

机器学习加速的变分蒙特卡洛选择组态相互作用方法*

张金哲 刘金德 秦毅 刘仁杰 蒋刚†

四川大学, 原子与分子物理研究所, 成都 610065

精确的原子激发态数据对天体物理学、等离子体物理学和粒子物理学的发展至关重要。针对传统组态相互作用方法(Configuration Interaction, CI)在高精度计算中面临的组态空间爆炸问题。本文基于前期建立的变分蒙特卡洛组态相互作用方法 (Variational Monte Carlo configuration-interaction, VMCCI) 框架^[1], 引入机器学习 (Machine Learning, ML) 解决 VMCCI 方法使用蒙特卡洛采样效率较低的问题, 开发了 ML-VMCCI 方法框架。本文通过测试不同机器学习模型与算法的表现, 发现神经网络(Artificial Neural Network, ANN)、卷积神经网络(Convolutional Neural Network, CNN)、随机森林 (Random Forest, RF) 这三个机器学习模型有潜力加速 VMCCI 方法。同时, 将 ML 在迭代过程中的作用机制由对组态重要性的二分类, 改进为二分类与重要性排序相结合, 使用二者混合采样策略来提高重要组态的采样效率、采样质量和组态覆盖率, 加速组态筛选的收敛进程。进一步在 Ge-like Nd^{28+} 离子激发态计算的能级计算上验证了该方法在计算原子激发态上的有效性和加速效果。结果表明 ML-VMCCI 方法在保持与参考组态空间计算能级相当精度的前提下, 显著加速了组态筛选的收敛进程, 提高了重要组态采样效率, 大幅降低了计算资源消耗, 为超大基组的高精度原子结构参数计算提供了可行的解决方案。

关键词: 相对论原子结构计算, 选择组态相互作用方法, 变分蒙特卡洛方法, 机

* 国家自然科学基金(批准号: 12474253)资助的课题

† 通信作者. E-mail: gjiang@scu.edu.cn

1 引言

精确的原子结构参数和光谱信息对于天体物理中的光谱测量、等离子体物理中的特性诊断以及粒子物理和宇宙学中的实验验证至关重要^[2-5]。特别是激发态原子结构参数在开发高精度的原子光钟、探索超精细常数 α 的变化以及暗物质探测等领域发挥着重要作用^[6-9]。理解激发态过程能够帮助本文准确地解释各种物理现象，从而提高我们对物理学的理解。

精确的激发态光谱参数获得，需要依赖高精度的原子结构计算。目前高性能从头计算原子结构代码提供了原子和离子的广泛电子性质，如能级、辐射跃迁率、 g 因子或超精细结构常数等。然而精确的激发态原子结构参数计算的实际困难在于原子序数 Z 高、电子多的原子或离子中的多体效应。现有方法在描述波函数时各有优劣，其中多体微扰方法（Møller–Plesset perturbation theory, MBPT）在处理高激发态或强电子关联系统时难以准确描述复杂的电子关联效应^[10,11]，耦合簇（Coupled cluster, CC）方法和 CI 方法精确计算激发态的成本会随着系统规模增大而迅速增加（其中 CC 方法成本 $\sim O(N^{5-10})$ ，全组态相互作用（Full Configuration Interaction FCI）方法成本 $\sim O(\exp(N))$ ）^[12-18]。特别是对于高 Z 元素激发态，在使用 CI 方法进行精确计算时所需组态波函数（Configuration State Function, CSF）的数目往往膨胀到 10^7 以上。研究者们目前主要通过 CI 方法处理复杂的电子关联效应^[19-24]。得益于 CI 方法中的哈密顿矩阵特征向量具有极端稀疏性，使得在可接受精度范围内压缩 CSF 数量、节省计算资源成为可能。选择组态相互作用

方法 (Selected Configuration Interaction SCI) 被开发出来, 显著推动了计算方法的研究进展^[25-27]。

SCI 方法旨在迭代过程中通过不同的方式高效筛选出对整体波函数有显著贡献的组态, 通过保留显著贡献组态并丢弃低贡献组态的方式得到 FCI 问题的近似解^[28-31]。该类方法基于微扰理论或蒙特卡洛方法, 通过迭代筛选重要组态来构建紧凑波函数, 从而实现高精度近似。这些方法的核心都是依据固定组态贡献截断标准, 通过不同采样策略迭代构建紧凑波函数^[32-37]。随着大数据时代的到来, 机器学习算法为提高组态筛选效率提供了新的方法^[38-41]。基于机器学习的 SCI 方法主要分为两类: 监督学习二分类方法和智能体随机优化算法。在监督学习二分类方法中, 模型关注的是组态重要性概率和分类边界。此类方法的计算效率收益受机器学习成本与从头算成本的叠加制约^[42-45]。智能体随机优化算法无需构建数据集, 如强化学习, 但训练时间较长, 在复杂的激发态问题上可能遇到收敛困难, 难以快速收敛到最优策略^[46]。到目前为止, 该类方法的研究主要集中于分子基态的计算, 并提出了不同的扩展方法来计算分子激发态^[47,48]。然而, SCI 方法在原子和固体物理计算上的研究较少。

在原子物理计算方面, Pavlo Bilous 等人的研究发展出了原子组态描述符, 利用机器学习方法迭代构建原子波函数, 在多组态 Dirac-Hartree-Fock (Multi-configuration Dirac-Hartree-Fock, MCDHF) 方法下成功应用 SCI 方法, 验证了原子结构领域进行组态筛选的可行性和 ML 有效性^[49,50]。Pavlo Bilous 等人的研究对原子结构计算做出了重要贡献, 但组态筛选流程的稳定性和收敛性方面尚未得到充分验证, 不同机器学习模型和算法的表现仍有待系统研究。在之前的工作中, 我们通过引入 VMC 方法对能量的评估实施拒绝-接受采样机制, 以提高收敛稳

定性，确保能量收敛到固定组态空间的能量上界，构建了基于 VMC 的原子结构组态筛选框架。同时在激发态计算方面，根据不同 J 值划分组态空间，使用 Extended Optimal Level (EOL)方法同时计算相同 J 值的激发态能级。这种处理方法有效减弱了引入新组态对激发态收敛性的影响，能够稳定的输出多条高精度激发态能级^[1]。然而 VMCCI 方法采用蒙特卡洛 (Monte Carlo, MC) 进行组态采样，其重要组态的采样效率受限于待采样空间中真正重要组态的占比。这导致 VMCCI 方法在迭代后期出现筛选乏力、收敛缓慢的问题。

在本文中，为在确保组态筛选方法收敛稳定性和可靠性的前提下提高重要组态的筛选效率、加速收敛进程，本文将 ML 方法引入 VMCCI 框架，开发了 ML-VMCCI 组态筛选方法。首先，在 LuI 体系上系统测试了经典机器学习模型 ANN、RF 等，以及深度学习模型 CNN 替代 MC 采样进行组态重要性分类采样的可行性与加速效果。在此基础上，为平衡整体算法的精度与效率，本文探索了三种采样策略：监督学习的组态重要性二分类、组态重要性排序以及二者的混合使用。结果表明，单纯分类受限于纯度下降，单纯排序则面临覆盖不足的问题，而混合采样通过同时兼顾"优先级排序"与"类别判别"，更好地满足了组态筛选中质量与覆盖度并重的物理需求。为验证基于混合采样策略的 ML-VMCCI 方法的有效性，本文计算了 Ge-like Nd^{28+} 离子的激发态能级，并与 VMCCI 方法进行了效率与精度对比。结果表明，ML-VMCCI 方法在保持与参考组态空间计算相当精度的前提下，显著加速了组态筛选的收敛进程，提高了重要组态采样效率，大幅降低了计算资源消耗，为超大基组的高精度原子结构参数计算提供了可行的解决方案。

2 理论方法与计算框架

2.1 MCDHF 方法

VMCCI 方法的实现基于 MCDHF 方法的 GRASP2018 程序包^[51]。Grant 和 Fischer 等人深入解释了 MCDHF 方法的理论细节，其实现包含在 GRASP2018 程序包中。在这项工作中，本文只简要介绍该方法的突出特征。在 MCDHF 方法中，组态构成的原子波函数（Atomic State Function, ASF）表示为^[52]：

$$\Psi_0(\Gamma JM_J \pi) \equiv |\Gamma JM_J \pi\rangle = \sum_{i=1}^{N_0} c_i^{\Gamma J} \Phi(\gamma_i JM_J \pi) \quad (1)$$

上式中，对于系数 c_i 有 $\sum_{i=1}^{N_0} c_i^2 = 1$ ， γ_i 表示完整描述 CSF 所必需的量子数， π 表示宇称， J 是总角量子数， $c_i^{\Gamma J}$ 为展开系数。CSF 可以用单电子轨道波函数的耦合反对称乘积表示：

$$\psi_{n\kappa m}(r, \theta, \varphi) = \frac{1}{r} \begin{pmatrix} P_{n\kappa}(r) \Omega_{\kappa m}(\theta, \varphi) \\ i Q_{n\kappa}(r) \Omega_{-\kappa m}(\theta, \varphi) \end{pmatrix} \quad (2)$$

其中 $P_{n\ell}(r)$ 和 $Q_{n\ell}(r)$ 分别表示径向轨道波函数的大、小分量， $\Omega_{lsm}(\theta, \varphi)$ 是单电子波函数的角向分量。在 MCDHF 方法中，原子态能量 E 和系数 c_i 是采用变分法通过自洽场迭代得到的。在自洽场迭代过程中首先求解 Dirac-Hartree-Fock 方程，得到一组径向波函数。随后采用相对论自洽场(Relativistic Self-Consistent Field, RSCF)程序来优化狄拉克轨道和系数 c_i ，以实现计算中的自洽性。随后，相对论组态相互作用(relativistic configuration interaction, RCI)计算考虑了 Breit 相互作用和量子电动力学(Quantum Electrodynamics, QED)相关性，包括自能和真空极化。

2.2 VMCCI 方法

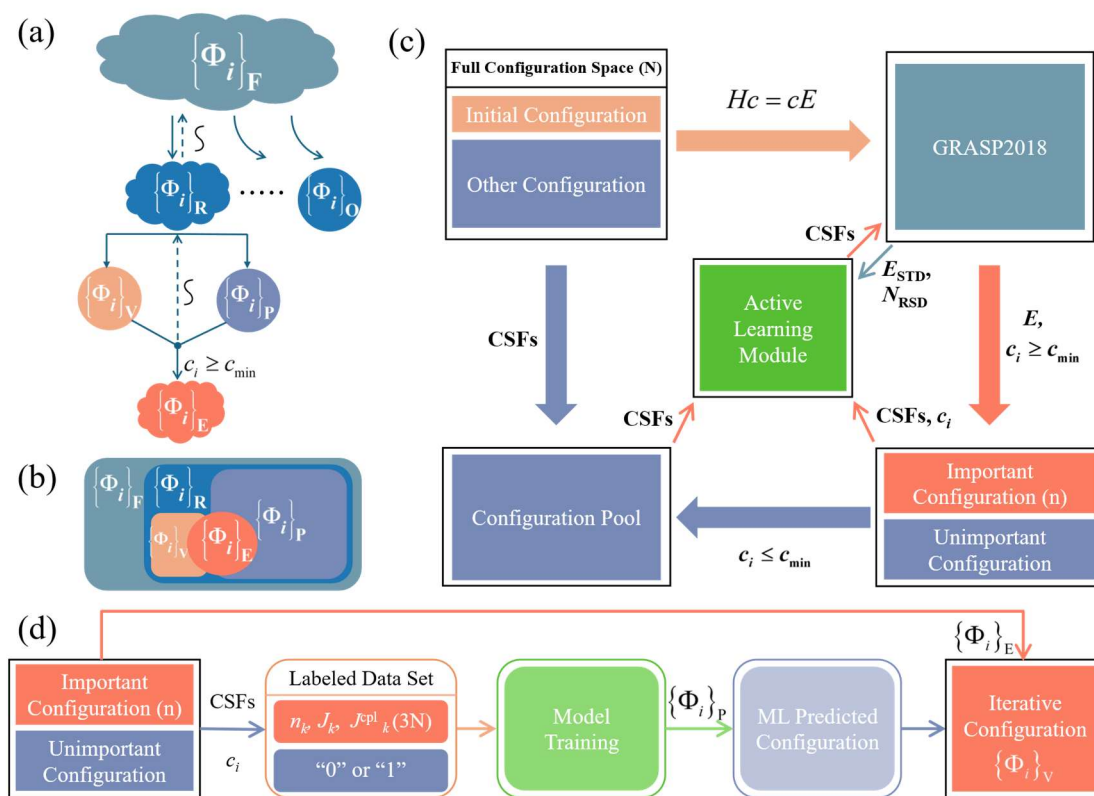


图 1 原子结构组态筛选框架示意图，其中(a)-(c)改编自文献^[1]中 VMCCI 方法的基本流程；(a) VMCCI 中涉及的集合变化过程；(b) VMCCI 中涉及的集合关系；(c) SCI 计算的原子结构 ML-VMCCI 框架示意图；(d) 主动学习模块的机器学习训练和采样具体过程

Fig. 1. Schematic diagram of the atomic-structure configuration-screening framework: (a) evolution of the sets involved in VMCCI; (b) relations among the sets involved in VMCCI; (c) schematic diagram of the atomic-structure ML-VMCCI framework for SCI calculations; (d) specific process of ML training and sampling for active learning modules.

VMCCI 方法的基本流程和主要步骤已在文献^[1]中详细介绍。为保持本文方法叙述的完整性，图 1(a)-(c)保留了 VMCCI 框架中的关键流程，并在图题中注明

其改编自文献[1]的图 1。本文的主要改进集中在图 1(d)所示主动学习模块：在原 VMCCI 迭代筛选流程中，机器学习采样被引入组态筛选，用于替代 MC 采样策略提高重要组态的筛选效率。因此，本节仅对 VMCCI 方法的主要特征作简要介绍，并在 2.3 节中重点阐述本文在原有 VMCCI 框架基础上引入的 ML 改进。

在 CI 方法中，ASF 表示为 $\Psi = \sum_{i=1}^{\infty} c_i \Phi_i$ ，其中无限个行列式波函数 Φ_i 构成的完备集表示为 $\{\Phi_i\}_F$ [53]。由于在计算中无法直接使用 $\{\Phi_i\}_F$ 进行计算。因此在本工作中采用多组态 Dirac-Hartree-Fock 方法中的活性空间方法。首先选取目标态的主导组态及与其强相互作用的近简并组态作为多参考组态集（multireference configurations, MR）。随后将 MR 中的占据轨道按价层和芯层划分，并根据所需关联模型逐步开放相应壳层：价-价关联（valence-valence correlation, VV）由价层电子的激发描述，芯-价关联（core-valence correlation, CV）由芯层与价层电子的联合激发描述，芯-芯关联（core-core correlation, CC）则通过开放更深芯层电子的激发加以考虑。在给定的活性轨道 nl 中，允许参考组态中被开放的占据轨道电子经单、双以及必要时三、四重等激发进入活性轨道集 nl 中，从而生成满足给定 J 对称性的有限个行列式函数集合 $\{\Phi_i\}_R \subseteq \{\Phi_i\}_F$ 。随后筛选出影响显著的行列式函数集合 $\{\Phi_i\}_E$ ，实现 $\{\Phi_i\}_E \sim \{\Phi_i\}_R$ 。图 1(a)展示了 VMCCI 中涉及的集合变化过程，图 1(b)展示了 VMCCI 中涉及的集合关系。

在组态筛选过程中，VMCCI 中引入了拒绝-接受采样机制，根据变分原理可知，能量的本征值越小，则对应的原子态波函数的性质越好。当 E 降低时，波函数得到改进，马尔科夫链蒙特卡洛(Markov Chain Monte Carlo, MCMC)接受这一变化 $\Psi_{m-1} \xrightarrow{p} \Psi_m$ [54]。随后对 $\{\Phi_i\}_V$ 中的组态进行检查，将 $c_i \geq c_{\min}$ 对应的组态添加到集合 $\{\Phi_i\}_E$ 中。同时自 $\{\Phi_i\}_R$ 中通过 ML 采样选择新的组态与 $\{\Phi_i\}_E$ 一起构建新的

$\{\Phi_i\}_V$ 。这种策略能够逐步优化波函数，并消除由初始组态选择引入的偏差。随着迭代的进行，对波函数影响显著的组态构成的集合 $\{\Phi_i\}_E$ 计算得到的结果逐步近似 $\{\Phi_i\}_R$ 的结果，直至达到收敛标准，实现 $\{\Phi_i\}_E \sim \{\Phi_i\}_R$ 。当 $c_{\min} \rightarrow 0$ 时，参考集合 $\{\Phi_i\}_R$ 的能量和波函数得到复现，表示为 $\{\Phi_i\}_E = \{\Phi_i\}_R$ 。

2.3 ML-VMCCI 算法流程

本工作选用了 8 种基于不同机理的机器学习分类模型，涵盖了从简单线性到深度非线性的方法论谱系，包括 ANN、RF、K 近邻算法（K-Nearest Neighbors, KNN）、朴素贝叶斯（Naive Bayes, NB）、逻辑回归（Logistic Regression, LR）、决策树（Decision Tree, DT）、梯度提升（Gradient Boosting, GB）以及 CNN。以便全面评估各范式在原子组态筛选任务中的适用性，使其在面对不同规模的组态数据时呈现显著的行为差异，为系统性对比提供了方法论基础^[55-61]。本文分别测试了利用机器学习模型进行组态重要性二分类任务以及组态重要性排序任务来进行组态采样。对于组态的特征化，使用利用文献^[49]中提到的原子结构 CSF 描述符处理方法。基于 N 轨道的 CSFs $|\Phi_\alpha\rangle$ 的唯一特征是 $3N$ 个量子数，统称为 α 。对于轨道 k ，它们由轨道占据电子数 n_k 、电子的总角动量 J_k 和耦合角动量 J^{pl}_k （由 J_k 和 J^{pl}_{k-1} 耦合）组成。同时，将 n_k 按轨道最大容纳电子数归一化， J_k 和 J^{pl}_k 按目标能级总角动量归一化，这一行为有助于模型关注特征本身而非特征数值。以上 $3N$ 个轨道量子数构成模型输入特征。每个 CSF 的重要性标签由 Grasp2018 通过哈密顿矩阵求解给出的 c_i 决定， $c_i \geq c_{\min}$ 的 CSF 被认定为是重要组态（标签为“1”）， $c_i \leq c_{\min}$ 的 CSF 被认定为是非重要组态（标签为“0”）。最终得到样本标签以后构成完整的输入描述符为 $3N+1$ 的一维数组。

在迭代过程中，可以根据 SCI 计算的结果更改标记组态池中的 CSF 重要性标签。来自标记组态池的数据子集用于训练 ML 模型(70%训练集,30%验证集)，然后使用训练好的 ML 模型来预测未标记组态池的 CSFs 是否重要或做 CSFs 的重要性排序。同时关注模型在迭代过程中每一步采样时的重要组态采样效率来反映模型真实作用。最终通过对模型进行超参数调优以及 5 折交叉验证 (5-fold cross-validation, cv5)，去除超参数对于模型测试结果的影响^[43]。详细的机器学习超参数调优搜索范围和结果见补充材料附录 I。

本研究的计算流程包括初始化、MCDHF 计算、波函数优化、机器学习训练与采样以及收敛性检查五个阶段。整体算法框架如图 1(c)所示，其中本文的主要改进体现在图 1(c)的主动学习模块。主动学习模块详细流程见图 1(d)：在 VMCCI 迭代筛选过程中引入机器学习模型，以替代原有 MC 采样策略提高重要组态的筛选效率。算法步骤如下（详细算法伪代码图见补充材料附录 II）：

(1) 利用 GRASP2018 中的 CSF 生成器构造 $\{\Phi_i\}_R$ 。自 $\{\Phi_i\}_R$ 中根据初始化比例初始化 $\{\Phi_i\}_V$ 。同时初始化 c_i 截断值，拓展比例（指当次迭代保留的重要组态数目与新采样组态数目的比例），收敛标准和目标谱项。

(2) 根据 $\{\Phi_i\}_V$ 构造哈密顿矩阵并对角化，通过自洽场迭代过程求解 Dirac-Hartree-Fock 方程和本征值方程 $Hc = cE$ ，得到 E, c_i 。在求解 Dirac-Hartree-Fock 方程过程中采用 EOL 方案优化能级。

(3) 根据 E, c_i 以及光谱项，判断目标波函数是否得到改进。若目标波函数未得到改进则丢弃该轮新加入的组态，保留当前已筛选出的重要组态，重新采样并回到步骤(2)继续迭代；若连续 L 次波函数未得到改进，则筛选过程能量和重要组态数量的收敛标准满足，迭代终止。

(4) 若目标波函数得到改进则利用文献^[49]中提到的原子结构 CSF 描述符处理方法将 $\{\Phi_i\}_V$ 中的组态按照 $c_i \geq c_{\min}$ 的标准处理为带重要性标签的数据集，随后按照主动学习方法对 ML 模型进行训练。

(5) 将 $\{\Phi_i\}_V$ 中满足条件 $c_i \geq c_{\min}$ 的组态添加到 $\{\Phi_i\}_E$ 中。随后根据设定的拓展比例，从候选组态池 $\{\Phi_i\}_p$ 中利用 ML 模型预测并采样新组态和 $\{\Phi_i\}_E$ 一起构建新的 $\{\Phi_i\}_V$ 。

(6) 判断是否满足收敛准则。如果满足收敛准则则提取收敛后的重要组态池 $\{\Phi_i\}_E$ ，并对 $\{\Phi_i\}_E$ 重新进行一次包含 Breit/QED 效应的 RCI 计算后输出 E , ci , $\{\Phi_i\}_E$ 。

(7) 若不满足收敛准则则回到步骤 (2)，重复步骤 (2) - (6) 直至满足收敛标准。

本文所使用的收敛标准沿用参考文献[1]中公式(7)、(8)所定义的收敛标准。具体而言，迭代过程能量标准差 E_{STD} 表示为：

$$E_{\text{STD}} = \sqrt{\frac{\sum_{i=1}^L (E_i - \bar{E})^2}{L-1}} \quad (3)$$

类似的迭代过程中的重要组态数量的相对标准差 N_{RSD} 表示为：

$$N_{\text{RSD}} = \frac{1}{\bar{N}} \sqrt{\frac{\sum_{i=1}^L (N_i - \bar{N})^2}{L-1}} \quad (4)$$

其中平均能量表示为 $\bar{E} = (\sum_i^L E_i) / L$ ，平均重要组态数量表示为 $\bar{N} = (\sum_i^L N_i) / L$ ， L 表示当此迭代判断收敛时所考虑的近 L 步迭代步数。在本文中如非特殊说明，VMCCI 与 ML-VMCCI 方法的收敛标准都将采用 $L=3$ 、 $E_{\text{STD}}=10^{-4}$ 、 $N_{\text{RSD}}=1\%$ 。需要指明的是参考文献[1]采取更为严苛的 $L=5$ 、 $E_{\text{STD}}=10^{-5}$ 、 $N_{\text{RSD}}=1\%$ 收敛标准目的是为了说明组态选择方法所能够达到的能量精度极限，而本文计算过程中的收

敛阈值调整并未牺牲能级精度。我们将在 3.3 节通过 Ge-like Nd^{28+} 体系 $J=0-5$ 的能级实验值、两种阈值下的 VMCCI 筛选能级以及本文阈值设置下的 ML-VMCCI 筛选能级对比来说明在当前收敛判据下，ML-VMCCI 仍能保持与参考组态空间计算相当的能级精度，同时保留显著的计算加速收益。

3 结果与讨论

3.1. 方法可行性验证

在本研究中，为了验证将机器学习引入 VMCCI 方法中进行组态筛选加速的可行性，并对多种不同原理模型在组态筛选任务上的有效性进行初步筛选。本文将多种模型融入到 VMCCI 方法中进行小基组原子 Lu I($[\text{Xe}]4f^{14}5d6s^2$)体系的迭代实验测试并与 VMCCI 方法进行了对比，在此过程中机器学习模型实行对原子组态的重要性的二分类筛选任务。在构造组态空间时以 $5d6s^2$ 作为价电子， $[\text{Xe}]4f^{14}$ 作为核电子，并在此过程中考虑了最重要的 VV 相关和 CV 相关的电子关联效应。通过将电子从参考组态中单、双激发到一系列 nl ($n=10, l \leq 4$) 轨道组成的活动空间中，生成了包含约 2.1×10^6 个 CSF 的组态空间。在 3.1 节的 Lu I 体系测试中，基于不同模型的 ML-VMCCI 和 VMCCI 方法统一设置如下，初始化比例为 2%，扩展比例为 1:2，截断阈值设置为 10^{-7} 。

表 1 VMCCI 及不同模型驱动的 ML-VMCCI 筛选流程参数；其中 Step 表示达到收敛所需的迭代步数；Energy 为基态能量的绝对值，单位为 Hartree；CSFs 表示以 10^{-7} 为截断阈值时，筛选流程收敛后保留的重要组态数。 S_E 表示平均采样效率。 A_{times} 表示迭代过程累计总耗时，包括机器学习耗时和从头算计算耗时，单位为 s。 M_{times} 表示迭代过程累计机器学习耗时，包括训练时间和预测时间，单位为 s。Recall, Precision, F1, PR-AUC 均为

收敛过程中的平均值，分别表示召回率，精确率，F1 分数，精确率-召回率曲线下面积。

Table I Screening workflow parameters for VMCCI and model-specific ML-VMCCI variants. Step denotes the number of iterations required for convergence. Energy is the absolute value of the energy in Hartree. CSFs indicates the number of important configuration state functions retained after convergence using a cutoff threshold of 10^{-7} . S_E denotes the average sampling efficiency. A_{times} denotes the total cumulative time spent during the iteration process, including both machine-learning and ab initio calculation times, in seconds; M_{times} denotes the total machine-learning time spent during the iteration process, including both training and prediction, in seconds. Recall, Precision, F1 and PR-AUC are averaged over the convergence process and correspond to recall, precision, F1 score and the area under precision-recall curve, respectively.

	Step	Energy	CSFs	S_E	A_{times}	M_{times}	Recall	Precision	F1	PR-AUC
VMCCI	28	14547.871494	4608	3.23	1050.60					
ANN	7	14547.871657	4618	18.94	298.14	85.99	0.80	0.79	0.79	0.84
CNN	7	14547.871782	4629	20.43	441.30	240.55	0.87	0.82	0.84	0.89
RF	10	14547.871441	4607	11.62	453.06	120.51	0.82	0.83	0.83	0.88
GB	14	14547.871489	4614	6.20	901.74	621.37	0.85	0.84	0.85	0.89
DT	11	14547.871390	4627	5.79	369.12	2.68	0.78	0.79	0.79	0.71
NB	18	14547.871259	4619	5.47	649.62	1.44	0.58	0.81	0.60	0.78
LR	18	14547.871361	4609	5.97	687.12	1.63	0.82	0.80	0.81	0.85
KNN	10	14547.871298	4598	10.69	337.50	1.84	0.82	0.72	0.77	0.77

为了验证将机器学习引入 VMCCI 方法中进行组态筛选加速的可行性，本文统计了迭代过程中的收敛情况和机器学习参数，汇总结果如表 1 所示，迭代收敛

过程如图 2 所示。由表 1 可见，8 种机器学习模型加速的 ML-VMCCI 方法在 Lu I 体系中均可收敛至与 VMCCI 方法相当的能量结果，最大偏差不超过 0.000514 Hartree，但所需迭代步数明显减少：其中，ANN 和 CNN 收敛最快，仅需要 7 步。RF、GB、KNN 和 DT 需要 10-14 步，NB 和 LR 需要 18 步，而 VMCCI 方法需要 28 步。这表明引入机器学习后，VMCCI 迭代筛选过程能够在保证收敛结果相当的前提下显著减少迭代次数。

同时，如表 1 所示计入模型训练、推理开销以及从头算耗时后，不同 ML-VMCCI 方法的总收敛时间约为 298.14 - 901.74 s，均短于 VMCCI 方法的 1050.60 s，表现出较好的综合加速效果。各 ML-VMCCI 方法相较于 VMCCI 的整体加速比约为 ANN: 3.5 倍，CNN: 2.4 倍，RF: 2.3 倍，GB: 1.2 倍，DT: 2.9 倍，NB: 1.6 倍，LR: 1.5 倍，KNN: 3.1 倍。这种加速主要来源于 ML 采样效率相对于 MC 采样的显著提升。由图 2(c)和(f)可见，在迭代初期，ML 采样效率普遍可达 40% - 60%，显著高于 MC 采样约 20%的水平；随着迭代进行，两者均逐渐下降，这是待选组态空间中真正重要组态的比例不断降低所致。但 ML 采样的平均采样效率较 MC 采样提升了 1.7-6.3 倍，其中各模型平均采样效率分别提升了 ANN: 5.9 倍，CNN: 6.3 倍，RF: 3.6 倍，GB: 1.9 倍，DT: 1.8 倍，NB: 1.7 倍，LR: 1.8 倍，KNN: 3.3 倍。上述结果表明以 ML 采样替代 VMCCI 中的 MC 采样，能够在保持收敛正确性的同时显著提高采样效率，实现对 VMCCI 方法的有效加速。

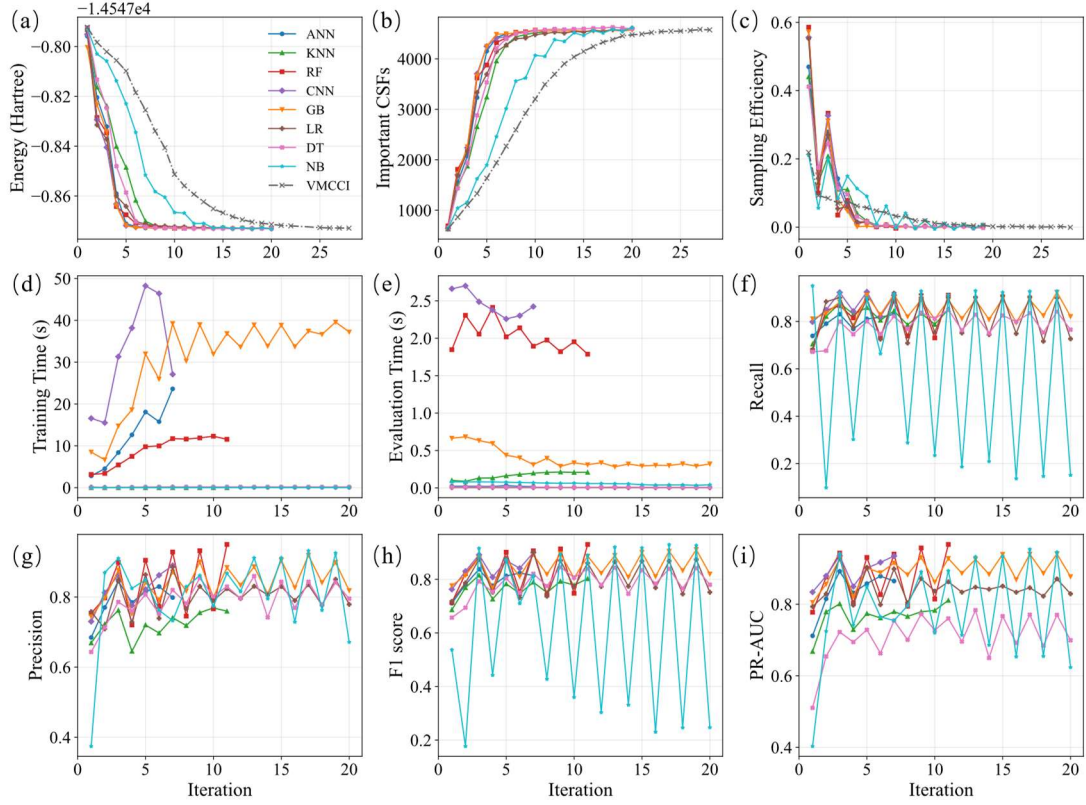


图 2 基于不同 ML 模型的 ML-VMCCI 方法迭代过程 ML 参数对比；(a)迭代过程中的能量变化；(b)迭代过程中的重要组态数目变化；(c)迭代过程中的采样效率变化；(d)迭代过程中的训练时间变化；(e)迭代过程中的预测时间变化；(f)迭代过程中的召回率变化；(g)迭代过程中的精确率变化；(h)迭代过程中的 F1 分数变化；(i)迭代过程中的 PR 曲线下面积变化

Fig. 2. Comparison of ML-related parameters in the iterative processes of ML-VMCCI

methods based on different ML models: (a) energy variation during iteration; (b) variation in the number of important configurations during iteration; (c) variation in sampling efficiency during iteration; (d) variation in training time during iteration; (e) variation in prediction time during iteration; (f) variation in recall during iteration; (g) variation in precision during iteration; (h) variation in F1 score during iteration; (i) variation in the area under the PR curve during iteration.

由于 ML 模型在迭代过程中主要承担组态筛选器的角色，因此其在未知体系中的稳定性和有效性是需要重点关注的维度。理想情况下，模型应在保持较低 ML 计算成本的同时，尽可能识别并保留更多正样本，降低对重要组态的漏判和

误判风险。因此，本文重点考察各模型在迭代过程中的 Recall、Precision 和 PR-AUC 表现。其中，Recall 反映模型保留正样本的能力，Precision 反映模型筛选结果的准确性；考虑到组态筛选任务中重要与非重要样本比例极度不均衡，PR-AUC 相比 ROC-AUC 更能反映模型在少数正样本识别中的有效性。

基于此，本文对各模型的稳定性和有效性指标进行了分析。如图 2(f)、(g)所示，NB 模型的 Recall 和 Precision 波动最为剧烈。Precision 标准差高达 0.1252，取值范围达到 0.5577；Recall 标准差达到 0.3671，在不同迭代中从 0.9504 急剧下降至 0.0985，最大单步变化达 0.8518；受此影响，F1 也在 0.1770 至 0.9295 之间大幅摆动，标准差为 0.2957。这表明 NB 模型对组态重要性的识别稳定性较差，容易受到迭代样本分布变化的影响。据表 1 所示，KNN 模型的 Precision 均值仅为 0.72，低于其他模型 ANN (0.79)、DT (0.79)、LR (0.80)、NB (0.81)、CNN (0.82)、RF (0.83) 和 GB (0.84)。说明其筛选结果中包含较多非重要组态，识别准确性相对不足。DT 模型虽然 Recall 和 Precision 均值分别达到 0.78 和 0.79，但其 PR-AUC 均值仅为 0.71，在所有模型中处于最低水平，表明在正负样本极度不平衡的条件下，DT 对重要组态的整体排序和区分能力较弱。同时，DT 平均采样效率也仅为 5.79%，其分类性能难以稳定支撑高效采样。相比之下，GB 和 LR 模型的整体分类指标并不低。然而，其平均采样效率仅为 6.20% 和 5.97%，明显低于 ANN、RF、CNN 等模型。进一步分析可知，GB 和 LR 在前 5 次迭代中的平均采样效率分别为 23.31% 和 21.59%，而第 6 步之后迅速下降至 0.11% 和 0.39%，后期几乎无法持续带来新的重要组态。这说明 GB 和 LR 虽然能够在测试集上保持较好的分类表现，但其线性判别能力难以充分覆盖复杂、稀疏且非线性的组态空间，导致分类指标未能有效转化为迭代采样中的实际收益。因此，GB

和 LR 作为筛选器的后期探索能力和持续有效性不足。

NB、KNN、DT、GB 和 LR 表现较弱，可能源于模型机制与 ML-VMCCI 迭代中数据分布变化的共同影响。NB 依赖特征条件独立假设，难以处理 CSF 描述符中轨道占据数、轨道角动量和逐级耦合角动量之间的物理相关性；KNN 依赖距离度量，易受高维特征尺度、局部样本密度和类别不平衡影响；DT 对局部样本扰动敏感，在迭代数据变化时易出现过拟合或边界不稳定；GB 依赖梯度提升式逐轮修正残差的迭代机制，在 ML-VMCCI 迭代过程中正负样本比例持续变化、数据分布不断漂移的情况下，容易将模型容量集中于少数难分样本和近阈值边界样本，从而削弱对低价值 CSF 的整体压缩效率。LR 受限于线性判别边界，难以刻画 CSF 特征与组态重要性之间的非线性关系。随着迭代推进，重要组态比例逐渐降低，正负样本比例和特征分布持续变化，进一步放大了这些模型对边界样本和类别不平衡的敏感性。相比之下，ANN、CNN 和 RF 具有更强的非线性表达和复杂边界建模能力，因此在识别稳定性、采样效率和收敛表现之间取得了更好的平衡。综上，ANN、CNN 和 RF 在 LuI 基准体系中同时满足能量收敛一致、整体收敛时间缩短、采样效率提升以及分类指标稳定等要求。因此，本文选择 ANN、CNN 和 RF 作为后续 Ge-like Nd^{28+} 大规模体系中的代表性 ML-VMCCI 模型，以进一步验证 ML 加速组态筛选在更大 CSF 空间中的加速有效性。

3.2. 采样策略研究

SCI 方法的实际效率瓶颈不仅取决于组态筛选本身，还涉及从头算求解 $Hc=cE$ 方程得到能量本征值 E_0 和展开系数 c_i 的过程。这意味着在 ML-SCI 方法中，最终精度取决于所选组态空间是否充分覆盖参考组态空间的重要贡献，而效

率收益则由机器学习成本与从头算成本共同决定。若 ML 采样策略低效，一方面可能遗漏重要组态导致精度受损，另一方面引入冗余组态会增加不必要的从头算开销、拖累整体效率。因此，为平衡 ML-SCI 方法的精度与效率收益，本文探索了机器学习在组态重要性二分类与组态重要性排序两种任务上的表现与差异。为此我们将在 Qin 等人计算的 Ge-like Nd^{28+} 奇宇称 $J=2$ 体系 (CSFs 数量为 5.9×10^5) 上进行测试^[21]。对 Ge-like Nd^{28+} 体系进行参考组态空间生成时与文献^[21]保持一致，即以 $[\text{Ar}]3d^{10}$ 为芯电子，外层电子作为价电子，选取 $4s^24p^2$ 、 $4s4p^3$ 、 $4s^24p4d$ 和 $4s4p^24d$ 作为 MR。随后，将电子从这些 MR 组态中单或双激发到一系列 nl 轨道活动空间中以生成扩展 CSFs，其中 $n=8$ 且 $l \leq 4$ 。在该过程中考虑 VV 和 CV 关联，允许价电子的单激发和双激发，以及两个最外层核电子的单激发和另一个价电子的单激发进入活动空间。在 3.2 的 Ge-like Nd^{28+} 奇宇称 $J=2$ 体系测试中，基于不同模型的 ML-VMCCI 和 VMCCI 方法统一设置如下，初始化比例为 2%，扩展比例为 1:2，截断阈值设置为 10^{-8} 。

表 2，基于不同 ML 模型、不同采样策略的 ML-VMCCI 方法与 VMCCI 方法在 Ge-like Nd^{28+} $J=2$ 体系计算中的表现对比。其中 MCDHF 代表使用 Grasp 对参考组态空间进行直接计算得到的结果。收敛时间单位为 min。能量为奇宇称 $J=2$ 的能级中能量最低的组态对应的绝对能量大小，单位为 Hartree。CSFs 表示使用 10^{-8} 为截断值时，对参考组态空间直接进行计算以后截取的组态数目和使用 ML-VMCCI 进行组态筛选迭代收敛以后的组态数目。模型后缀 M 代表模型进行组态重要性排序和组态二分类任务的混合采样方式，C 代表模型进行组态二分类任务的采样方式，S 代表模型进行组态重要性排序任务的采样方式。

Table II. Comparison of ML-VMCCI methods using different machine-learning

models and sampling strategies with the original VMCCI method for the Ge-like $\text{Nd}^{28+} J = 2$ system. MCDHF denotes the direct GRASP calculation performed in the reference configuration space. The convergence time is given in minutes. Energy is the absolute energy, in Hartree, of the lowest level in the odd-parity $J = 2$ manifold. The CSF column gives the number of configurations retained with a cutoff of 10^{-8} : for MCDHF, this number is obtained by applying the cutoff after direct diagonalization in the reference configuration space; for ML-VMCCI, it corresponds to the number of CSFs retained after iterative configuration screening has converged. The suffixes M, C, and S denote the hybrid strategy combining configuration-importance ranking and binary classification, the binary-classification strategy, and the configuration-importance-ranking strategy, respectively.

方法	收敛步数	收敛时间	能量	CSFs	平均采样效率
MCDHF			-9230.488929	41668	
VMCCI	45	150.18	-9230.488179	41403	2.48%
ANN-M	7	37.19	-9230.488654	41025	19.40%
ANN-C	9	47.12	-9230.488907	41195	13.32%
ANN-S	11	50.86	-9230.488861	41049	12.93%
CNN-M	6	45.51	-9230.488741	41079	21.01%
CNN-C	7	52.68	-9230.488655	41034	15.64%
CNN-S	9	94.55	-9230.488979	41132	10.32%
RF-M	9	57.45	-9230.488968	41056	14.34%
RF-C	14	91.86	-9230.488872	41260	9.50%

RF-S	13	88.90	-9230.488644	40313	10.23%
------	----	-------	--------------	-------	--------

Ge-like Nd^{28+} 奇宇称 $J=2$ 体系的筛选结果对比如图 3 和表 2 所示。根据表 2 显示,VMCCI 方法在该体系中需要 45 步、150.18 min 收敛至 -9230.488486 Hartree。引入机器学习后,基于三个不同模型的 ML-VMCCI 方法均显著缩短了收敛时间并减少了迭代步数。其中,三个模型混合采样策略(后缀 M)整体表现最优:ANN-M、CNN-M 和 RF-M 分别仅需 7、6 和 9 步即可收敛,平均采样效率分别达到 19.40%、21.01%和 14.34%(VMCCI 仅为 2.48%),收敛时间分别降至 37.19、45.51 和 57.45 min,对应加速比分别为 4.04、3.29 和 2.61 倍。单纯基于二分类的采样方式(后缀 C)ANN-C、CNN-C 和 RF-C 的收敛时间分别为 47.12、52.68 和 91.86 min,总体慢于混合采样。单纯基于排序的采样方式(后缀 S)在不同模型上表现更不稳定:ANN-S 和 CNN-S 仍能维持较高精度,但 RF-S 需要 13 步、88.90 min,明显弱于对应的混合采样版本。

图 3 进一步显示,基于排序策略的各模型在迭代前期的能量下降更快,并在较少重要组态数下获得了较低能量。然而排序策略的 CNN 和 RF 模型在迭代后期出现了重要组态数收敛较慢的情况。在 RF 测试中,排序策略收敛后虽然绝对能量接近 VMCCI,但保留的重要组态数明显减少,说明单纯依赖排序可能遗漏部分贡献较小但仍处于重要性阈值附近的边界组态。为解释这一现象,同时探究不同采样策略的机理,我们在补充材料附录 III 中分析了模型在全待筛选空间中的排序能力,以及在给定扩展比例、即计算成本受限条件下的实际采样效果。结果表明,排序策略能够在小采样比例下优先富集 c_i 系数较大的高贡献组态,从而提高采样质量并加快早期收敛;但随着迭代推进,待筛选空间中重要组态比例降

低，部分 c_i 较小但仍满足重要性阈值的边界组态容易被排在较后位置。二分类策略虽然对高 c_i 组态的优先富集能力较弱，但其基于分类边界筛选组态，因而在一定程度上保留了对重要性阈值附近边界组态的覆盖能力。混合采样策略结合了二者优势：一方面利用排序优先选取高贡献组态，提高采样质量和早期收敛速度；另一方面利用二分类补充边界组态覆盖，增强后期搜索稳定性。因此，混合采样在收敛步数、总计算时间和平均采样效率之间取得了更好的平衡。我们将在后续测试中采样混合采样策略验证 ML-VMCCI 方法的有效性和加速效果。

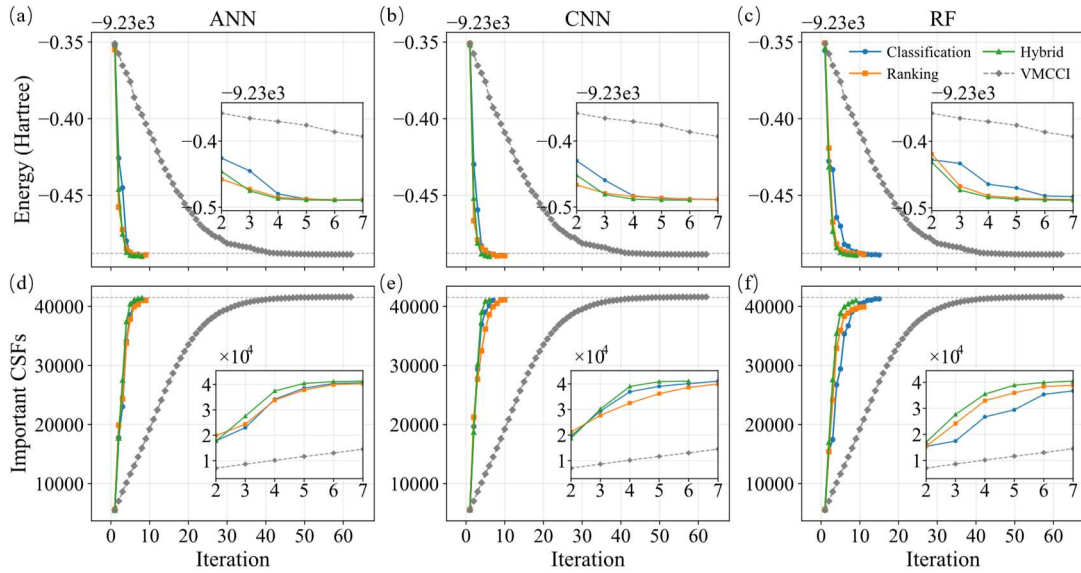


图 3 Ge-like Nd^{28+} 体系 VMCCI 与基于不同 ML 模型的 ML-VMCCI 方法迭代过程对比；(a)-(c) 能量随步数的迭代过程对比；(d)-(f) 截断阈值取 10^{-8} 时的重要组态数目随步数的迭代过程对比；(a)-(f) 中的插图展示了第 2-7 次迭代期间的局部收敛过程。其中，(a)-(c) 的插图以能量为纵坐标，单位为 Hartree；(d)-(f) 的插图以截断阈值为 10^{-8} 时的重要组态数目为纵坐标

Fig. 3. Comparison of the iterative processes of VMCCI and ML-VMCCI methods based on different ML models for the Ge-like Nd^{28+} system: (a)-(c) comparison of the energy convergence processes as functions of iteration step; (d)-(f) comparison of the

number of important configurations at a truncation threshold of 10^{-8} as a function of iteration steps. The insets in panels (a)-(f) highlight the local convergence behavior from iterations 2 to 7. In panels (a)-(c), the inset y-axis represents the energy in Hartree, whereas in panels (d)-(f), it represents the number of important configurations obtained with a truncation threshold of 10^{-8} .

3.3.方法有效性验证

为了验证基于混合采样方式的 ML-VMCCI 方法有效性和加速效果，本文将其与基于不同模型的 ML-VMCCI 方法结合应用于 Ge-like Nd^{28+} 体系 $J=0-5$ 的计算中，并与 VMCCI 方法进行了对比。其中 Ge-like Nd^{28+} 体系的组态空间生成方式以及组态筛选参数设置均与 3.2 节保持一致，使用 3.2 节所述组态空间生成方式最终生成 2.46×10^6 个奇宇称 ($J=0-4$) CSFs 和 4.67×10^6 个偶宇称 ($J=0-5$) CSFs，共 7.14×10^6 个 CSFs 的参考组态空间。迭代收敛速度和采样效率对比汇总结果如表 3 所示，迭代收敛能级结果与参考文献[62]中的实验值以及文献[21]使用 MCDHF 对参考组态空间直接计算得到的能级误差统计结果如图 4 所示。详细能级对比结果见补充材料附录 IV。

表 3. 基于不同机器学习模型的 ML-VMCCI 方法与 VMCCI 方法在 Ge-like Nd^{28+} 体系偶宇称与奇宇称计算中的表现对比。VMCCI-2 为收敛阈值采用 $L=3$ 、 $E_{\text{STD}}=10^{-4}$ 、 $N_{\text{RSD}}=1\%$ 时的 VMCCI 计算结果。 A_{times} 表示在相应宇称空间内完成迭代收敛所需的累计计算时间，单位为 min；CSFs 表示在相应宇称空间内采用 10^{-8} 截断值后保留的累计重要组态数目； S_E 表示在相应宇称空间内迭代过程的平均采样效率。模型后缀 M 表示模型采用组态重要性排序和组态二分类任务

的混合采样方式。

Table III. Comparison of the performance of ML-VMCCI methods based on different machine-learning models with that of the VMCCI method for even- and odd-parity calculations of the Ge-like Nd^{28+} system. VMCCI-2 represents the calculation result of VMCCI when the convergence threshold is set as $L=3$, $E_{\text{STD}}=10^{-4}$, and $N_{\text{RSD}}=1\%$. A_{times} denotes the cumulative computational time required to achieve iterative convergence within the corresponding parity space, in min. CSFs denotes the total number of important configuration state functions retained in the corresponding parity space after applying a cutoff threshold of 10^{-8} . S_E denotes the average sampling efficiency during the iterative process within the corresponding parity space. The suffix M indicates that the model adopts a hybrid sampling strategy combining configuration-importance ranking and configuration binary classification.

Method	Even			Odd		
	A_{times}	CSFs	S_E	A_{times}	CSFs	S_E
VMCCI-2	1549.55	295683	2.65%	390.26	129878	4.30%
ANN-M	253.79	296018	24.09%	84.74	129942	26.32%
CNN-M	314.69	295997	27.32%	98.54	129701	27.26%
RF-M	386.62	295436	22.41%	120.21	129025	18.32%

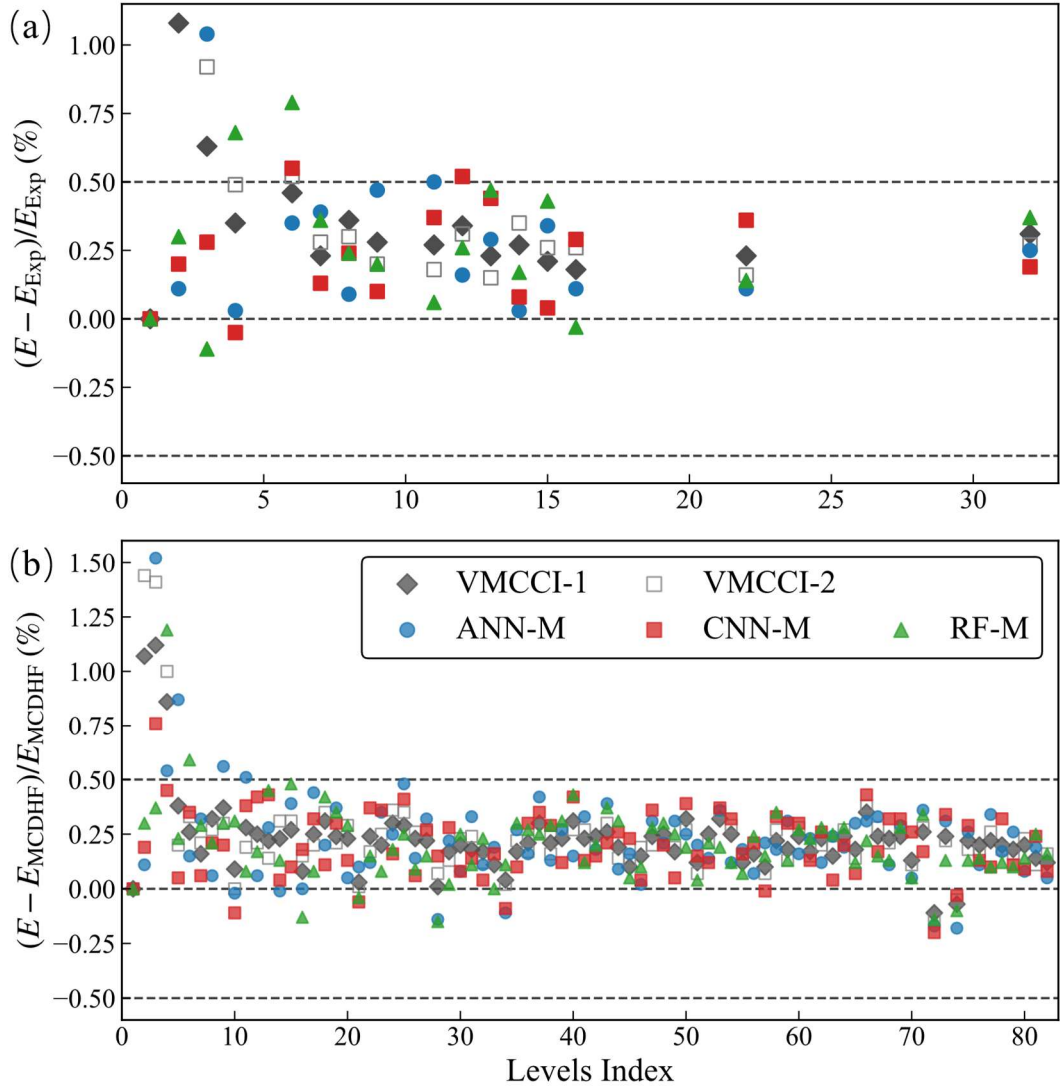


图 4 Ge-like Nd^{28+} 体系的 VMCCI 与 ML-VMCCI 筛选结果能级相对于实验值和 MCDHF 计算值的能级误差分布；其中横坐标为能级序号，纵坐标为相对误差百分比。灰色虚线分别表示 0 和 $\pm 0.5\%$ 误差范围；(a) VMCCI 方法与基于不同模型的 ML-VMCCI 方法筛选结果能级相对于实验能级的相对误差分布；(b) VMCCI 方法与基于不同模型的 ML-VMCCI 方法筛选结果能级相对于 MCDHF 能级的相对误差分布；其中 VMCCI-1 和 VMCCI-2 的收敛阈值分别采用 $L=5$ 、 $E_{\text{STD}}=10^{-5}$ 、 $N_{\text{RSD}}=1\%$ 以及 $L=3$ 、 $E_{\text{STD}}=10^{-4}$ 、 $N_{\text{RSD}}=1\%$

Fig. 4. Energy error distributions of the VMCCI and ML-VMCCI selected levels for the Ge-like Nd^{28+} system relative to experimental and MCDHF calculated values. The

level index is shown on the horizontal axis, and the relative error in percent is shown on the vertical axis. The gray dashed lines mark 0 and $\pm 0.5\%$ error levels. (a) Error distributions of VMCCI and ML-VMCCI predictions based on different machine-learning models relative to experimental energy levels. (b) Error distributions of VMCCI and ML-VMCCI predictions based on different machine-learning models relative to MCDHF energy levels. The convergence thresholds for VMCCI-1 and VMCCI-2 are respectively set as $L=5$, $E_{STD}=10^{-5}$, $N_{RSD}=1\%$, and $L=3$, $E_{STD}=10^{-4}$, $N_{RSD}=1\%$.

在加速效率方面，从表 3 可以看出，基于机器学习模型的 ML-VMCCI 方法在保持 CSFs 数目与传统 VMCCI 方法基本一致的情况下，显著降低了偶宇称和奇宇称空间内的累计迭代收敛时间。对于偶宇称计算，VMCCI 的收敛时间为 1549.55 min，而 ANN-M、CNN-M 和 RF-M 分别缩短至 253.79、314.69 和 386.62 min，对应加速比分别为 6.11、4.92 和 4.01；同时，三种 ML-VMCCI 方法的平均采样效率分别由 VMCCI 的 2.65% 提升至 24.09%、27.32% 和 22.41%，采样效率提升约 9.09、10.31 和 8.46 倍。对于奇宇称计算，VMCCI 的收敛时间为 390.26 min，而 ANN-M、CNN-M 和 RF-M 分别缩短至 84.74、98.54 和 120.21 min，对应加速比分别为 4.61、3.96 和 3.25；平均采样效率则由 4.30% 分别提升至 26.32%、27.26% 和 18.32%，提升约 8.42、8.98 和 6.13 倍。上述结果表明，ML-VMCCI 方法能够通过机器学习模型对组态空间进行更高效的筛选，在保留相当 CSFs 数量的前提下显著提高采样效率并缩短迭代收敛时间。

在能级精度方面，图 4 (a) 结果显示四种方法相对于实验值的偏差整体集中在 0-0.5% 区间内，仅 VMCCI 与 ANN-M 在第 2-3 个实验能级附近均出现约 1%

的误差峰值。而 CNN-M 和 RF-M 的实验误差分布相对更平缓，其中 CNN-M 的大部分点位于 0-0.5% 范围内，最大偏差也低于 VMCCI 和 ANN-M。相对于实验值，VMCCI-1、VMCCI-2、ANN-M、CNN-M 和 RF-M 的平均绝对误差分别为 0.339%、0.381%、0.267%、0.240% 和 0.288%，均方根误差 (root mean square error, RMSE) 分别为 0.411%、0.511%、0.368%、0.292% 和 0.361%，标准差分别为 0.239%、0.340%、0.261%、0.181% 和 0.246%。上述结果表明，相对于实验能级各 ML-VMCCI 方法的误差分布与 VMCCI 处于相当水平。特别是 CNN-M 在平均绝对误差、RMSE 和标准差上均低于 VMCCI。图 4 (b) 结果相对于 MCDHF 的误差分布中，各方法前几个低能级的波动较明显，随后大多数能级的误差收敛到 0-0.35% 附近，且绝大部分点位于 $\pm 0.5\%$ 辅助线之间。说明主要偏差集中在少数低能级，对整体能级序列并未形成持续放大的系统误差。定量统计也支持这一趋势。相对于 MCDHF，VMCCI-1、VMCCI-2、ANN-M、CNN-M 和 RF-M 的对应平均绝对误差分别为 0.235%、0.243%、0.239%、0.215% 和 0.217%，RMSE 分别为 0.291%、0.330%、0.315%、0.254% 和 0.269%，标准差分别为 0.179%、0.228%、0.222%、0.154% 和 0.176%。其中 CNN-M 在两组参照下均表现出较低的平均误差和离散度，RF-M 也与 VMCCI 保持相当水平。以上 VMCCI-1 和 VMCCI-2 两种收敛阈值下的筛选结果与实验值以及 MCDHF 能级精度对比结果说明两种收敛阈值下能级精度处于相当水平，收敛阈值的改变并未显著改变筛选能级精度。ML-VMCCI 方法筛选结果与实验值与 MCDHF 的对比结果说明 ML-VMCCI 模型的误差分布与 VMCCI 基本处于相当水平，并且在部分指标上优于原始 VMCCI，机器学习模型并未引入明显额外偏差或更大的误差离散性。由此可见，ML-VMCCI 在提高建模效率的同时，并未以损失精度为代价。上述结果验证了基于

混合采样策略的 ML-VMCCI 方法的有效性和可靠性。

4 结论

本文围绕大组态空间下高精度原子激发态计算中效率与精度的矛盾,发展了基于 VMCCI 框架的机器学习加速方案 ML-VMCCI。首先在 Lu I 体系上验证了以机器学习采样替代 MC 采样的可行性,系统比较了多种经典与深度学习模型的表现,发现 ANN、CNN 和 RF 在综合采样效率、额外计算成本和稳定性上最具潜力。在此基础上,评估了组态重要性二分类、排序以及二者混合三种采样策略的效果。结果表明,排序策略建立与物理贡献对应的优先级序列,二分类策略提供额外覆盖,二者结合的混合采样在采样质量、覆盖范围与收敛稳定性之间取得了最优平衡。将基于混合策略的 ML-VMCCI 方法应用于 Ge-like Nd²⁸⁺离子激发态计算,收敛速度提升约 3–6 倍,重要组态识别效率提升约 6–10 倍,且相对于实验值和 MCDHF 的整体误差统计表明,ML-VMCCI 方法的加速并未引入明显额外偏差或更大的误差离散性。

需要指出的是,本文方法仍存在一定适用边界。原子能级的绝对精度主要取决于参考组态空间和活动空间的构造,而 ML-VMCCI 的作用是在给定参考组态空间内提高重要组态筛选效率,并保持与参考计算一致的精度。此外,ML 模型的训练效果依赖于数据集构建,包括初始样本代表性、正负样本比例等。未来可进一步发展更稳健的数据集构建方式和自适应采样策略等,以提高对不同原子体系的适用性。

总体而言,本文工作表明机器学习与 VMCCI 的结合进一步打通了“组态空间压缩”与“采样过程加速”两个层面的优化路径,使得高精度原子激发态计算

能够在更可控的成本下实现。本文提出的 ML-VMCCI 方法不仅适用于相对论原子结构问题，也为将机器学习用于更广泛的原子、分子及其他强电子关联体系中的组态筛选与高精度电子结构计算提供了可推广的思路。该方法有望在高 Z 离子、复杂激发态结构、精细结构分析以及实验光谱解释等研究中发挥更大的作用。

参考文献

- [1] Zhang J Z, Liu J D, Qin Y, Liu R J, Yang C, Jiang G 2024 *Phys. Rev. A* **110** 062817
- [2] Flambaum V V, Kozlov M G 2007 *Phys. Rev. Lett.* **99** 150801
- [3] Rosenband T, Hume D B, Schmidt P O, Chou C W, Brusch A, Lorini L, Oskay W H, Drullinger R E, Fortier T M, Stalnaker J E, Diddams S A, Swann W C, Newbury N R, Itano W M, Wineland D J, Bergquist J C 2008 *Sci.* **319** 1808
- [4] Hees A, Guéna J, Abgrall M, Bize S, Wolf P 2016 *Phys. Rev. Lett.* **117** 061301
- [5] Qiu H, Hu Z, Wu S, Chen J, Zhong C, Wu J, Li X, Xiao D, Shi C, Liu J, Xiong W, Hu T, Cai Q, Yuan Y 2023 *Sci. Rep.* **13** 9551
- [6] Berengut J C, Flambaum V V, Ong A, Webb J K, Barrow J D, Barstow M A, Preval S P, Holberg J B 2013 *Phys. Rev. Lett.* **111** 010801
- [7] Porsev S G, Safronova M S, Safronova U I, Kozlov M G 2018 *Phys. Rev. Lett.* **120** 063204
- [8] Safronova M S, Dzuba V A, Flambaum V V, Safronova U I, Porsev S G, Kozlov M G 2014 *Phys. Rev. Lett.* **113** 030801
- [9] Derevianko A, Pospelov M 2014 *Nat. Phys.* **10** 933
- [10] Bachau H, Cormier E, Declewa P, Hansen J E, Martín F 2001 *Rep. Prog. Phys.* **64** 1815
- [11] Lindgren I, Morrison J 1986 *Atomic Many-Body Theory* (Berlin: Springer) pp56–89
- [12] Bartlett R J, Musiał M 2007 *Rev. Mod. Phys.* **79** 291
- [13] Koch H, Jørgensen P 1990 *J. Chem. Phys.* **93** 3333
- [14] Gershgorin Z, Shavitt I 1968 *Int. J. Quantum Chem.* **2** 751
- [15] Knowles P J, Handy N C 1989 *J. Chem. Phys.* **91** 2396
- [16] Ohtsuka Y, Hasegawa J Y 2017 *J. Chem. Phys.* **147** 034102
- [17] Zhang N, Liu W, Hoffmann M R 2020 arXiv:1911.12506v2 [physics.chem-ph]
- [18] Coe J P 2023 *J. Chem. Theory Comput.* **19** 874
- [19] Safronova M S, Dzuba V A, Flambaum V V, Safronova U I, Porsev S G, Kozlov M G 2014 *Phys. Rev. A* **90** 42513
- [20] Silwal R, Dreiling J M, Sanders S C, Takacs E, Ralchenko Y 2020 *J. Phys. B: At. Mol. Opt. Phys.* **53** 145002
- [21] Qin Y, Shi B L, Dou X L, Jiang G 2022 *At. Data Nucl. Data Tables* **147** 101517
- [22] Yu Y S, Yang C, Jiang G 2023 *Chinese Phys. B* **32** 033101
- [23] Song C X, Zhang C Y, Wang K, Si R, Godefroid M, Jönsson P, Dang W, Zhao X H, Yan J, Chen C Y 2021 *At. Data Nucl. Data Tables* **138** 101377
- [24] Karpušienė R, Kisieliū R 2025 *At. Data Nucl. Data Tables* **161** 101693
- [25] Coe J P, Taylor D J, Paterson M J 2012 *J. Chem. Phys.* **137** 194111

- [26] Evangelista F A 2014 *J. Chem. Phys.* **140** 124114
- [27] Coe J P, Paterson M J 2014 *Mol. Phys.* **112** 733
- [28] Bytautas L, Ruedenberg K 2009 *Chem. Phys.* **356** 64
- [29] Greer J C 1995 *J. Chem. Phys.* **103** 1821
- [30] Mitrushenkov A O, Dmitriev Y Yu 1995 *Chem. Phys. Lett.* **235** 410
- [31] Harrison R J 1991 *J. Chem. Phys.* **94** 5021
- [32] Greer J C 1998 *J. Comput. Phys.* **146** 181
- [33] Tong L, Nolan M, Cheng T, Greer J C 2000 *Comput. Phys. Commun.* **131** 142
- [34] Coe J P, Paterson M J 2012 *J. Chem. Phys.* **137** 204108
- [35] Coe J P, Taylor D J, Paterson M J 2013 *J. Comput. Chem.* **34** 1083
- [36] Coe J P, Murphy P, Paterson M J 2014 *Chem. Phys. Lett.* **604** 46
- [37] Kelly T P, Perera A, Bartlett R J, Greer J C 2014 *J. Chem. Phys.* **140** 084114
- [38] Silver D, Huang A, Maddison C J, Guez A, Sifre L, Van Den Driessche G, Schrittwieser J, Antonoglou I, Panneershelvam V, Lanctot M, Dieleman S, Grewe D, Nham J, Kalchbrenner N, Sutskever I, Lillicrap T, Leach M, Kavukcuoglu K, Graepel T, Hassabis D 2016 *Nat.* **529** 484
- [39] Bansak K, Ferwerda J, Hainmueller J, Dillon A, Hangartner D, Lawrence D, Weinstein J 2018 *Sci.* **359** 325
- [40] Shen D, Wu G, Suk H I 2017 *Annu. Rev. Biomed. Eng.* **19** 221
- [41] Yang X, Wang Y, Byrne R, Schneider G, Yang S 2019 *Chem. Rev.* **119** 10520
- [42] Coe J P 2018 *J. Chem. Theory Comput.* **14** 5739
- [43] Jeong W, Gaggioli C A, Gagliardi L 2021 *J. Chem. Theory Comput.* **17** 7518
- [44] Qu C, Houston P L, Yu Q, Conte R, Pandey P, Nandi A, Bowman J M 2023 *J. Chem. Phys.* **159** 071101
- [45] Rano M, Ghosh D 2023 *J. Phys. Chem. A* **127** 3705
- [46] Goings J J, Hu H, Yang C, Li X 2021 *J. Chem. Theory Comput.* **17** 5482
- [47] Györfy W, Bartlett R J, Greer J C 2008 *J. Chem. Phys.* **129** 064103
- [48] Murphy P, Coe J P, Paterson M J 2018 *J. Comput. Chem.* **39** 319
- [49] Bilous P, Pálffy A, Marquardt F 2023 *Phys. Rev. Lett.* **131** 133002
- [50] Bilous P, Cheung C, Safronova M 2024 *Phys. Rev. A* **110** 042818
- [51] Froese Fischer C, Gaigalas G, Jönsson P, Bieroń J 2019 *Comput. Phys. Commun.* **237** 184
- [52] Dyall K G, Grant I P, Johnson C T, Parpia F A, Plummer E P 1989 *Comput. Phys. Commun.* **55** 425
- [53] Grant I P 2007 *Relativistic Quantum Theory of Atoms and Molecules: Theory and Computation* (New York: Springer) pp587–626
- [54] Andrieu C, Thoms J 2008 *Stat. Comput.* **18** 343
- [55] Cox D R 1958 *J. R. Stat. Soc. Ser. B* **20** 215
- [56] Maron M E 1961 *J. ACM* **8** 404
- [57] Cover T, Hart P 1967 *IEEE Trans. Inf. Theor.* **13** 21
- [58] Quinlan J R 1986 *Mach. Learn.* **1** 81
- [59] Lecun Y, Bottou L, Bengio Y, Haffner P 1998 *Proc. IEEE* **86** 2278
- [60] Breiman L 2001 *Mach. Learn.* **45** 5
- [61] Rumelhart D E, Hinton G E, Williams R J 1986 *Nat.* **323** 533
- [62] Suzuki C, Dipti, Yang Y, Gall A, Silwal R, Sanders S, Naing A, Tan J, Takacs E, Ralchenko Y

录用稿件，非最终出版稿

Machine-learning-accelerated variational Monte Carlo selected configuration interaction method*

ZHANG Jinzhe LIU Jinde QIN Yi LIU Renjie JIANG Gang[†]

Institute of Atomic and Molecular Physics, Sichuan University, Chengdu 610065, China

Abstract

High-precision relativistic atomic-structure calculations for many-electron heavy atoms and highly charged ions are often limited by the rapid growth of the configuration space. Selected configuration-interaction methods alleviate this problem by retaining only configuration state functions (CSFs) that contribute appreciably to the target wave function, but their efficiency depends on how rapidly important configurations can be identified. The previously developed variational Monte Carlo configuration-interaction method (VMCCI) improves the stability of configuration selection through a variational rejection-acceptance sampling mechanism^[1]. However, as the configuration space expands, Monte Carlo sampling becomes increasingly inefficient and converges slowly because the probability of sampling physically important configurations is governed by their typically small fraction of the candidate space. To address this bottleneck, we develop a machine-learning-accelerated VMCCI method (ML-VMCCI), in which model-guided sampling replaces random Monte Carlo sampling.

* Project supported by the National Natural Science Foundation of China (Grant No. 12474253).

[†] Corresponding author. E-mail: gjiang@scu.edu.cn

In ML-VMCCI, each CSF is represented by an atomic-structure descriptor containing orbital occupations, orbital angular momenta, and coupled angular momenta. Importance labels are assigned from CI coefficients obtained with GRASP2018 and updated during the iterative screening process. We first evaluate eight machine-learning models in the Lu I system, whose reference configuration space contains about 2.1×10^6 CSFs. Under identical screening parameters, all ML-VMCCI variants converge to energies consistent with VMCCI, with a maximum deviation below 0.000514 Hartree. The number of iterations decreases from 28 to 7-18, the total convergence time is reduced from 1050.60 s to 298.14-901.74 s, and the average sampling efficiency increases by a factor of 1.7-6.3. ANN, CNN, and RF show the best balance among sampling efficiency, machine-learning cost, and iterative stability.

We then compare binary classification, configuration-importance ranking, and a hybrid strategy in the odd-parity $J = 2$ subspace of Ge-like Nd^{28+} , which contains about 5.9×10^5 CSFs. The results indicate that ranking enriches high-contribution configurations in early iterations, whereas classification improves coverage near the importance threshold. Their combination gives the most balanced sampling behavior. In this benchmark, VMCCI requires 45 iterations and 150.18 min to converge. By contrast, ANN-VMCCI, CNN-VMCCI, and RF-VMCCI, all employing the hybrid sampling strategy, converge in 7, 6, and 9 iterations, respectively,

yielding speedups of 4.04, 3.29, and 2.61 relative to VMCCI.

Finally, we apply the ML-VMCCI method with the hybrid sampling strategy to $J = 0-5$ excited-state calculations of Ge-like Nd^{28+} in a reference configuration space containing 7.14×10^6 CSFs. For even-parity states, ANN-M, CNN-M, and RF-M reduce the convergence time from 1549.55 min to 253.79, 314.69, and 386.62 min, respectively. For odd-parity states, the corresponding time decreases from 390.26 min to 84.74, 98.54, and 120.21 min. Overall, ML-VMCCI accelerates configuration screening by about 3-6 times and improves important-configuration sampling by about 6-10 times.

The acceleration does not introduce a noticeable loss of accuracy. Relative to experimental energies and direct MCDHF calculations in the reference configuration space, the ML-VMCCI error distributions remain comparable to VMCCI, with no systematic increase in bias or dispersion^[21,62]. These results show that ML-VMCCI connects configuration-space compression with sampling-process acceleration, substantially reducing the cost of high-precision excited-state calculations while maintaining reference-space accuracy.

Keywords: Relativistic atomic structure; Configuration interaction; Variational Monte Carlo;

Machine learning