arXiv:2601.00503v1 [physics.chem-ph] 1 Jan 2026

用于人工传感材料性质预测的可解释量子信息机器学习

LiChen[1], Leonardo Medrano Sandonas[1,*], Shirong Huang[1], Alexander Croy[2], 和 Gianaurelio Cuniberti[1,3,4,5,*]

[1]德累斯顿工业大学材料科学研究所和马克斯·贝格曼生物材料中心，德国德累斯顿 01062[2]耶拿弗里德里希·席勒大学物理化学研究所，德国耶拿 07737[3]德累斯顿工业大学德累斯顿计算材料科学中心（DCMS），德国德累斯顿 01062[4]CARE 卓越集群，德累斯顿工业大学和亚琛工业大学，德国[5]CeTI 卓越集群，德累斯顿工业大学，德国[*]通讯作者：Leonardo Medrano Sandonas (leonardo.medrano@tu-dresden.de)，Gianaurelio Cuniberti (gianaurelio.cuniberti@tu-dresden.de)

摘要

数字传感面临着开发可持续方法的挑战，以拓展定制化电子鼻在复杂人体气味挥发组（BOV）中的适用性。为应对这一挑战，我们开发了 MORE-ML，这一计算框架将电子鼻分子构建单元的量子力学（QM）性质数据与机器学习（ML）方法相结合，用于预测传感相关性质。在该框架中，我们通过采样 BOV 分子与黏蛋白衍生受体之间相互作用的更大构象空间，将先前的数据集 MORE-Q 扩展为 MORE-QX。该数据集提供了在 BOV 吸附过程中计算得到的大量电子结合特征（BFs）。对 MORE-QX 性质空间的分析表明，构件的量子力学性质与由此产生的结合特征之间相关性较弱。基于这一发现，我们将构件的电子描述符定义为基于树的机器学习模型的输入，以预测结合特征。基准测试表明，CatBoost 模型优于其他模型，尤其是在向未见化合物迁移方面。可解释人工智能方法还揭示了哪些量子力学性质对结合特征预测的影响最大。总体而言，MORE-ML 将量子力学洞见与机器学习相结合，为 BOV 传感中的分子受体提供机理理解和理性设计原则。这一方法为开发能够分析复杂气味混合物的人工传感材料奠定了基础，弥合了分子尺度计算与实际电子鼻应用之间的差距。

引言

人工智能（AI）的飞速发展显著加快了人工智能驱动技术的发展，使其能够精准识别物体、面孔、声音和触觉[1,2]。尽管取得了这些进展，如何有效解读和预测人类周围的化学环境，仍存在相当大的技术差距。为弥合这一差距，定制化电子鼻应运而生，并展现出检测挥发性有机化合物（VOCs）的卓越能力[3–6]。具体而言，人体释放的挥发性有机化合物（称为体味挥发组（BOV））可作为独特的化学指纹，在医疗保健应用方面展现出巨大潜力[7]，例如，可作为阿尔茨海默病和帕金森病的生物标志物[8,9]。然而，人们仍迫切需要能够快速、可靠地检测生物标志物的传感材料[10] 例如，数字嗅觉系统中的 BOV 分子，尤其是用于医学诊断的场景。

受人类嗅觉系统[12]，的灵敏度[11] 和辨别能力启发，近年来，人们合成了多种分子嗅觉受体（例如，黏蛋白衍生受体[13–15]）。这项进展推动了实验方案的发展：通过在 glaco 偶联[16]结构上引入具有不同化学特性的特定官能团，以控制气体传感中受体对 BOV 分子的亲和力。然而，若依赖经验性的试错筛选，获取 BOV–受体相互作用的详细信息——并据此指导受体优化——仍然既耗费成本又耗费时间。这表明，当前原型受体的一项关键局限，在于缺乏对其在广阔 BOV–受体系统化学空间中的灵敏度和选择性的机理认知。这一瓶颈凸显了采用可持续策略合理设计高性能基于受体的仿生传感器的必要性。与几十年前分子电子学带来的变革性影响相似[17]，，量子力学（QM）

方法有望通过加深对决定恢复时间、电荷转移和肖特基势垒电势等关键性能指标的物理和化学相互作用的理解，彻底改变化学传感领域[18]。此外，将量子力学导出的性质数据与人工智能技术相结合，有望构建可靠高效的计算框架，为高灵敏度和高选择性传感器材料的设计提供指导——这一方法近期已在药物发现研究中取得成功[19–22]。

在此背景下，我们近期推出了 MORE-Q 数据集[23]，，首次提供了与人工嗅觉分子传感器原子级构筑单元相对应的大规模量子力学性质数据：BOV 分子、黏蛋白衍生嗅觉受体以及 BOV-受体二聚体系统。MORE-Q 还包含电子结构数据，用于描述最稳定二聚体系统与石墨烯表面之间的分子间相互作用。总之，该数据集支持研究由 BOV吸附诱导的关键结合特征（BFs），例如吸附能[24]，电荷转移[25]，以及功函数变化[26]。这批结合特征的集合为 BOV–受体系统的理性设计与优化迈出了重要一步，因为它全面描述了传感性能的电子特征；然而，在建立可持续的 BOV–受体设计框架之前，仍有一些挑战需要应对。例如，与配体-口袋基序[27,28]，类似，传感过程本质上是动态的，并受到较弱的非共价相互作用（静电作用和氢键）支配，这表明其结构具有柔性，因而形成崎岖的能量景观，其中存在大量局部极小值和多样的结合构型[29]。另一方面，当前理论模型缺乏定量刻画性质-性质关系和结构-性质关系所需的严谨性，阻碍了我们清晰理解传感构筑单元在结合特征行为中的作用。

一种有前景的途径是采用机器学习（ML）方法，以阐明原子结构与结合特征之间复杂的映射关系。例如，Ulissi 等人近期提出了 AdsorbML框架[30]，，该框架将启发式搜索与机器学习势相结合，以加速气体-金属吸附能计算；与传统密度泛函理论（DFT）相比，该框架不仅预测准确率高，还显著提升了计算速度。同样，GAME-Net[31]，这一图神经网络模型旨在预测有机分子在催化表面上的吸附能，其准确度接近密度泛函理论；对于大型生物质和塑料片段，误差为 0.18 电子伏特（每个原子 0.016 电子伏特）。近期，Chen 等人提出了 AdsMT[32]，，这是一种多模态Transformer，通过交叉注意力机制将催化剂表面图表示与吸附质特征向量相结合，无需枚举吸附位点即可预测全局最小吸附能。尽管各种基于机器学习的研究[33–37] 主要关注小型、简单吸附物（例如，O_2，、CO_2，和 H_2）在平坦金属表面上的吸附，但对电荷转移和功函数变化等其他电子结合特征的研究仍较少。此外，对于 BOV-受体系统等存在相互作用的大型分子，由于相互作用形态复杂且具有构型多态性，结合位点和吸附距离的概念变得难以明确定义，这使得预测模型的开发更具挑战性。此外，这些研究大多优先追求较高的预测准确率，往往以牺牲模型可解释性为代价，从而限制了人们从这些复杂映射关系中获得的物理和化学见解。缺乏可解释性也会影响对结合特征空间的探索：优化一种特征并不能保证其他特征也同时改善，进而使传感材料的理性设计更加困难。

为应对这些挑战，我们开发了 MORE-ML 框架，将量子力学衍生分子性质与机器学习方法相结合，以研究电子鼻构建单元的结构、整体和原子级特征如何影响 BOV吸附。为此，我们旨在阐明人工传感材料中 BOV–受体系统的传感机制，并制定设计原则。为近似热力学系综，我们通过采样多个吸附在石墨烯上的低能 BOV–受体二聚体（DM）构象异构体，将 MORE-Q 数据集扩展[23] 为 MORE-QX。此过程将 BOV–受体–石墨烯复合物的数量从 1,836 增至 10,441（见图 1）。对 MORE-QX 的全面分析显示，结合能相近的 DM 构象异构体，其结合特征（BFs）仍可能存在显著差异，例如电荷转移和功函数变化。此外，尽管 DM 性质和结合特征是依据基本物理和化学原理选定的，两者之间仍仅呈弱至中度相关。尽管结合特征之间的相关性尤其弱，我们仍能灵活识别具有相似电子结合特征组合的体系——这清楚地证明 MORE-QX 性质空间中存在“设计自由度”[38]。为快速、准确地探索结合特征空间，进而支持实际设计用于传感的 BOV–受体复合物，我们开发了基于树的回归模型，将构件的量子力学衍生性质数据映射至其相关结合特征。在 MORE-ML 框架中，我们进一步利用这些模型的可解释性，通过 SHapley 加性解释（SHAP）[39] 提取传感过程的机理见解。总体而言，本研究提供了对吸附机制的量子力学启发式认识，并推动 BOV–受体系统的理性设计，为可控且稳健地发现人工传感材料铺平道路。

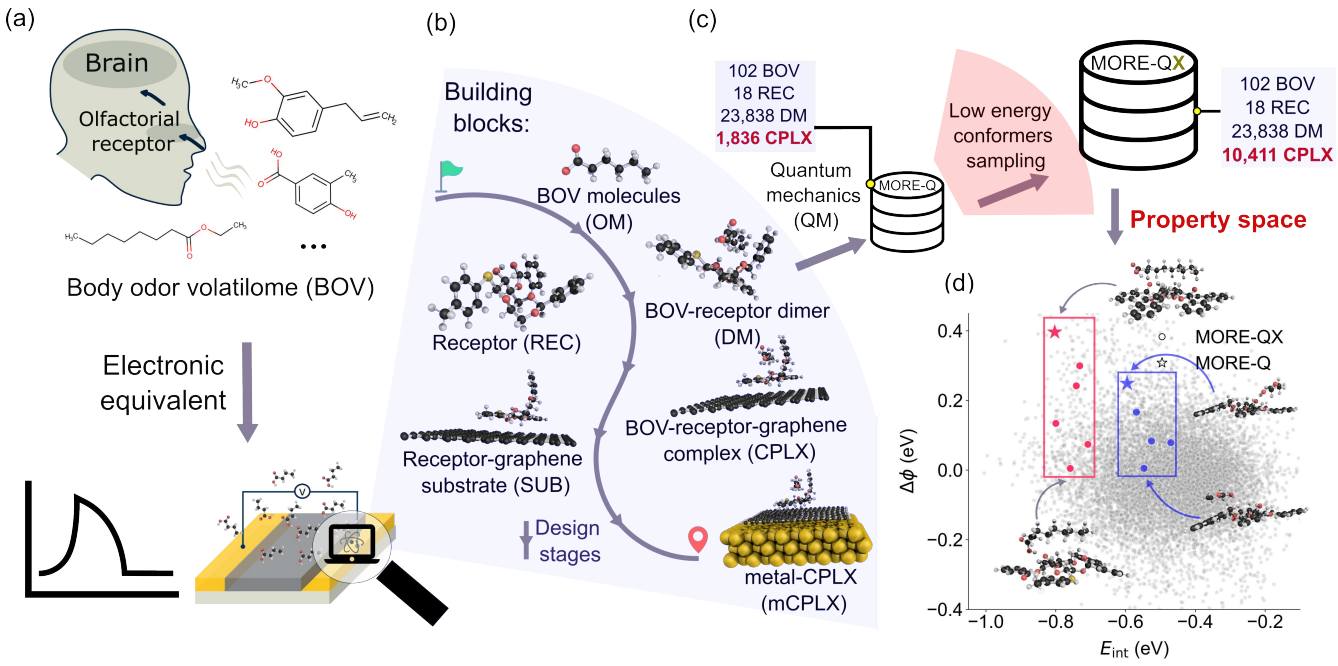


图 1. MORE-Q 数据集扩展为 MORE-QX 数据集的示意工作流程**M**olecular **O**lfactorial **R**eceptor **E**ngineering by **Q**uantum mechanics (MORE-Q)[23]数据集扩展为 MORE-QX 数据集。(a) 生物电子鼻（右上图）被设计为嗅觉系统（左上图）的电子等效物，例如，用于感知BOV分子（或气味分子，即OM）。(b) 从量子力学视角展示仿生传感器在不同设计阶段的构件，包括OM分子、分子受体（REC）、OM-REC二聚体分子（DM）、REC-石墨烯基底（SUB）、OM-REC-石墨烯复合体系（CPLX），以及最终沉积在金电极上的这些体系（mCPLX）。本手稿中通篇使用这些缩写。(c) 计算了相关构件的量子力学性质，并将其纳入MORE-Q数据集。该数据集包括102个OM分子和18个REC分子的单体体系、23,838个DM体系，以及从最稳定DM构型导出的1,836个CPLX体系。对多个低能DM构象体进行采样，扩展了CPLX子集，得到包含10,411个CPLX体系的MORE-QX数据集。(d) 由功函数变化 $\Delta\phi$ 和DM相互作用能E_{int}定义的高维MORE-QX性质空间的二维投影。图中标出了两个DM体系（红色和蓝色）的构象，其中最稳定的构象（MORE-Q）以星号标记，其他低能构象体（MORE-QX）以圆圈标记。图中还展示了每个DM体系中取值最大和最小的对应原子结构。

结果

评估构象效应对结合特征的影响

大型分子受体（REC）的内在柔性，是理解人工嗅觉传感器传感机制时必须考虑的关键因素。与配体进入蛋白质口袋的结合过程类似[27,28]，，气味分子（OM）与分子受体之间的相互作用本质上是动态的。事实上，二聚体体系（DM体系）会在多种构象之间持续相互转化，即在势能面（PES）的不同极小值之间跃迁。以往研究表明，大分子在气相中[40] 以及沉积在石墨烯纳米带上时[41]，均可能表现出相近的量子力学（QM）性质；这引出了一个问题：这类效应是否也会出现在这些 DM体系中。因此，我们没有像许多 DFT研究中常见的做法那样，仅考虑每种 DM构型中最稳定的构象体，而是采样了更广泛的低能构象体系，并将其吸附到石墨烯上。此过程将 MORE-Q 数据集扩展[23] 为 MORE-QX，并将 CPLX体系的数量从 1，836 增至 10，441（见图 1(a-c)）。相应地，我们也对额外的 DM构象计算了 MORE-Q 中 CPLX体系所计算的同一组 QM 性质（见方法部分）。需要注意的是，采样构象的数量会根据 DM体系的形态复杂性进行自适应调整，以确保采样稳健。这意味着形态越灵活的体系，采样得到的构象数量越多。平均而言，我们为每个 DM体系考虑了六种构象。关于 DM体系构象采样的更多详情，见参考文献[23]。

图中展示了 MORE-QX 中 CPLX 体系所跨越的高维性质空间的二维（2D）投影，见图 1(d)。此处，我们展示由功函数变化（$\Delta\phi$）和二聚体相互作用

表 1. BOV 受体（二聚体系统，DM）和 BOV 受体-石墨烯（复合系统，CPLX）相互作用的相关物理化学性质列表。每种性质均列出名称、符号和单位。a_0 和 D 分别表示玻尔半径和德拜的原子单位。

#	性质	符号	Unit
1	相互作用能	E_{int}	eV
2	各向同性分子极化率	$\alpha_{\text{s,DM}}$	a_0^3
3	标量偶极矩	μ_{DM}	D
4	沿薄层（z）方向的偶极矩分量	$\mu_{z,\text{DM}}$	D
5	HOMO 能量	$\varepsilon_{\text{H,DM}}$	eV
6	LUMO 能量	$\varepsilon_{\text{L,DM}}$	eV
7	HOMO-LUMO 能隙	ε_{gap}	eV
8	吸附能	ε_{gap}	eV
9	功函数变化	$\Delta\phi$	eV
10	电荷转移	ΔQ	e

能量（E_{int}）定义的性质空间，即（$\Delta\phi$,E_{int}）。总体而言，可以看出这两种性质之间缺乏相关性，这表明在寻找具有给定（$\Delta\phi$, E_{int}）数值组合的二聚体构象时，存在一定的灵活性。为更好地理解构象采样的影响，我们选取了两个示例构型，见图 1(d) 中的矩形。在红色矩形中，$\Delta\phi$ 与 E_{int} 同样不相关，且相对于最稳定构象对应的数值，其大小变化范围很大，从 0.0 到 0.4 电子伏特。相比之下，E_{int} 的变化幅度小得多，仅从 -0.8 降至 -0.75电子伏特（即 ~ 0.05 电子伏特）。类似地，在第二组研究的构象中（蓝色矩形所示），由于 OM 分子换成了更小的分子，E_{int} 有所降低，但 $\Delta\phi$ 仍覆盖了更大的性质范围（~ 0.25电子伏特）。这种灵活性在整个（$\Delta\phi$,E_{int}）性质空间中始终存在，与所选 DM 构型无关，这凸显了从二聚体构象的量子力学性质推断结合特征的复杂性。此外，这一结果也已表明，在 MORE-QX 性质空间中同时优化多种性质，难以确定简单的物理和化学规则（下文详述）。尽管如此，图 1(d) 还传达了设计人工嗅觉系统的另一条重要信息：在 E_{int}固定的情况下，我们或许能够在较大的范围内找到多个具有理想 $\Delta\phi$ 数值的 CPLX 体系。反过来，也有可能在较大的 $\Delta\phi$ 范围内找到具有理想 E_{int} 数值的不同 DM 构型。这些初步观察首次证明了 MORE-QX 性质空间中存在内在的“设计自由度”[38]，，下一节将结合结合特征空间对此展开讨论（见图2）。补充信息（SI）的图 S1 中还展示了更多体现构象采样影响的性质分布。

MORE-QX 性质空间中的“设计自由度”

为了深入理解构件的量子力学性质与由此产生的结合特征（BFs）之间的关系，我们考察了 MORE-QX 所覆盖的高维性质空间中的部分成对相关性。具体而言，我们分析了 DM 系统的性质与相应结合特征之间的相关性（完整性质列表见表1）。之所以选择 DM 性质，是因为它们与分子–表面相互作用所引发的物理化学效应密切相关，例如轨道杂化、极化效应、电荷密度重新分布和电荷转移，而这些效应最终会影响结合特征[42]。总体而言，图2(a)显示，45组独特的成对投影（即二维相关图）中，几乎全部都呈现为无明显结构的“斑块”，表明这些量子力学性质大多彼此不相关。为量化相关程度，我们计算了斯皮尔曼相关系数的绝对值，$|\rho_s|$（见公式4）。图 $|\rho_s|$ 2（b）的上、下两幅图分别展示了 DM 性质和结合特征的分布，其中成对相关性依据其 $|\rho_s|$ 数值进行分类。若 $|\rho_s| > 0.8$，则认为性质间存在强相关；若 $0.5 < |\rho_s| \leq 0.8$，则为中等相关；若 $|\rho_s| \leq 0.5$，则为弱相关。因此，在 DM 性质中，21组成对相关性里有1组（$\approx 4.8\%$）呈强相关，4组（$\approx 19\%$）呈中等相关，其余16组（$\approx 76\%$）呈弱相关。相比之下，与结合特征相关的24组相关性中，没有一组呈强相关；2组（$\approx 8\%$）呈中等相关，其余22组（$\approx 92\%$）呈弱相关。这一比较表明，DM 性质和结合特征总体上都表现出较弱的相关性，而结合特征之间的相关性甚至弱于 DM 性质之间的相关性。这种现象反映出，与单个二聚体（OM-REC 系统）相比，OM–REC–石墨烯（CPLX）体系中的原子间相互作用更为复杂且难以预测。

在分析的二维性质空间中，有少数值得关注的情况呈现中等至强相关性（图2(a) 中以黄色边框标出）。例如，DM 体系的 HOMO-LUMO 能隙（$\varepsilon_{\text{gap,DM}}$）与最低未占分子轨道能量（$\varepsilon_{\text{L,DM}}$）的线性相关性强于其与 HOMO 能量（$\varepsilon_{\text{H,DM}}$）的相关性。这一观察表明，$\varepsilon_{\text{H,DM}}$ 可用于区分具有相似$\varepsilon_{\text{gap,DM}}$的 DM 体系，而这对于构建高效的电子描述符至关重要。

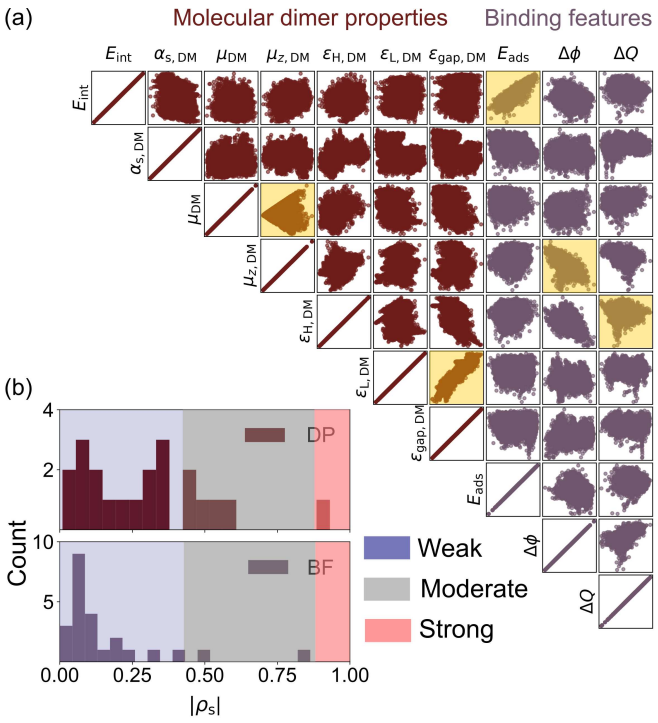


图 2。 (a) MORE-QX 数据集所覆盖的高维性质空间的二维（2D）投影。我们展示了 MORE-QX 中七种二聚体性质（DP，棕色）和三种结合特征（BF，紫色）的相关性图。性质的详细描述见表 1。一些有趣的投影以黄色框标出，并在手稿中进行了讨论。(b) 斯皮尔曼相关系数绝对值的计数测量 $|\rho_s|$，其中 DP（上图）和 BF（下图）为二维投影。 $|\rho_s|$ 这些 $|\rho_s| \leq 0.5$值形成三个明显的簇：弱相关 $0.5 < |\rho_s| \leq 0.8$、中等相关 $|\rho_s| > 0.8$ 和强相关，分别由蓝色、灰色和红色框覆盖。

关于与结合特征的相关性，DM 体系的相互作用能（E_{int}）与相应的吸附能（E_{ads}）呈现强相关性， $|\rho_s| = 0.86$。这一结果表明，可将 OM 与 REC 体系之间的相互作用机制迁移至 CPLX 体系，以描述 E_{ads} 的变化趋势。然而，二者的相关性并非完全线性，这说明表面相互作用诱导的几何结构和电荷分布变化会造成波动。另一个相关的结合特征是吸附物诱导的电荷转移（ΔQ），通常采用轨道混合理论进行解释；该理论描述了基底费米能级（本研究中的 DM 体系）与吸附物前线轨道能量之间的对齐关系[43]。通过计算 ΔQ 与 DM 体系轨道能量之间的 $|\rho_s|$，我们发现 HOMO 和 LUMO 能量与 ΔQ 的相关性都较弱，分别为 $|\rho_s| = 0.02$ 和 $|\rho_s| = 0.11$。类似地，OM 体系的轨道能量也与 ΔQ 不相关，得到 $|\rho_s| < 0.3$。这种缺乏相关性的现象揭示了仅凭轨道能量来制定调控 ΔQ 的设计原则十分复杂。同时，这也体现了结合特征空间中一定程度的“设计自由度”，使我们能够识别具有目标轨道能量的 DM 体系，并将其用作电子描述符的组成部分，以预测结合特征（下文详述）。

在对功函数变化（$\Delta\phi$）进行相关性分析时，我们发现，二聚体（DM）（$\mu_{z,\,DM}$）和体味挥发组分子（OM）（$\mu_{z,\,OM}$）体系中偶极矩的 z 分量与 $\Delta\phi$呈中等相关性，其对应的 $|\rho_s| = 0.51$ 和 $|\rho_s| = 0.68$分别为此结果。如我们在先前的工作中所讨论的[44–46]，$\Delta\phi$ 遵循亥姆霍兹关系：

$$\Delta\phi = -e/\varepsilon_0 \cdot \Delta P_{tot}, \tag{1}$$

其中，表面偶极矩变化（ΔP_{tot}）可分解为若干分量，表示为：

$$\Delta\phi = -e/\varepsilon_0 \cdot (\Delta p_{cplx} + p_a + p_s - p_0), \tag{2}$$

其中，组分 Δp^{cplx}、p_a 和 $p_s - p_0$ 分别表示空间电荷重分布导致的吸附物诱导表面偶极矩变化、吸附物偶极矩和表面形变。$\mu_{z,\,DM}$ 本身包含与 $\mu_{z,\,OM}$相关的信息，而后者与 p_a 项紧密相关，因此两者呈中等相关性（$|\rho_s| = 0.51$）。然而，其他

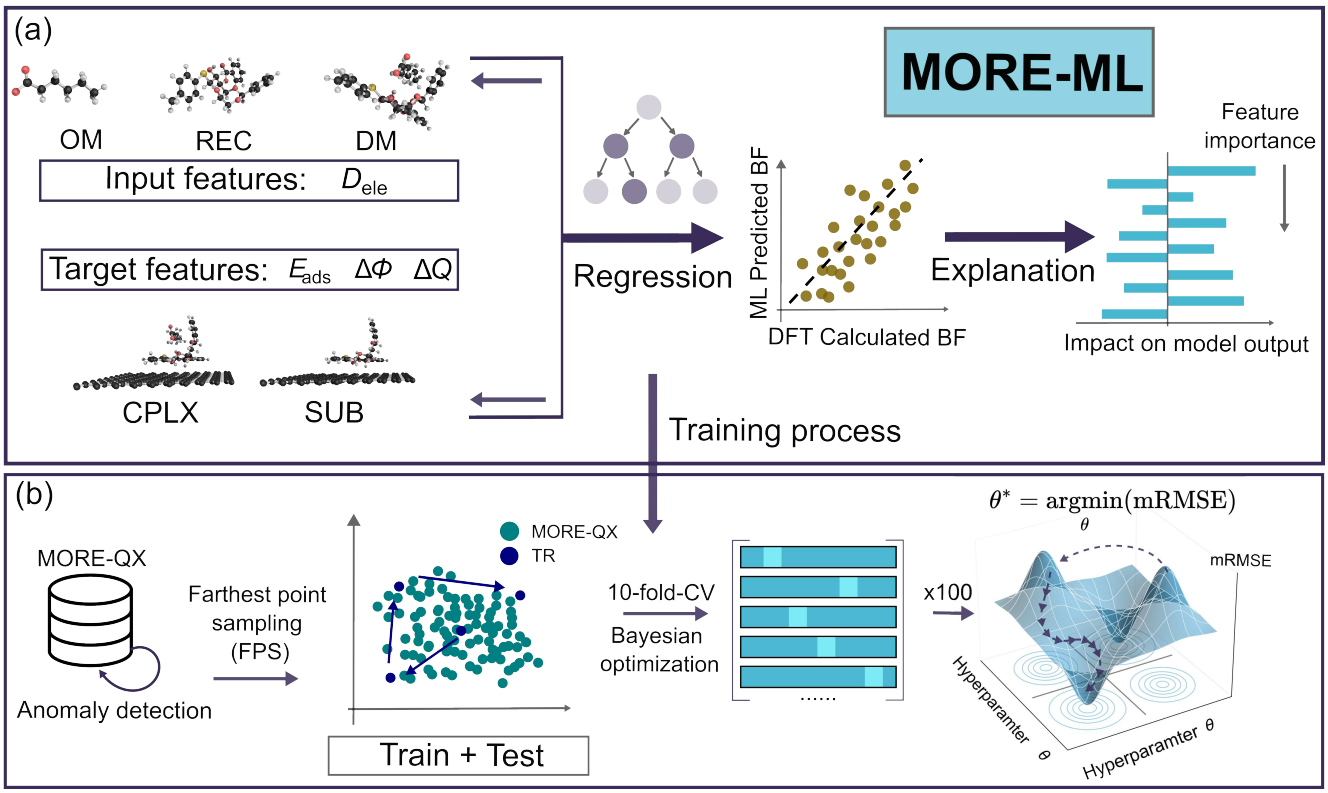


图3。 (a) MORE-ML 框架示意图。MORE-ML 代表基于机器学习的分子嗅觉受体工程，将分子构建单元的量子力学性质（D_{ele}）与机器学习技术相结合，用于预测结合特征（BFs），例如 E_{ads}、$\Delta\phi$ 和 ΔQ。MORE-ML 框架旨在完成回归和模型解释任务。(b) MORE-ML 中的机器学习模型训练始于异常检测（见补充信息），随后采用最远点采样（见方法）构建训练集和测试集。在训练集上进行 100 次迭代的贝叶斯优化和 10 折交叉验证，用于超参数调优。最终的模型性能在测试集上进行评估。

贡献——尤其是描述空间电荷重分布的 Δp^{cplx}——无法通过 $\mu_{z,DM}$ 或任何其他 DM 性质（例如极化率 $\alpha_{S,DM}$）得到充分表征； $\Delta\phi$ 与 α_S之间的相关性极低（$|\rho_s| = 0.07$），足以证明这一点（DM）。这些发现凸显了 $\Delta\phi$ 固有的复杂性，也表明当前物理化学启发式方法不足以对其进行定制。尽管 DM 性质与结合特征之间弱至中等的相关性可基于物理化学直觉提供一些理论指引，但并未显现出能指导结合特征空间探索的明确规律。此外，结合特征彼此之间的相关性也很弱。例如，E_{ads} 与 $\Delta\phi$ 的相关性较弱（$|\rho_s| = 0.14$）。同样， ΔQ 与 E_{ads} 的相关性较弱（$|\rho_s| = 0.05$），与 $\Delta\phi$ 则呈中等相关性（$|\rho_s| = 0.40$）；后者源于吸附过程中发生的空间电荷重分布[44,46]。综合来看，这些观察表明，限制 DM 体系同时呈现图2(a)中所考察的任意一对 DM 性质与结合特征的条件很少，这有力地证明了结合特征空间中存在“设计自由度”。基于这一概念，我们还分析了结合特征之间的弱相关性如何帮助识别针对特定目标性质定制的 DM 构象（详情见补充信息）。因此，大量电子特征都可能成为结合特征预测的有效分子描述符；然而，鉴于相关性强度和潜在物理化学见解各不相同，某些描述符可能比其他描述符更为重要。

请注意，尽管原则上可以使用玻尔兹曼加权性质来构建更准确的系综[47]，但为了探究构象异构体特异性效应，并比静态玻尔兹曼平均更全面地探索势能面，我们对每个低能二聚体构象一视同仁。由于这些低能构象的玻尔兹曼权重相近，采用加权或直接平均都会掩盖细微的构象异构体间差异。鉴于我们的首要目标是考察结合特征在各个表面结合二聚体中的变化，因此不采用玻尔兹曼加权。

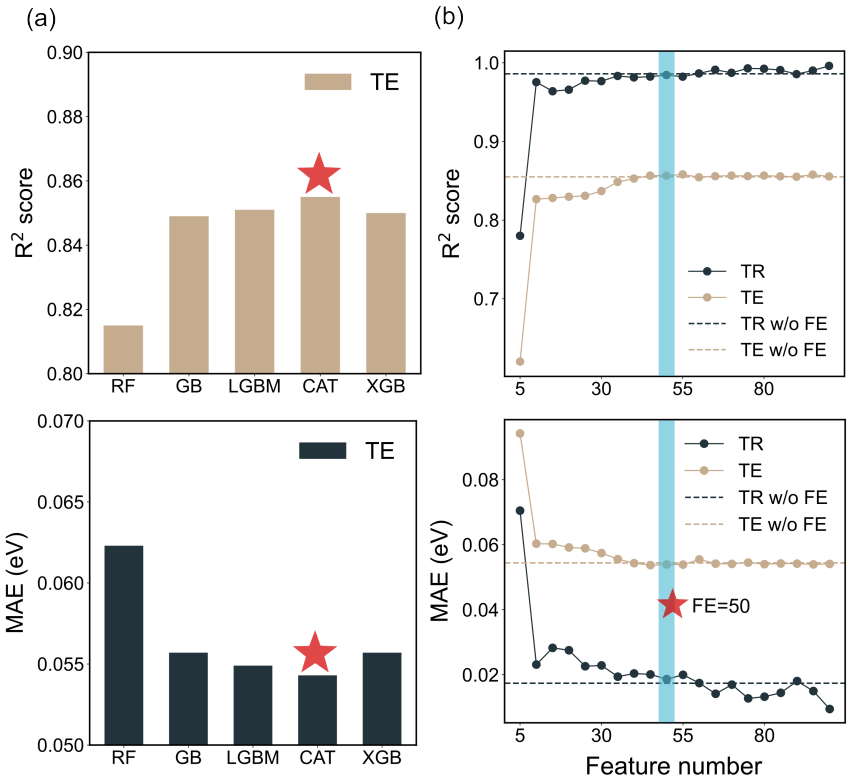


图4. 用于预测吸附能（E_{ads}）的模型基准测试与特征工程。(a) 在测试（TE）集上评估基于树的模型的决定系数（R^2）和平均绝对误差（MAE）：随机森林（RF）、梯度提升决策树（GB）、LightGBM（LGBM）、CatBoost（CAT）和XGBoost（XGB）。性能最佳的模型以红色星号标示。(b) 在CatBoost模型的特征工程（FE）过程中，预测E_{ads}时R^2（上图）和MAE（下图）的变化。根据SHAP分析进行排名（见方法），特征数量以每次增加五个的步长逐步增加。模型性能分别展示了训练（TR，黑色）集和测试（TE，棕色）集上的结果。虚线表示使用全部QM性质作为特征时获得的性能。

通过机器学习探索结合特征空间

尽管结合特征之间缺乏相关性，为设计具有所需传感相关性质的 CPLX体系提供了灵活性，但确定结合特征与分子构建单元（体味挥发组分子、REC 和二聚体体系）的电子性质之间的关系仍颇具挑战。为解决这一问题，我们采用了机器学习（ML）框架 MORE-ML（见图 3），旨在利用机器学习回归技术和可解释人工智能方法（见方法部分），建立这些性质空间之间定量且可解释的映射。为确定最适合结合特征预测的回归模型，我们对几种基于树的方法进行了基准测试：随机森林（RF）[48], 、梯度提升决策树（GB）、XGBoost（XGB）[49], 、CatBoost（CAT）[50], 和 LightGBM（LGBM）[51]。随后，我们将使用 SHapley 加性解释（SHAP）[39], 分析性能最佳的模型；这是一种高效的可解释人工智能框架，非常适用于基于树的模型。在训练策略上，我们优先采用电子结构衍生描述符（D_{ele}），其由体味挥发组分子、REC 和二聚体体系的量子力学性质构成，具有计算开销小且物理化学意义清晰的优点（见补充信息中的表 S3）。此外，受 QUED 框架[21], 发展的启发，我们考察了将 D_{ele} 与几何描述符 D_{geo}（例如，键袋[52], 、SOAP[53], 和 MACE[54]）及相应的 Mulliken 原子电荷 q（D_q）结合，能否进一步提升模型性能。然而，如补充信息中的图 S7 和 S8 所示，加入这些额外描述符并未提升机器学习模型的性能。这种性能未见提升，可能归因于多个二聚体结构中存在相似的体味挥发组分子和 REC 体系，从而产生了冗余的几何信息（见上文）。因此，我们仅使用D_{ele}对机器学习模型的准确率进行了更深入的分析。

图 4(a) 展示了使用完整 R^2）和平均绝对误差（MAE）预测吸附能 E_{ads}时的决定系数（D_{ele} 描述符（130 个特征）。MORE-QX 数据集按照固定的 9:1 比例划分为训练集（TR）和测试集（TE）。有关数据集划分流程和训练样本选择的更多细节，请参见方法部分。通过比较 E_{ads} 与其他结合特征所对应的结果（见补充信息中的图

S7），我们发现，所有基准方法在不同特征上的性能趋势相似。因此，此处采用 E_{ads}作为代表性案例。在基于树的方法中，随机森林在测试集中的表现最差，R^2 为 0.818，平均绝对误差为 0.062 电子伏特。这表明，与随机森林的装袋策略相比，梯度提升方法在捕捉 D_{ele} 与结合特征之间的潜在相关性方面表现更优。一种可能的解释是，在提升法中，每棵后续树都会校正其前一棵树的残差，而随机森林则依赖于独立树集成。在梯度提升系列中，CatBoost 的性能最佳，R^2 为 0.86，平均绝对误差为 0.058 电子伏特。这一优势可能源于其采用了对称树，即在给定深度上的所有节点都基于同一特征进行划分。这种结构对树模型的复杂度施加了较强的正则化，从而改善了结合特征预测的泛化能力。因此，我们将 CatBoost 作为后续分析的最终机器学习回归模型。

随后，我们着重从D_{ele} 中筛选出信息量最大的量子力学性质子集，以缓解高维度问题并减少模型噪声。通过识别并剔除冗余且高度相关的特征，我们旨在避免过拟合并提高机器学习模型的泛化能力。为此，我们利用CatBoost模型，采用迭代式SHAP驱动的特征选择流程。在每次迭代中，模型都会使用完整D_{ele}的一个精简子集重新训练，该子集由按照SHAP重要性排序的前列特征构成。选取的特征数量从5逐步增加至105，每次递增5个，并在各阶段评估模型性能。所得R^2 和平均绝对误差指标的学习曲线如图4(b)所示。根据TR集和TE集的结果，特征学习行为可分为增长阶段和饱和阶段。在D_{ele} 较小的阶段，模型性能逐步提升，但仍不及使用完整描述符集合时的表现（见虚线），这表明量子力学性质数量不足，无法准确捕捉吸附机制。当D_{ele} 的规模超过某一临界阈值后，性能曲线开始趋于平稳：TE得分不再提升，而TR得分仅有轻微波动。这表明增加特征并不能进一步加深模型对吸附机制的理解，说明存在一个能够兼顾准确率和效率的最优量子力学性质子集。基于这一全面分析，我们选取排名前50的电子特征（以星号标记）作为E_{ads}的有效描述符集合。采用相同流程，我们分别为$\Delta\phi$ 和 ΔQ选取排名前60和前80的特征（见补充信息中的图S7）。

事实上，最终用于预测 E_{ads}, $\Delta\phi$ 和 ΔQ 的机器学习回归模型，是基于优化后的量子力学特征子集并采用 CatBoost 方法构建的（见图5）。为评估模型的整体学习能力，我们首先考察 R^2 指标，该指标用于量化密度泛函理论计算值与机器学习预测值之间的方差。对于训练集，E_{ads} 和 ΔQ 的 R^2 均达到 0.99，而 $\Delta\phi$的数值略低，为 0.93。这一差异也体现在橙色数据点围绕 $y = x$ 参考线（虚线）的分散程度更大。根据平均绝对误差指标，训练集中 E_{ads}和 ΔQ 的对应值分别为 0.017 电子伏特和 0.001e，而 $\Delta\phi$的平均绝对误差较高，为 0.026 电子伏特，$_{DFT}$鉴于结合特征空间的离散性以及 MORE-QX 数据集覆盖范围有限，我们进一步采用相对误差 $\varepsilon = \frac{|y_{ML} - y_{\Delta}|}{y \times 100}$ 评估模型性能，其中 y_{ML} 和 y_{DFT} 分别表示性质 y 的机器学习值和密度泛函理论值。 Δy 表示整个数据集中该性质取值范围的大小。E_{ads}, $\Delta\phi$ 和 ΔQ 的相对误差分别为 1.7%、2.6% 和 1%。这些较小的数值表明，模型能够准确复现训练数据。

接下来，我们通过评估机器学习模型在测试集（TE set）中未见体系上的表现，考察其泛化能力。与预期一致，模型的准确率相较于训练集（TR set）有所下降，E_{ads}、 $\Delta\phi$和 ΔQ的R^2 值分别为0.86、0.76和0.81。相对误差也有所增加，但仍低于6%：E_{ads}为5.4%， $\Delta\phi$为4.8%， ΔQ为5.6%。训练集与测试集之间的性能差距适中，表明模型对新体系仍具有较高的预测准确率，凸显了其在更大范围的构型空间和构象空间中实现泛化的潜力。训练集和测试集的预测结合特征分布高度吻合（见图5(a-c)右侧面板），进一步支持了这一结论。各模型的完整评估指标汇总于表S6。为阐明预测 $\Delta\phi$的模型准确率略低的原因，我们使用相同的训练集和测试集，分别预测复合物（ϕ_{CPLX}）和基底（ϕ_{SUB}）体系的 ϕ 值。如图S10所示，ϕ_{CPLX}的预测结果为$R^2 = 0.87$，平均绝对误差为 $= 0.045$电子伏特；而 ϕ_{SUB}的预测结果为$R^2 = 0.89$，平均绝对误差为 $= 0.029$ 电子伏特。尽管这些指标表现良好， ϕ_{SUB} 的奇偶图在训练集和测试集中均呈现出意料之外的锯齿状，而模型未能重现 ϕ_{SUB}的双峰分布。这一缺陷可能源于数据集中 ϕ_{SUB} 值的多样性有限：构件的量子力学描述符，尤其是从18种REC结构中提取的描述符，不足以捕捉控制 ϕ_{SUB}的细微构象变化。因此，尽管ϕ_{CPLX} 的建模准确， ϕ_{SUB}的预测仍会引入噪声。

结合特征预测模型的基于人工智能的解释

为更好地解读用于结合特征预测的基于树的机器学习模型，我们结合其内在可解释性和 SHAP方法进行了可解释性分析（见方法部分）。图5(d-f)中的蜂群图总结了各预测任务中影响最大的特征所对应的 SHAP值分布。在这些图中，特征按重要性从上到下排序，其对应的 SHAP 值显示在x轴上。正 SHAP 值表示某个

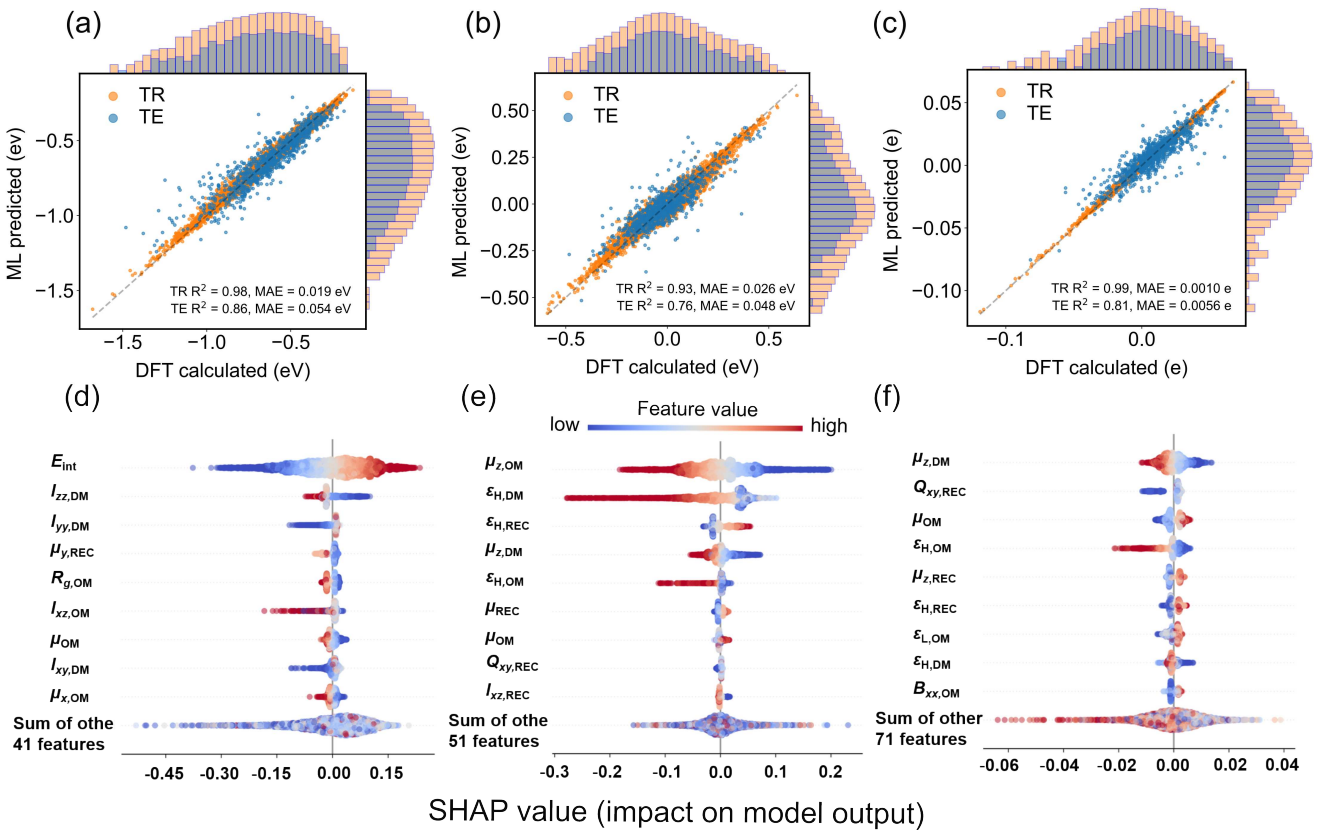


图5. 对于表现最佳的模型，图中展示了密度泛函理论计算值与机器学习预测值之间的相关性，这些模型用于预测 (a) E_{ads}、(b) $\Delta\phi$ 和 (c) ΔQ。橙色和蓝色的柱/点分别表示训练集（TR）和测试集（TE）。侧边面板展示了各结合特征的分布。面板 (d–f) 展示了相应的 SHAP 蜂群图（见方法），分别对应 (d) E_{ads}、(e) $\Delta\phi$ 和 (f) ΔQ。在每张蜂群图中，特征按重要性从上到下升序排列，SHAP 值分布在零基线两侧。每个点的颜色根据相应的特征值确定。图中仅展示排名前九的特征；其余所有特征的累积 SHAP 值列于最后一列（10th 位置）。

特征会使预测结果增大，而负值则表示预测结果会降低。颜色渐变表示特征的大小，其中红色代表较高值，蓝色代表较低值。

图5(d)中的 SHAP 值分布清楚地表明，二聚体（DM）的相互作用能 E_{int} 在决定 E_{ads}方面起着最主要的作用。颜色梯度表明，较小的 E_{int} 值会导致较小的 E_{ads} 值，反之亦然，因为这两个量均为负值。E_{int} 与其 SHAP 值之间的这种强耦合体现在较高的斯皮尔曼相关系数 $|\rho_s| = 0.86$ 上，这表明 E_{int} 可作为描述石墨烯表面 E_{ads} 的有效描述符。尽管 E_{int} 包含了预测 E_{ads}的大部分信息，但模型仍需考虑这两种能量之间较小的残差差异，以提高准确率。这一差异由形态描述符捕捉，例如 DM 系统的惯性张量（I）分量以及体味挥发组分子（OM）体系的回转半径（R_g），这凸显了分子结构也发挥着关键作用。此外，OM 和 REC 系统的偶极矩（μ）也位列前十特征，表明电荷重分布与描述吸附过程中的非共价相互作用有关。这些附加特征的 SHAP 值分布比 E_{int}窄得多，这解释了它们整体重要性较低。因此，这些特征与其余描述符一起，主要起到微调作用，用于捕捉少量离群值并进行细微修正，而 E_{int}则贡献占主导地位。

在预测 ΔQ的SHAP分析中也能观察到类似趋势（见图5(f)）。与形貌相关的结合特征E_{ads}, Δ不同，Q主要与电荷相关性质相关。尤其是偶极矩和四极矩成为最相关的特征，而分子轨道能量的排名较低；例如，$\varepsilon_{H,\ OM}$ 和$\varepsilon_{H,\ REC}$ 分别位列 4th 和 6th 。这表明，多种性质协同作用，共同预测 ΔQ，并不存在单一的主导特征。此外，SHAP分析凸显了纯定性轨道混合描述在解释电荷转移[43], 方面能力有限，因为前线轨道能量并非主导预测因子。

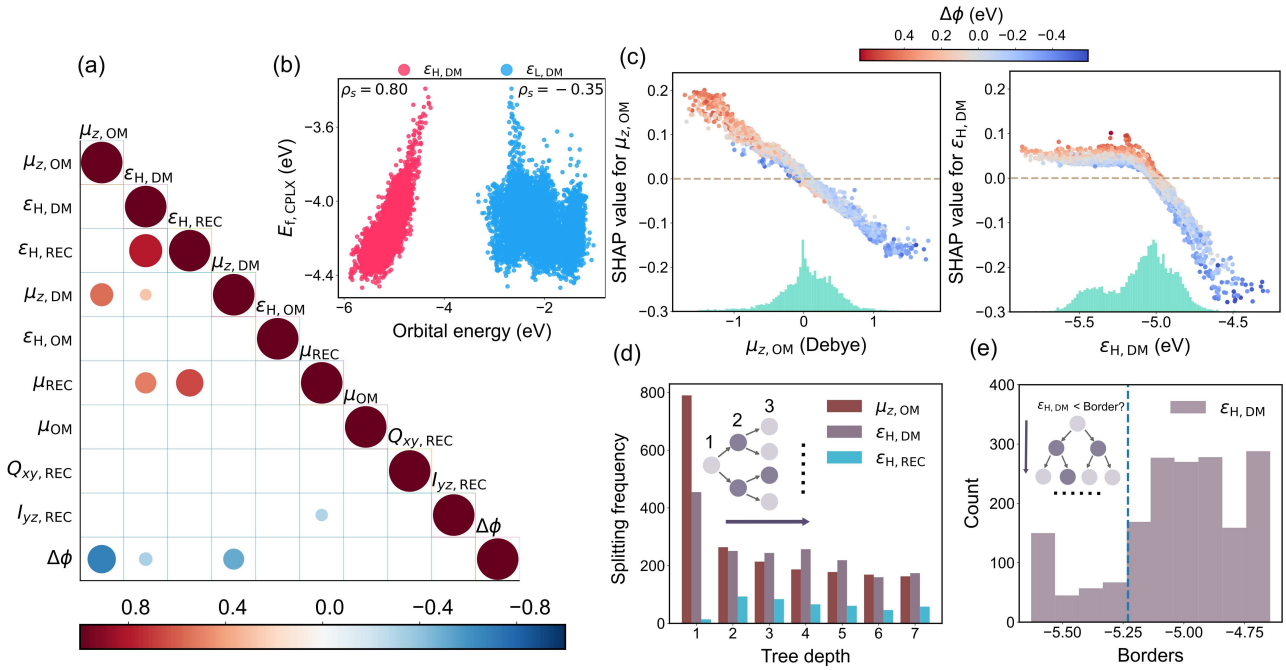


图 6. 对 $\Delta\phi$ 预测的机器学习模型进行解释。(a) 前九个特征与 $\Delta\phi$ 之间的两两斯皮尔曼相关系数 ρ_s。圆的大小和颜色分别表示 ρ_s 的大小和正负。(b) CPLX 体系的费米能级 $E_{f,CPLX}$ 与二聚体最高占据分子轨道 $\varepsilon_{H,DM}$（红色）和最低未占分子轨道 $\varepsilon_{L,DM}$（蓝色）能量之间的相关性图。
图中显示了相应的 ρ_s 值。(c) 两个最重要特征的 SHAP 值依赖图：体味挥发组分子 OM 的垂直偶极矩 $\mu_{z,OM}$（左图）和二聚体 HOMO 能量$\varepsilon_{H,DM}$（右图）。SHAP 值根据相应的特征值绘制；数据点按 $\Delta\phi$ 着色，特征值分布显示在 x 轴上。(d) 三个最重要特征 $\mu_{z,OM}$、$\varepsilon_{H,DM}$ 和受体最高占据分子轨道能量 $\varepsilon_{H,REC}$ 参与节点分裂的频率随树深度的变化。(e) 分裂频率随 $\varepsilon_{H,DM}$ 边界值变化的分布。

预测因子。不同于E_{ads} 和 ΔQ，$\Delta\phi$ 的预测由两个主导特征决定：OM的垂直偶极矩，$\mu_{z,\ OM}$以及DM体系的HOMO 能量，$\varepsilon_{H,\ DM}$，这两个特征在图5(e)中呈现出较宽的分布。μ_z的重要性，$_{OM}$ 可由公式2直接解释，因为它直接影响表面偶极矩的总变化。有趣的是，$\varepsilon_{H,\ OM}$（排名为 5th）的SHAP分布与$\varepsilon_{H,\ DM}$的趋势相似，后者是第二重要的特征。前线轨道能量（HOMO/LUMO）排名靠前的一种定性解释是，$\Delta\phi$ 部分源于空间电荷重分布。在此背景下，HOMO和LUMO能量分别代表主要的供体轨道和受体轨道，从而在相关原子上及其附近引发电荷密度变化。

为了进一步理解对 $\Delta\phi$的预测所揭示的物理机制，我们首先分析了前10个QM特征与 $\Delta\phi$ 之间的斯皮尔曼相关系数 $|\rho_s|$（见图6(a)）。在这些特征中，只有少数性质表现出明显的相关性。例如，$\mu_{z,\ OM}$ 和$\mu_{z,\ DM}$ 之间存在很强的相关性，因为$\mu_{z,\ DM}$ 包含了来自$\mu_{z,\ OM}$的信息。这些特征也与 $\Delta\phi$ 相关，因为它们部分出现在式2中。相比之下，前10个特征中的大多数相关性较弱（$|\rho_s| < 0.05$），这表明基于SHAP的特征排名有效缓解了QM描述符之间的多重共线性。该过程筛除了高度相关、因而包含噪声的特征，最终通过利用多样且非冗余的描述符提高了泛化能力。此外，$\varepsilon_{H,\ DM}$ 的对应特征$\varepsilon_{L,\ DM}$ 并未出现在前10个特征中，而$\varepsilon_{H,\ DM}$ 则排名第二。这表明表面费米能级主要与$\varepsilon_{H,\ DM}$相互作用，这与轨道混合理论[43]一致。因此，$\varepsilon_{H,\ DM}$ 倾向于与表面费米能级匹配，而复合系统的费米能级$E_{f,\ CPLX}$与$\varepsilon_{H,\ DM}$ 的关联强于与$\varepsilon_{L,\ DM}$的关联，分别对应 $\rho_s = 0.8$ 和 $\rho_s = -0.35$（见图6(b)）。为进一步探究最重要QM特征（例如$u_{z,\ OM}$ 和$\varepsilon_{H,\ DM}$）在调节 $\Delta\phi$时的协同机制，我们通过将SHAP值与性质分布相关联，分析这些特征的贡献行为（见图6(c)）。左图中，$\mu_{z,\ OM}$ 的SHAP值与其特征值呈现明显的线性相关：负的$\mu_{z,\ OM}$ 值对 $\Delta\phi$产生正向贡献，反之亦然，其符号取决于表面偶极矩的方向。总体而言，$\mu_{z,\ OM}$ 的绝对值越大，对 $\Delta\phi$的贡献越强，这与其 ρ_s 值一致。相比之下，右图中$\varepsilon_{H,\ DM}$ 的SHAP值几乎保持不变，直至$\varepsilon_{H,\ DM}$ 从最小值增大到约 −5.25电子伏特。

超过这一转折点后，随着特征值进一步增大，SHAP值与$\varepsilon_{H、DM}$之间开始呈现线性相关。这些行为可以通过基于树的模型的内在可解释性来理解。由于层次化分裂过程，在较浅树深度使用的特征比在更深处使用的特征更重要，因为早期分裂通常能通过划分数据集中的更大部分带来更多信息增益。如图6(d)所示，我们量化了$\mu_{z、OM}$、$\varepsilon_{H、DM}$和$\varepsilon_{H、REC}$（在图5中排名 4th ）在各树深度参与分裂的频率。在第一个树层级，代表$\mu_{z、OM}$和$\varepsilon_{H、DM}$的柱明显高于$\varepsilon_{H、REC}$，且$\mu_{z、OM}$也显著高于$\varepsilon_{H、DM}$。在更深的层级中，$\mu_{z、OM}$和$\varepsilon_{H、DM}$仍频繁参与分裂，但由于剩余数据点较少，信息增益有所降低。值得注意的是，在较深层级，$\varepsilon_{H、DM}$的分裂频率略高于$\mu_{z、OM}$，对应于能够分离出少量特殊离群值的分裂。这弥补了$\mu_{z、OM}$在早期更强的贡献，因此这两个特征的总SHAP贡献相近，这也反映在图5的重要性排名中。总体而言，这些观察结果证实了$\mu_{z、OM}$和$\varepsilon_{H、DM}$具有主导且广泛的重要性，其中$\mu_{z、OM}$的总体排名略高。$\varepsilon_{H、REC}$的柱高较低，也与其SHAP值分布范围狭窄、因而重要性较低相符。此外，$\varepsilon_{H、DM} \approx -5.25$电子伏特处的转折点，可以通过性质值（对应于基于树的模型中的决策边界）在分裂节点上的参与情况加以说明（见图6(e)）。值为 > -5.25 电子伏特的边界数量显著多于值为 ≤ -5.25电子伏特的边界，这表明高于该阈值的值更频繁地出现在分裂决策中。在这一范围内，性质值更为连续，因此会产生更广泛的贡献值；而值为 ≤ -5.25电子伏特的边界参与分裂过程的频率要低得多。事实上，值为 ≤ -5.25电子伏特的边界所带来的累积信息增益仅产生微小贡献，使 $\Delta\phi$在0至0.1电子伏特之间波动。相反，值为 > -5.25 电子伏特的边界会产生较大的贡献，且分布范围较广，这与图6(c)左图所示的阈值效应一致。这种行为可归因于CPLX体系中分子前线轨道与表面费米能级之间的能量匹配。具体而言，$\varepsilon_{H、DM}$起着关键作用，其与$E_{f、CPLX}$之间更强的相关性便证明了这一点（见图6(b)）。因此，我们推测，当$\varepsilon_{H、DM}$远低于表面费米能级时，最高占据分子轨道在能量上难以参与作用，不会在表面引起显著的电荷重分布，因而对功函数调制的影响极小。反之，当$\varepsilon_{H、DM}$高于表面费米能级时，可能发生显著的电荷重分布，而 $\Delta\phi$ 则取决于最高占据分子轨道与表面费米能级之间的能量差。此外，这种阈值效应还可能受到分子[55], 上前线轨道空间局域化的影响，进而影响其与表面的耦合。然而，对这些空间效应的详细分析超出了本研究的范围。

讨论

在本研究中，我们提出了 MORE-ML，这是一种计算框架，将电子鼻分子构件的量子力学（QM）性质数据与机器学习（ML）方法相结合，用于预测和阐释支配传感相关性质的物理化学机制。为应对这一挑战，我们将先前生成的 MORE-Q 数据集扩展为 MORE-QX；该数据集涵盖了由体味挥发组（BOVs）与黏蛋白衍生受体（REC）组合构成的相互作用体系，其构象空间和性质空间均显著扩大。基于 MORE-QX，我们通过计算吸附能（E_{ads}）、功函数变化（$\Delta\phi$）和电荷转移（ΔQ），构建了一组结合特征（BFs）。这些量化指标刻画了 BOV–REC 相互作用对 REC–石墨烯体系的能量、功函数（ϕ）和电荷分布的影响。对 MORE-QX 所涵盖性质空间的分析表明，在结合特征空间中存在明显的“设计自由度”，即能够识别出具有目标结合特征组合、化学多样的 OM–REC–石墨烯（CPLX）构象。这种灵活性源于大多数量子力学性质之间观察到的弱相关性。此外，性质间相关性分析凸显了若干电子特征在区分相似二聚体（DM）与 CPLX 构象方面的潜力，而这正是构建高效分子描述符的关键要求。MORE-QX 中的大多数电子特征不受平移、旋转和原子排列变化的影响，因此满足了适用于基于机器学习的预测建模的完备分子表示的一项核心要求。

借助 MORE-ML 框架中的这些见解，我们为分子构建单元（例如 OM、REC 和 DM 系统）的电子特征与结合特征（BF）定义了确定性映射。这些映射旨在降低确定传感相关性质的计算成本，因为计算各个构建单元的量子力学性质，远比直接计算结合特征所需的成本低。为此，我们进行了特征工程，并对多种机器学习回归技术进行了基准测试，以确定适用于每种结合特征、可构建准确可靠回归模型的最优电子特征集。与以往主要强调预测性能的机器学习研究不同，我们通过结合基于树的模型的内在可解释性与 SHapley 加性解释（SHAP）分析，着重提升模型可解释性。事实上，我们发现，E_{ads} 主要由 OM 与 REC 体系之间的相互作用能决定，而 ΔQ 主要受偶极矩和四极矩等电荷相关性质的影响。对于 $\Delta\phi$，OM 的垂直偶极矩与 DM 体系的最高占据分子轨道能量相互作用，体现了决定功函数 ϕ 的物理机制。这项深入研究揭示了

决定各结合特征的关键物理化学因素，并由此为探索尚未充分研究的结合特征空间，开辟了一条更透明、更易于探索的路径。

从电子鼻传感材料设计的角度来看，结合特征空间中所展现的“设计自由度”尤为宝贵，因为这表明可以通过受体工程半独立地调控传感器灵敏度、基线稳定性和选择性，而无需依赖反复试错的材料筛选。吸附能、电荷转移和功函数调制由彼此不同且相关性较弱的电子描述符所支配，这一发现与传感器阵列中的实际观察高度一致：在传感器阵列中，信号幅度、恢复行为以及器件间差异往往彼此解耦。重要的是，MORE-ML 框架的可解释性为选择或合成针对特定换能机制的受体分子提供了具有实验可操作性的指导方针，从而减少经验性优化循环。这一量子力学—机器学习—实验反馈循环代表着朝向下一代数字嗅觉系统实现理性化、可扩展设计的关键一步。

基于本研究所呈现的结果，我们成功展示了一种可持续的基于人工智能的框架，从计算角度揭示了多种传感机制。尽管 MORE-QX 仅限于研究石墨烯表面上的传感相关性质，但这一综合分析通过调控 DM二聚体性质，阐明了控制结合特征（BFs）的基本机制——这些性质与传感性能密切相关。这些见解为定义高性能、高灵敏度和高选择性分子受体的新型设计原则铺平了道路，并可通过生成式人工智能方法或实验测量加以验证。此外，本研究获得的认识还可直接迁移至更具实际意义的传感材料，例如二维MXene[56,57], 过渡金属硫族化物（TMDs）[58,59] 或金属有机框架（MOFs）[60], ，它们可为气体传感应用提供更丰富的化学空间和更强的电子可调性。我们注意到，要全面理解电子鼻器件中的传感机制，还需要研究电极与传感表面之间的接触电势（即肖特基势垒效应[26]），因为它可能在传感性能中发挥主导作用。因此，我们预计本研究将推动未来的相关研究，通过结合物理和化学见解，以及借助量子科学与可解释机器学习回归模型的整合所实现的确定性性质映射，促进传感材料的发展。

方法

密度泛函理论计算细节

BOV、分子受体及其二聚体构象的量子力学性质分别采用 GFN2-xTB+D4 方法和 PBE+D3 方法结合 def2-TZVPP 基组，分别使用 xTB（6.6.0 版）[61] 和 ORCA（5.0.3 版）[62] 软件包计算。在 MORE-QX 中，二聚体相互作用能E_{int}定义为二聚体构象的总能量减去各组成部分的能量，即，

$$E_{\mathrm{int}} = E_{\mathrm{DM}} - E_{\mathrm{REC}} - E_{\mathrm{OM}}. \tag{3}$$

BOV 受体-石墨烯复合物（CPLX）体系使用 DFTB+[63] 软件包进行几何结构优化，采用 GFN2-xTB哈密顿量和 D4色散校正。优化结构时，我们固定石墨烯层中的原子位置，因为 OM分子吸附不会显著影响石墨烯的几何结构，而且在化学电阻式传感器件中，电极会限制石墨烯的形变程度。为构建 SUB体系（或 REC-石墨烯体系），我们从 CPLX体系中移除 BOV分子，且不对结构进行优化，以研究结合特征的纯电子效应。

与 MORE-Q 数据集类似[23], ，MORE-QX 提供了单个 BOV/受体分子（MORE-QX-G1）、BOV-受体分子二聚体（MORE-QX-G2）以及复杂体系（MORE-QX-G3）的一系列丰富的 QM 全局和局部性质（最多 39 项）。MORE-QX-G1 子集包含 102 个 BOV 分子和 18 个分子受体的 QM 性质数据。在 39 项分子和原子性质中，我们计算了 D3 能量、偶极矩、极化率和穆利肯电荷。MORE-QX-G2 子集基于 MORE-QX-G1 中的几何结构，通过搜索 BOV 分子与受体的分子对接构象构建而成。因此，MORE-QX-G2 包含 23,838 个在 GFN2-xTB+D4 级别下计算的二聚体构象的 QM 性质数据，以及 10,411 个在 PBE+D3 级别下结合能最低的二聚体数据（性质列表见补充信息中的表 S2 和 S3）。MORE-QX-G3 子集包含从 MORE-QX-G2 中筛选出的 10,411 个位于石墨烯表面的二聚体。因此，MORE-Q-G3 包含 CPLX 和 SUB 体系在 PBE+D3 级别下的 QM 性质数据，以及反映 BOV 分子吸附导致单个体系性质变化的结合特征。此次扩展涉及 GFN-xTB+D4 几何弛豫和密度泛函理论计算，共耗费 ∼25 Mio CPU小时。

为衡量 MORE-QX 中量子力学性质之间的相关性，我们采用了斯皮尔曼相关系数，其定义为

计算如下：

$$|\rho_s| = |1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}|, \tag{4}$$

其中，每对观测值 (X_i,Y_i) 的对应秩分别表示为 $R(X_i)$ 和 $R(Y_i)$，随后将 d_i 定义为 $d_i = R(X_i) - R(Y_i)$。选择斯皮尔曼相关系数，是因为与对应的皮尔逊相关系数相比，它对离群值更稳健，且具有更强的非线性捕捉能力。

结合特征计算

SUB体系和CPLX体系的电子结构计算均在严格收敛的PBE+D3理论水平上，采用维也纳从头算模拟软件包（VASP[64,65], 6.3.1版）进行。平面波基组的能量截断值和自洽场收敛阈值分别设为600和 $1 \cdot 10^{-4}$ 电子伏特。所有模拟均在伽马点进行。为获得平坦的静电势，沿薄层方向（50.68埃）开启了偶极修正。

为计算结合特征，我们开展了不同类型的密度泛函理论计算。吸附能（E_{ads}）由单点计算的总能量获得，其定义如下：

$$E_{\text{ads}} = E_{\text{CPLX}} - E_{\text{SUB}} - E_{\text{OM}}. \tag{5}$$

而功函数变化（$\Delta\phi$）定义为BOV吸附前后功函数（ϕ）之差，即 ϕ 对于CPLX体系和SUB体系：

$$\Delta\phi = \phi_{\text{CPLX}} - \phi_{\text{SUB}}. \tag{6}$$

此处，各体系的 ϕ 通过以下方式计算：

$$\phi = E_{\text{V}} - E_{\text{F}}, \tag{7}$$

其中，E_F 为费米能级，E_V 为真空能级。E_V 通过分析沿薄层方向的静电势 P(z) 的平坦区域得出。P(z) 按以下公式计算：

$$P(z) = \int n(z)dz, \tag{8}$$

其中，平面平均电荷密度 n(z) 定义如下：

$$n(z) = 1/A \iint n(x,y,z)dxdy \tag{9}$$

最后，电荷转移 ΔQ 通过 BOV分子与 SUB体系之间转移的总 Bader电荷[66] 计算得出。

构象采样

使用自动化相互作用位点筛选（aISS）程序包[67]搜索了初始的 83,916 种二聚体构型（每种组合 50 种构型）。随后，我们进行了基于几何均方根偏差（RMSD）的层次聚类，并将 RMSD 距离截断值设为阈值，以筛除全部 83,916 种构型中的几何冗余构型，最终得到 23,838 种构型。结果表明，在剩余的 23,838 种二聚体构型中，简单几何结构的结合构型较少，而复杂几何结构的构型则较多。因此，在本研究中将低能构象体沉积到石墨烯表面时（根据相互作用能 E_{int}评估），复杂构象的采样频率高于简单构象，每种二聚体组合平均有六种构象。更多计算细节见参考文献[23]。

MORE-ML 框架

我们设计了分子嗅觉受体工程机器学习（MORE-ML）框架，以同时执行结合特征回归和模型解释任务，如图3（a）所示。在机器学习算法中，线性模型具有最高的可解释性，但容量不足；神经网络则具备卓越的表征能力，却存在可解释性不足的问题[68]。为在预测性能与透明性之间取得平衡，我们采用基于树的模型。这类模型不仅具有稳健的准确率，还能通过层次划分实现内在可解释的决策过程[69]。此外，当与可解释人工智能（XAI）工具——例如，**SH**apley

加性解释（SHAP）[39]结合使用时，这些模型不仅能够准确预测结合特征，还能帮助提取其背后的物理洞见[70]。

基于已定义的机器学习任务，我们现在介绍单个机器学习模型的训练流程，如图3(b)所示。在最初的训练循环中，我们发现部分体系的主导相互作用发生在体味挥发组分子（OM）与石墨烯表面之间，而非与受体之间。通过将数据投影到 UMAP 空间，并依据 SHAP 值进行聚类（以更好地形成聚类[71,72]），我们发现了一个对应于这些离群值的独特聚类。随后，我们移除了全部 932 个体系，因为它们不属于 DM 对设计的研究范围，否则会降低模型的有效性（更多详情请参见补充信息中的图 S4）。

随后，我们在结合特征的 t-SNE 空间中通过最远点采样（FPS）将剩余数据点划分为训练集和测试集，因为 t-SNE 能够捕捉非线性关系，并将具有相似结合机制的体系聚为一类——与 PCA 或 UMAP 等替代方法相比，它能更好地保持局部一致性；而在该空间中进行均匀采样，则可最大限度地减少两组之间的分布差异。数据集按固定的 9:1 比例划分为训练集和测试集。相应的学习曲线如图 S5 所示。在整个训练过程中，这个固定的测试集用于评估中间模型和最终模型。随后，我们进行了 100 次贝叶斯优化（BO）迭代，以确定最优超参数；每次 BO 迭代均以 10 折交叉验证得到的均方根误差均值（mRMSE）为优化目标。接着，我们使用找到的最佳超参数在训练集上重新训练模型，并在固定测试集上评估最终性能。

基于树的回归模型的可解释性策略

在这项工作中，我们采用 SHapley 加性解释（SHAP）来解释用于预测结合特征的机器学习回归模型。SHAP 是一种用于解释机器学习模型输出的博弈论框架，以合作博弈论为基础，并基于 Shapley值，用于量化每个输入特征对单次预测偏离模型预期/平均输出的影响。因此，该方法能够更透明地解释学得的相关性，突出特征的相对重要性，以及特征之间如何相互作用并影响预测结果。SHAP 通过考量特征 j 对模型 f 输出的边际贡献，将特征 j 的值转换为 SHAP 值 ϕ_j，因此，特征 j 的 SHAP 值定义如下：

$$\phi_j(f) = \sum_{S \subseteq N \setminus \{j\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [f(S \cup \{j\}) - f\{S\}], \tag{10}$$

其中，S 表示不包含特征 j 的特征子集，N 是特征全集，f 是机器学习模型。该公式将 SHAP 值定义为特征 j 在每个子集 S 上的边际贡献之和，每一项均按如下概率加权：在所有特征的各种排列组合中，恰好 S 中这些特征排在 j 之前的概率。

在相同背景下，我们还利用基于决策树的模型所固有的可解释性。这类模型将预测表示为嵌套的规则结构。从根节点开始，模型通过应用依赖于特征的阈值条件（例如 CatBoost 中的边界值），递归地细分特征空间，从而形成一系列逐步受限的决策子空间。划分过程在叶节点处终止，每个叶节点都关联一个固定的预测值或一组分布参数。通过沿着其唯一的划分路径回溯，可以明确还原任一叶节点的预测机制，从而将模型表示为定义在输入空间互不相交区域上的分段常数函数，使其具有可解释性。

参考文献

1. Zhang, J., Yin, Z., Chen, P. & Nichele, S. 使用多模态数据和机器学习技术进行情绪识别：教程与综述。 *Inf. Fusion* **59**, 103–126 (2020).**2.** Wang, J. 等。 人工感官技术：模拟并拓展生物感官。 *ACS nano***15**, 18671–18678 (2021).**3.** Ali, M. M., Hashim, N., Abd Aziz, S. & Lasekan, O. 用于农业和食品质量检测的电子鼻：原理与最新进展。 *Trends FoodSci. Technol.* **99**, 1–10 (2020).**4.** Huang, S. 等。 采用本征石墨烯和生物相容性稳定剂实现高灵敏度室温氨气传感器：稳定剂的作用。 *Carbon***173**, 262–270 (2021).**5.** Huang, S. 等。 基于机器学习的智能气体传感平台，用于识别工业气体。 *Adv.Intell.Syst.***4**, 2200016 (2022).**6.** Li, Y. 等。 用于检测和区分挥发性有机化合物的电子鼻：应用、挑战与展望。 *TrACTrends Anal. Chem.* **180**, 117958 (2024).

7. Drabi´nska, N. 等。健康人呼吸和体液中所有挥发性物质的文献综述：人类挥发组。J. Breath Res. 15, 034001 (2021)。8. Trivedi, D. K. 等。发现帕金森病的皮脂挥发性生物标志物。ACS Cent. Sci. 5, 599–606 (2019)。9. Tisch, U. 等。利用基于纳米材料的传感器，通过呼出气体检测阿尔茨海默病和帕金森病。Nanomedicine 8, 43–56 (2013)。10. Karnaushenko, D. 等。轻便、柔性、高性能诊断平台。Adv. Heal. Mater. 4, 1517–1525 (2015)。11. Firestein, S. 嗅觉系统如何感知气味。Nature 413, 211–218 (2001)。12. Morgan, J. 超级嗅觉者的乐趣：皮脂为帕金森病诊断提供线索。Lancet Neurol. 15, 138–139 (2016)。13. Bakhatan, Y. 等。加速固相聚糖合成：Asgs。Chem. Eur. J. 29, e202300897 (2023)。14. Sukhran, Y. 等。在空间拥挤且构象受限的半乳糖苷酰基转移中意外出现的亲核试剂屏蔽。J. Org. Chem. 88, 9313–9320 (2023)。15. Huang, S. 等。基于机器学习的石墨烯电子嗅觉传感器及其嗅觉性能评估。Appl. Phys. Rev. 10, 021406 (2023)。16. Shitrit, A. 等。单糖衍生的对映选择性在单壁碳纳米管化学电阻式挥发性有机化合物传感中的应用。Chem. Eur. J. e02553 (2025)。17. Cuniberti, G., Fagas, G. & Richter, K. 分子电子学导论：简要概述。分子电子学导论 1–10 (2005)。18. Bhati, V. S., Kumar, M. & Banerjee, R. 二维纳米材料/金属氧化物纳米复合材料的气体传感性能：综述。J. Mater. Chem. C 9, 8776–8808 (2021)。19. Ginex, T., Vázquez, J., Estarellas, C. & Luque, F. 药物发现中基于量子力学的策略：为药物设计的新挑战寻找步调。Curr. Opin. Struct. Biol. 87, 102870 (2024)。20. Vargas-Rosales, P. A. & Caflisch, A. 药物设计中的物理学与人工智能对话。RSC Med. Chem. 16, 1499–1515 (2025)。21. Hinostroza Caldas, A., Kokorin, A., Tkatchenko, A. & Medrano Sandonas, L. 评估量子力学描述符在物理化学和生物学性质预测中的表现。ChemRxiv 10.26434/chemrxiv-2025-hj4dc (2025)。22. Manathunga, M., Götz, A. W. & Merz, K. M. 计算机辅助药物设计：用于生物学问题的量子力学方法。Curr. Opin. Struct. Biol. 75, 102417 (2022)。23. Chen, L. 等。MORE-Q：一个用于基于量子力学的分子嗅觉受体工程的数据集。Sci. Data 12, 324 (2025)。24. Zhang, Y.-H. 等。通过引入掺杂剂和缺陷改善石墨烯的气体传感性质：第一性原理研究。Nanotechnology 20, 185504 (2009)。25. Wehling, T. 等。石墨烯的分子掺杂。Nano letters 8, 173–177 (2008)。26. Mathew, M. & Rout, C. S. 基于二维材料的肖特基二极管用于环境气体监测：新兴趋势、近期进展与未来展望综述。J. Mater. Chem. C 9, 395–416 (2021)。27. Ryde, U. & Söderhjelm, P. 采用量子力学方法支持的配体结合亲和力估算。Chem. Rev. 116, 5520–5566 (2016)。28. Puleva, M. 等。将量子力学基准精度拓展至生物配体-口袋相互作用。Nat. Commun. 16, 8583 (2025)。29. Jeindl, A., Hörmann, L. & Hofmann, O. T. 表面多晶型对有机物/金属界面功函数的影响有多大？Appl. Surf. Sci. 575, 151687 (2022)。30. Lan, J. 等。AdsorbML：利用可泛化的机器学习势，高效计算吸附能。npj Comput. Mater. 9, 172 (2023)。31. Pablo-García, S. 等。通过图神经网络快速评估有机分子在金属表面的吸附能。Nat. Comput. Sci. 3, 433–442 (2023)。32. Chen, J., Huang, X., Hua, C., He, Y. & Schwaller, P. 用于预测全局最小吸附能的多模态变换器。Nat. Commun. 16, 3232 (2025)。

33. Tran, K. & Ulissi, Z. W. 主动学习跨越金属间化合物空间，指导用于CO_2 还原和H_2 析出的电催化剂发现。Nature Catalysis 1, 696–703 (2018). 34. Zhong, M. et al. 利用主动机器学习加速发现CO_2 电催化剂。Nature 581, 178–183 (2020). 35. Fung, V., Hu, G., Ganesh, P. & Sumpter, B. G. 利用态密度的机器学习特征准确预测吸附能。Nature Communications 12, 88 (2021). 36. Xu, W., Reuter, K. & Andersen, M. 利用受物理学启发的图表示和机器学习预测复杂吸附物的结合基元。Nature Computational Science 2, 443–450 (2022). 37. Li, Z. et al. 利用基于自动化机器学习的特征删除实验阐释化学吸附强度。美国国家科学院院刊 121, e2320232121 (2024). 38. Medrano Sandonas, L. et al. 化学化合物空间中的“设计自由度”：迈向针对目标量子力学性质的分子理性计算机辅助设计。Chemical Science 14, 10702–10717 (2023). 39. Lundberg, S. M. & Lee, S.-I. 一种阐释模型预测的统一方法。载于 Guyon, I. et al.（编）神经信息处理系统进展 30, 4765–4774 (Curran Associates, Inc., 2017). 40. Medrano Sandonas, L. et al. 用于量子力学探索大型类药分子构象和溶剂效应的数据集。Scientific Data 11, 742 (2024). 41. Ravera, F., Medrano Sandonas, L., Gutierrez, R., Graziano, M. & Cuniberti, G. 非平衡效应是否与手性分子辨别相关？Journal of Chemical Physics 163, 014702 (2025). 42. Nørskov, J. K., Abild-Pedersen, F., Studt, F. & Bligaard, T. 表面化学与催化中的密度泛函理论。美国国家科学院院刊 108, 937–943 (2011). 43. Zhou, C., Yang, W. & Zhu, H. 单层WS_2上吸附气体分子的电荷转移机制及其对费米能级钉扎的影响。The Journal of Chemical Physics 142 (2015). 44. Chen, L. et al. 掺杂石墨烯基底与挥发性有机化合物相互作用的电子响应计算设计。Nanomaterials 14, 1778 (2024). 45. Khazaei, M. et al. 羟基终止的二维过渡金属碳化物和氮化物：超低功函数材料。Physical Review B 92, 075411 (2015). 46. Leung, T.-C., Kao, C., Su, W., Feng, Y. & Chan, C. T. 表面偶极、功函数与电荷转移之间的关系：对既定规律的一些例外。Physical Review B 68, 195408 (2003). 47. Kim, S., Schroeder, C. M. & Jackson, N. E. 可合成聚合物的功能单体设计。Chemical Science 16, 4755–4767 (2025). 48. Breiman, L. 随机森林。Machine Learning 45, 5–32 (2001). 49. Chen, T. & Guestrin, C. Xgboost：一种可扩展的树提升系统。载于第22届ACM SIGKDD知识发现与数据挖掘国际会议论文集，785–794 (2016). 50. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V. & Gulin, A. Catboost：具有类别特征的无偏提升法。神经信息处理系统进展 31 (2018). 51. Ke, G. et al. Lightgbm：一种高效的梯度提升决策树。神经信息处理系统进展 30 (2017). 52. Hansen, K. et al. 分子性质的机器学习预测：准确的多体势以及化学空间中的非局域性。Journal of Physical Chemistry Letters 6, 2326–2331 (2015). 53. De, S., Bartók, A. P., Csányi, G. & Ceriotti, M. 比较结构与嬗变空间中的分子和固体。Physical Chemistry Chemical Physics 18, 13754–13769 (2016). 54. Kovács, D. P. et al. Mace-off：用于有机分子的短程可迁移机器学习力场。Journal of the American Chemical Society 147, 17598–17611 (2025). 55. Egger, D. A. & Zojer, E. 分子偶极矩变化与诱导的功函数变化之间的反相关性。Journal of Physical Chemistry Letters 4, 3521–3526 (2013). 56. Cai, Z. & Kim, H. MXene气体传感器的最新进展：合成、复合材料与机制。npj 2D Materials and Applications 9, 66 (2025). 57. Yu, S., Li, P., Ding, H., Liang, C. & Wang, X. 二维MXene基气体传感器：进展、应用与挑战。Small Methods 2402179 (2025).

58. Mirzaei, A., Kim, J.-Y., Kim, H. W. & Kim, S. S. 基于二维过渡金属硫族化合物和MXene的电阻式气体传感器。Accounts Chem. Res. 57, 2395–2413 (2024). 59. Jana, D. et al. 作为多属性传感平台的二维材料。Adv. Funct. Mater. n/a, e16728 (2025). 60. Wang, W. et al. 用于痕量硫化氢检测的高灵敏度、高选择性锌基金属有机框架衍生物气体传感器。《ACS传感器》10, 7584–7598 (2025). 61. Bannwarth, C., Ehlert, S. & Grimme, S. GFN2-xTB——一种精确且参数化范围广的自洽紧束缚量子化学方法，具有多极静电学和密度依赖色散贡献。《化学理论与计算》15, 1652–1671 (2019). 62. Neese, F. ORCA程序系统。《Wiley跨学科评论：计算与分子科学》2, 73–78 (2012). 63. Hourahine, B. et al. Dftb+：用于高效近似密度泛函理论原子模拟的软件包。Journal of Chemical Physics 152 (2020). 64. Kresse, G. & Furthmüller, J. 使用平面波基组进行从头算总能量计算的高效迭代方案。Physical Review B 54, 11169 (1996). 65. Kresse, G. & Furthmüller, J. 使用平面波基组对金属和半导体进行从头算总能量计算的效率。《计算材料科学》6, 15–50 (1996). 66. Henkelman, G., Arnaldsson, A. & Jónsson, H. 一种快速稳健的电荷密度Bader分解算法。《计算材料科学》36, 354–360 (2006). 67. Plett, C. & Grimme, S. 自动且高效地生成通用分子聚集体结构。《德国应用化学国际版》62, e202214477 (2023). 68. Adadi, A. & Berrada, M. 窥探黑箱内部：可解释人工智能（XAI）综述。《IEEE Access》6, 52138–52160 (2018). 69. Lundberg, S. M. et al. 利用面向树模型的可解释人工智能，从局部解释迈向全局理解。《自然·机器智能》2, 56–67 (2020). 70. Oviedo, F., Ferres, J. L., Buonassisi, T. & Butler, K. T. 面向材料科学与化学的可解释且可阐释的机器学习。《材料研究述评》3, 597–607 (2022). 71. Cooper, A., Doyle, O. & Bourke, A. 用于子群发现的监督聚类：其在新冠肺炎症状学中的应用。载于欧洲机器学习与数据库知识发现联合会议，408–422 (施普林格, 2021). 72. Usuga, A., Praveen, C. & Comas-Vives, A. 基于局部描述符的机器学习模型经聚类分析优化，可准确预测双金属合金上的吸附能。《材料化学A辑》12, 2708–2721 (2024). 73. Chen, L., Medrano Sandonas, L., Huang, S., Croy, A. & Cuniberti, G. More-qx：扩展版数据集，适用于“MORE-Q：基于量子力学的分子嗅觉受体工程数据集”（版本2.0） [数据集]。ZENODO 10.5281/zenodo.14720508 (2025).

致谢

作者衷心感谢欧盟地平线欧洲 EIC Pathfinder Open 项目“用于体味诊断的智能电子嗅觉”（SMELLODI，资助协议编号：101046369）、大众基金会“嗅觉感知电子学”资格培养计划（项目编号：9B396）、德国研究基金会（DFG）研究培训组项目“B13：通过二维材料-纳米颗粒网络实现的神经形态感知”（NeuroSense-2D，资助协议编号：RTG2767）、德国研究基金会（DFG）资助的 CeTI 卓越集群“人机协同触觉互联网中心”（EXC 2050）、德国研究基金会（DFG）资助的卓越集群 REC2“气候变化时代的负责任电子学”（EXC 3035），以及德国研究基金会（DFG）资助的 CARE 卓越集群“气候中和与资源高效型建筑”（EXC 3115，项目编号：533767731）。我们还感谢德累斯顿工业大学信息服务与高性能计算中心（ZIH）提供计算资源。

作者贡献

该研究最初由 LC 和 LMS 构思，并在 SH、AC 和 GC 的贡献下完成设计。LC 创建了 MORE-QX 数据集，并开发了机器学习回归模型。MORE-ML 框架由 LC 和 LMS 实现，他们还起草了手稿初稿。所有作者均讨论了研究结果，并为手稿最终版本作出了贡献。

数据可用性

MORE-QX 数据集可在 ZENODO.ORG与本研究相关联的数据仓库中获取[73]。MORE-ML 框架中用于预测结合特征的代码和机器学习回归模型可在 GitHub 仓库 MORE-Q 中找到。

利益冲突

作者声明不存在财务利益冲突

以下为补充信息（SI）：

人工传感材料中基于量子信息的性质预测：可解释机器学习

Li Chen[1], Leonardo Medrano Sandonas[1, *], Shirong Huang[1], Alexander Croy[2], Gianaurelio Cuniberti[1, 3, 4, 5, *],

[1] 材料科学研究所和 Max Bergmann 生物材料中心，德累斯顿工业大学，德国德累斯顿 01062。

[2] 耶拿弗里德里希·席勒大学物理化学研究所，德国耶拿 07737。

[3] 德累斯顿计算材料科学中心（DCMS），德累斯顿工业大学，德国德累斯顿 01062。

[4] CARE 卓越集群，德累斯顿工业大学和亚琛工业大学，德国 y

[5] CeTI 卓越集群，德国德累斯顿工业大学

[*] 通讯作者：Leonardo Medrano Sandonas（leonardo.medrano@tu-dresden.de），Gianaurelio Cuniberti（gianaurelio.cuniberti@tu-dresden.de）

1 性质和缩写表

表 S1 手稿中使用的缩写列表。

缩写	定义
BOV	体味挥发组
OM	BOV分子
REC	受体
DM	BOV–受体二聚体
BD	结合特征
OM–REC	BOV–受体复合物
CPLX	BOV–受体–表面复合物
SUB	受体–表面底物系统
TR	训练集
TE	测试集
UMAP	用于降维的统一流形逼近与投影
t-SNE	t 分布随机邻居嵌入
RF	随机森林
GB	梯度提升决策树
CAT	CatBoost
XGB	XGBoost
LGBM	LightGBM
HOMO	最高占据分子轨道
LUMO	最低未占分子轨道
SHAP	Shapley 加性解释
MORE-Q	基于量子力学的分子嗅觉受体工程
MORE-QX	扩展的基于量子力学的分子嗅觉受体工程

表 S2 本研究分析的 MORE-QX 数据集中所采用的量子力学（QM）性质（及其对应符号）列表。对于所列各项 QM 性质的单位，a_0 表示原子长度单位（玻尔半径）。性质类型根据构件分为：单体（OM、REC）、二聚体（DM）、复合系统（CPLX）和结合特征（BD）。MORE-QX 数据集中 QM 性质的完整表征见 MORE-Q 手稿[1]。

符号	性质描述	单位	Type
$\mu_{\mathrm{z,OM}}$	体味挥发组分子偶极矩 z 分量	德拜	单体
$\epsilon_{\mathrm{H,REC}}$	REC HOMO 轨道能量	eV	单体
$\epsilon_{\mathrm{H,OM}}$	OM HOMO 轨道能量	eV	单体
μ_{REC}	REC 总偶极矩	德拜	单体
μ_{OM}	OM 标量总量	德拜	单体
$Q_{xy,\mathrm{REC}}$	REC 四极矩张量 xy 分量	Buckingham	单体
$I_{xy,\mathrm{REC}}$	惯性矩张量 xy 分量	原子质量单位 · Å	单体
$\epsilon_{\mathrm{H,DM}}$	OM-REC HOMO 轨道能	eV	二聚体
$\epsilon_{\mathrm{L,DM}}$	OM-REC 最低未占分子轨道能量	eV	二聚体
$\mu_{z,\mathrm{DM}}$	OM-REC 偶极矩 z 分量	德拜	二聚体
$\alpha_{\mathrm{s,DM}}$	OM-REC 分子各向同性极化率	a_0^3	二聚体
μ_{DM}	OM-REC 总偶极矩	德拜	二聚体
$\epsilon_{\mathrm{gap,DM}}$	OM-REC HOMO-LUMO 能隙	eV	二聚体
E_{int}	OM-REC 结合能	eV	二聚体
$E_{f,\mathrm{CPLX}}$	OM-REC-石墨烯 费米能级	eV	复合物
E_{ads}	吸附能	eV	结合特征
$\Delta\phi$	功函数变化	eV	结合特征
ΔQ	Bader电荷转移	e	结合特征

表 S3 用作输入电子特征的量子力学（QM）性质完整列表（及其对应符号）D_{ele}。所列单位和性质类型类别与表 S2 相同。性质类型根据构件划分如下：单体（OM、REC）、二聚体（DM）、复合系统（CPLX）和结合特征（BD）。一项性质可能同时适用于不同体系。因此，机器学习模型以 130 个特征作为原始特征。MORE-QX 数据集中的 QM 性质完整表征见 MORE-Q 手稿[1]。

#	性质	符号	单位	维度	体系	HDF5 键
1	PBE+D3总能量	E_{tot}	eV	1	体味挥发组分子、REC、二聚体	’ePBE+D3’
2	核排斥能	E_{nuc}	eV	1	体味挥发组分子、REC、二聚体	’eNUC’
3	电子排斥能	E_{ele}	eV	1	体味挥发组分子、REC、二聚体	’eELE’
4	单电子能	E_{1e}	eV	1	体味挥发组分子、REC、二聚体	’e1E’
5	双电子能	E_{2e}	eV	1	体味挥发组分子, REC, 二聚体	’e2E’
6	维里势能	E_{pe}	eV	1	体味挥发组分子, REC, 二聚体	’ePE’
7	维里动能	E_{ke}	eV	1	体味挥发组分子、REC、二聚体	’eKE’
8	交换能	E_{x}	eV	1	体味挥发组分子、REC、二聚体	’eX’
9	相关能	E_{c}	eV	1	体味挥发组分子、REC、二聚体	’eC’
10	交换-相关能	E_{xc}	eV	1	体味挥发组分子、REC、二聚体	’eXC’
11	D3总能量	E_{D3}	eV	1	体味挥发组分子、REC、二聚体	’eD3’
12	E6色散能	E_{6}	eV	1	体味挥发组分子、REC、二聚体	’eE6’
13	色散 E8 能量	E_{8}	eV	1	体味挥发组分子、REC、二聚体	’eE8’
14	HOMO 能量	ϵ_{H}	eV	1	体味挥发组分子、REC、二聚体	’eH’
15	LUMO 能量	ϵ_{L}	eV	1	体味挥发组分子、REC、二聚体	’eL’
16	HOMO-LUMO 能隙	ϵ_{gap}	eV	1	体味挥发组分子、REC、二聚体	’HLgap’
17	各向同性分子 C_6 系数	C_{6}	$E_{\mathrm{h}} \cdot a_0^6$	1	体味挥发组分子、REC、二聚体	’mC6’
18	总偶极矩	μ	D	3	体味挥发组分子，REC，二聚体	’vDIP’
19	标量总偶极矩	μ_{s}	D	1	体味挥发组分子，REC，二聚体	’DIP’
20	转动光谱常数	B	MHz	3	体味挥发组分子、REC、二聚体	’vRS’
21	转动偶极矩	μ_{B}	d	3	体味挥发组分子、REC、二聚体	’vRSDIP’
22	总四极矩张量	Q	Buckingham	6	体味挥发组分子、REC、二聚体	’TQP’
23	各向同性分子四极矩	Q_{s}	Buckingham	1	体味挥发组分子、REC、二聚体	’mQP’
24	分子极化率张量	α	a_0^3	6	体味挥发组分子、REC、二聚体	’mTPOL’
25	分子各向同性极化率	α_{s}	a_0^3	1	体味挥发组分子、REC、二聚体	’mPOL’
26	回转半径	R_{g}	Å	1	体味挥发组分子、REC、二聚体	’RG’
27	惯性矩张量	I_{TS}	原子质量单位·埃²	6	体味挥发组分子、REC、二聚体	’IM’
28	原子化能	E_{at}	eV	1	体味挥发组分子、REC、二聚体	’eAT’
29	结合能	E_{int}	eV	1	DM	’eBIND’

2 MORE-QX 数据分布

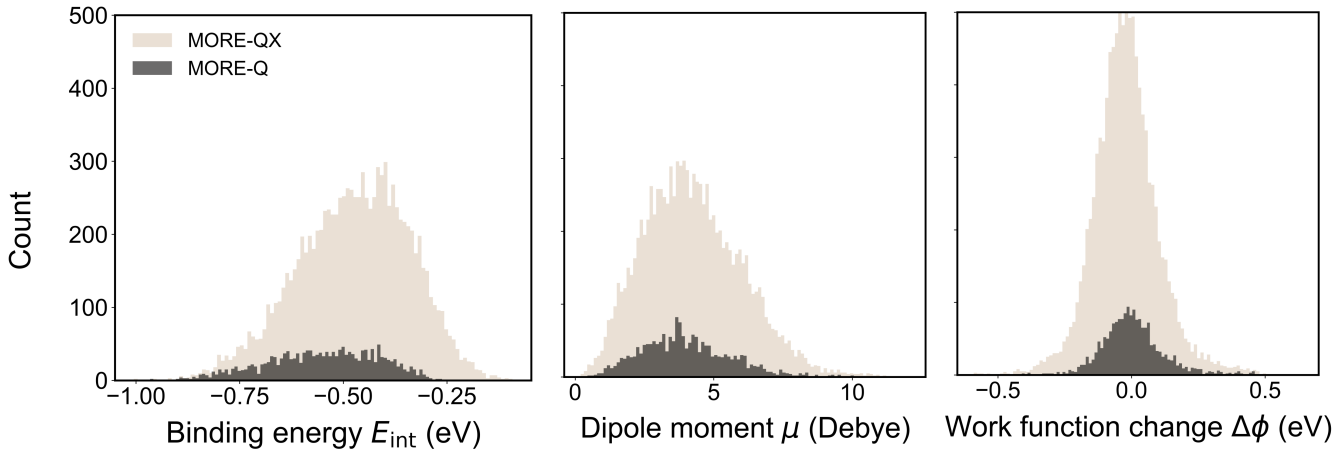


图 S1 三个示例　　用于比较 MORE-QX（棕色）与 MORE-Q（黑色）性质分布的示例　　)

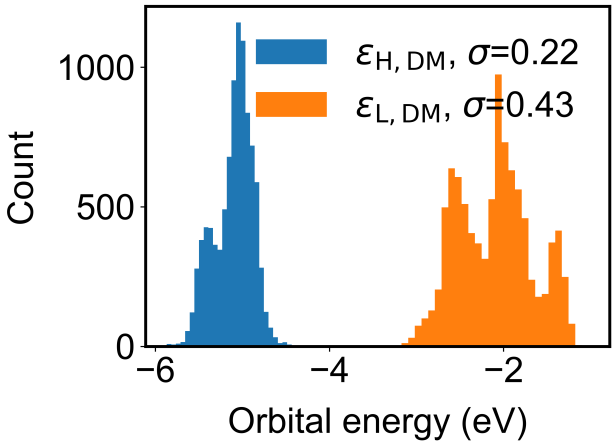


图 S2 二聚体最高占据分子轨道（$\epsilon_{H,\,DM}$，蓝色）和最低未占分子轨道（$\epsilon_{L,\,DM}$，橙色）轨道能级的分布。我们展示了相应的方差 σ。

3 基于树的机器学习模型

基于树的机器学习模型属于集成学习类别，可分为装袋法和梯度提升方法，其在不同回归任务中的表现已在图 S7 中进行基准比较。本节介绍各模型在回归任务方面的主要特征。统一定义如下：

$$F(x) = \sum_{m=1}^{M} w_m \, h_m(x;\, \theta_m), \tag{1}$$

其中，$h_m(x;\theta_m)$ 是 m^{th} 树（弱学习器），其结构和叶节点值由超参数 θ_m 决定。w_m 是 m^{th}的权重。训练过程的目标函数以及每棵树的损失函数定义如下：

$$\theta_m = \arg\min_{\theta} \sum_{i=1}^{n} \ell\big(y_i,\ F_{m-1}(x_i) + w_m \, h_m(x_i;\theta)\big) \ + \ \Omega\big(h_m(x;\theta)\big), \tag{2}$$

随后，模型的输出更新如下：

$$F_m(x) = F_{m-1}(x) + w_m \, h_m(x;\theta_m), \quad F_0(x) = \bar{y}, \tag{3}$$

其中，$\ell(y, \hat{}y)$ 是真实值与预测值之间的损失函数，$\Omega(h)$ 表示树 h 的复杂度，并对训练过程进行正则化。初始残差，即 0th 的输出结果，设为输出 $\bar{y}$ 的平均值。

随机森林（RF）

随机森林通过简单地对所有 M 单棵树的预测结果取平均值，独立地训练弱学习器，从而将 w_m 转化为$\frac{1}{M}$ 公式 S1。模型复杂度通过树深度、叶节点最小样本数等超参数进行控制。通常会省略显式的正则化项 $\Omega(h)$。

梯度提升决策树（GB）

在梯度提升决策树方法中，将固定学习率分配给 $w_m = \nu \in (0, 1]$。在每次训练迭代中，伪残差计算如下：

$$r_{im} = -\frac{\partial \ell(y_i, F_{m-1}(x_i))}{\partial F_{m-1}(x_i)} \tag{4}$$

然后将新树 $h_m x$ 拟合到残差上，并通过公式 S3更新模型。正则化项明确表示为：

$$\Omega(h) = \gamma T + \frac{1}{2}\lambda \Sigma_j w_j^2, \tag{5}$$

其中，T 表示叶节点数，γ、λ用于惩罚树结构和叶权重的复杂度。

XGBoost（XGBoost）

为提高训练速度并增强模型稳定性，损失函数部分的公式 S2 在训练样本的预测值处进行泰勒展开 m^{th} 训练样本的预测值 i 因此公式 S2 变为：

$$\theta_m = \arg\min_{\theta} \sum_{i=1}^{n} \big(g_i h(x_i;\theta) + \frac{1}{2} h_i h(x_i)^2\big) \ + \ \Omega\big(h_m(x;\theta)\big), \tag{6}$$

其中，$g_i = \frac{\partial \ell(y_i, \hat{}y)}{\partial \hat{}y}\Big|_{\hat{y}=F_{m-1}(\hat{}x_i)}$和 $h_i = \frac{\partial^2 \ell(y_i, \hat{}y)}{\partial \hat{}y}\Big|_{\hat{y}=F_{m-1}(\hat{}x_i)}$分别是损失函数 ℓ 在 m^{th} 预测处的一阶和二阶导数（梯度和海森矩阵）。

LightGBM（LGBM）和 CatBoost（CatBoost）

LGB 基于 XGBoost，采用直方图和按叶生长策略来提升训练速度并减少内存占用；而 CatBoost 则通过在每一层的节点分裂时强制使用 ical 特征，采用对称树以进一步克服过拟合问题。

4 关于结合 特征空间中“设计自由度”的更多结果

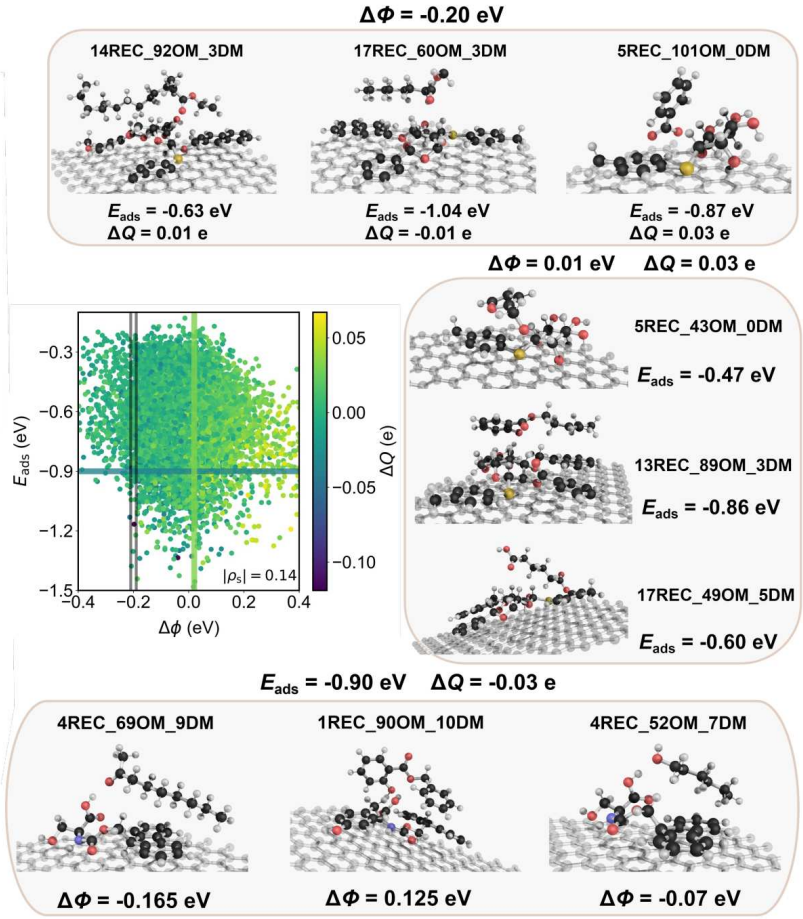


图 S3 设计自由度假设下设计任务的示例。每个系统根据 n受体 m体味挥发组分子 l二聚体命名，其中 n、 m和 l分别表示受体、BOV 分子及其二聚体构象的数量。BOV 分子和受体分子的完整列表及其数量可在 MORE-Q 中查看[1]。上图：约束 $\Delta\phi = -0.2 \pm 0.01$ 电子伏特的设计任务。右图：约束 $\Delta\phi = 0.01 \pm 0.01$ 电子伏特的设计任务。下图：约束 $E_{\text{ads}} = -0.9 \pm 0.01$ 电子伏特和 $\Delta Q = -0.03 \pm 0.001$ e。中图：吸附能与功函数变化的散点图，按电荷转移着色。

图 S3 中间面板的散点图展示了吸附能与功函数变化之间的相关性，相关系数为 $|\rho_s| = 0.14$。这种极弱的相关性为探究设计自由度猜想提供了理想示例。因此，我们首先从一个简单的约束设计任务入手，仅给定 $\Delta\phi = -0.20 \pm 0.01$电子伏特，对应于 $\Delta\phi$负半部分分布的 50th ，如图 S3 散点图中的深色空心线所示。沿着这些深色线， E_{ads} 在大致从 -0.3 到 -1.1电子伏特的良好范围内变化；其数值可能与 DM 相互作用面积相关，尤其是在由静电作用或范德华相互作用驱动的弱相互作用体系中[2]。与图 S3 顶部面板中突出显示的三个体系不同，后者均满足相同的$\Delta\phi$，却表现出截然不同的吸附能：最大的体味挥发组分子（ID 92）为 $E_{\text{ads}} = -0.63$电子伏特，而另外两个分子尺寸较小，分别为 -1.04 和 -0.87电子伏特。 E_{ads} 标度定律的偏离可能归因于表面 DM 相互作用中含氧口袋的形成。此外，这些 DM 相互作用产生了几乎相同的 $\Delta\phi$ ，但 ΔQ 在数值和符号上各不相同，展现了在单一简单 $\Delta\phi$约束下，寻找具有低相关性 E_{ads} 和 ΔQ 的复杂结构时所体现的设计自由度猜想。

约束，同时也反映出要将 $\Delta\phi$ 与体系的形貌和组成方面联系起来的复杂性。接下来，我们施加更严格的设计约束，目标为 $\Delta\phi = 0.01 \pm 0.01$ 电子伏特，即 $\Delta\phi$ 非主导的情况，以及相对较大的 $\Delta Q = 0.03$电子伏特，如散点图中的黄色条带所示。在更严格的条件下，如中间面板所示，我们仍能找到具有定制 E_{ads}的体系。在这些情况下，标度定律适用于从 5REC–43OM、经 13REC–89OM 到 17REC–49OM 的范围。作为最后的示例，我们施加 $E_{\mathrm{ads}} = -0.9 \pm 0.01$电子伏特和 $\Delta Q = -0.03 \pm 0.001$电子的约束，从而确保恢复时间相同且遵循电荷转移主导机制，如底部面板中的灰色水平线所示。在这些复合物中，OM 与 REC分子之间的氢键补偿了相互作用面积大小的差异，使长链体系、基于芘环的体系以及小尺寸体系都呈现相同的 E_{ads} 值。有趣的是，在相同的 ΔQ下，这三个体系呈现出数值和符号均不同的 $\Delta\phi$ ，从另一个角度再次体现了设计自由度。

5 MORE-ML 的完整工作流程

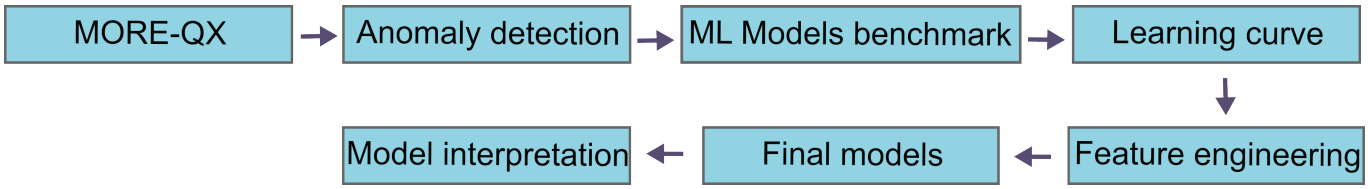


图 S4 MORE-ML 中完整机器学习工作流程概览。

6 异常检测

在最初使用 XGBoost（XGB）预测吸附能（E_{ads}）的基准测试中，经过多次尝试后，模型性能仍不理想，如图 S5 (a) 所示。随后，我们检查了图 S5 (a) 中离群值对应的几何构型。在图 S5 (b) 所示的示例中，体味挥发组分子（OM）的主要相互作用对象是石墨烯，而非受体。因此，我们统计了每个体系中与表面距离为 $>3.5^{\circ}$ 埃的原子数，该距离为 $\pi-\pi$ 堆叠距离，并将这一数量定义为描述符 $N_{d_{o-s<3.5\text{Å}}}$。$N_{d_{o-s<3.5\text{Å}}}$ 的分布如图 S5 (c) 所示，表明确实存在少数特殊体系，其中有不可忽略数量的原子主要与石墨烯相互作用。图 S5 (c) 所示的$N_{d_{o-s<3.5\text{Å}}}$ 分布揭示了一小部分特殊体系，其中有不可忽略数量的 OM 原子主要与石墨烯相互作用。为了识别最有可能降低模型性能的异常样本，我们将 $N_{d_{o-s<3.5\text{Å}}}$ 作为“诊断描述符”加入输入特征中。如图 S5 (d) 所示，与图 S5 (a) 相比，模型性能大幅提升，这表明加入 $N_{d_{o-s<3.5\text{Å}}}$ 有助于基于树的模型更好地对与新诊断描述符相关的数据点进行分类。因此，$N_{d_{o-s<3.5\text{Å}}}$ 具有很高的信息量，对预测 E_{ads}非常重要。SHAP 分析（图 S5 (e)）也证实了这一点，$N_{d_{o-s<3.5\text{Å}}}$ 被列为最重要的特征。有趣的是，尽管大多数 $N_{d_{o-s<3.5\text{Å}}}$ 较低的体系对模型的贡献微乎其微，但其中一部分对预测产生了不成比例的巨大影响，例如红色数据点。接下来，我们检查了这些体系的聚类情况，以识别其中 $N_{d_{o-s<3.5\text{Å}}}$ 是唯一异常因素的离群值。为此，我们使用 UMAP 进行降维，因为它能保留与簇形成相关的全局结构。此外，我们嵌入的是 SHAP 值，而非原始特征值，因为 SHAP 值能捕捉每个特征的贡献，提供更具区分性的表示，使贡献特征谱相似的样本能够更清晰地聚类。如图 S5 (g) 所示，$N_{d_{o-s<3.5\text{Å}}}$ 较高的主要离群值聚集在相邻区域，而图 S5 (h) 中 $N_{d_{o-s3.5<\text{Å}}}$ 较高的数据点则分布广泛，在基于特征值的 UMAP 空间中没有呈现共同的相似性结构。因此，离群值的 SHAP 值比特征值形成了更好的聚类。由此，我们将图 S5 (g) 中标出的 932 个数据点识别为异常样本，并将其从数据集中移除，因为它们不参与 OM-REC 相互作用，因此对于受体设计任务并不重要。此外，我们还列出了 XGBoost 的超参数表，以便复现研究结果。

表 S4 用于异常检测的 XGBoost 超参数列表。

超参数	值
lambda	0.603225906844846
alpha	0.007555374299051771
按列采样按树采样	0.6000000000000001
子采样	0.9
学习率	0.016
估计器数量	2000
最大深度	13
最小子节点权重	62
随机状态	20240815
作业数	1

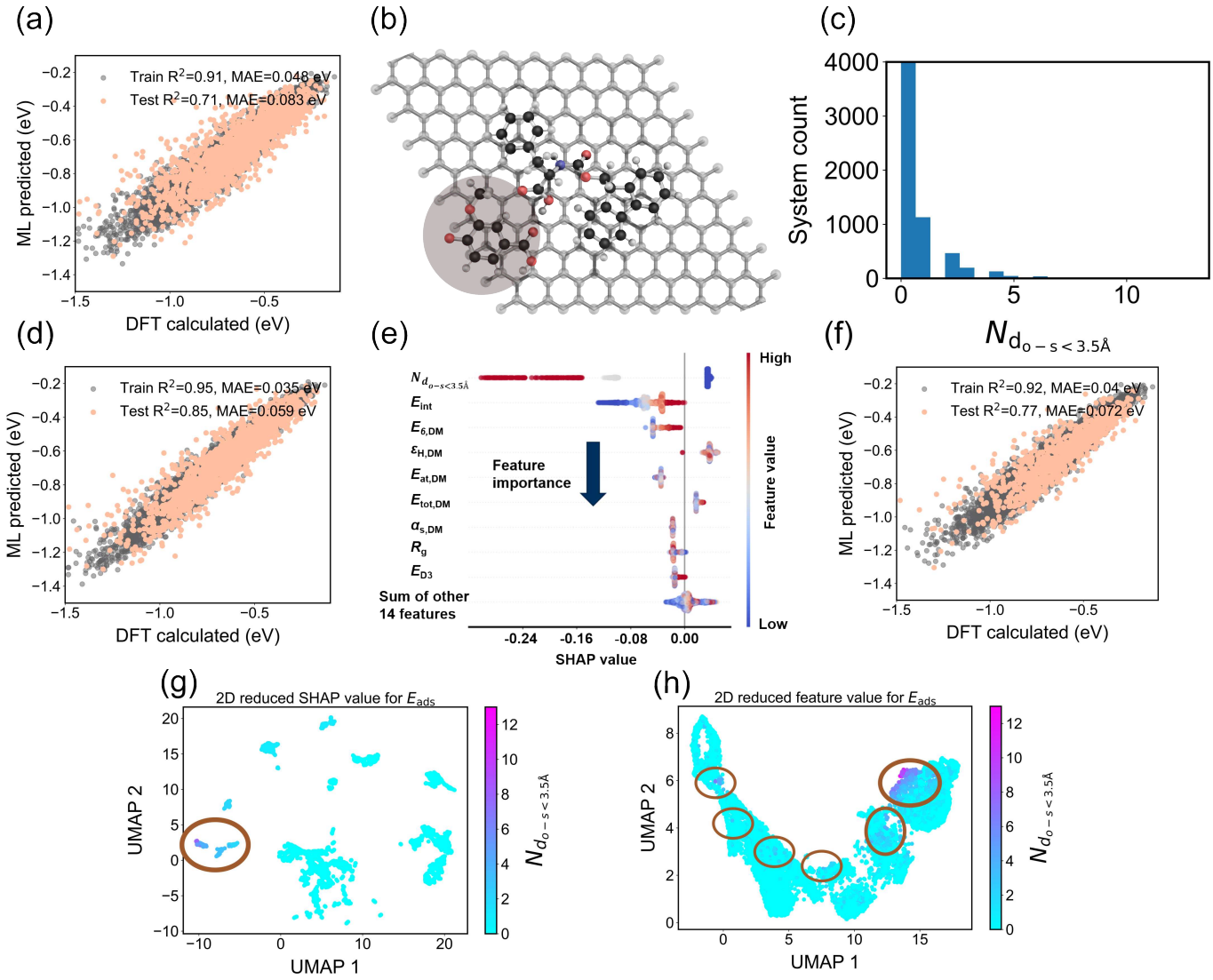


图 S5 MORE-QX 的异常检测工作流程。(a) 使用二聚体性质进行吸附能 E_{ads} 预测。(b) 异常案例的原子尺度示意图，其中体味挥发组分子主要暴露于石墨烯。(c) 10, 411 个体系中 $N_{d_{o-s<3.5\text{Å}}}$ 的分布。(d) 在相同超参数下，将 $N_{d_{o-s<3.5\text{Å}}}$ 添加到输入特征中后的 E_{ads} 预测。(e) 基于 (d) 中预测结果的 SHAP 分析蜂群图。(f) 移除 932 个离群值后的 E_{ads} 预测结果。(g) 使用 SHAP 值对离群值进行聚类的 UMAP 图。(h) 使用特征值对离群值进行聚类的 UMAP 图。(g) 和 (h) 中的绿色圆圈标出了离群值的位置。

7 学习曲线

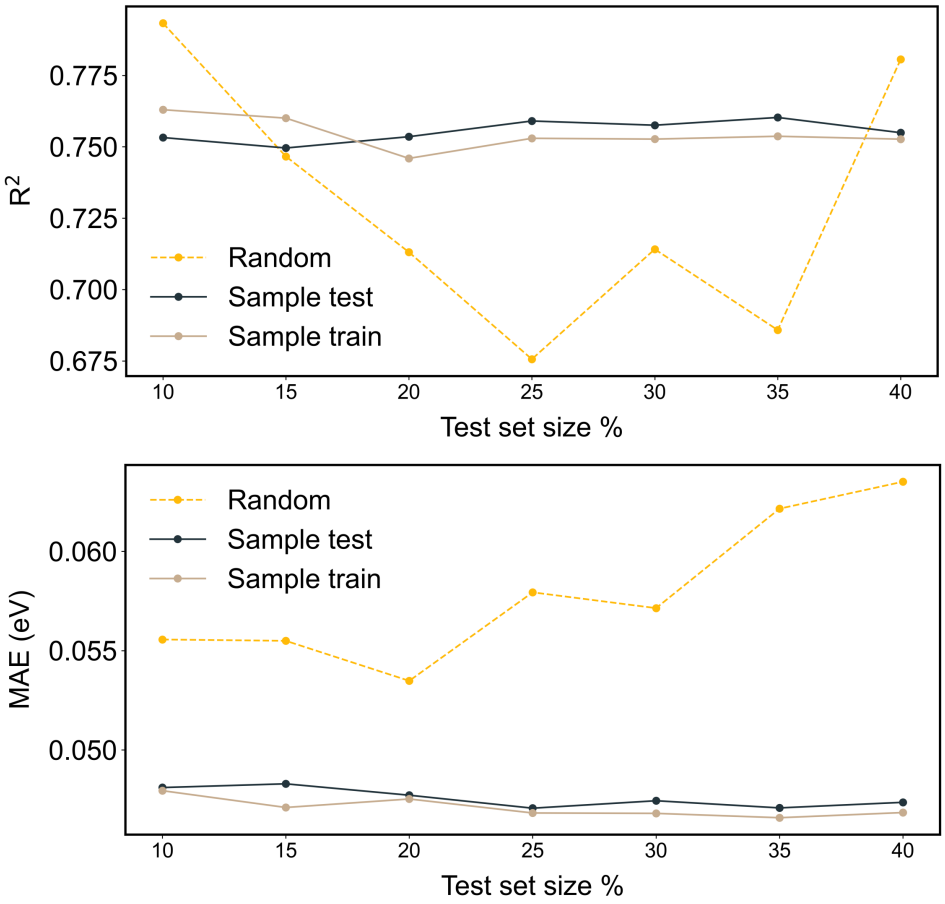


图 S6 使用 CatBoost（CAT）预测功函数变化 $\Delta\phi$ 时，测试集占整个数据集的比例从 10% 变化到 $\sim$ 40% 对应的学习曲线。上图：R平方分数。下图：平均绝对误差（MAE）。黄色虚线由随机划分生成。黑色和绿色点线由最远点采样法生成，分别在 t-SNE 空间中采样训练集（黑色）和测试集（绿色）。对于每个数据点，均通过贝叶斯优化对模型超参数进行优化。

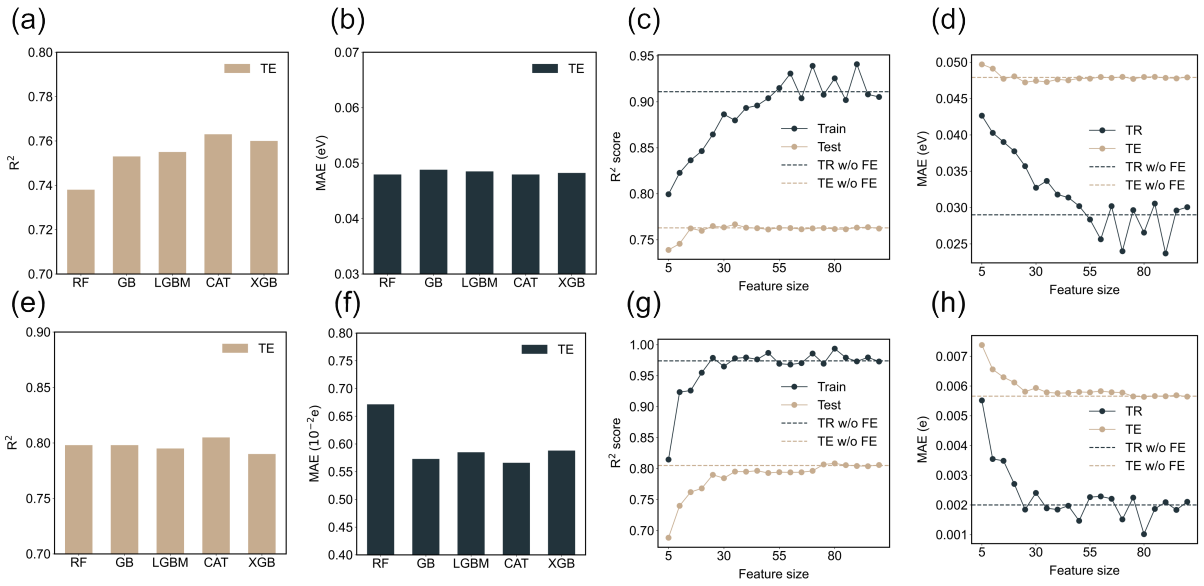
为了确定最佳训练集与测试集比例，我们进行了学习曲线研究，如图 S6所示。首先，我们可以观察到，与随机划分相比，使用最远点采样（FPS）方法采样时模型性能更加稳定；随机划分会随测试数据集大小变化，且无法保证训练集与测试集的分布相似，而最远点采样可确保采样数据集与源数据集之间呈均匀分布。因此，我们选择最远点采样作为划分方法。其次，使用最远点采样时，我们可以选择训练集或测试集，另一部分则为采样后的剩余数据。我们注意到，这一微小差异会导致模型性能略有差异。如图 S6所示，选择测试集时的模型性能通常略差于选择训练集时的表现，因为选择最具代表性的测试集也意味着选择了最具挑战性的测试集。因此，考虑到我们的不平衡数据集和模型性能，我们选择 9 : 1 的划分比例，并选取训练集。

8 超参数搜索空间

表 S5 本研究中用于贝叶斯优化的树模型超参数搜索空间。

模型	超参数	搜索空间
XGBoost	lambda，alpha	$[10^{-3}, 10^{-2}, ..., 10.0]$ (对数均匀)
	按列采样 按树采样，子采样	$[0.1, 0.2, \ldots, 1.0]$
	学习率 _	$[0.008, 0.010, \ldots, 0.020]$
	估计器数量	$[500, 1000, 3000, 5000, 7000]$
	最大深度	$[2, 4, 6, 8, 10, 12, 14]$
	最小子节点权重	$[1, 2, ..., 300]$ (整数)
RF	n_估计器	$[500, 1000, 3000, 5000, 7000]$
	最大深度	$[2, 4, 6, 8, 10, 12, 14]$
	最小分裂/叶节点样本数	$[2, 4, 6, 8, 10, 12, 14, 16, 18, 20]$
	最大特征数	$[0.1, 1.0]$ (浮点数)
GB	n_估计器	$[500, 1000, 3000, 5000, 7000]$
	学习率 _	$[0.008, 0.010, \ldots, 0.020]$
	最大深度	$[2, 4, 6, 8, 10, 12, 14]$
	最小分裂/叶节点样本数	$[2, 4, 6, 8, 10, 12, 14, 16, 18, 20]$
	子采样	$[0.1, 0.2, \ldots, 1.0]$
LightGBM	叶节点数	$[20, 150]$ (整数)
	最大深度	$[2, 4, 6, 8, 10, 12, 14]$
	学习率 _	$[0.008, 0.010, \ldots, 0.020]$
	估计器数量	$[500, 1000, 3000, 5000, 7000]$
	最小子节点样本数	$[5, 50]$ (整数)
	子采样，按树列采样 _	$[0.1, 0.2, \ldots, 1.0]$
CatBoost	迭代次数	$[500, 1000, 3000, 5000, 7000]$
	学习率 _	$[0.008, 0.010, \ldots, 0.020]$
	深度	$[2, 14]$ (整数)
	L2_叶节点正则化	$[1.0, 10.0]$ (浮点数)
	边界数量	$[32, 255]$ (整数)

9 模型基准测试信息



图S7 功函数变化 $\Delta\phi$ (a)~(d) 和电荷转移 ΔQ (e)~(h) 的模型基准测试。对于这两种结合特征，最佳模型仍为 CatBoost（CatBoost）。$\Delta\phi$ 和 ΔQ 的特征数量分别为 60 和 80。

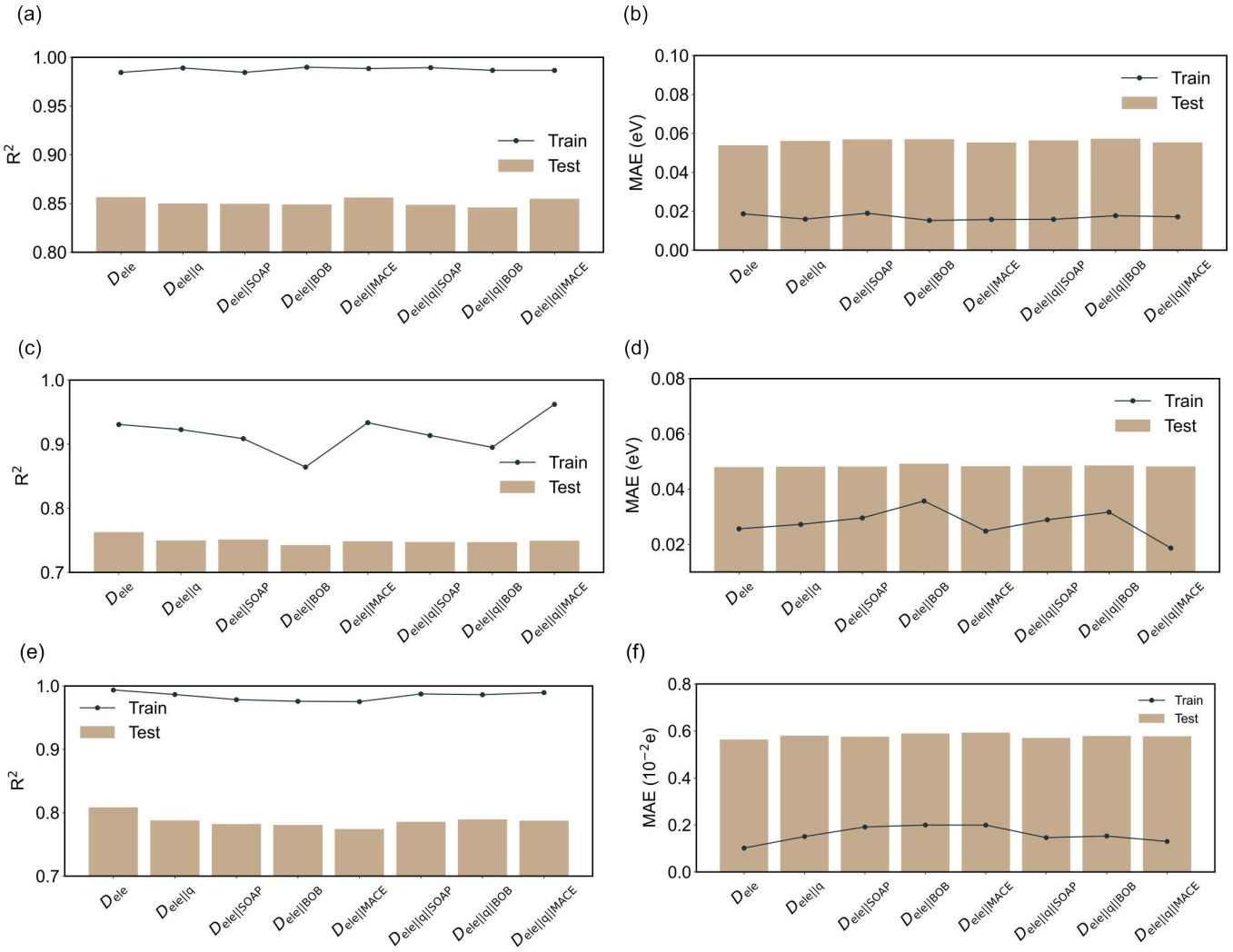


图 S8 结合 D_{BOB}、 D_{SOAP} 、D_{MACE}、 D_q 和量子力学性质所得的模型性能，其中：(a)~、(b)吸附能 E_{ads}、(c)~、(d)功函数变化 $\Delta\phi$、(e)~和(f)电荷转移 ΔQ

10 几何描述符 D_{geo} 和Mulliken原子电荷 q的性能

如图 S8 和 S9所示，我们系统地将训练好的模型与几何描述符 D_{geo} 和 Mulliken 原子电荷 q 相结合，以评估向量特征的能力。具体而言，我们对几何描述符进行主成分分析，以获得信息量最大的表达，相应地，它们记为 D_{BOB}， D_{SOAP} 长度分别为 63 和 100。我们还引入了包含多体色散相互作用信息的 MACE 描述符 D_{MACE}[3] 。此外，我们以类似于 BOB 描述符的方式，对原子 Mulliken 电荷 D_q 进行了人工构造。Mulliken 原子电荷 q 和 MACE 描述符的长度分别为 87 和 256，在总数据量中占比合理，为 $\sim$ 4%。我们加入这些输入特征，并再次重新训练模型。使用 D_{geo} 和 q 进行的所有结合特征预测结果如图 S8和 S9所示。

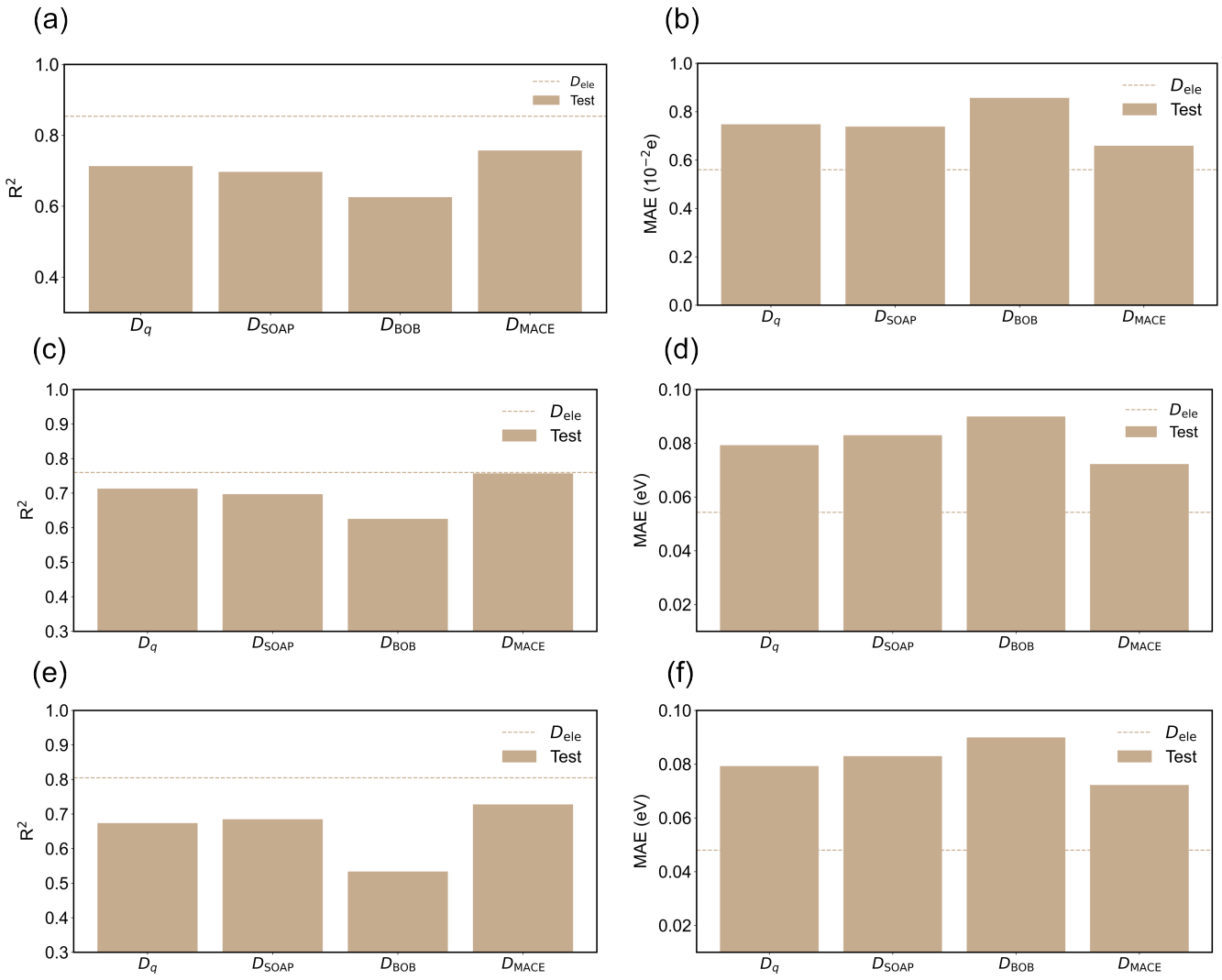


图 S9 与图 S8 但仅使用这些不包含电子性质的特征。

11 最终模型指标

表 S6 最终 CatBoost模型的结合特征预测指标。

	E_{ads}	$\Delta\phi$	ΔQ
超参数			
迭代次数	7000	5000	7000
学习率	0.02	0.01	0.02
深度	7	7	7
L2_叶节点正则化	9.240	2.387	1.341
边界数量	178	147	169
训练集			
R^2	0.984	0.931	0.994
MAE	0.019	0.026	0.001
RMSE	0.024	0.033	0.001
测试集			
R^2	0.857	0.763	0.808
MAE	0.054	0.048	0.006
RMSE	0.078	0.064	0.008

12 功函数变化 $\Delta\phi$ 预测的解释

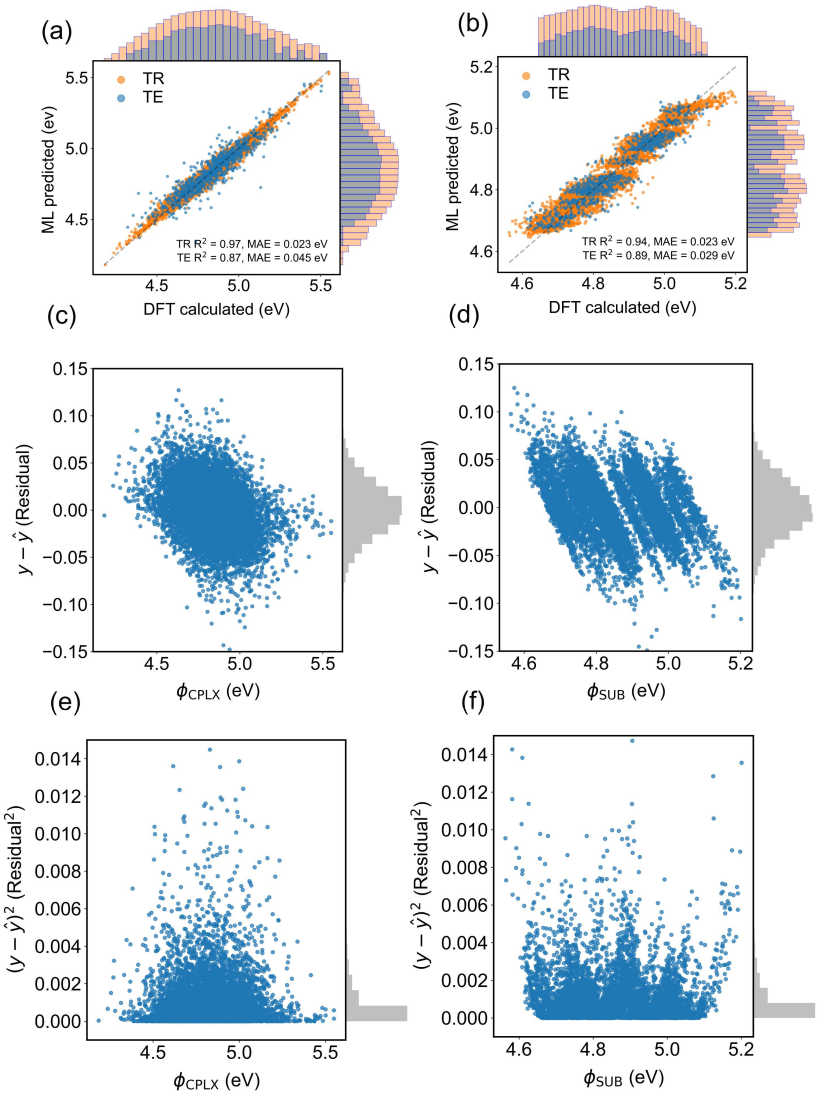


图 S10 （a）～（b） ϕ_{CPLX} 和 ϕ_{SUB} 预测的奇偶图。（c）～（d）残差与 ϕ_{CPLX} 和 ϕ_{SUB}的散点图。（e）～（f）残差平方与 ϕ_{CPLX} 和 ϕ_{SUB}的散点图。

如图 S10 （a）和（b）所示， ϕ_{CPLX} 和 ϕ_{SUB} 的单独预测表现良好，而 ϕ_{SUB} 的奇偶图呈现出异常模式。为探究 ϕ_{SUB}异常的预测行为，我们绘制了功函数与机器学习预测的功函数值之间的残差。在图 S10 （c）～（d）中，残差与 ϕ_{SUB} 呈现出令人意外的大量杆状形状，并具有线性相关性，而 ϕ_{CPLX} 未显示出任何特定模式。

从 $-0.1 \sim 0.1$电子伏特偏移到计算残差平均值时的极小总误差，呈现出振荡变化。因此，残差平方也表现出异常模式，如图 S10（f）所示。这种 ϕ_{SUB} 残差模式可能是由于受体表面缺乏多样性而导致系统误差，进而造成对功函数变化 $\Delta\phi$预测不完善。未来的工作可以着重于扩大受体的多样性。

参考文献

[1] Li Chen、Leonardo Medrano Sandonas、Philipp Traber、Arezoo Dianat、Nina Tverdokhleb、Mattan Hurevich、Shlomo Yitzchaik、Rafael Gutierrez、Alexander Croy 和 Gianaurelio Cuniberti。MORE-Q：一个基于量子力学的分子嗅觉受体工程数据集。Scientific Data，12(1)：324，2025年。

[2] Hans J¨urgen Kreuzer 和 Zbigniew W Gortel。《物理吸附动力学》，第 1 卷。施普林格科学与商业媒体，2012 年。

[3] D´avid P´eter Kov´acs、J Harry Moore、Nicholas J Browning、Ilyes Batatia、Joshua T Horton、Yixuan Pu、Venkat Kapil、William C Witt、Ioan-Bogdan Magdau、Daniel J Cole 等。《Mace-off：适用于有机分子的短程可迁移机器学习力场》。美国化学会志，147(21):17598–17611，2025年。