声明：本稿件由 UT-Battelle, LLC 根据与美国能源部签订的合同（编号：DE-AC0500OR22725）撰写。美国政府保留相关权利；出版商接受本文发表，即承认美国政府保留一项非独占、已付费、不可撤销、全球范围的许可，可为美国政府之目的发表或复制本稿件的已发表版本，或允许他人这样做。能源部将根据能源部公共访问计划（http://energy.gov/downloads/doe-public-access-plan），向公众开放这些由联邦政府资助的研究成果。

通过对化学空间进行假设驱动的主动学习，发现分子的结构—性质关系

Ayana Ghosh，[1] Sergei V. Kalinin[2] 和 Maxim A. Ziatdinov[1,3]

[1] 橡树岭国家实验室计算科学与工程部，美国田纳西州橡树岭，邮编 37831

[2]诺克斯维尔大学材料科学与工程系，美国田纳西州诺克斯维尔，邮编 37996

[3]橡树岭国家实验室纳米相材料科学中心，美国田纳西州橡树岭，邮编 37831

发现可用于药物靶点、生物分子系统、催化剂、光伏、有机电子学和电池的分子候选物，需要开发能够快速探索化学空间、以寻找具备所需功能的分子的机器学习算法。在此，我们提出一种基于假设学习、用于探索化学空间的主动学习新方法。我们利用一小部分数据构建关于结构与目标功能之间可能关系的假设，并将其作为高斯过程的（概率）均值函数。该方法将 SISSO 等符号回归方法的要素与主动学习融入同一框架。构建这一框架的主要目标是在主动学习机制下近似物理定律，以提升预测性能的稳健性；这是因为机器学习中传统的留出集评估无法考虑分布外效应，可能导致模型在未见化学空间中完全失效。本文以 QM9 数据集为例展示该方法，但它也可广泛应用于分子材料科学和固态材料科学两个领域的数据集。

电子邮箱：ghosha@ornl.gov

引言

化学发现[1-4] 植根于定量构效关系/定量构性关系（QSAR/QSPR）。[5-12] 这些研究主要依赖于寻找合适的分子表示，继而建立结构与分子表现出的活性之间的关系。人们利用这些模型探索化学空间，以筛选感兴趣的分子[13-15]，用于药物靶点、[16-20]抗生素、[21] 催化剂、[22-23] 光伏、[24-27] 有机电子学、[28] 氧化还原液流电池。[29] 此外，化学发现还包括理解化学过程，例如反应能量路径、[30-32] 优化反应条件、[33] （例如催化活性[34]）、结晶、[35-36] 对接[37] 和合成。[38-39]

定量构效关系/定量构性关系技术已被证明在所有此类场景中（但不限于此）都十分有用。其受欢迎程度[40-42] 源于其简便性：能够结合结构信息与物理化学性质；能够可靠地捕捉性质景观；能够识别现有的化学模式，并在数据中发现活性悬崖；同时计算成本也较低。这些描述符或特征可以是多维描述符，用于捕捉电子或拓扑特征；也可以是分子指纹，即通过基于图或字符串的表示（简化分子线性输入规范、[43] SELFIES[44]）对分子进行有效表征。定量构效关系/定量构性关系模型的历史可追溯至近 60 年前，Hansch 等人[45] 开展的开创性工作使用少量简单描述符来捕捉二维结构—活性关系。此后，该领域中用于性质/过程预测的各种传统机器学习算法（例如朴素贝叶斯、支持向量机、随机森林）的应用迅速增长，并通过实验合成进行验证。其成功也得益于易于访问的公共数据库（例如 PubChem、[46-48] ZINC、[49] ChEMBL、[50-51] QM9,[52] ANI-1x、[53] 和 QM7-X[54]）的建立，其中包含数千至数百万种分子的结构及物理化学性质（通过量子力学计算得出，或通过实验观测获得）。总而言之，这些进展为化学设计与发现铺平了道路，并在日常应用中发挥作用。

如果说定量构效关系/定量构性关系研究掀起了计算机模拟设计领域的革命，那么深度神经网络（NN）算法[55-56] 的应用则进一步加速了这一进程。这些算法不仅能够高效利用大数据来寻找感兴趣的分子，还能量化 [57-60] 分子相互作用和化学键合、针对目标逆向设计分子，并加深对机制的认识。然而，这些模型的性能都高度依赖于数据集的质量、数量和可建模性。此外，发现过程需要将已学到的相关关系外推至化学空间中此前未曾探索的区域。

相应地，如何突破标准（人工）设计规则或解决方案来生成[61] 新分子这一任务，已受到广泛关注。已有多项研究[62] 探索采用不同的基于神经网络的算法来应对这一挑战。自编码器是常见模型之一，可通过复杂的潜在表示对分子[63] 进行编码，以优化特定性质，并通过解码将其映射回分子结构。另一些实例则包括循环神经网络[64] 算法的应用，其中分子生成被视为序列任务，算法可以根据模型提供的信息，在

每个阶段生成样本。另有研究采用由自注意力驱动的 Transformer 模型[65-66] 进行目标结构生成。此外，在这一领域也已实施强化学习策略[67-68]，以寻找具有多个目标性质的分子。现代分子生成模型已将传统的分子字符串表示转变为嵌入空间[69]，其中包含关于整个分子骨架的信息。

然而，大多数生成模型背后的学习方法都是在潜在嵌入空间中进行探索的。潜在嵌入通常并不平滑，因此无法直接使用基于梯度的优化方法。在完整的分子库上训练后，具有实用功能的分子通常只占据化学空间的一小部分，因此发现这类分子颇为复杂。化学空间不可微，因此无法采用基于梯度的下降方法或简单的高斯过程（GP）[70-71] 方法。包括变分自编码器在内的机器学习方法旨在为化学空间构建合适的低维、理想情况下可微的潜在嵌入，从而支持贝叶斯优化（BO）[72-74] 类流程。然而，迄今为止，这些方法要么基于为整个候选分子库构建嵌入空间，要么基于为这些分子寻找相似的表示核，以用于目标探索。

从历史上看，高斯过程（GP）一直用于主动学习和贝叶斯优化，使这些流程本质上完全由数据驱动且非参数。它能够在相对低维的参数空间中插值拟合函数行为。在经典高斯过程中，会使用核函数（例如径向基函数核）来定义参数空间中的相关程度。核参数根据探索-利用步骤中获得的可用数据进行推断。该过程并未纳入系统物理或化学行为的任何先验信息，而是通过核函数从数据本身学习系统的物理规律，因此可能导致次优结果。因此，即使扫描低维参数空间，重建函数行为所需的优化步骤也会很多。在此类情况下，更先进的物理信息核函数[75-77] 或许有所帮助，这也是一个活跃的研究领域。然而，要在化学或物理科学领域恰当应用这一技术，就不能局限于在贝叶斯优化框架中使用传统高斯过程来建模感兴趣的功能。[78-81] 一系列分子核[82] 可用于高斯过程框架，包括指纹核（例如标量积核、Tanimoto 核）、字符串核（例如基于简化分子线性输入规范（SMILES）、SELFIES 的核）以及图核（通常使用分子片段）。选定性能最佳的核后，即可进行贝叶斯优化，以应用于现实场景，例如优化化学反应。

在近期的一项研究中，Ziatdinov 等人[83-85] 提出了一种用于主动学习和贝叶斯优化的物理增强算法。该算法将高斯过程模型的灵活性与物理先验相结合，从而实现机器学习中的假设驱动的发现。迄今为止，该算法已应用于扫描探针显微镜实验[86]，为浓度诱导相变提供了新的见解，并确定了铁电材料中的畴生长规律。

在此，我们将假设学习的概念扩展到分子发现领域，并将其与压缩感知方法相结合，以识别相关的结构描述符，并

通过奖励驱动的采集（类似于强化学习）评估自动生成的多个假设，从而选出下一个评估点。具体而言，我们对非线性泛函化特征的各种组合进行压缩感知，以筛选出最相关的组合，并据此选择假设。随后，我们通过平衡维度（即遵循物理约束），结合目标性质将这些假设构造成可行方程。在此，我们只考虑少数易于计算且与目标性质相关的特征，以使假设形式简单、易于计算。最后，在主动学习循环中，我们在广阔的参数空间内评估这些假设，以预测所关注的功能。我们采用孤立分子的 QM9 数据集作为用例，来构建本研究的这一部分。

我们注意到，构建此类框架的主要重点仍在于学习或近似主动学习机制中的广义物理

图1展示了一个通用框架示意图，详细说明了利用假设学习在设计与发现之间进行主动探索的过程。结合估计不确定性的结果表明，这是一种具有普适性且经济高效的近似结构-性质关系的方法，适用于各种材料体系。

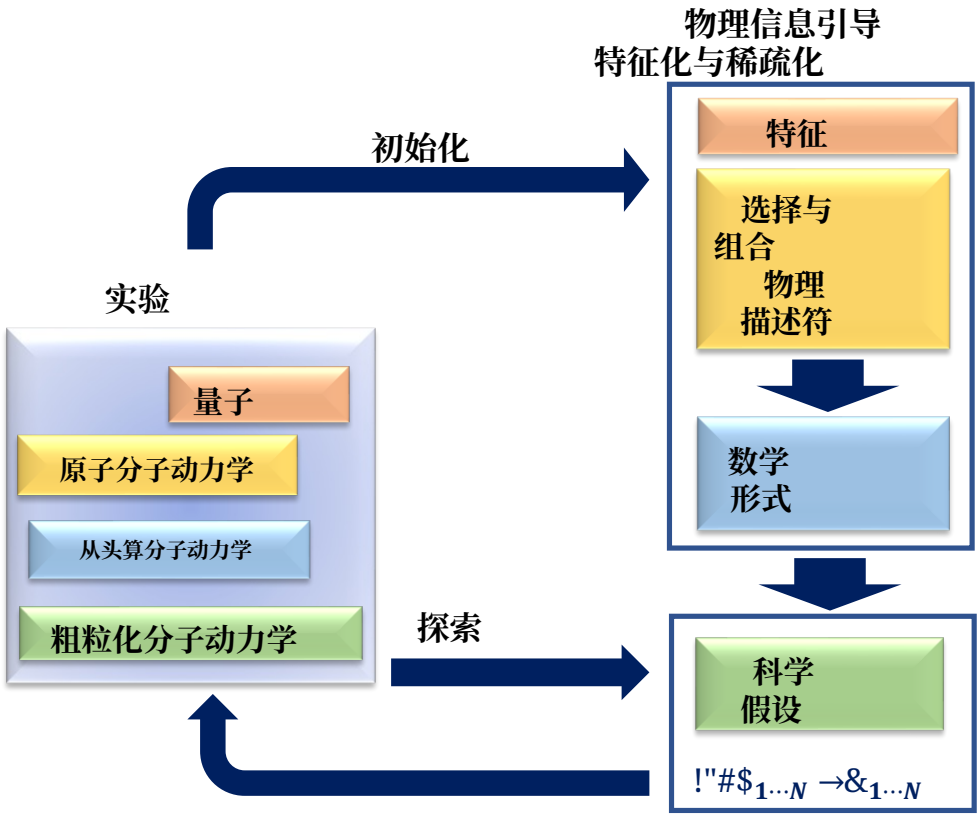

图1：从设计到发现的工作流程示意图。图（左侧面板）展示了常用的模拟技术，用于为各种材料体系生成可靠数据。右侧面板展示了主动学习循环：结合特征，提出数学形式，并将其作为通过统计方法推导出的科学假设，以评估和发现结构—性质关系。

结果与讨论

一般性考量：

我们设计的物理信息引导的特征化方案建立在 Ghiringhelli 等人用于特征选择的压缩感知方法之上[87]，并在此作为主动学习过程发现循环的种子步骤。与原始 SISSO 实现类似，我们的物理信息引导的特征化和稀疏化方案通过使用最小绝对收缩与选择算子（LASSO），实现对最相关描述符的选择。作为特征选择方案一部分的 LASSO 算法利用l_1 范数的稀疏性，有效地将描述符集缩减为完整集合（D）中最相关的描述符（d_i）。它选择目标性质（P）的$l1$正则化线性最小二乘近似中的非零项。目标性质可近似表示为$P(d) = dc$，其中c是与 Ω 维描述符d对应的系数（或权重）。该方程的解可通过最小化$argmin\ (||P - Dc||_2^2) + \lambda||c||_1$确定。

对于所有特征化描述符，系数 c 均非零；随后对这些描述符进行排序，以确定相应的重要性。

通过对每个特征应用多个数学算子（例如 $1/x$, $\sqrt{x}$, x^2, x^3, $\log(x)$, $1/\log(x)$, 以及 $\exp(x)$），生成描述符的非线性组合（d_i），从而在 D 与 P 之间建立非线性映射。此外，将这些函数化描述符求和组合所形成的复杂度，有助于生成更有效的 D 与 P 之间的映射。本研究考虑了由 2 个或 3 个项构成的函数化描述符。需要指出的是，纳入更多项可能会提高与终点的相关性，但也会引入由各项及其组合所用数学算子带来的额外不确定性。物理信息引导的特征化与稀疏化方法使我们能够通过线性组合整合多个特征，从而建立与终点目标的直接相关性。在此方法中，我们在庞大的组合空间中进行搜索，将特征组合后平衡单位/量纲（通过系数），以将其转化为可行方程。随后，通过为参数引入适当的先验，将这些数学形式转化为概率模型（假设），适用于所有使用场景。

所提出方法的第二个组成部分是基于 SISSO 推导出的函数形式开展假设驱动的主动学习。在假设学习中，我们使用结构化高斯过程（sGP），而非标准的零均值高斯过程作为代理模型，以便在高斯过程/贝叶斯优化框架中引入物理信息引导的先验。为说明这一方法，我们指出，在传统的高斯过程/贝叶斯优化过程中，高斯过程定义为

$$y = f(x) + \varepsilon,$$
$$f \sim MVNormal\,(m, K) \quad (1)$$

其中，多元正态分布表示多元正态分布，m为先验均值函数，通常设为 0，K 为先验协方差函数（核），而 e 为服从正态分布的观测噪声。

高斯过程模型的训练过程包括使用贝叶斯推断技术，根据现有的一组观测值（x，y）推断核参数。训练完成后，可以通过从某个分布中采样，获得函数在尚未测量的参数空间上的概率预测：

$$f_* \sim MVNormal\left(\mu_\theta^{post}, \Sigma_\theta^{post}\right) \quad (2)$$

$$\mu_\theta^{post} = m(X_*) + K(X_*, X|\theta)K(X, X|\theta)^{-1}\left(y - m(X)\right),$$
$$\Sigma_\theta^{post} = K(X_*, X_*|\theta) - K(X_*, X|\theta)(X, X|\theta)^{-1}K(X, X_*|\theta) \quad (3)$$

这里用 X_* 表示新的输入。对于每组核参数，我们都可以得到一个后验预测分布。接下来要评估的点由以下方式确定：

$$x_{next} = \arg max_x \frac{1}{M}\sum_{i=1}^{M} \alpha\left(\mu_{\theta^i}^{post}, \Sigma_{\theta^i}^{post}\right) \quad (4)$$

这里，α 是预定义的采集函数，M 是具有核参数的后验样本数量。

在 *s*GP 中，方程 1 中的先验均值函数被一个物理信息概率模型所替代，该模型的参数与核参数一同进行推断。方程 3 中的后验均值函数随后变为

$$\mu_{\theta^i\phi^i}^{post} = m\left(X_*|\phi^i\right) + K\left(X_*, X|\theta^i\right)K\left(X, X|\theta^i\right)^{-1}\left(y - m\left(X|\phi^i\right)\right)$$

(5)

因此，采用 *s*GP 的主动学习方案纳入了与模型参数相关的不确定性。该模型的参数化特性捕捉了基于物理的知识，用于确定下一步的探索点。

假设驱动学习利用了 *s*GP 模型的集成，这些模型有效地相互竞争，以在大型参数空间中重建物理行为。先验均值函数由 SISSO 方法导出。我们为所有 *s*GP 模型采用 Matérn 核，并将预测不确定性用作采集函数，α =diag(Σ_θ^{post}) 在每个探索步骤中，使用一种基础的强化学习策略（ε-贪心）来采样一个假设。采样得到的假设被封装到 *s*GP 中，其参数通过哈密顿蒙特卡洛进行推断。利用 *s*GP 参数的后验分布，可以对参数空间中尚未测量的部分进行概率预测。随后，将总预测不确定性与前一步的总预测不确定性进行比较；如果不确定性降低（升高），则对采样模型给予奖励（惩罚）。

分子的生成焓:

预测分子能量学以理解孤立分子的可形成性和稳定性，已成为当今机器学习模型最重要的应用之一。这些分子在材料设计和药物发现中都发挥着关键作用。在各种近似下进行的量子力学（QM）计算，是研究潜在化学原理、寻找活性位点以及理解化学反应的基础。在众多免费在线获取的、包含此类信息的数据集中，QM9 因数据同质性高、纯度高且缺乏噪声而受到研究人员的欢迎。因此，它已成为测试各种机器学习模型的经典基准。
QM9 数据集包含 ~130,000 种有机化合物的几何、能量、电子和热力学性质信息，这些化合物由 C、H、O、N 和 F 元素组成，并通过分子密度泛函理论，在 B3LYP/6- 31G (2df, p) 理论水平下计算得到。线性回归、核岭回归（KRR）、随机森林回归和深度神经网络等常见机器学习算法[88] 已被用于预测原子化能、[89] HOMO-LUMO 能量、[90] 生成能。[91]

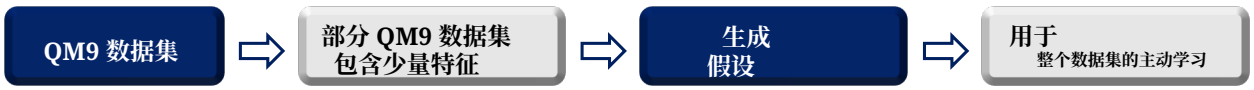

图 2：**QM9 数据集的流程图**。该图展示了框架应用于 QM9 数据集时的关键阶段。关键步骤包括：从原始 QM9 数据集中组合少量特征，生成部分数据集；随后通过物理信息引导的特征化方法形成科学假设，并在奖励驱动的假设驱动学习方案中对其进行评估。

文献中也有适当的基准测试研究，[92-99] 比较从传统机器学习到最先进神经网络（NN）的计算精度、耗时和算法收敛性。大多数此类研究依赖高维特征空间或分子表示来获得目标性质。通常，训练集和测试集按 80/20 的比例划分，用于执行基准测试任务，并据此对现有的结构-性质关系得出结论。

尽管付出了大量努力，人们仍未能清楚地了解如何在未知的化学空间（即使是在 QM9 中）建立结构—性质关系。换言之，要学习新分子及其相应的性质——这正是检验机器学习工具能力的好机会——我们还需要以数据高效的方式，在学习过程中即时发现将变量与感兴趣目标联系起来的物理或唯象定律。

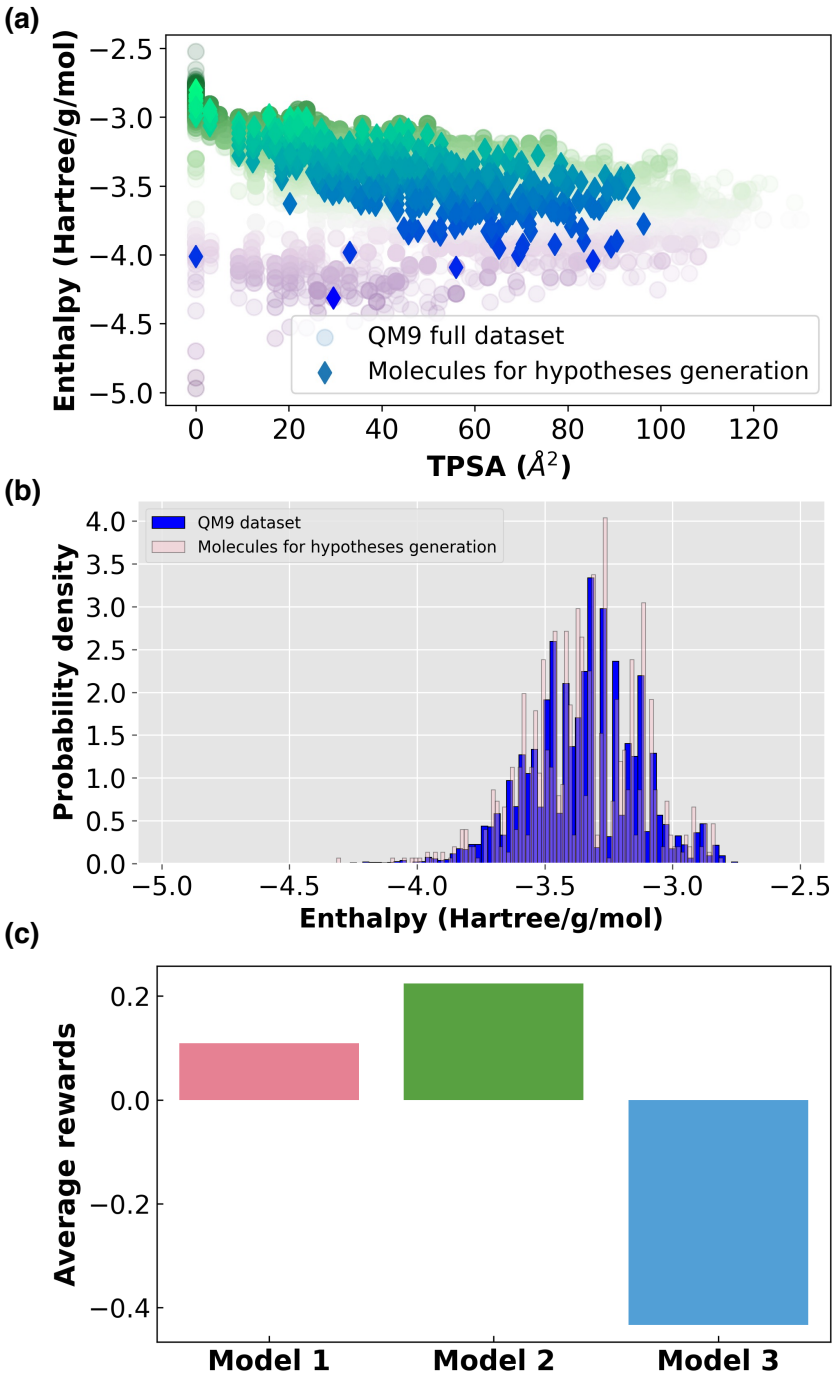

图 3：生成焓（Hartree/g/mol）的分布与预测。(a) 比较图；(b) QM9 数据集整体生成焓概率密度的直方图，并叠加用于生成假设的 1000 个分子的分布。(b) 主动学习方案中各个生成焓预测模型的平均奖励。

在这个例子中，我们将阐述如何运用假设驱动学习，重建 298.15 开尔文下生成焓（FE，Hartree/g/mol）随拓扑或

QM9 数据集的极性表面积（TPSA，Å²）。图 2 展示了本示例中所实现框架的示意图。我们从 QM9 数据集中选取前 1,000 个分子，并创建一个特征-目标数据集，其中包含若干特征，例如分子量（MW）、极性表面积（TPSA）、摩尔对数 P（molelogP）、空间范围（SP）和内能（IE）。MW、TPSA、molelogP 等特征由 RdKit 软件包的 Python 实现计算得出（易于获取且计算成本低）。SP 和 IE 等化学性质可直接从 QM9 数据集中获取；若采用其他方式计算，则需要使用计算成本高昂的密度泛函理论方法。

QM9 中前 1,000 个分子的化学概况（由利用结构信息计算出的特征所描述）与整个数据集所涵盖的情况存在很大差异。因此，面临的挑战是：判断使用 QM9 数据集中一小部分数据生成的简单数学表达式，能否推广至其余数据，从而对 FE 作出有意义的预测。图 3 展示了整个数据集（包含 133885 个分子）的 FE 分布，并将其与用于生成假设的数据集部分（占数据集的 0.74%）进行了比较。（a）图 3（b）中的 FE 归一化直方图展示了前 1,000 个分子对整个数据集的代表性。我们采用物理信息引导的特征化，提出以下三种相互竞争的假设。

$$FE = IE * \left(1 + \left(\frac{TPSA}{SP}\right)^2\right) \quad (6)$$

$$FE = IE * \left(1 + \left(\frac{TPSA}{SP}\right)^2 + molelogP^2\right) \quad (7)$$

$$FE = IE * \left(1 + \frac{molelogP}{1 + TPSA}\right) \quad (8)$$

我们已考虑下表 1 所列的所有参数的先验。接下来，我们将公式 6–8 视为封装在s高斯过程中的独立概率模型，以此实现假设驱动学习方案。

表 1: 此表列出了用于预测激发能的各公式中所有参数所考虑的先验。

参数	先验
IE（Hartree/g/mol）	均匀分布 [-4, 2]
SP（Å²）	均匀的 [2, 0.05]

用于生成假设阶段的分子不纳入训练和测试。我们从数据集中随机选取 300 个样本作为种子或起始点。以这种方式初始化五组不同的样本。每个模型在所有初始化设置下的平均奖励

经过 200 次探索后，如图 3 (c) 所示。模型2似乎累积了最高的奖励，能够重构 QM9 整个化学空间的 FE。

然而，若观察图4中绘制的单次运行结果，可以明显看出，模型1和模型2在探索过程中相互竞争。图4绘出了三种不同随机初始化下，各模型的中位数不确定性（a、c、e）和奖励（b、d、f）随探索步骤的变化情况

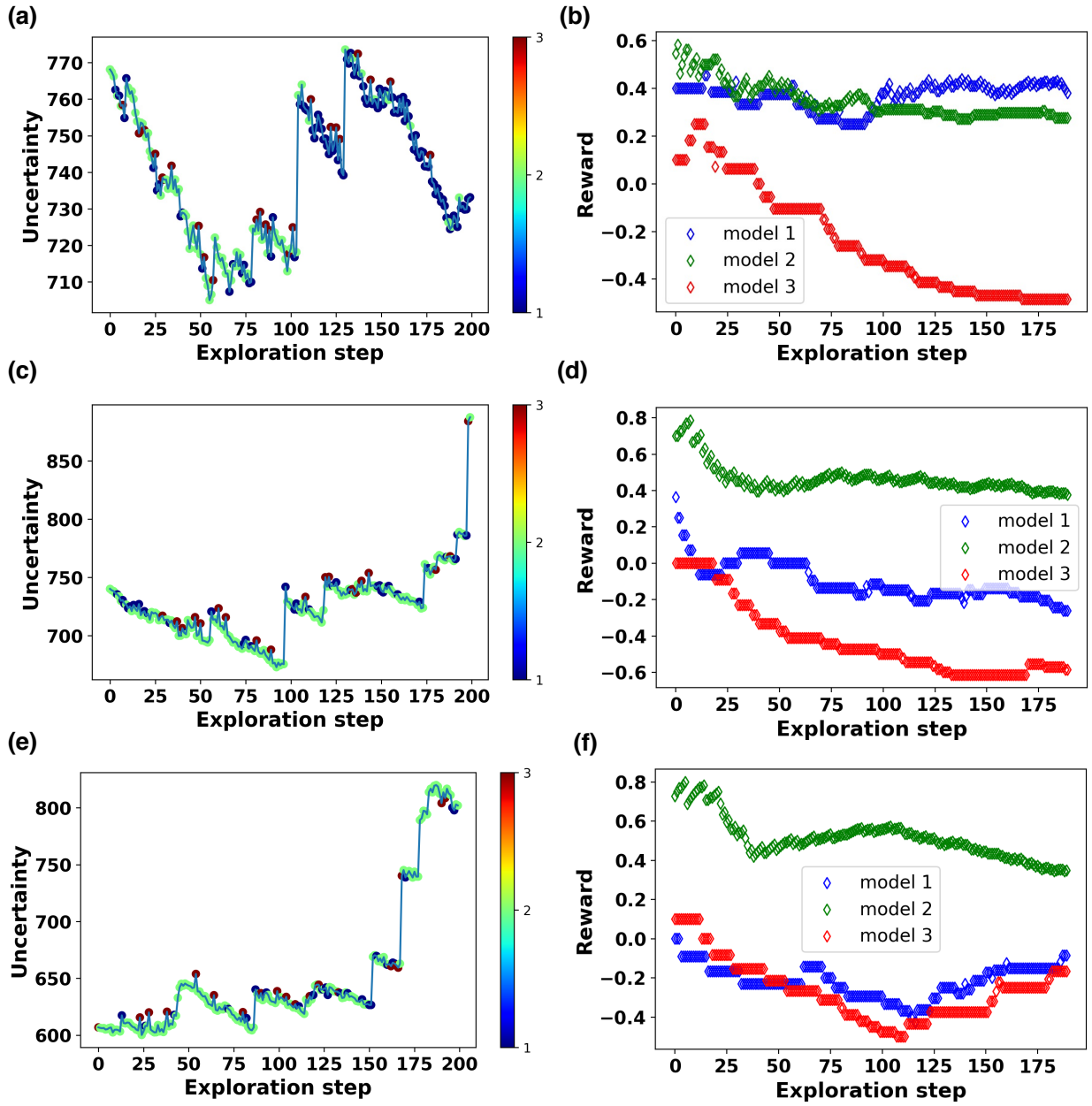

图4：针对整个 QM9 数据集，使用结构化高斯过程（sGPs）进行假设驱动的主动学习。此处展示了三种不同随机初始化下，各模型随探索步数变化的中位数不确定性（a、c、e）和奖励（b、d、f）。

。探索步骤中的某一阶段出现额外的不确定性跃升，这可归因于结构复杂性，以及学习过程中幂律的演变。此外，模型1和模型2给出的近似具有相似的函数形式，因此二者都能得出合理的预测。模型3获得的奖励始终最少，这表明这种特定近似未能捕捉整个数据集上 FE 的分布。模型给出的近似可能会因生成假设所依据的初始分子集合不同而略有差异。在构建模型表达式时采用高阶近似，仍可捕捉这些差异；预计这将纳入当前模型中已使用项的组合。

这个示例展现了该技术重建分子性质行为及其在广阔化学空间中变化规律的能力。它仅使用少量易于计算的特征和简单（更易解释）的数学形式，从而得出有意义的预测。

需要强调的是，构建此类框架的主要目标是在主动学习环境中学习或估计广义物理定律。通常，机器学习模型根据其在测试集/验证集上的表现进行评估，但这种方法忽略了真实世界数据的分布外特征所带来的影响。因此，依赖留出验证集的平均误差，可能无法准确反映机器学习框架对未见数据的预测能力。相反，通过观测数据近似广义物理定律，可以使框架更好地预测尚未探索的化学空间中的行为。

总之，我们以 QM9 数据集中的分子为例，展示了物理信息引导的特征化与假设驱动的主动学习相结合，在重建材料性质方面的能力。这有助于理解和近似属于不同材料类别的系统的功能行为，即使这些系统可能没有可用的模拟或实验数据。本文提出的框架为物理科学和机器学习界带来了若干重要进展。首先，它有望成为一种模板，用于构建易于解释的模型，以表征从既有观测值中获得的功能。该框架还尝试同时满足设计和发现的需求，而这对于推动物理科学的发展至关重要。

致谢

本研究由橡树岭国家实验室实验室定向研究与开发计划下的 INTERSECT倡议资助。该实验室由 UT-Battelle, LLC 管理，并根据合同 DE-AC05-00OR22725 为美国能源部开展工作。本研究的部分工作在纳米相材料科学中心完成，该中心是美国能源部科学办公室用户设施。

作者贡献

A.G. 提出构想、设计了工作流程并撰写了稿件初稿。S.V.K. 参与讨论了概念并参与写作。M.Z. 开发了结构化高斯过程和假设学习方法，并监督了稿件的准备工作。

数据可用性

本研究使用的所有数据集均可通过 GitHub代码库获取。https://github.com/aghosh92/SISSO_sGP

代码可用性

工作流程可通过 GitHub代码库获取。https://github.com/aghosh92/SISSO_sGP

利益冲突

作者声明不存在竞争性利益。

参考文献：

1. Tkatchenko, A.，用于化学发现的机器学习。Nature Communications 2020年，11。2. von Lilienfeld, O. A.，Müller, K.-R. 和 Tkatchenko，通过基于量子的机器学习探索化合物空间。Nat. Rev. Chem. 2020年，4。3. Bartók, A. P. 等，用机器学习统一材料与分子的建模。Sci. Adv. 2017年，3。4 2013年，12。Cutarolo, S.，通往计算材料设计的高通量之路。Nat. Mater.。5. Paula Carracedo-Reboredo, J. L.-B., Nereida Rodríguez-Fernández, Francisco Cedrón, Francisco J. Novoa, Adrian Carballal, Victor Maojo, Alejandro Pazos, Carlos Fernandez-Lozano，药物发现中机器学习方法与趋势综述。Computational and Structural Biotechnology Journal 2021年，19，4538-4558。6. J. Phillip Kennedy, L. W., Thomas M. Bridges, R. Nathan Daniels, David Weaver, and Craig W. Lindsley，组合化学科学在现代药物发现中的应用。Journal of Combinatorial Chemsitry 2008年，10，345-354。7. Jun Xu, A. H.，化学信息学与药物发现。Molecules 2002年，7，566-600。8. Muratov, E. N., Jürgen Bajorath, Robert P. Sheridan, Igor V. Tetko, Dmitry Filimonov, Vladimir Poroikov, Tudor I. Oprea 等，跨越边界的定量构效关系。Chemical Society Reviews 2020年，49，3525-3564。9. Tropsha, A.，定量构效关系模型开发、验证与探索的最佳实践。Molecular Informatics 2010年，29，476-488。10. A.Lill, M.，药物发现中的多维定量构效关系。Drug Discovery Today 2007年，12 (23-24)，1013-1017。11. Tropsha, A. G., Alexander，预测性定量构效关系建模工作流程、模型适用域与虚拟筛选。Current pharmaceutical design 2007年，13，3494-3504。12. Sheridan, R. P. F., B. P.; Maiorov, V. N.; Kearsley, S. K.，与训练集分子相似是衡量定量构效关系预测准确度的良好判别指标。J. Chem. Inf. Comput. Sci. 2004年，44，1912-1928。13. Butler, K. T., Davies, D. W., Cart wright, H., Isayev, O. & Walsh, A，分子与材料科学中的机器学习。Nature 2018年，559。14. Noé, F., Tkatchenko, A., Müller, K.-R. & Clementi, C.，用于分子模拟的机器学习。Ann. Rev. Phys. Chem. 2020年，71。15. Takahashi, S., Masamichi Takahashi, Shota Tanaka, Shunsaku Takayanagi, Hirokazu Takami, Erika Yamazawa, Shohei Nambu 等，多组学分析与机器学习开创的神经肿瘤学研究新时代。Biomolecules 2021年，11 (4)。16. José Jiménez-Luna , F. G. a. G. S.，借助可解释人工智能进行药物发现。Nature Machine Intelligence 2020年，2，573-584。

17. 等，B. N. e.，QSAR与保形预测方法的大规模比较及其在药物发现中的应用。J. Cheminf. 2019, 11. 18. Vamathevan, J., Dominic Clark, Paul Czodrowski, Ian Dunham, Edgardo Ferran, George Lee, Bin Li 等，机器学习在药物发现与开发中的应用。Nature Reviews drug discovery 2019, 18 (6), 463-477. 19. Ekins, S., Ana C. Puhl, Kimberley M. Zorn, Thomas R. Lane, Daniel P. Russo, Jennifer J. Klein, Anthony J. Hickey, and Alex M. Clark，利用机器学习实现端到端的药物发现与开发。Nature Materials 2019, 18 (5), 435-441. 20. Lo, Y. C., Rensi, S.E., Torng, W. and Altman, R.B.，化学信息学与药物发现中的机器学习。Drug Discovery Today 2018, 23 (8), 1538-1546. 21. Stokes, J. M. 等，利用深度学习发现抗生素。Cell 2020, 180. 22. Toyao, T., Zen Maeno, Satoru Takakusagi, Takashi Kamachi, Ichigaku Takigawa, and Ken-ichi Shimizu，催化信息学中的机器学习：近期应用与前景。ACS Catalysis 2019, 10 (3), 2260-2297. 23. Wenhong Yang, T. T. F., and Wen-Hua Sun，催化中的机器学习：从提出到实践。ACS Omega 2020, 5 (1), 83-88. 24. Sun, W., Yujie Zheng, Ke Yang, Qi Zhang, Akeel A. Shah, Zhou Wu, Yuyang Sun 等，机器学习辅助的高性能有机光伏材料分子设计与效率预测。Science Advances 2019, 5 (11), eaay4275. 25. B. Cao, L. A. A., A. O. Oliynyk, E. J. Luber, B. C. Olsen, A. Mar, J. M. Buriak，如何通过实验设计和机器学习优化材料与器件：以有机光伏为例。ACS纳米 2018, 12, 7434-7444. 26. J. Hachmann, R. O.-A., S. Atahan-Evrenk, C. Amador-Bedolla, R. S. Sánchez-Carrera, A. Gold-Parker, L. Vogt, A. M. Brockway, A. Aspuru-Guzik，哈佛清洁能源项目：在世界社区网上大规模计算筛选和设计有机光伏材料。J. Phys. Chem. Lett. 2011, 2, 2241-2251. 27. K. Sun, Z. X., S. Lu, W. Zajaczkowski, W. Pisula, E. Hanssen, J. M. White, R. M. Williamson, J. Subbiah, J. Ouyang, A. B. Holmes, W. W. H. Wong, D. J. Jones，一种用于高性能有机光伏的分子向列相液晶材料。《自然·通讯》2015, 6. 28. Gómez-Bombarelli, R. 等，通过高通量虚拟筛选与实验相结合的方法设计高效分子有机发光二极管。Nature Materials 2016, 15. 29. Er, S., Suh, C., Marshak, M. P. & Aspuru-Guzik, A.，用于全醌氧化还原液流电池的分子计算设计。Chem. Sci. 2015, 6, 885. 30. Segler, M. H. S., Preuss, M. & Waller, M. P.，利用深度神经网络和符号人工智能规划化学合成。Nature 2018, 555. 31. Fedorov, D. V., Sadhukhan, M., Stöhr, M. & Tkatchenko A.，原子偶极极化率与范德华半径之间的量子力学关系。. Physical Review Letters 2018, 121. 32. Wilkins, D. M. 等，结合耦合簇理论和机器学习准确预测分子极化率。Proc. Natl Acad. Sci. 2019, 116. 33. Zhenpeng Zhou, X. L., and Richard N. Zare，利用深度强化学习优化化学反应。《美国化学学会中心科学》2017, 3 (12), 1337-1344. 34. Jiayan Xu, X.-M. C. a. P. H.，多相催化中利用机器学习方法进行计算反应预测的观点。Phys. Chem. Chem. Phys. 23, 11155-11179. 35. Christos Xiouras*, F. C., Gustavo Lunardon Quilló, Mihail E. Kavousanakis, Dionisios G. Vlachos, and Georgios D. Stefanidis，人工智能与机器学习算法在结晶中的应用。Chem. Rev. 2022, 122 (15).

36. Ayana Ghosh, L. L., Kapildev K Arora, Bruno C Hancock, Joseph F Krzyzaniak, Paul Meenan, Serge Nakhmanson, Geoffrey PF Wood，评估机器学习方法对活性药物成分结晶倾向的预测能力。CrystEngComm 2019, 8, 1215-1223. 37. Shen, C., Junjie Ding, Zhe Wang, Dongsheng Cao, Xiaoqin Ding, and Tingjun Hou，Shen, Chao 等。《从机器学习到深度学习：蛋白质–配体对接评分函数的进展》。Computational Molecular Science 2020, 10 (1). 38. Coley, C. W., William H. Green, and Klavs F. Jensen.，计算机辅助合成规划中的机器学习。Accounts of chemical research 2018, 51 (5), 1281-1289. 39. Gromski, P. S., Jarosław M. Granda, and Leroy Cronin，借助“The Chemputer”实现通用化学合成与发现。Trends in Chemistry 2020, 2 (1), 4-12. 40. Sheridan, R. P.，领域适用性指标在估算QSAR预测误差中的相对重要性随训练集多样性而变化。J. Chem. Inf. Model 2015, 1098-1107. 41. Hanser, T. B., C.; Marchaland, J. F.; Werner, S.，适用域：迈向更正式的定义。. SAR QSAR Environ Res. 2016, 27, 865-871. 42. Baumann, D. B., K.，在模型不确定性下使用双重交叉验证可靠估算QSAR模型的预测误差。J. Cheminf. 2014, 6. 43. D, W.，简化分子线性输入规范（SMILES），一种化学语言与信息系统。1. 方法学与编码规则简介。J. Chem. Inf. Comput. Sci. 1988, 28, 31-36. 44. Krenn M, H. F., Nigam A, Friederich P and Aspuru-Guzik，自引用嵌入字符串（SELFIES）：一种100%稳健的分子字符串表示。Machine Learning: Science and Technology 2020, 1. 45. Hansch C, F. T.，p-σ-p分析：一种用于研究生物活性与化学结构相关性的方法。. J Am Chem Soc 1964, 86, 1616. 46. Wang, Y. X., J.; Suzek, T. O.; Zhang, J.; Wang, J.; Bryant, S. H.，PubChem：用于分析小分子生物活性的公共信息系统。Nucleic Acids Res. 2009, 37, W623-W633. 47. Li, Q. C., T.; Wang, Y.; Bryant, S. H.，PubChem作为药物发现的公共资源。. Drug Discovery Today 2010, 15, 1052-1057. 48. Kim, S. T., P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H.，PubChem物质与化合物数据库。Nucleic Acids Res. 2016, 44, D1202-D1213. 49. John J. Irwin*, T. S., Michael M. Mysinger, Erin S. Bolstad, and Ryan G. Coleman，ZINC：发现化学与生物学关联的免费工具。J. Chem. Inf. Model 2012, 52 (7), 1757-1768. 50. Bento, A. P. G., A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Krüger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; Nowotka, M.; Papadatos, G.; Santos, R.; Overington, J. P.，ChEMBL生物活性数据库：最新进展。Nucleic Acids Res. 2014, 42, D1083-1090. 51. Gaulton, A. B., L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P.，ChEMBL：用于药物发现的大型生物活性数据库。Nucleic Acids Res. 2012, 40, D1100–D1107. 52. Ramakrishnan, R., Dral, P., Rupp, M. & vonLilienfeld, O. A.，13.4万种分子的量子化学结构与性质。Sci. Data 2014, 1, 140022. 53. 等，J. S. e.，ANI-1ccx和ANI-1x数据集：分子的耦合簇与密度泛函理论性质。Sci. Data 2020, 7. 54. Hoja, J., Leonardo Medrano Sandonas, Brian G. Ernst, Alvaro Vazquez-Mayagoitia, Robert A. DiStasio Jr, and Alexandre Tkatchenko，QM7-X：涵盖小型有机分子化学空间的综合量子力学性质数据集。Sci. Data 2021, 8 (1), 1-11. 55. Mohit Pandey, M. F., Francesco Gentile , Olexandr Isayev, Alexander Tropsha, Abraham C. Stern and Artem Cherkasov，GPU计算与深度学习在药物发现中的变革性作用。Nature Machine Intelligence 2022, 4, 211-221.

56. Justin S. Smith, B. N., Nicholas Lubbers 等，《少即是多：利用主动学习对化学空间进行采样》。J. chem. Phys. 2018年，148。57. Chmiela, S., Sauceda, H. E., Müller, K.-R. & Tkatchenko, A,《利用机器学习力场实现精确分子动力学模拟》。Nat. Commun. 2018年，9。58. Schütt, K. T., Arbabzadah, F., Chmiela, S., Müller, K. R. & Tkatchenko, A,《深度张量神经网络带来的量子化学洞见》。Nat. Commun. Nat. Commun. 2017年，8。59. Deringer, V. L. 等，《利用机器学习驱动的分子动力学获得非晶硅的真实原子结构》。J. Phys. Chem. Lett. 2018年，9。60. Zubatyuk, R., Smith, J. S., Leszczynski, J. & Isayev, O,《利用原子分子神经网络准确且可迁移地进行化学性质多任务预测》。Sci. Adv. 2019年，5。61. Maria Korshunova, N. H., Stephen Capuzzi, Dmytro S. Radchenko; Yuriy S. Moroz, C. I. W., Timothy M. Willson, Alexander Tropsha and Olexandr Isayev,《用于生物活性化合物自动化从头设计的生成式与强化学习方法》。Communications Chemistry 2022年，5 (1)，1-11。62. Barzilay, W. P. W. a. R.,《深度学习在分子生成与分子性质预测中的应用》。Accounts of chemical research 2021年，54，263-270。63. Gomez-Bombarelli, R. W., J.N.;Duvenaud,D.;Hernańdez- Lobato, J. M.; Sań chez-Lengeling, B.; Sheberla, D.; Aguilera- Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A,《利用分子的数据驱动连续表示实现自动化学设计》。ACS Cent. Sci. 2018年，4，268-276。64. EJ Bjerrum, R. T.,《利用循环神经网络（RNNs）进行分子生成》。arXiv预印本 2017 年，arXiv:1705.04612。65. Viraj Bagal, R. A., P. K. Vinod, and U. Deva Priyakumar*,《MolGPT：利用Transformer解码器模型进行分子生成》。J. Chem. Inf. Model 2022年，62，2064-2076。66. Grechishnikova, D.,《作为机器翻译问题的蛋白质特异性药物从头生成Transformer神经网络》。Sci. Rep. 2021年，11。67. Popova, M., Olexandr Isayev, and Alexander Tropsha,《用于药物从头设计的深度强化学习》。Science Advances 2018年，4。68. Olivecrona, M., Thomas Blaschke, Ola Engkvist, and Hongming Chen,《通过深度强化学习进行分子从头设计》。Journal of Cheminformatics 2017年，9 (1)，1-14。69. Jin, W. B., R.; Jaakkola, T.,《用于分子图生成的连接树变分自编码器》。Proc. Mach. Learn. Res. 018，80，2323-2332。70. Kanagawa, M. a. H., P. and Sejdinovic, D. and Sriperumbudur, B. K.,《高斯过程与核方法：对其关联与等价性的综述》。arXiv预印本 2018年，arXiv:1805.08845v1。71. Ping Li, S. C.,《高斯过程潜变量模型综述》。CAAI Transactions on Intelligence Technology 2016年，1 (4)。72. Shahriari B, S. K., Wang Z, Adams R P and Freitas N D,《让人类退出闭环：贝叶斯优化综述》。Proc. IEEE 2016年，104，148-75。73. J, K. H.,《一种未知且随时间变化的函数形式的通用随机模型》。Math. Anal. Appl. 1962年，5，150-67。74. J, K. H.,《一种在噪声环境下确定任意多峰曲线最大值的新方法》。J. Basic Eng. 1964年，86，97-106。75. Wilson, A. G.; Nickisch, H.,《用于可扩展结构化高斯过程的核插值（KISS-GP）》。载于第32届国际机器学习大会论文集——第37卷，JMLR.org：法国里尔，2015年；第1775–1784页。76. Wilson, A. G.; Hu, Z.; Salakhutdinov, R.; Xing, E. P.,《深度核学习》，载于《人工智能与统计学》，PMLR：2016年；第370-378页。

77. Wilson, A.; Adams, R.，《用于模式发现与外推的高斯过程核》，载于国际机器学习大会，PMLR：2013年；第1067-1075页。78. Grégoire Ferré, T. H., and Kipton Barros，《利用局域图核学习分子能量》。Journal of Chemical Physics 2017年，146。79. Ramakrishnan, R., and O. Anatole von Lilienfeld.，《从化学空间中的一个核获得多种分子性质》。CHIMIA International Journal for Chemistry 2015年，69 (4)，182-186。80. Johannes A. Mohr, B. J. J., and Klaus Obermayer†，《分子核：一种无描述符和无对齐的定量构效关系方法》。Journal of Chemical Information and Modeling 2008年，48 (9)，1868-1881。81. Hiroshi Yamashita, T. H., and Ryo Yoshida，《分子上的原子环境核》。Journal of Chemical Information and Modeling 2014年，54 (5)，1289-1300。82. Griffiths, R.-R., Leo Klarner, Henry Moss, Aditya Ravuri, Sang T. Truong, Bojana Rankovic, Yuanqi Du 等，Griffiths, R. R., Klarner, L., Moss, H., Ravuri, A., Truong, S. T., Rankovic, B., ... & Tang, J. (2022)，《GAUCHE：化学中的高斯过程库》。载于 ICML 2022 第2届人工智能促进科学研讨会，2022年。83. Maxim Ziatdinov, A. G., Sergei V Kalinin，《物理带来不同：通过增广高斯过程实现贝叶斯优化与主动学习》。Machine Learning: Science and Technology 2021年，3 (1)，015022。84. Maxim Ziatdinov, Y. L., Kyle Kelley, Rama Vasudevan, Sergei V Kalinin，《扫描探针显微镜的贝叶斯主动学习：从高斯过程到假设学习》。ACS nano 2022年，16 (9)，13492-12512。85. Maxim A Ziatdinov, Y. L., Anna N Morozovska, Eugene A Eliseev, Xiaohang Zhang, Ichiro Takeuchi, Sergei V Kalinin，《自动化实验中的假设学习：在组合材料库中的应用》。Advanced Materials 2022年，22013345。86. Maxim A. Ziatdinov, Y. L., Anna N. Morozovska,Eugene A. Eliseev,Xiaohang Zhang,Ichiro Takeuchi,Sergei V. Kalinin，《自动化实验中的假设学习：在组合材料库中的应用》。Advanced Materials 2022年，2201345。87. Ouyang, R. C., S.; Ahmetcik, E.; Scheffler, M.; Ghiringhelli, L. M.，《SISSO：一种压缩感知方法，用于从海量候选项中识别最佳低维描述符》。Phys. Rev. Mater. 2018年，2。88. Gabriel A. Pinheiro*, J. M., Marinalva D. Soares, Ronaldo C. Prati, Juarez L. F. Da Silva, and Marcos G. Quiles*，《基于QM9量子化学数据集的简化分子线性输入规范表示预测九种分子性质》。Journal of Physical Chemistry A 2020年，124，9854-9866。89. Logan Ward, B. B., Ian Foster, Rajeev S. Assary, Badri Narayanan and Larry Curtiss，《基于低保真度量子化学计算，通过机器学习预测有机分子的准确原子化能》。MRS Communications 2019年，9，891-899。90. Gagliardi*, O. R. a. A.，《利用分子指纹混合，通过深度学习预测大型有机分子的总能量和轨道能量》。Journal of Chemical Information and Modeling 2020年，60，5971-5983。91. Dominik Lemm, G. F. v. R. O. A. v. L.，《基于机器学习的无能量分子、过渡态和固体结构预测》。Nature Communications 2021年，12。92. Zhenqin Wu, B. R., Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing and Vijay Pande，《MoleculeNet：分子机器学习基准》。Chem. Sci. 2018年，9，513-530。93. Zhang, S., Yang Liu, and Lei Xie.，《用于分子结构的、由分子力学驱动的多重图图神经网络》。arXiv预印本 2020年，(arXiv:2011.07457)。

94. Shui, Z., & Karypis, G., 用于预测分子性质的异构分子图神经网络。载于*2020*年 *IEEE*国际数据挖掘会议（*ICDM*）2020年；第492-500页。95. Unke, O. T., and Markus Meuwly, PhysNet：一种用于预测能量、力、偶极矩和部分电荷的神经网络。《化学理论与计算杂志》*15*（6），3678-3693。96. Schütt, K., Pieter-Jan Kindermans, Huziel Enoc Sauceda Felix, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller, Schnet：一种用于模拟量子相互作用的连续滤波卷积神经网络。神经信息处理系统进展**2017年，** *30*。97. Faber, F. A., Luke Hutchison, Bing Huang, Justin Gilmer, Samuel S. Schoenholz, George E. Dahl, Oriol Vinyals, Steven Kearnes, Patrick F. Riley, and O. Anatole von Lilienfeld., 机器学习的预测误差优于密度泛函理论精度。*arXiv*预印本**2017年，** （arXiv:1702.05532）。98. Choudhary, K., and Brian DeCost, 用于提升材料性质预测效果的原子线图神经网络。*npj*《计算材料》**2021年，** *7*（1），1-8。99. Chen, C., Weike Ye, Yunxing Zuo, Chen Zheng, and Shyue Ping Ong, Chen, Chi 等。“图网络：用于分子和晶体的通用机器学习框架。”《材料化学》**2019年，** *31*（9），3564-3572。