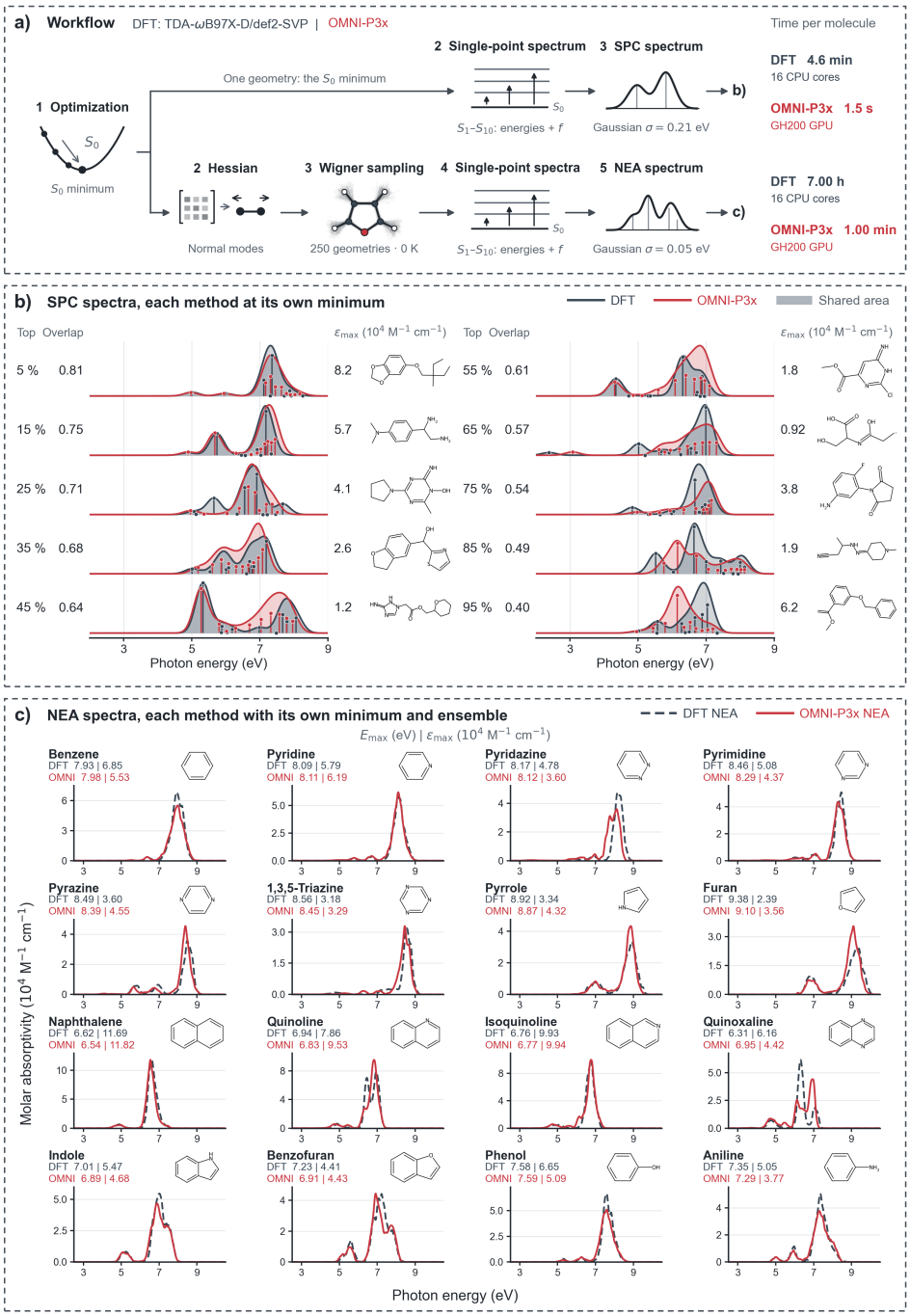



图7：使用 OMNI-P3x 模拟紫外-可见光谱。(a) 单点卷积（SPC）和核系综方法（NEA）光谱工作流程示意图，以及 OMNI-P3x 和参考含时密度泛函理论的计算成本。(b) PubChemQC数据集中测试分子的模拟 SPC 光谱，展示了按光谱重叠评分不同百分位排名随机选取的示例。(c) 16 种中等大小有机分子的模拟 NEA 光谱，其中参考含时密度泛函理论以黑线表示，OMNI-P3x 以红光线表示。

核系综方法（NEA）是一种模拟分子紫外-可见吸收光谱的方法，可将振动效应纳入光谱，从而再现峰分裂或振动允许跃迁等二阶特征。[68,69] 其缺点是需要对基态势能面进行大量采样，并对数百或数千个分子执行能量和振子强度的单点计算。机器学习能够加速此类模拟，[52, 70–72] 本文展示了 OMNI-P3x 无需针对特定任务进行微调即可用于这一目的。图 7a 展示了一个 NEA 工作流程，包括优化系统的基态几何结构、计算其频率（海森矩阵）、进行维格纳采样以获得集成、对所有集成成员进行单点计算，并汇总得到光谱。

我们在 16 种精选的中等大小芳香族和杂芳香族有机分子上测试 OMNI-P3x，计算其 NEA 光谱，并与参考含时密度泛函理论进行比较。每种方法都会生成各自优化得到的极小值、正规模和核系综，因此该比较评估了结构弛豫、频率计算和激发态性质预测的综合影响。光谱采用高斯函数展宽，展宽参数为 $\sigma$=0.05（见方法）。所得光谱如图7c所示。OMNI-P3x 对吸收峰值的再现效果非常好，光谱峰值之间的平均绝对差为 0.14 eV。在大多数情况下，峰的相对强度也得到了很好的再现，但喹喔啉除外，OMNI-P3x 错误地预测能量较高的峰更强。值得注意的是，使用参考方法在 16 个 CPU 核心上进行模拟，每个分子平均需要 7 小时；而使用 OMNI-P3x 在单个 GH200 GPU 上仅需 1 分钟，加速比达到 420 倍。这些结果表明，OMNI-P3x 能以极低的成本替代整个工作流程中的含时密度泛函理论计算，涵盖从几何结构到光谱的各个环节。

## 可解释激发态电荷

迄今讨论的所有性质都源自 OMNI-P3x 直接接受训练的量。神经网络部分电荷 $q_i^n$ 则有所不同：预训练模型不使用任何参考原子电荷，而是仅通过电荷所贡献的量间接学习电荷。这些量是永久偶极矩和能量的长程库仑项，两者都纳入了方程 1 的损失函数。先前研究表明，即使没有显式波函数或电子密度，学习得到的原子量也能揭示激发态特征，SchNarc[52] 和 OMNI-P2x.[48] 的研究便展示了这一点。在此，我们表明，神经电荷会随激发时电子密度的变化而变化，因此不同状态之间的电荷差 $\Delta q_i^n = q_i^n - q_i^0$能够直观地呈现电子在不同分子单元之间的移动方式。为此，我们将神经电荷与电子密度本身进行比较，方法是对每个状态下弛豫含时密度泛函理论差分密度在各原子上的积分。作为基准，我们采用 QUEST-6[73] 集中已发表几何结构下的分子内电荷转移（CT）激发数据，共包含 16 个不同体系中的 26 个电子跃迁，不包括 HCL 和两个二肽体系。CT 状态是一个自然的测试案例，因为它涉及电荷在定义明确的分子单元之间定向位移。由于原子电荷并无唯一的定义，我们主要比较这些单元上的电荷总和。

如图8a所示，神经电荷能够跟随 TD-DFT密度变化；对于从供体转移到受体的电荷量，OMNI-P3x 与 TD-DFT 的相关系数 R=0.91，且在 26 个状态中有 23 个状态的转移方向判断正确。与弱电荷转移特征的体系相比，强电荷转移态的结果更为一致。图8b所示的基态与 CT 状态之间的偶极矩变化，其平均绝对误差为 2.56 D，并且与 TD-DFT预测具有良好的相关性（R=0.91）。图8c展示了弱、中、强三种典型 CT 案例，分别为氨基苯甲腈中从氨基到氰基的电荷转移、硝基苯中从苯环到硝基的电荷转移，以及扭曲*N*-苯基吡咯中从吡咯到苯环的电荷转移。在所有案例中，OMNI-P3x 都正确识别了供体和受体，并预测了 CT 状态的定性特征，但系统性地低估了转移电荷量。因此，尽管预训练模型的神经电荷未能完整反映转移电荷量，却能很好地描述这些 CT 状态的定性特征，尽管其训练过程中没有使用任何电荷参考。

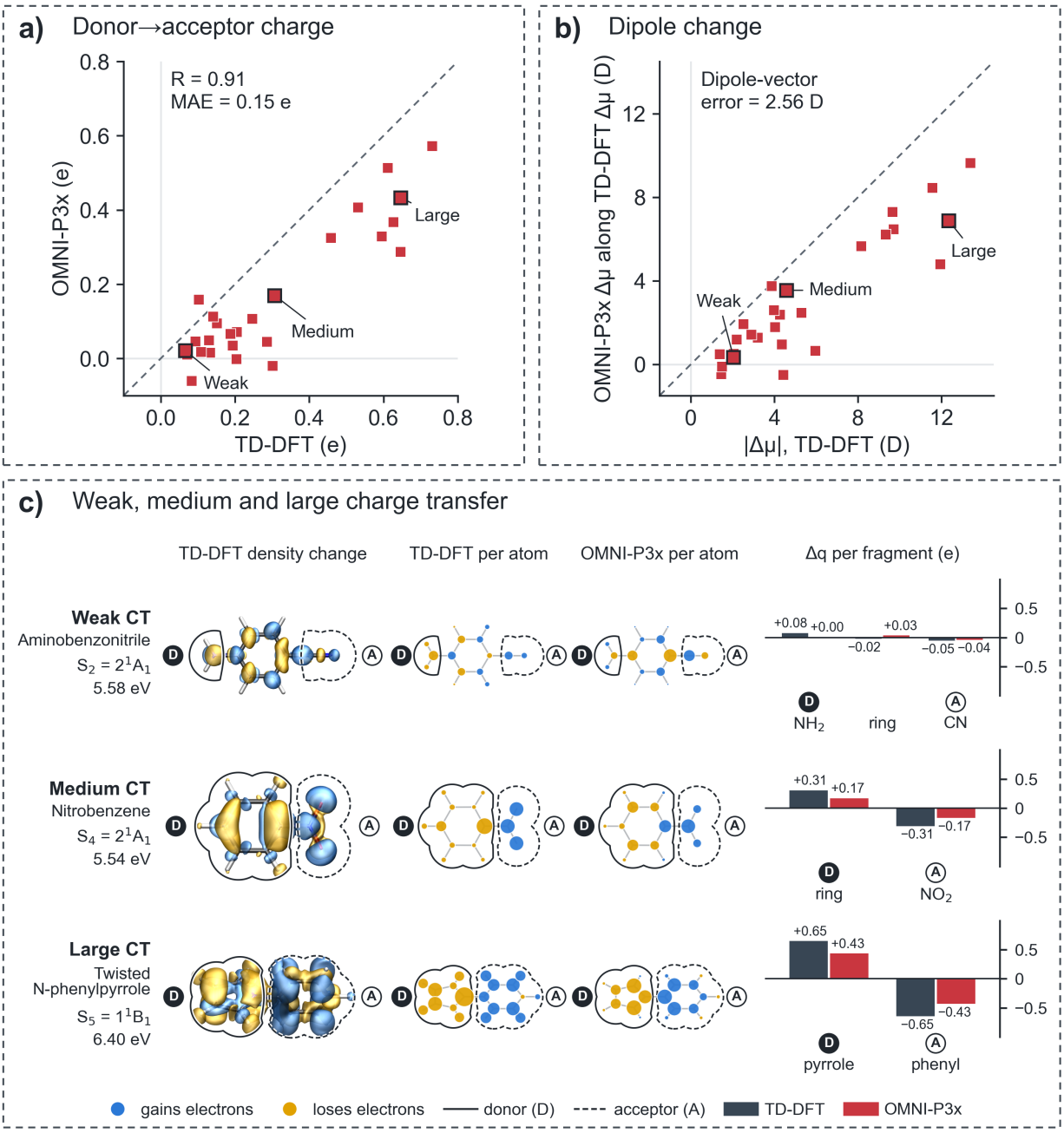



图 8：针对 QUEST 6 CT 数据集中电荷转移态的 OMNI-P3x 神经电荷与含时密度泛函理论的比较。a）从弛豫差分密度积分得到的供体向受体转移的电荷量，以及从 OMNI-P3x 原子电荷得到的电荷量。b）激发时偶极矩的变化，其中 OMNI-P3x 向量在含时密度泛函理论向量上的投影。c）具有弱、中等和强电荷转移的激发态示例：弛豫差分密度（等值面值为 0.002 a.u.），按原子积分后的相同数据及 OMNI-P3x 原子电荷的变化，以及供体和受体电荷变化的总和。

## 黄烷鎓染料的高通量筛选

高通量筛选是通用机器学习势的另一个特别有前景的应用领域，因为它们能够快速处理庞大的分子结构库。在这里，我们通过一个有机染料吸收性质筛选的示例，证明了 OMNI-P3x 可以直接用于此任务。为此，我们选取了黄烷鎓阳离子，它是众多天然产物染料（如花青素）的核心结构，而相关染料还可应用于光致变色材料和用于生物成像的荧光探针。[74,75]

图9a展示了该结构及所研究的取代模式，涵盖5种具有不同给受体特性的选择。利用 OMNI-P3x，我们得以研究涵盖234,375个体系的完整设计空间。重要的是，机器学习模型被用于整个工作流程：既包括基态极小值优化，也包括在该极小值处的紫外-可见光谱预测。与密度泛函理论计算相比，多层感知机将速度提升了500倍以上：计算一个基准染料的性质，在单个 GH200 GPU 上平均耗时4.7秒，而使用16个 CPU 核心进行时变密度泛函理论计算则几乎需要45分钟。总体而言，完成整个设计空间的筛选耗时435 GPU 小时；若使用密度泛函理论，则约需270万 CPU 小时。

OMNI-P3x 的性能在图9b中进行了评估，测试集包含509个测试分子，其中包括从设计空间中随机选取的500个体系，以及基于黄烷鎓阳离子的9个已知参考化合物。令人鼓舞的是，就 $S_0/S_1$ 能隙而言，OMNI-P3x 达到0.17 eV的平均绝对误差；在预测染料的最低吸收峰方面，其平均绝对误差为41 nm（带符号误差中位数为8 nm）。优化结构与参考结构吻合良好，重原子均方根偏差的中位数仅为0.170 °A。光谱结果同样显示出极佳的一致性：对于给定体系，参考含时密度泛函理论与 OMNI-P3x 模拟结果之间的斯皮尔曼相关系数（SCC）中位数为0.96。需要特别指出的是，这些指标反映的是完整工作流程之间的差异，比较的是两种方法所得的优化最终结果和光谱模拟最终结果。

对黄烷鎓阳离子完整组合库的分析，使我们能够直接了解每个取代位点上各取代基的影响，以及各取代位点的相对重要性；后者以该位点所解释的 $S_0/S_1$ 跃迁能方差占比衡量。具体而言，8、 $3'/5'$ （由于体系具有对称性，二者合并分析）、7 和 4’ 位最为重要，其余三个位置对体系最终 $S_0/S_1$ 能隙的影响较小。就具体取代基而言，在 8 位和 $3'/5'$ 位引入电子给体，由于它们与氧上的正电荷共轭，会产生显著红移；其中，二甲氨基的作用尤为强烈。相反，在 $4'$ 位和 7 位引入供体会导致蓝移。吸电子的氰基官能团平均会降低体系的吸收能。对于 $3'/5'$ 位和 8 位，二甲氨基取代会产生显著红移；而在 7 位和 $4'$ 位，这种效应则会反转。

图9d展示了所研究染料的模拟单点卷积吸收光谱。上排将 OMNI-P3x 预测值与三种已知天然产物——天竺葵素、矢车菊素和飞燕草素——的含时密度泛函理论结果进行了比较，显示多层感知机与参考结果高度一致。此外，我们的筛选成功识别出新型黄烷鎓衍生物，其吸收显著红移。对于 OMNI-P3x 预测的激发能最低且取代基不超过 4 个的 100 种候选物，

我们进行了参考含时密度泛函理论/ωB97X-D 计算。图9d 下排展示了三个示例结构及其光谱。

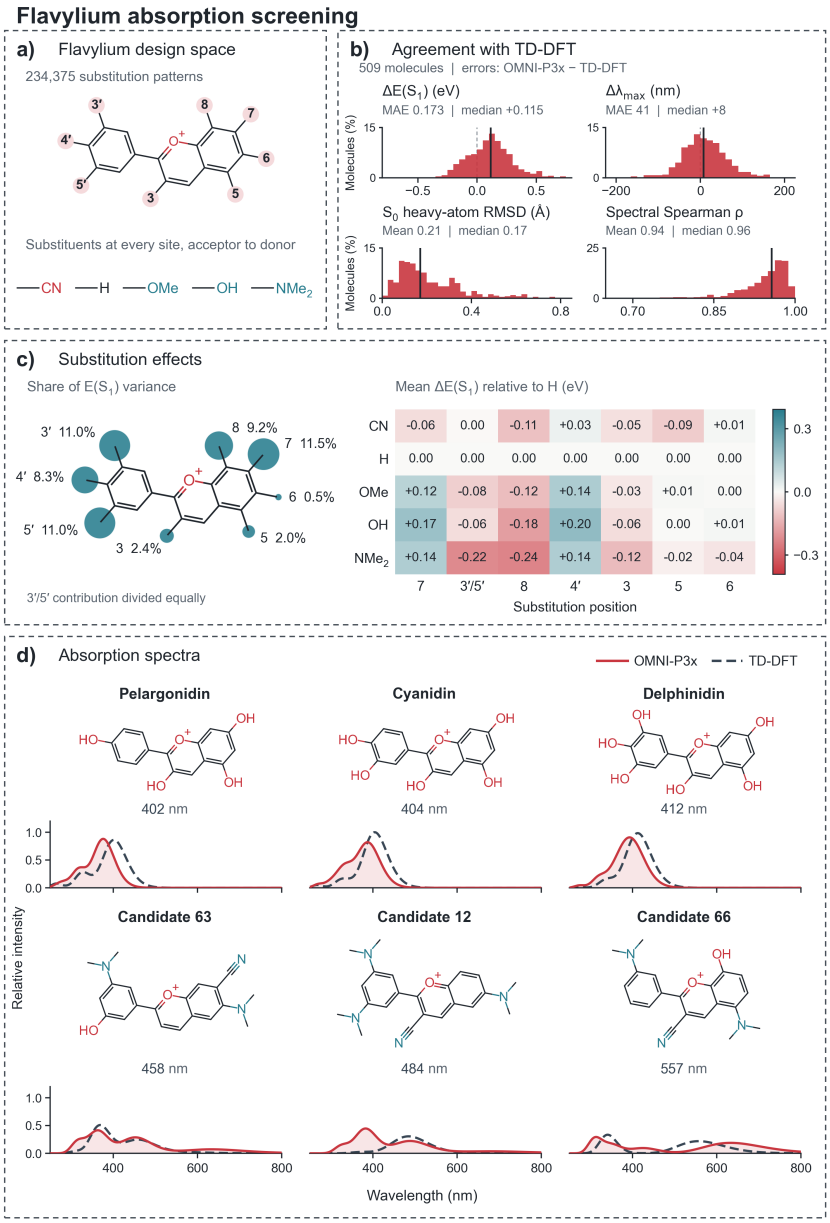



图9：使用 OMNI-P3x 筛选黄烷鎓设计空间。面板 a 显示分子骨架、8 个可能的取代位点以及筛选的 5 种取代基。面板 b 展示 OMNI-P3x 与参考含时密度泛函理论筛选结果的一致性，依据 $S_0/S_1$ 能隙误差、第一吸收峰的波长、重原子均方根偏差和斯皮尔曼秩相关系数进行描述。面板 c 展示取代基对 $S_0/S_1$ 能隙的影响，即各位点所解释的方差比例（见“方法”）。圆的面积与各位点所占的方差比例成正比。右侧展示各位置上每种取代基的影响，以相对于氢取代体系的 $S_0/S_1$能隙平均变化量进行量化。面板 d 展示模拟光谱：上方为三种天然产物花色素苷元，下方为三个新的候选体系；实线红色表示 OMNI-P3x，黑色虚线表示含时密度泛函理论。

## 基于呫吨的发射体的高通量筛选

在证明 OMNI-P3x 开箱即用即可用于可靠筛选吸收性质后，我们转而开展更具挑战性的发射性质预测任务。除了优化体系的基态结构外，还需要在其 $S_1$ PES 上定位极小值，并计算该处的激发态性质。为此，我们选取了呫吨鎓阳离子；它是派洛宁、罗丹明和罗丹醇染料中常见的分子骨架，并广泛应用于荧光标记[76] 和生物成像等领域。[77,78]

开发具有定制光物理性质的新型衍生物仍是一个活跃的研究领域，这主要受到活细胞成像和超分辨率成像进展的推动。[79,80] 图10a 展示了所选的取代模式及其涵盖的设计空间；该模式基于成熟的合成文献，表明发射性质可以在整个可见光和近红外范围内进行调节。[81–85] 具体修改包括在桥位进行杂原子取代（10）、改变$meso$ （9）位的官能团、在助色基团（3,6）位引入电子给体，以及在其余环位（2,4,5,7）进行给体/受体取代。两个半部独立进行取代，共得到 162,960 种候选染料。

为提高模型预测的质量，我们使用筛选分子中的一个小型子集对 OMNI-P3x 模型进行了微调。这些分子是随机抽取的，并在参考理论水平（$\omega$B97X-D/def2-SVP/TDA）下进行标注；OMNI-P3x 预训练所用的激发态数据也采用了这一理论水平。采样点来自 $S_1$ 优化轨迹中的帧、极小值附近的正则模采样、机器学习与含时密度泛函理论预测结构之间的插值，以及单点计算。微调集的构建详情见方法部分。最终数据集涵盖约 3.2 万个构象。重要的是，仅需对约 1,000 个分子执行成本高昂的 $S_1$优化流程，不到筛选体系的 1%。

在包含 513 种染料的独立测试集上（其中 10 种因在含时密度泛函理论或机器学习理论水平下优化最终停留在 $S_0/S_1$ 交叉点而被排除），微调后的OMNI-P3x实现了 0.21 eV 的发射能量平均绝对误差，同时能很好地再现参考 $S_1$ 极小值几何结构，中位均方根偏差为 0.23 $^{\circ}$A，如图10b所示。与吸收基准测试类似，此比较针对各方法优化后的几何结构进行，衡量完整工作流程所得结果的差异。单张 GPU 上对每种染料进行完整计算平均耗时 12.6 秒；相比之下，使用 16 个 CPU 核心进行密度泛函理论计算则需 5 小时以上，速度约提升 1,500 倍。对整个设计空间进行发射筛选总计耗费 570 GPU 小时；若采用含时密度泛函理论，则需约 1,400 万 CPU 小时。我们根据达到极小值时 $S_1$ 状态的性质对筛选结果分类：振子强度 $\geq 0.05$ 的标记为发射体，低于该值的则标记为暗态。假定 $S_0$/$S_1$ 能隙低于 0.5 eV 的分子可能发生向基态的内转换，因此将其排除在荧光分析之外。另有 2,746 个分子因优化到达 $S_0$/$S_1$ 交叉点而被排除在分析之外；在此处，参考含时密度泛函理论的结果并不可靠。[86] 被分类为发射体的体系中，吸收波长与发射波长之间的关系如图10c所示。通过拟合以取代基身份作为分类输入的加性模型，确定 (3,6) 助色基团位置对 $S_1$的发射特性影响最强。通过省略某个取代位点组并重新拟合模型后，评估各组的贡献，即考察 $R^2$ 的下降幅度

（见方法）。如图10d所示，两个 -$NH_2$ 基团对应的预测发射体比例最高，达 95%；而 N-甲基苯胺基（N(Me)Ph）会显著抑制发射。在预测为发射体的类别中，桥连取代和 meso 位取代是发射能的最强预测因素。具体而言，$SO_2$ 和 P(=O)Ph 桥连基团，以及强吸电子取代基（$CF_3$ 和 CN基团）会产生红移幅度最大的荧光，如图10e所示。

此外，我们还筛选出了3个在红光、近红外（NIR）和短波红外（SWIR）波段具有良好发射性质的候选体系，如f面板所示。为验证含时密度泛函理论（TD-DFT）结果，我们使用经TD-DFT优化的几何结构，在ADC(2)水平上重新计算了相关性质。在23个已知实验结果的体系上进行测试后，该工作流程预测的发射能与实验发射波长符合得非常好（平均绝对误差为0.1 eV），而TD-DFT则系统性地高估了这些发射能，平均有符号误差为0.9 eV。因此，我们将OMNI-P3x 预测值平移0.89 eV后筛选候选项；该偏移量是从随机抽取的100个发光体样本中，在完成两个理论水平计算的98个发光体上得到的ADC(2)/TD-DFT平均差值。有关校准以及与实验结果进行验证的流程详情，请参见补充信息中的S1节。在平移后的尺度上，我们定义了三个窗口：红光（600–700 nm）、近红外（700–1000 nm）和短波红外（1000–1300 nm）。每个窗口根据预测的发射振子强度选取5个分子，每个桥接基团最多选取2个，并使用ADC(2)对TD-DFT优化的 $S_1$ 几何结构进行单点计算重新评估（表S2）。ADC(2)得到的发射波长可能落在原始筛选窗口之外。

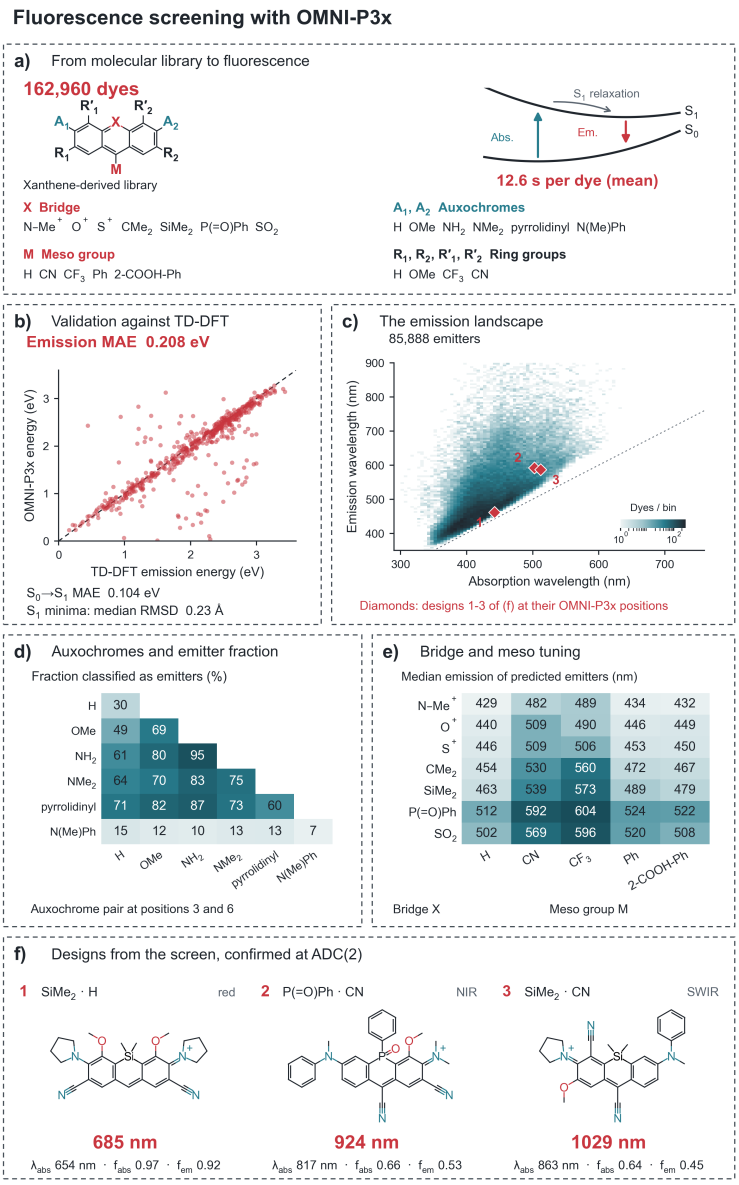



图10：使用微调后的OMNI-P3x对基于砧吨的分子库进行荧光筛选。c–e面板中的波长是OMNI-P3x@TDA-ωB97X-D在参考含时密度泛函理论尺度上的预测值，该尺度比ADC(2)和实验值高约0.9–1.0 eV。（a）分子库设计和荧光工作流程示意图。（b）测试染料的发射能，比较机器学习与含时密度泛函理论在各自弛豫几何结构下的结果。（c）筛选体系中预测的吸收-发射关系。626个发射波长超过900 nm的发光体位于图示范围之外。带编号的菱形标记出f面板中三个设计在其预测位置上的位置。（d）对于每个助色团对，被分类为发光体的比例，相对于具有该助色团对且成功完成分析的所有结构。（e）对于每种桥基-介观组合，分类发光体的发射波长中位数。（f）三个选定设计及其ADC(2)/def2-SVP发射波长（较大字号的红色数字）、吸收波长和振子强度。

# 结论

我们介绍了 OMNI-P3x，这是一种通用机器学习原子间势，可用于模拟中性和带电有机分子的基态及单重激发态。该模型将 AIMNet2 架构与多状态和一体化学习相结合，并通过扩大元素覆盖范围以及新生成的基态和激发态构象数据集，拓展了其适用性。与前代模型 OMNI-P2x 相比，该模型的准确度有所提高，无论是复现参考方法，还是拟合采用更准确的密度泛函理论泛函采样的训练数据，均是如此。该模型能够完成典型光化学工作流程所需的所有步骤，例如在单一模型中模拟吸收光谱（包含或不包含振动效应）和发射光谱、进行基态及激发态几何优化、开展频率计算，以及预测光学性质；其准确度接近 TDA-$\omega$B97X-D。我们通过比较完整工作流程的结果对其进行基准测试，证明了它可作为含时密度泛函理论建模的替代模型。除了预测能量学和光学性质，OMNI-P3x 还通过学习得到的、特定于状态的神经网络部分电荷，揭示了激发时电子电荷的重新分布通过。我们表明，尽管训练数据中没有参考原子电荷，这些电荷仍能再现分子内电荷转移的特征和方向。

我们通过两种高通量筛选工作流程展示了新模型的能力。首先，我们使用预训练的 OMNI-P3x 模型筛选包含 234,375 种黄烷鎓染料的组合库，在不进行针对特定任务的微调的情况下，完成基态几何优化并预测吸收光谱。由此，我们能够识别具有红移吸收的衍生物，并量化取代基种类及其位置对光学性质的影响。随后，我们将微调后的 OMNI-P3x 用于筛选 162,960 种基于氧杂蒽的候选发射体，其中包括弛豫至第一激发态极小值。筛选结果揭示了哪些结构修饰会影响该状态的发射特性及预测的发射波长。通过校准以及后续的 ADC(2)计算验证，我们得以选出有望实现红光、近红外和短波红外发射的候选项。

与此同时，模型的准确性仍受训练数据的限制。底层的 TDA-TD-DFT 无法可靠地描述多参考体系[86]，而且在描述某些体系和激发类型时存在系统性误差。由于计算成本的限制，采用了规模适中的 def2-SVP基组，因此对里德伯激发的描述仍不理想。然而，本文提出的一体化学习方案提供了一条直接途径，可在合适的训练数据可用时纳入额外且更准确的参考层级。

# 方法

## 量子化学计算

所有激发态参考数据均采用范围分离杂化泛函$\omega$B97X-D[87]，在 Tamm–Dancoff近似（TDA）下，[88] 使用 def2-SVP基组[89] 和 RI-J近似[90,91]，并通过 TURBOMOLE 7.9 实现计算。[92] 对于每个

结构，我们计算了闭壳层单重基态（$S_0$）和最低的十个单重激发态的能量，以及全部十个 $S_0 \rightarrow S_n$ 激发的跃迁偶极矩，进而计算其振子强度。能量梯度始终针对 $S_0$ 和 $S_1$ 计算；只要高阶激发态的梯度能够纳入每个结构 10 分钟的固定计算时间预算，就会将其加入计算，该预算根据基于基函数数量拟合的成本模型估算。$S_0$ 以及所有已计算梯度的激发态均有偶极矩数据。所有计算均由 MLatom[93,94]统一编排。

图 4 中展示的 ADC(2) 结果使用 TUR-BOMOLE 7.$9,^{92}$ 的 ricc2 模块[95] 计算，采用 aug-cc-pVTZ 基组[96,97] 、带匹配辅助基组的单位分解近似[98] 以及冻结核心近似。与比较中的其他所有方法一样，单点计算均在已发表的 QUEST 几何结构上进行，未施加对称性约束，并将最低单重态激发取为 $S_1$。对于每个分子，均要求计算 20 个单重态激发；如果根数较少，求解器可能会漏掉被弥散基组引入的里德伯态掩盖的最低价层 $\pi \rightarrow \pi^*$ 态。由此得到的全部 40 个分子的 $S_0 \rightarrow S_1$ 激发能与 QUEST[61,62] 中报告的 ADC(2)/aug-cc-pVTZ 数值一致，差异在 0.001 eV 以内。

## 数据采样

初始结构取自 Joung等人[54] 建立的实验表征发色团数据库。去重后，将 4,657 个独特分子从 SMILES 转换为三维结构，并使用 RDKit[55] 进行优化。通过 GFN2-xTB 元动力学[99] 在 300 K 下对基态构象进行采样，时间步长为 1 fs，持续 10 ps，每 200 fs 保存一个结构，从而为每个分子获得 50 个构象体。使用 AM1 哈密顿量[100] （MNDO 程序[101]）在 $S_1$ 表面上通过元动力学对激发态构象进行采样，采用相同的偏置参数，最长持续 2 ps；当 $S_1$–$S_0$ 能隙降至 0.3 eV 以下时，轨迹即停止。基态元动力学传播使用 xtb 程序，而 $S_1$ 元动力学则使用 MLatom 程序中的内部实现。

对于后续采样，不确定性量化为两个模型之间的分歧：一个是在能量和梯度上训练的主模型，以及一个仅在能量上训练的辅助模型。[47,57] 候选结构根据这种分歧进行排序，并选取其中不确定性最高的结构进行标注。为避免采样到无关且化学上不合理的结构，排序最靠前的 1% 被剔除。

使用基于发色团数据训练的初始模型，按不确定性量化（UQ）值对 Pub- Che mQC[102] 中的分子进行排序。选取不确定性最高的 10,000 个分子，并通过 GFN2-xTB 元动力学为每个分子生成 20 个构象体。由此，训练集规模增至 515,528 个结构。利用基于该数据集训练的模型进行不确定性筛选，进一步选取了 100,000 个 PubChemQC 数据点，并在其平衡几何结构下为其标注；其中 73,243 个被纳入数据集。进一步重新训练后，按照 ANI-1ccx协议通过正则模采样生成非平衡 $S_0$ 和 $S_1$ 结构。[60] 使用 GeomeTRIC 分别优化每个分子在 $S_0$和 $S_1$ 中的结构；[103] 利用模型计算相应状态的解析 Hessian 矩阵，并通过 PySCF[104] 对其进行对角化，以获得谐正则模。对于

每个分子及其状态，通过沿所有正规模位移原子，生成五种构象，使总谐能 $(3/2)\,N_{\mathrm{at}}k_{\mathrm{B}}T$ ($N_{\mathrm{at}}$ 个原子，$T = 300$ K) 随机分配至各个模，最多可产生一百万个候选项。根据中间模型主版本与辅助版本之间各状态能量差异的最大值对这些候选项进行排序，并对不确定性最高的候选项进行标注；其中 23,102 个被纳入数据集。

进一步从 AIMNet2 训练集中抽取候选几何结构，[28] 并根据基于截至当时收集的全部数据训练的模型进行 UQ 排名，选取不确定性最高的 500,000 个几何结构并进行标注，其中 436,530 个进入数据集。最后，为提升模型在小分子上的性能，将 QM9 数据集[59]（133,885 个分子）在 TDA-DFT/$\omega$B97X-D/def2-SVP 理论水平下重新标注，并加入训练数据，同时保留 10% 作为测试集。

## 模型训练

OMNI-P3x 是由四个模型组成的集成，这些模型采用相同的训练流程，唯一差异在于训练数据的随机划分方式。新采样的 TDA-DFT 和 ANI-1ccx 训练集被随机划分为子训练集（90%）和验证集（10%），每个集成成员使用不同的随机划分。对于 AIMNet2 重放集，留出 20,000 个随机结构用于验证（约 1%），其余数据用于训练。所有模型均从已发表的 AIMNet2 $\omega$B97M-D3/def2-TZVPP 参数初始化，[28] 作用于新的状态和理论水平输入的权重设为零，振子强度输出头则随机初始化。通过对能量与元素计数进行线性回归，为每个理论水平重新拟合自原子能；拟合前会剔除位于第 2–98 百分位范围之外的离群值。

模型使用 RAdam 优化器[105]（权重衰减 $10^{-8}$）进行训练，在两个节点上运行，每个节点配备四个 GH200 GPU，每个 GPU 的批次大小为 512 条数据。上述设备上每个集成成员的训练耗时为 30–37 小时。学习率在五个训练轮次的预热阶段线性提升至$3 \times 10^{-4}$；此后，如果验证损失连续 10 个训练轮次未改善，则将学习率乘以 0.75，直至降至低于 $10^{-5}$，这一过程耗用了 314–401 个训练轮次；最终保留验证损失最低时的参数。OMNI-P3x 预测值是四个集成成员的能量、梯度、偶极矩和振子强度的平均值。

## 紫外-可见吸收光谱

紫外/可见吸收光谱采用标准的单点卷积近似获得，该近似基于文献 68 推导的展宽函数：

$$L(x) = 0.619 \sum_i \frac{f_i}{\Gamma} e^{-(x-x_i)^2/\Gamma^2}, \tag{2}$$

其中，$x_i$ 是预测的跃迁能量，$f_i$ 是相应的振子强度，$\Gamma$ 是预先设定的高斯展宽参数，设为 0.3 eV。

用于评估参考光谱与机器学习光谱之间相似度的光谱重叠定义如下：

$$O = \frac{\int \min\left[L^{\mathrm{ref}}(x), L^{\mathrm{ML}}(x)\right] \mathrm{d}x}{\int \max\left[L^{\mathrm{ref}}(x), L^{\mathrm{ML}}(x)\right] \mathrm{d}x}, \tag{3}$$

其中，$L^{\mathrm{ref}}(x)$ 和 $L^{\mathrm{ML}}(x)$ 分别是公式 2 中以光子能量 $x$ 为自变量的参考光谱和机器学习光谱。$O$ 对于完全相同的光谱，其值为 1；对于没有重叠面积的光谱，其值为 0。

NEA 光谱根据每个分子的 250 个几何结构计算得出，这些结构是在 $S_0$ 几何优化和谐性正规模分析后，通过 0 K 下的维格纳采样生成的。采用 OMNI-P3x 和 TDA-$\omega$ B97X-D/def2-SVP 分别独立执行该工作流程，每种方法均使用其自身优化得到的极小值和正规模。在每个采样几何结构上计算最低的十个单重激发态的激发能和振子强度，并使用 MLatom 中的 NEA 实现进行汇总，[93,94] 同时采用标准差为 $\sigma = 0.05$ eV 的高斯展宽。

## 黄烷鎓染料的高通量筛选

通过枚举黄烷鎓骨架上的取代基（H、CN、OMe、OH 和 $NMe_2$），生成初始分子结构。位置 3、5、6、7、8 和 $4'$ 分别独立变化，而对称等价的 $3'/5'$ 位置则以 15 种无序取代基对表示。由此得到 $N = 5^6 \times 15 = 234,375$ 种不同的取代模式。使用 RDKit 生成三维结构。

使用 OMNI-P3x 在 TDA-$\omega$B97X-D 理论级别描述符下进行基态优化。优化使用 GeomeTRIC 执行。[103] 针对每个优化后的几何结构，计算了前四个单重激发态的激发能和振子强度。参考密度泛函理论计算采用了与数据生成步骤相同的设置。

紫外/可见吸收光谱采用与图 7 所示基准相同的设置和方法获得。第一个吸收峰定义为能量最低、且强度至少达到光谱最大强度 2% 的局部最大值。如果不存在符合条件的局部最大值，则使用全局最大值。

图9b 中误差统计所依据的基准由随机选取的 500 种衍生物和九种参考化合物构成。所有误差均以 OMNI-P3x 预测值与参考值之间的差异计算，所用几何结构由各方法独立优化得到。光谱一致性通过光谱之间的斯皮尔曼秩相关系数进行量化，定义如下：

$$r_{\mathrm{SCC}} = 1 - \frac{6}{n(n^2-1)} \sum_i d_i^2, \tag{4}$$

其中，$d_i$ 是两个光谱之间的秩差，$n$ 是光谱的总长度。

为评估筛选候选物，从最多含有四个非氢取代基的结构中，选取预测的 $S_0 \to S_1$ 激发能最低的 100 个分子。随后，在相应的 DFT 优化几何结构上，使用 TDA-$\omega$ B97X-D/def2-SVP 对这些候选项进行评估。

取代效应的统计分析采用组合库中预测的垂直 $S_0 \to S_1$ 激发能。将六个独立变化的位置视为各有五个水平的分类因子，并将联合 $3'/5'$ 对视为具有 15 个水平的因子。每种取代模式的权重相同。因子 $j$ 解释的方差占比计算如下：

$$P_j = \frac{\sum_\ell n_{j\ell}\left(\overline{E}_{j\ell} - \overline{E}\right)^2}{\sum_{i=1}^{N}\left(E_i - \overline{E}\right)^2} \times 100\%, \tag{5}$$

其中，$\overline{E}_{j\ell}$ 是因子 $j$在水平 $\ell$ 下的平均激发能，$n_{j\ell}$ 是具有该水平的模式数量，$E$ 是总体平均值。该分析用于描述枚举库中的方差，其中56%可由此解释，其余44%归因于取代基间相互作用。为便于可视化，联合 $3'/5'$ 因子的贡献在两个位置之间平均分配。取代基引起的能量位移计算为：给定取代基的平均 $S_0/S_1$ 跃迁能与H对应平均值之差，并对其他位置的所有取代基进行平均。对于汇总的$3'/5'$ 比较，将在任一位置含有该取代基的分子与在任一位置含有H的分子进行比较。对于缺失的23个预测值（其计算失败），采用用于方差分解的迭代加性模型进行估算；取代基平均值仅使用计算成功的数据。

## 基于氧杂蒽的发射体的高通量筛选

为在发射筛选工作流程中对OMNI-P3x进行微调，从设计空间中选取了两组分子；根据规范 SMILES，这两组分子彼此不重叠，且均不属于测试集。第一组的目的在于提升模型描述所研究染料的$S_1$ 极小值的能力。该组包含1,000个分子，并按照筛选中采用的相同四步工作流程进行计算：$S_0$ 优化、垂直激发（三个状态）、$S_1$优化（两个状态）以及在 $S_1$ 极小值处的发射。通过对TD-DFT $S_1$ 极小值进行正则模式位移，进一步扩大了数据覆盖率；计算在50 K和200 K下进行，并使用OMNI-P3x频率。为改善对尤其困难的系统类别的描述，对于317个分子，当基线OMNI-P3x模型的 $S_1$ 极小值与TD-DFT结果相差超过 0.3 $^\circ$Å时，在参考极小值与机器学习极小值之间取线性插值点，每个分子取6个。对于 $S_1$ 能量高于该分子TD-DFT $S_1$ 极小值超过3 eV的构象，则从训练集中剔除。

第二个数据集旨在提高模型的化学空间覆盖度。该数据集由随机抽取的 15k 个体系组成；对这些体系进行了成本较低的 OMNI-P3x 基线优化，随后在所得几何结构上进行了一次含时密度泛函理论计算，以求得四个最低激发及其振子强度，但未计算能量梯度。完整的微调集包含 31765 个构象。

微调过程在经过五个训练轮次的预热后，采用了较低的学习率 $5 \times 10^{-5}$。与预训练时相同的学习率衰减调度器被用于驱动优化过程，直至学习率降至 $5 \times 10^{-6}$ 以下。能量、力、振子强度和偶极矩的损失权重分别设为 1、1、0.5 和 0.1。为

防止灾难性遗忘，从原始 OMNI-P3x 在各个理论层级的训练数据中随机抽取 1%，作为回放数据加入训练集；同时，每个训练轮次将微调数据呈现两次，使其约占训练样本的一半。使用了 OMNI-P3x 的一个集成成员（编号 0）。

ADC(2)计算采用了与QUEST基准中相同的ricc2程序设置，并在 $S_1$ 极小值处使用 TDA-$\omega$B97X-D/def2-SVP $S_0$进行单点计算，采用def2-SVP基组。[89]

### 激发态电荷

含时密度泛函理论计算采用了训练数据的理论水平，在已发表的QUEST几何结构（CC3或CCSD(T)）上进行单点计算，且不考虑对称性。每个QUEST跃迁均根据跃迁偶极矩的方向，分配给具有相同不可约表示、且与TBE/aug-cc-pVTZ值相差不超过 $\pm 0.8$eV的含时密度泛函理论状态。在有多个状态符合条件的情况下（26个跃迁中有4个属于此情况），选择偶极矩变化最大的状态。OMNI-P3x在相同几何结构上使用TDA-$\omega$B97X-D理论级别描述符进行评估，并读取与所分配含时密度泛函理论状态相同的状态索引。

在 OMNI-P3x 中，原子电荷的变化直接取自模型输出，即对四个集成成员中每个原子在状态 $S_n$ 和 $S_0$ 下的神经电荷之差取平均：

$$\Delta q_i^n = q_i^n - q_i^0 \tag{6}$$

对于含时密度泛函理论，将弛豫差分密度 $\Delta\rho^n(\mathbf{r}) = \rho^n(\mathbf{r}) - \rho^0(\mathbf{r})$ 在原子元胞上积分：

$$\Delta q_i^{n,\mathrm{ref}} = -\int_{V_i} \Delta\rho^n(\mathbf{r})\,\mathrm{d}\mathbf{r} \approx -\Delta V \sum_{k:\,\mathbf{r}_k \in V_i} \Delta\rho^n(\mathbf{r}_k), \tag{7}$$

其中，$V_i$ 是原子 $i$的 Voronoi元胞，即所有比起其他原子核更靠近该原子核的点；$\mathbf{r}_k$ 表示网格中的点，$\Delta V$ 是单个网格元胞的体积，由此得到可进行比较的逐原子电荷变化。

在每个体系中，均识别出供体（失去电子，正电荷增加）和受体（获得电子，负电荷增加）片段，并可能存在连接二者的桥联单元。通过对这些片段求和，计算从供体转移至受体的电荷：

$$q^n_{\mathrm{D}\to\mathrm{A}} = \frac{1}{2}\left(\sum_{i\in\mathrm{D}} \Delta q_i^n - \sum_{i\in\mathrm{A}} \Delta q_i^n\right). \tag{8}$$

## 代码和数据可用性

OMNI-P3x数据集可在 `https://doi.org/10.6084/m9.figshare.34937758`获取，此外还提供支持研究结果的数据：黄烷鎓和氧杂蒽筛选的结果，包括微调数据和模型。

该代码以 MIT许可证 在开源程序 MLatom 中提供，

详见 https://github.com/dralgroup/mlatom，模型权重托管在 Zenodo 上：https://doi.org/10.5281/zenodo.24344107由程序自动下载。在 MLatom 中运行 OMNI-P3x 计算的教程见 http://mlatom.com/docs/tutorial_omnip3x.html。用户也可以通过我们的智能体系统 Protomia 使用自然语言指令运行这些计算，[106] 它是 Ait-omia 的后继系统。[107] Protomia 可在线通过 Aitomistic Hub[108] 和 Aitomistic Lab@XMU 使用。[109]用户也可以在本地通过 MLatom Skills 库运行这些计算。[110]

# 作者贡献

M.M. 设计并开发了用于训练 OMNI-P3x 模型的代码，生成了训练数据，设计并执行了文中所述的基准测试，构思并实施了文中所述的筛选流程，还撰写了论文初稿并制作了图表。J.J. 和 P.O.D. 参与了结果分析与解读，共同监督了这项研究，并修订了论文。

# 致谢

P.O.D. 感谢中国国家自然科学基金委员会国际资深科学家项目（项目编号：W2531013）以及新一代人工智能国家科技重大项目 2026ZD0128501 提供的资助。我们感谢波兰高性能计算基础设施 PLGrid 授予本项目使用 LUMI 超级计算机的权限。该超级计算机由欧洲高性能计算联合 undertaking 所有，由 CSC（芬兰）和 LUMI 联盟托管，项目编号为 PLL/2026/10/019484。本研究得到 PIAST-AI Factory 项目——PIAST AIF 的支持，该项目由欧洲高性能计算联合 undertaking 通过欧盟数字欧洲计划以及作为参与国的波兰共同资助，资助协议编号为 101250730。本文所表达的观点和意见仅代表作者本人，并不一定反映欧盟、欧洲高性能计算联合 undertaking 或波兰的观点。欧盟、欧洲高性能计算联合 undertaking 和波兰均不对此承担责任。

M.M. 和 J.J. 感谢波兰国家科学中心通过“PRELUDIUM”项目（资助编号：2025/57/N/ST4/03587）为本研究提供资助，并衷心感谢波兰高性能计算基础设施 PLGrid（HPC 中心：ACK Cyfronet AGH）在计算资助项目 PLG/2026/019355 下提供计算设施与支持。

作者感谢在本手稿准备过程中使用大语言模型，应用于编码、数值实验设计和数据分析，以及图表制作和手稿校对。所使用的模型包括：Claude Opus 4.6- 5.5 和 Fable 5.0-5.2通过 Claude Code 工具，以及 GPT-6 Astra通过 Codex 工具。所有科学内容和最终文稿均已由作者进行全面审阅和批准。

# 参考文献

(1) Mori, T. 塑造双螺烯圆偏振发光的振电耦合的重要性。《德国应用化学国际版》**2024年**, *63*, e202319702。

(2) Palombo, R.; Barneschi, L.; Pedraza-Gonz´alez, L.; Padula, D.; Schapiro, I.; Olivucci, M. 视黄醛发色团的电荷离域与限域解释了新视紫红质极端的光物理特性。*Nature Communications* **2022**,*13*, 6652.

(3) Chibani, S.; Le Guennic, B.; Charaf-Eddin, A.; Laurent, A. D.; Jacquemin, D. 借助从头算工具重新审视 BODIPY 的光学特征。《化学科学》**2013**,*4*, 1950–1963.

(4) Adrion, D. M.; Karunaratne, W. V.; Lopez, S. A. 多组态光动力学模拟揭示了环丙烯酮在显式水环境中的光脱羰基化机制。《化学科学》**2023**, *14*, 13205–13218.

(5) Barbatti,M.; Aquino,A.J.A.; Szymczak,J.J.; Nachtigallov´a,D.; Hobza,P.; Lis- chka, H. 紫外光激发DNA和RNA核碱基的弛豫机制。《美国国家科学院院刊》**2010年**,*107*, 21453–21458。

(6) Curtolo, F.; Dong, S. S. 非天然光酶促自由基反应中化学选择性、对映选择性和底物选择性同时产生的分子起源。《美国化学会志》**2025年**, *147*,41639–41649。

(7) Feng,Z.; Guo, W.; Kong,W.-Y.; Chen, D.; Wang,S.; Tantillo,D.J.光化学反应与基态过渡态后分叉之间的类比揭示了选择性动力学起源。《自然·化学》**2024年**,*16*, 615–623。

(8) Sumita, M.; Terayama, K.; Suzuki, N.; Ishihara, S.; Tamura, R.; Chahal, M. K.; Payne, D. T.; Yoshizoe, K.; Tsuda, K. 基于量子化学计算和机器学习，从头创制肉眼可检测的荧光分子。《科学进展》**2022年**, *8*。

(9) Wang, C.; Qiao, Q.; Chi, W.; Chen, J.; Liu, W.; Tan, D.; McKechnie, S.; Lyu, D.; Jiang, X.; Zhou, W.; Xu, N.; Zhang, Q.; Xu, Z.; Liu, X. 通过准确预测扭曲分子内电荷转移（TICT），定量设计高亮度荧光团和聚集诱导发光分子。《德国应用化学国际版》**2020年**, *59*, 10160–10172。

(10) Gerstner, N. C.; McCann, J. T.; Martin, J. G.; Henn, K. M.; Riske, K.; Anantakrishnan, S.; Graham, T. G.; Darzacq, X.; Miller, E. W. 双（三氟甲基）碳硼罗丹明：用于活细胞荧光显微镜、基于活性的传感和单分子显微镜的高荧光性远红光至近红外染料。《美国化学会志》**2025年**, *147*, 21950–21960。

(11) Hartinger, M.; Dube, H.; M¨uller, C. 从激发态动力学到产物选择性：关于氮杂二芳基乙烯光诱导化学的理论见解。《物理化学化学物理》2026年, 28, 15728–15738. (12) Hartinger, M.; Herm, M.; Sch¨ußlbauer, C.; K¨ottner, L.; Guldi, D.; Dube, H.; M¨uller, C. 从三重态到扭转：近红外光开关peri-蒽硫靛的光化学E/Z异构化路径。《德国应用化学国际版》2025年, 64. (13) Chen, Z.; Zhao, W.; Gong, X.; Zong, S.; Li, J.; Chen, P.; Sun, W.; Du, J.; Fan, J.; Peng, X. 面向窄带纯紫色有机发光二极管的多重共振分子生成式筛选。《先进功能材料》2025年, 36. (14) Aizawa, N.; Pu, Y.-J.; Harabuchi, Y.; Nihonyanagi, A.; Ibuka, R.; Inuzuka, H.; Dhara, B.; Koyama, Y.; Nakayama, K.-i.; Maeda, S.; Araoka, F.; Miyajima, D. 来自倒置单重态和三重态激发态的延迟荧光。Nature 2022年, 609, 502–506. (15) Omar, O. H.; Xie, X.; Troisi, A.; Padula, D. 通过高通量虚拟筛选鉴定未知的倒置单重态–三重态核心。《美国化学会志》2023年, 145, 19790–19799. (16) Gillett, A. J. et al. 有机太阳能电池中电荷复合形成三重态激子的作用。Nature 2021年, 597, 666–671. (17) Hart, L. J. F.; Medranda, D. G.; Yuan, S. W.; Lindh, L.; M¨uller, J. S.; Yang, H.; Gerard, H.; Zhao, T.; Quesada-Ramirez, A.; Campoy-Quiles, M.; Azzouzi, M.; Eis- ner, F. D.; Nelson, J. 控制有机光伏材料中电荷对生成的分子因素。《自然·材料》2026年, 25, 1209–1218. (18) Ganguly, G.; Gonzalez, L. 揭示用于分子太阳能热能储存的第8族富瓦烯桥联双金属配合物的光异构化机制。《美国化学会志》2025年, 147, 41855–41866. (19) Ganguly, G.; Gonzalez, L. 富瓦烯桥联双金属分子太阳能热能燃料的预测性设计。《JACS Au》2026年, 6, 5648–5662. (20) Hillers-Bendtsen, A. E.; Elholm, J. L.; Obel, O. B.; H¨olzel, H.; Moth-Poulsen, K.; Mikkelsen, K. V. 在双环二烯化学空间中寻找分子太阳能热能储存候选项。《德国应用化学国际版》2023年, 62. (21) Runge, E.; Gross, E. K. U. 含时系统的密度泛函理论。《物理评论快报》1984年, 52, 997–1000. (22) Casida, M. E. 收录于《密度泛函方法最新进展》第一部分；Chong, D. P., 编；世界科学出版社：新加坡，1995年；第155–192页。

(23) Schirmer, J. 超越随机相位近似：一种新的极化传播子近似方案。《Physical Review A》1982年，26，2395–2416。（24）Stanton, J. F.; Bartlett, R. J. 运动方程耦合簇方法。分子激发能、跃迁概率和激发态性质的系统性双正交方法。《The Journal of Chemical Physics》1993年，98，7029–7039。（25）H´egely, B.; Szirmai, ´A. B.; Mester, D.; Tajti, A.; Szalay, P. G.; K´allay, M. 激发态多层方法的性能。《The Journal of Physical Chemistry A》2022年，126，6548–6557。（26）Wu, X.; Tegner, B.; Troisi, A. 有机电子学的高通量虚拟筛选：从能级到高级性质与物理洞见。《Chemical Physics Reviews》2026年，7，021304。（27）Kov´acs, D. P.; Moore, J. H.; Browning, N. J.; Batatia, I.; Horton, J. T.; Pu, Y.; Kapil, V.; Witt, W. C.; Magd˘au, I.-B.; Cole, D. J.; Cs´anyi, G. MACE-OFF：面向有机分子的短程可迁移机器学习力场。《Journal of the American Chemical Society》2025年，147，17598–17611。（28）Anstine, D. M.; Zubatyuk, R.; Isayev, O. AIMNet2：满足中性、带电、有机及元素有机体系需求的神经网络势。《Chemical Science》2025年，16，10228–10244。（29）Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1：具有 DFT 精度且计算成本达到力场水平的可扩展神经网络势。《Chemical Science》2017年，8，3192–3203。（30）Wood, B. M. 等。UMA：原子通用模型家族。Advances in Neural Information Processing Systems。2025年；第143528–143564页。（31）Yuan, E. C.-Y.; Liu, Y.; Chen, J.; Zhong, P.; Raja, S.; Kreiman, T.; Vargas, S.; Xu, W.; Head-Gordon, M.; Yang, C.; Blau, S. M.; Cheng, B.; Krishnapriyan, A.; Head-Gordon, T. 用于化学与材料原子模拟的基础模型。Nature Reviews Chemistry 2026年，10，212–230。（32）Li, J.; Reiser, P.; Boswell, B. R.; Eberhard, A.; Burns, N. Z.; Friederich, P.; Lopez, S. A. 利用纳秒机器学习光动力学模拟自动发现光异构化机制。《Chemical Science》2021年，12，5302–5314。（33）Mausenberger, S.; M¨uller, C.; Tkatchenko, A.; Marquetand, P.; Gonz´alez, L.; Wester-mayr, J. SPAINN：用于激发态非绝热分子动力学的等变消息传递。《Chemical Science》2024年，15，15880–15890。（34）Westermayr, J.; Gastegger, M.; Marquetand, P. 结合 SchNet 与 SHARC：用于激发态动力学的 SchNarc 机器学习方法。《The Journal of Physical Chemistry Letters》2020年，11，3828–3834，PubMed唯一标识符：32311258。

（35）Westermayr, J.; Marquetand, P. 分子的电子激发态机器学习。《Chemical Reviews 》2021年，121，9873–9926。（36）Westermayr, J.; Gastegger, M.; V¨or¨os, D.; Panzenboeck, L.; Joerg, F.; Gonz´alez, L.; Marquetand, P. 酪氨酸的深度学习研究表明，漫游可能导致光损伤。《Nature Chemistry 》2022年，14，914–919。（37）Li, J.; Reiser, P.; Boswell, B. R.; Eberhard, A.; Burns, N. Z.; Friederich, P.; Lopez, S. A. 利用纳秒机器学习光动力学模拟自动发现光异构化机制。《Chemical Science 》2021年，12，5302–5314。（38）Hu, D.; Xie, Y.; Li, X.; Li, L.; Lan, Z. 在即时非绝热分子动力学模拟中引入机器学习核岭回归势能面。《The Journal of Physical Chemistry Letters 》2018年，9，2725–2732。（39）Axelrod, S.; Shakhnovich, E.; G´omez-Bombarelli, R. 利用化学可迁移机器学习势研究大型光开关分子的激发态非绝热动力学。《Nature Communications 》2022年，13。（40）Axelrod, S.; Shakhnovich, E.; G´omez-Bombarelli, R. 偶氮苯衍生物的热半衰期：基于系间窜越和机器学习势的虚拟筛选。《ACS中心科学 》2023年，9，166–176。（41）Tiefenbacher, M. X.; Dietschreit, J. C. B.; Axelrod, S.; Goomez-Bombarelli, R.; Gon-zalez, L. 调控芳基偶氮吡唑光开关热寿命的设计规则。《ACS中心科学 》2026年，12，1246–1257。（42）Barrett, R.; Ortner, C.; Westermayr, J. 使用集成DeepSets 的激发态可迁移机器学习势 X-MACE。《arXiv预印本 arXiv:2502.12870》2025年，（43）Barrett, R.; Wesely, S.; Westermayr, J. 可迁移的激发态动力学实现荧光蛋白发色团的筛选。2026年； https://arxiv.org/abs/2604.12699。（44）Friederich, P. 面向通用激发态机器学习势：用于尺寸一致性与可迁移性的局域激发哈密顿量。2026年； https://arxiv.org/abs/2611.11019。（45）Juergens, D.; St¨ohr, M.; Hillers-Bendtsen, A. E.; Fajen, O. J.; Mart´ınez, T. J. 面向激发态化学的潜在统一光滑哈密顿量。2026年； https://arxiv.org/abs/2609.01871。（46）Malosso, C.; How, W. B.; D´ıaz Mir´on, G.; Hassanali, A.; Ceriotti, M. 利用极值池化实现激发态动力学的可迁移机器学习。2026年； https://arxiv.org/abs/2606.16859。

(47) Martyka, M.; Zhang, L.; Ge, F.; Hou, Y.-F.; Jankowska, J.; Barbatti, M.; Dral, P. O. 借助多态学习和通过高效、稳健的主动学习实现的能隙驱动动力学，绘制跨分子的电子态流形图谱。*npj Computational Materials***2025年**，*11*。(48) Martyka, M.; Tong, X.-Y.; Jankowska, J.; Dral, P. O. 用于激发态模拟的 OMNI-P2x 通用神经网络势。*Nature Communications* **2026年**，*17*。(49) Chen, Y.; Dral, P. O. 一统天下：跨量子化学层级学习的通用原子间势。*Journal of Chemical Theory and Computation***2025年**，*21*，8762–8772。(50) Alnoman, R. B.; Nabil, E.; Parveen, S.; Hagar, M.; Zakaria, M.; Hasanein, A. A. 用于透明染料敏化太阳能电池共敏化的有机紫外染料的合成与计算表征。*Molecules* **2021年**，*26*，7336。(51) Uoyama, H.; Goushi, K.; Shizu, K.; Nomura, H.; Adachi, C. 基于延迟荧光的高效有机发光二极管。*Nature* **2012年**，*492*，234–238。(52) Westermayr, J.; Marquetand, P. 使用 SchNarc 深度学习紫外吸收光谱：迈向化学化合物空间可迁移性的第一步。*The Journalof Chemical Physics* **2020年**，*153*，154112。(53) Dreuw, A. 计算光化学为何具有挑战性且可能始终如此：一位量子化学家的视角。*Advanced Science* **2026年**，*13*，e21012。(54) Joung, J. F.; Han, M.; Jeong, M.; Park, S. 有机化合物光学性质实验数据库。*Scientific Data* **2020年**，*7*，295。(55) RDKit：开源化学信息学工具。`https://www.rdkit.org`。(56) Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB——一种精确且参数覆盖广泛的自洽紧束缚量子化学方法，包含多极静电作用和密度依赖色散贡献。*Journal ofChemical Theory and Computation* **2019年**，*15*，1652–1671。(57) Hou, Y.-F.; Zhang, L.; Zhang, Q.; Ge, F.; Dral, P. O. 利用物理信息主动学习加速量子化学模拟。*Journal of Chemical Theory andComputation* **2024年**，(58) Nakata, M.; Shimazaki, T. PubChemQC 项目：面向数据驱动化学的大规模第一性原理电子结构数据库。*J. Chem. Inf. Model.***2017年**，*57*，1300–1308，doi：10.1021/acs.jcim.7b00083。(59) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. 134,000个分子的量子化学结构与性质。*Scientific Data* **2014年**，*1*，140022。(60) Smith, J. S.; Zubatyuk, R.; Nebgen, B.; Lubbers, N.; Barros, K.; Roitberg, A. E.; Isayev, O.; Tretiak, S. ANI-1ccx 和 ANI-1x 数据集：分子的耦合簇与密度泛函理论性质。*Scientific Data* **2020年**，*7*，134。

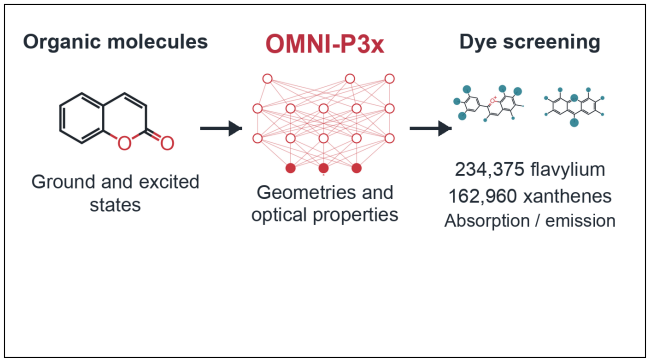
(61) Loos, P.-F.; Lipparini, F.; Boggio-Pasqua, M.; Scemama, A.; Jacquemin, D. 激发态的登山策略：中等大小分子的高精度能量与基准测试。*Journal of Chemical Theory and Computation* **2020年**，*16*，1711–1741。(62) V´eril, M.; Scemama, A.; Caffarel, M.; Lipparini, F.; Boggio-Pasqua, M.; Jacquemin, D.; Loos, P.-F. QUESTDB：面向电子结构研究领域的高精度激发能数据库。*WIREs Computational Molecular Science***2021年**，*11*，e1517。(63) Koch, H.; Christiansen, O.; Jørgensen, P.; Sanchez de Mer´as, A. M.; Helgaker, T. CC3模型：一种包含连接三重激发的迭代耦合簇方法。*TheJournal of Chemical Physics* **1997**，*106*，1808–1818。(64) Tuna, D.; Lu, Y.; Koslowski, A.; Thiel, W. 半经验量子化学正交化校正方法：电子激发态基准测试。*Journal of Chemical Theory and Computation* **2016**，*12*，4400–4422。(65) Goerigk, L.; Hansen, A.; Bauer, C.; Ehrlich, S.; Najibi, A.; Grimme. 借助先进的 GMTKN55 数据库审视密度泛函理论“动物园”：通用主族热化学、动力学和非共价相互作用。*Physical Chem-istry Chemical Physics* **2017年**，*19*，32184–32215。(66) Devereux, C.; Smith, J. S.; Huddleston, K. K.; Barros, K.; Zubatyuk, R.; Isayev, O.; Roitberg, A. E. 将 ANI 深度学习分子势的适用范围扩展至硫和卤素。*Journal of Chemical Theory and Computation* **2020年**，*16*，4192–4202。(67) Chen, Y.; Dral, P. O. AIQM2：超越密度泛函理论的有机反应模拟。*ChemicalScience* **2025年**，*16*，15901–15912。(68) Bai, S.; Mansour, R.; Stojanovi´c, L.; Toldo, J. M.; Barbatti, M. 分子光吸收中垂直激发与带峰值之间偏移的起源。《分子建模杂志》**2020年**，*26*，107。(69) Crespo-Otero, R.; Barbatti, M.*Marco Antonio Chaer Nascimento*；Springer Berlin Heidelberg，2012年；第89–102页。(70) Xue, B.-X.; Barbatti, M.; Dral, P. O. 机器学习吸收截面。*The Journal of Physical Chemistry A* **2020年**，*124*，7199–7210，PubMed唯一标识符：32786977。(71) Cerd´an, L.; Roca-Sanju´an, D. 使用概率机器学习重建核系综方法电子光谱。*Journal of Chemical Theoryand Computation* **2022年**，*18*，3052–3064，PubMed唯一标识符：35481363。(72) Dral, P. O.; Barbatti, M. 通过机器学习视角研究分子激发态。*Nature Reviews Chemistry* **2021年**，*5*，388–405。

(73) Loos, P.-F.; Comin, M.; Blase, X.; Jacquemin, D. 分子内电荷转移激发的参考能量。*Journal of Chemical Theory and Computation* **2021年**,*17*, 3666–3686. (74) Cruz, L.; Bas´ılio, N.; Mateus, N.; de Freitas, V.; Pina, F. 天然和合成黄酮鎓基染料：色彩背后的化学。《化学评论》 **2022年**,*122*, 1416–1481. (75) Sun, Y.; Sun, P.; Li, Z.; Qu, L.; Guo, W. 受天然黄酮鎓启发的远红光至近红外-II染料及其作为生物医学传感荧光探针的应用。*ChemicalSociety Reviews* **2022年**,*51*, 7170–7205. (76) K¨uhn, S.; Nasufovic, V.; Wilhelm, J.; Kompa, J.; de Lange, E. M. F.; Lin, Y.-H.; Egoldt, C.; Fischer, J.; Lennoi, A.; Tarnawski, M.; Reinstein, J.; Vlijm, R.; Hiblot, J.; Johnsson, K. SNAP-tag2：实现更快速、更明亮的蛋白质标记。*Nature ChemicalBiology* **2025年**, *21*, 1754–1761. (77) Farrants, H. 等. 用于多重体内功能成像的模块化化学遗传学钙指示剂。*Nature Methods* **2024年**, *21*, 1916–1925. (78) Fiore, A.; Yu, G.; Northey, J. J.; Patel, R.; Ravenscroft, T. A.; Ikegami, R.; Kolk- man, W.; Kumar, P.; Dilan, T. L.; Ruetten, V. M. S.; Ahrens, M. B.; Shroff, H.; Wang, S.; Weaver, V. M.; Pedram, K. 使用聚糖结合荧光团对细胞外基质进行活体成像。*Nature Methods* **2025年**, *22*, 1070–1080. (79) Reja, S. I.; Hori, Y.; Takeda, Y.; Nishiura, M.; Minoshima, M.; Kikuchi, K. 利用五元杂环调控分子内旋转，促进高细胞渗透性氧杂蒽基荧光生成探针的设计。《美国化学会志》 **2025年**, *147*, 47997–48012. (80) Holland, K. L. 等. 一系列用于超分辨显微镜的自发闪烁染料。*Nature Methods* **2026年**, *23*, 909–913. (81) Wang, L.; Du, W.; Hu, Z.; Uvdal, K.; Li, L.; Huang, W. 可见光/近红外区域的杂化罗丹明荧光团及其在生物成像中的应用。*Angew. Chem. Int. Ed.***2019年**,*58*, 14026–14043. (82) Lavis, L. D. 让老染料焕发新用途：以荧光素和罗丹明为基础构建生物探针。*Annu. Rev. Biochem.***2017年**, *86*, 825–843. (83) Koide, Y.; Urano, Y.; Hanaoka, K.; Terai, T.; Nagano, T. 第14族罗丹明的发展：利用光诱导电子转移构建近红外荧光探针平台。*ACS Chem. Biol.* **2011年**,*6*, 600–608. (84) Grimm, J. B.; English, B. P.; Chen, J.; Slaughter, J. P.; Zhang, Z.; Revyakin, A.; Patel, R.; Macklin, J. J.; Normanno, D.; Singer, R. H.; Lionnet, T.; Lavis, L. D. 一种改进活细胞和单分子显微镜用荧光团的通用方法。*Nat. Methods* **2015年**,*12*, 244–250.

(85) Grimm, J. B.; Tkachuk, A. N.; Xie, L.; Choi, H.; Mohar, B.; Falco, N.; Schaefer, K.; Patel, R.; Zheng, Q.; Liu, Z.; Lippincott-Schwartz, J.; Brown, T. A.; Lavis, L. D. 一种优化和功能化红移罗丹明染料的通用方法。*Nat.Methods* **2020年**,*17*, 815–821. (86) Levine, B. G.; Ko, C.; Quenneville, J.; Mart´ınez, T. J. 含时密度泛函理论中的圆锥交叉与双激发。*Molecular Physics* **2006年**, *104*, 1039–1051. (87) Chai, J.-D.; Head-Gordon, M. 采用阻尼原子间色散校正的长程校正杂化密度泛函。《物理化学化学物理》**2008年**,*10*, 6615–6620. (88) Hirata, S.; Head-Gordon, M. Tamm–Dancoff近似下的含时密度泛函理论。*Chemical Physics Letters* **1999年**,*314*, 291–299. (89) Weigend, F.; Ahlrichs, R. 适用于从氢到氡、具有分裂价层、三ζ价层和四ζ价层精度的平衡基组：设计与精度评估。《物理化学化学物理》 **2005年**, *7*, 3297–3305.¨(90) Eichkorn, K.; Treutler, O.; Ohm, H.; H¨aser, M.; Ahlrichs, R. 用于近似库仑势的辅助基组。*Chemical Physics Letters* **1995年**, *240*, 283–290. (91) Weigend, F. 适用于从氢到氡的高精度库仑拟合基组。《物理化学化学物理》 **2006年**,*8*, 1057–1065. (92) TURBOMOLE V7.9 2024，由卡尔斯鲁厄大学和卡尔斯鲁厄研究中心有限公司开发，1989–2026年，TURBOMOLE有限公司，自2007年起。获取地址：`http://www.turbomole.com`，2024年。(93) Dral, P. O. 等. MLatom：用于原子尺度模拟的机器学习软件包，3.17.0版。厦门大学，中国厦门，2013–2024年，`http://mlatom.com`，2024年。(94) Dral, P. O. 等. MLatom 3：机器学习增强型计算化学模拟与工作流程平台。*Journal of Chemical Theory and Com-putation* **2024年**, *20*, 1193–1213. (95) H¨attig, C. 使用相关二阶方法对激发态进行结构优化：CC2和ADC(2)。*Advances in Quantum Chemistry* **2005年**, *50*, 37–60. (96) Dunning, T. H., Jr. 相关分子计算所用的高斯基组。I. 硼至氖原子和氢原子。*The Journal of Chemical Physics*1989**年**,*90*, 1007–1023. (97) Kendall, R. A.; Dunning, T. H., Jr.; Harrison, R. J. 重新审视第一周期原子的电子亲和能：系统基组与波函数。*The Journal of ChemicalPhysics* **1992年**, *96*, 6796–6806.

(98) Weigend, F.; K¨ohn, A.; H¨attig, C. 在恒等分辨 MP2 计算中高效使用相关一致基组。*The Journal of Chemical Physics* **2002年**,*116*, 3175–3183. (99) Grimme, S. 基于紧束缚量子化学计算，通过元动力学模拟探索化合物、构象异构体和反应空间。*Journal of Chemical Theory and Computation* **2019年**,*15*, 2847–2862. (100) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. 量子力学分子模型的开发与应用。76. AM1：一种新型通用量子力学分子模型。《美国化学会志》 **1985年**, *107*, 3902–3909. (101) Thiel, W. MNDO2020：一种半经验量子化学程序。马克斯·普朗克煤炭研究所, 2020年；参与贡献者包括 M. Beck、S. Billeter、R. Kevorkiants、M. Kolb、A. Koslowski、S. Patchkovskii、A. Turner、E.-U. Wallenborn、W. Weber、L. Sp¨orkel 和 P. O. Dral。(102) Nakata, M.; Shimazaki, T. PubChemQC 项目：面向数据驱动化学的大规模第一性原理电子结构数据库。*Journal of Chemical Informa-tion and Modeling* **2017年**, *57*, 1300–1308. (103) Wang, L.-P.; Song, C. 利用平移和旋转坐标简化几何优化。*The Journal of Chemical Physics* **2016年**,*144*, 214108. (104) Sun, Q. 等。PySCF 程序包的最新进展。*The Journal ofChemical Physics* **2020年**,*153*, 024109. (105) Liu, L.; Jiang, H.; He, P.; Chen, W.; Liu, X.; Gao, J.; Han, J. 关于自适应学习率的方差及其他问题。国际学习表征会议。2020年。(106) Protomia：面向计算化学的自主研究智能体。 `https://aitomistic.com/protomia`，2026年。(107) Hu, J.; Nawaz, H.; Hou, Y.-F.; Chi, L.; Tong, X.-Y.; Rui, Y.; Chen, Y.; Ullah, A.; Dral, P. O. Aitomia：用于人工智能驱动的原子级与量子化学模拟的智能体框架。*J. Chem. Theory Comput.* **2026年**, *22*, 6557–6571. (108) Aitomistic Hub：用于 Aitomistic 模拟的按需在线平台。 `https://www.aitomistic.xyz`，2025–2026年；访问日期：2026-11-24。(109) Aitomistic Lab@XMU：厦门大学的 Aitomistic 模拟平台。`https://atom.xmu.edu.cn`，2025–2026年；访问日期：2026-11-24。(110) Dral, P. O. MLatom 技能。 `https://github.com/dralgroup/mlatom-skills`，2026年。

# 目录图


Organic molecules
OMNI-P3x
Dye screening
Ground and excited states
Geometries and optical properties
234,375 flavylium
162,960 xanthenes
Absorption / emission