arXiv:2304.05046v1 [physics.chem-ph] 11 Apr 2023

ReaxFF 的敏感度分析
使用希尔伯特–施密特独立性准则进行重新参数化

MichaelFreitas Gustavo,[†,‡] Matti Hellström,[‡] 和Toon Verstraelen[*,†]

†分子建模中心，根特大学，比利时根特

‡*Software for Chemistry and Materials BV*，荷兰阿姆斯特丹

电子邮箱：toon.verstraelen@ugent.be

摘要

我们采用一种全局敏感性方法——希尔伯特–施密特独立性准则（HSIC）——对 Zn/S/H ReaxFF 力场进行重新参数化，以确定最适合重新参数化的参数。在此背景下，参数选择仍然是一项挑战，因为高维优化容易导致过拟合且耗时较长，而选择过少的参数又会导致力场质量较差。我们表明，HSIC 能够准确且快速地识别最敏感的参数，并且使用少量敏感参数进行的优化，其性能优于采用更高维合理用户参数选择所进行的优化。仅使用敏感参数的优化具有以下特点：1）收敛更快；2）损失值与采用朴素选择时得到的损失值相当；3）在验证测试中的准确性相近；4）不会出现过拟合问题。我们证明，HSIC全局敏感性分析是一种成本低廉的优化预处理步骤，能够带来定性和定量上的双重优势，从而显著简化并加速 ReaxFF 重新参数化。

1 引言

1.1 ReaxFF 重新参数化

ReaxFF（反应力场）[1,2] 是用于模拟反应化学的势能面（势能面），可在空间和时间尺度上进行建模，而这些尺度通常是其他更精确但成本更高的方法难以企及的。ReaxFF的速度优势是以精确形式体系被经验方程所取代为代价的，而这些经验方程通常可能包含数百个拟合参数。其中许多参数并不容易从物理角度解释，只有专家用户可能了解合适的取值。

这使得调整参数的任务相当困难。通常采用的做法是最小化某个代价函数，该函数用于衡量 ReaxFF 的预测结果与用户希望复现的某组训练集数值之间的偏差。[3–12] 然而，究竟应该优化哪些参数，一直是个棘手的问题。同时优化所有可能相关的参数容易产生过拟合结果，而且高维优化代价高昂，探索也不充分。[4] 我们此前的工作[4,13] 得出的一个关键结论是，谨慎地设定误差函数（包括参数选择和训练集项）对于获得合理结果至关重要。

通常，研究人员会采用基础的敏感性分析和专家知识来指导参数选择。然而，这些过程通常不够准确、耗费大量工作，或无法对参数空间进行充分采样。在本研究中，我们旨在通过引入一种系统的全局敏感性分析技术，以更稳健的方式指导参数选择，从而解决这一问题。该方法也已编入我们的ParAMS 软件包，[14]使其成为任何重新参数化工作流程中无缝衔接的一部分。

1.2 全局敏感性方法

不确定性分析关注不确定性从模型的输入传播到输出的过程，即根据输入的所有可能变化估计输出的分布。[15] 敏感性分析则将输出的不确定性归因于特定输入。通常通过衡量无条件输出分布$\mathbb{P}(Y)$与给定特定输入时的输出分布 $\mathbb{P}(Y|X_i)$之间的差异来实现。[16,17] 用于进行这种比较的统计算子类型将全局敏感性方法分为两类：基于方差的方法和基于分布的方法。

基于方差的方法假设，这些分布的方差足以描述它们。[16,18] 这种方法已相当成熟，也是我们首先尝试的方法。最受欢迎的基于方差的技术之一称为 Sobol 方法。它基于这样的假设：模型结果的总方差 Y 可以线性分摊到每个输入以及输入的各种组合上：[19]

$$\sum_i S_i + \sum_i \sum_{j>i} S_{ij} + \sum_i \sum_{j>i} \sum_{k>j} S_{ijk} + \cdots + S_{123\ldots d} = 1. \tag{1}$$

S_i 是参数 i 一阶效应的 Sobol 指数，计算如下：[19]

$$S_i = \frac{\mathbb{V}(\mathbb{E}(Y|X_i))}{\mathbb{V}(Y)}, \tag{2}$$

其中，$\mathbb{E}$ 和 $\mathbb{V}$ 分别表示期望值和方差。高阶效应的计算方式类似：

$$S_{ij} = \frac{\mathbb{V}(\mathbb{E}(Y|X_i, X_j))}{\mathbb{V}(Y)}. \tag{3}$$

一种特殊情形称为总效应，其计算方式如下：

$$S_{Ti} = 1 - \frac{\mathbb{V}(\mathbb{E}(Y|\mathbf{X}_{-i}))}{\mathbb{V}(Y)}, \tag{4}$$

其中，$\mathbf{X}_{-i}$ 表示除 X_i 之外的所有参数组成的向量。只有当维度数量极少时，才适合对方差进行完整分解，因为总共需要计算 $2^d - 1$ 个指数。不过，研究表明，一阶效应和总效应足以识别出最敏感的参数。[19]

这种方法的主要缺点是所需样本数量带来的成本。高效计算一阶和总效应 Sobol 指数的方法需要进行$\eta(d + 2)$次函数评估。[19] 虽然没有规定如何选择 η，但通常取几千左右。[20,21] 与获取样本的成本相比，计算指数本身的成本微不足道，因此文献中建议的流程是反复采样并计算，直到指数收敛。在本研究中，我们尝试将预处理步骤引入 ReaxFF 参数化，因此这一步骤的成本应显著低于优化；然而，文献资料表明，实际情况并非如此。[20,21]

Sobol 方法成本高得难以承受，这一点广为人知，因此人们投入了大量精力来降低其成本。[19] 一种常用的方法比 Sobol 方法便宜得多，即基本效应法，也称为 Morris 筛选。[22] Morris 筛选采用精心设计的采样策略来计算“估计效应”，研究表明，该效应可以很好地近似总效应 Sobol 指数。[19,22] 这种方法是文献中常用的降维筛选方法，尤其适用于计算成本高昂的模型。从理论上说，Morris 筛选应当适用于我们的目的；然而，在早期研究中，我们发现事实并非如此。

Sobol 指数和 Morris 筛选都会对样本的生成方式施加规则。对于 Sobol 指数，必须固定某些参数值，同时允许其他参数值变化。对于 Morris 筛选，样本沿轨迹收集，每次只改变一个参数。在我们的具体情境中，这些规则使得获取足够数量的有效样本非常令人沮丧且耗时。这是因为 ReaxFF 损失函数曲面上充满了未定义点，这些点对应的参数值会导致计算崩溃并产生非物理结果。

例如，如果生成莫里斯轨迹时遇到此类点，就必须放弃整条轨迹，因为这些方法不允许输入非有限值或未定义值。寻找一种允许使用这些值的方法，或者不设采样规则、从而可以直接丢弃此类样本的方法，会更简单。

除了与实际应用相关的现实问题外，基于方差的技术所作的假设并不适用于识别 ReaxFF 误差函数中的参数敏感性。这是因为该函数可能产生极大的值，而且其极小值往往位于狭窄的谷底，在采样过程中很可能无法捕捉到。[4] 成本较低的近似方法通常无法充分覆盖参数空间，或者依赖于对输入或输出结构的强假设。[15,23] 然而，对基于方差的技术最主要的批评在于，它们将从分布中提取的信息限制为单一指标。

基于分布的敏感性分析方法是较新的一类方法。与基于方差的方法不同，这些技术不会将分布简化为仅考察其方差，原则上可以通过考察整个分布来捕捉任意依赖关系。人们使用不同的统计量来衡量分布之间的差异，[16]并已开发出多种此类方法。[18,23,24] 我们关注希尔伯特–施密特独立性准则（HSIC），[25] 因为它简单、快速且准确。

HSIC 最初由 Gretton 等人提出[25]。借助核函数，它能够识别输入与输出之间任意的非线性依赖关系。[16] 与许多其他基于分布的技术相比，它更简单、收敛更快，而且无需显式正则化。[25] 此后，它已应用于特征选择、[23,24,26–29] 事件检测、[30,31] 机器学习[32,33] 以及全局优化中的降维[17] 等领域。这最后一种应用尤其符合我们的研究背景。除 Spagnol 等人[17], 之外，我们尚未发现其他将 HSIC 直接用作预优化步骤的研究。

在这项工作中，我们将 HSIC 分析应用于 ReaxFF 重新参数化问题。根据敏感性结果，我们将原始的高维优化简化为低维优化。随后，我们证明，在这一降维后的空间中进行优化，能够得到与高维优化所产生的 ReaxFF 力场相当的结果，同时收敛更快，并降低过拟合风险。

我们聚焦于 ReaxFF，因为文献中经常会出现这种力场的参数化问题。[10, 34–37] 不过，我们在此介绍的技术理论上也可应用于许多领域中常见的任何高维参数化问题，[38–46] 因为这些问题都具有相同的底层松弛数学行为。[47]

在下一节中，我们将介绍 ReaxFF、HSIC度量、所使用的核函数以及重新参数化损失函数。在第 4 节中，我们介绍一个应用敏感性技术的测试案例，其结果将在第 5 节中进行描述，并在第 6 节中讨论未来的研究方向。第 3 节详细介绍我们的计算实现，并在第 7 节中得出结论。

2 数学推导

2.1 ReaxFF 能量势能

全面阐明 ReaxFF能量势超出了本研究的范围，不过，为便于理解，我们在此作简要介绍。感兴趣的读者可查阅 Chenoweth 等人[5]和 Senftle 等人[2]的研究以了解更多详情。

ReaxFF能量势是不同能量贡献项的总和：[5]

$$\begin{aligned} E_{\text{sys}} =& E_{\text{bond}} + E_{\text{lp}} + E_{\text{over}} + E_{\text{under}} + E_{\text{val}} \\ &+ E_{\text{pen}} + E_{\text{coa}} + E_{\text{C2}} + E_{\text{triple}} \\ &+ E_{\text{tors}} + E_{\text{conj}} + E_{\text{H-bond}} + E_{\text{VdW}} \\ &+ E_{\text{charges}} + \dots \end{aligned} \tag{5}$$

主要贡献是键能 E_{bond}，它根据键级计算能量，而键级又是根据原子位置计算得出的（这是任何势能面最基本的输入）。其他贡献可视为对键能的修正。有些修正具有普遍性，例如 E_{over} 和 E_{under}，用于惩罚因成键而配位过多或配位不足的原子。有些修正则非常具体，例如 E_{C2}，用于修正 C_2 分子的能量。此外，还有一个描述电荷波动的能量项（E_{charges}）。ReaxFF 通常使用 EEM[48] 或 ACKS2[49] 电荷模型。本研究使用 EEM 电荷。

每个能量贡献通常都是键级和经验参数的复杂函数。例如：[5]

$$\begin{aligned} E_{\text{bond}} = & -D_e^{\sigma} \cdot BO_{ij}^{\sigma} \cdot \exp\left[p_{be1}\left(1-\left(BO_{ij}^{\sigma}\right)^{p_{be2}}\right)\right] \\ & - D_e^{\pi} \cdot BO_{ij}^{\pi} - D_e^{\pi\pi} \cdot BO_{ij}^{\pi\pi} \end{aligned} \tag{6}$$

其中，D_e^{σ}、D_e^{π}、$D_e^{\pi\pi}$、p_{be1}和 p_{be2}是参数，BO_{ij}是键级。

这些贡献的复杂性和数量很快就会使 ReaxFF 势能面包含大量参数，尤其是因为大多数参数都是一个或多个原子类型的函数。例如，每种待建模原子类型组合都对应一个 D_e^{σ} 参数。

为了便于理解这些参数，原作者将

它们分为以下几组或模块：常规项、原子、键、非对角项、键角、扭转角和氢键。这些分组与势能面的计算无关，但有助于组织参数，并且出现在大多数 ReaxFF 实现的输入格式规则中。当我们在第4.1节介绍一种新的 Zn/S/H 力场时，还会回到这些分组。

在本研究中，我们使用“力场”一词来指代一整套 ReaxFF参数及其取值。力场取决于其包含的原子类型，有时仅适用于特定条件。例如，包含氢和氧的 ReaxFF 力场通常被认为适用于水相或燃烧化学。

2.2 损失函数

力场的重新参数化需要构建一个训练集（$\mathbf{b}_{\mathrm{ref}} \in \mathbb{R}^m$），该训练集是由 m 标量化学性质训练点数值组成的向量，这些数值是希望 ReaxFF复现的对象。重新参数化的目标是最小化一个损失函数，该函数用一个数值量化 ReaxFF模型与训练数据之间的总体差异。这个损失函数可以采用多种形式，但最常见的是平方误差和（SSE）。我们将 SSE 定义为:

$$\mathrm{SSE}(\mathbf{x}) \coloneqq \sum_{i=1}^{m} w_i \left(\frac{b_{\mathrm{ref},i} - b(\mathbf{x}, \mathbf{G}_i)}{\sigma_i} \right)^2 . \tag{7}$$

训练集（$b_{\mathrm{ref},\ i}$）中的每一项都会与 ReaxFF 模型（b($\mathbf{x}$，$\mathbf{G}_i$）生成的结果进行比较。比较时使用一个可调整的参数向量（$\mathbf{x}$），以及一组系统输入和固定参数（$\mathbf{G}_i$）。我们将选定进行优化的参数称为活跃参数，并将活跃参数集记为 $\mathscr{D}$。训练集中的项目类型可以各不相同（能量、力、键角等）。为了使这些项目具有可比性且不受单位影响，损失函数中的每一项都要除以一个适当的 σ_i ，其单位与 $b_{\mathrm{ref},\ i}$相同。 σ_i 值通常被视为该项可接受的误差

。最后，可以根据训练值的重要性，通过权重向量（**w**）赋予不同的权重。例如，与复现几何构型极度夸张时的能量相比，准确得到能量极小值更为重要；因此，在损失函数中可以给前者赋予更高的权重。对于力等多维数据，$n \times 3$ 矩阵只需展平为训练数据向量即可。换句话说，$\mathbf{b}_{\mathrm{ref}}$ 中的单个元素并不是“分子 A 上的力”，而是“分子 A 中原子 1 所受力的 y 分量”。有关具体实现的更多详情，请参阅 ParAMS 文档。[14]

σ 和**w** 的功能基本相似，因为它们都是常数，理论上可以合并。使用其中一个、另一个或两者兼用，取决于不同从业者的偏好。传统上只使用 **σ** ，不过对一些用户而言，权重提供了更直观的数值。我们选择为 **σ**使用固定的值向量，并调整权重以平衡训练集。

2.3 希尔伯特–施密特独立性准则

本节从从业者的角度介绍希尔伯特–施密特独立性准则（HSIC）。我们的阐述不像其他文献那样一般化，[17,25,28] 我们只力求向对参数选择感兴趣的读者解释其基本思想。我们假设单个参数的取值空间为 $\mathcal{X} \subseteq \mathbb{R}$ ，损失值的空间为 $\mathcal{Y} \subseteq \mathbb{R}$。通常，参数有界且损失值为正，但这并非 HSIC 的严格要求。

设 $\mathbb{R}$ 值函数在 $\mathcal{X}$ 上构成的希尔伯特空间为 $\mathcal{H}_x : \mathcal{X} \to \mathbb{R}$，其上定义了内积 $\langle \cdot, \cdot \rangle_{\mathcal{H}_x}$。若存在核 $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ ，使得 $\forall x \in \mathcal{X}$ 且 $\forall f \in \mathcal{H}_x$:，则该空间称为再生核希尔伯特空间（RKHS）。

$$k(x, \cdot) \in \mathcal{H}_x \text{ and}$$

$$\langle f, k(x, \cdot) \rangle_{\mathcal{H}_x} = f(x).$$

第二个条件称为再生性质：函数值 $f(x)$ 可以通过将 f 与部分求值的核取内积来再现。类似地，对于损失值空间 $\mathcal{Y}$，我们定义核为 ℓ 的 RKHS $\mathcal{H}_y$。

核 k 可以用特征映射 ϕ 来定义： $\mathcal{X} \to \mathcal{F}_x$，其中 $\mathcal{F}_x$ 是一个新的希尔伯特空间，通常称为 x的特征空间。这意味着，每个值 x 都由函数 $\phi(x) \in \mathcal{F}_x$唯一表示。任何正定对称核 k 都可以表示为特征空间中的内积：

$$k(x, x') \coloneqq \langle \phi(x), \phi(x') \rangle_{\mathcal{F}_x} \tag{8}$$

在实际应用中，为了使计算易于处理，会避免直接使用 ϕ，而改用核 k。即使特征映射在计算上难以实现，核通常也很容易计算。同样，存在特征映射 ψ： $\mathcal{Y} \to \mathcal{F}_y$，使得 $\ell(y, y') := \langle \psi(y) \psi(y') \rangle_{\mathcal{F}_y}$。

现在考虑随机变量 $X \in \mathcal{X}$ 和 $Y \in \mathcal{Y}$，其联合分布为 $\mathbb{P}_{XY}$。基于方差的敏感性方法通过 $\mathrm{cov}[X, Y] = \mathbb{E}_{XY}[XY] - \mathbb{E}_X[X]\mathbb{E}_Y[Y]$ 刻画 Y 对 X 的依赖关系，但这只能捕捉线性相关性。原则上，可以改为在特征空间中计算协方差，从而克服这一局限：$\mathrm{cov}[\phi(X) \psi(Y)] = \mathbb{E}_{XY}[\phi(X)\psi(Y)] - \mathbb{E}_X[\phi(X)]\mathbb{E}_Y[\psi(Y)]$。然而，后一种协方差并不实用，因为它直接在特征空间中运算，而且该协方差是新希尔伯特空间 $\mathcal{F}_x \otimes \mathcal{F}_y$ 的一个元素，这使得对其进行直接解释变得困难。Gretton 等人[25] 展示了如何通过将后一种协方差的平方范数改写为以下形式来解决这两个问题。

核函数:

$$\mathrm{HSIC}(\mathcal{F}_x, \mathcal{F}_y, \mathbb{P}_{xy}) \coloneqq \left\|\mathrm{cov}[\phi(X), \psi(Y)]\right\|^2_{\mathcal{F}_x \otimes \mathcal{F}_y} \tag{9}$$

$$= \mathbb{E}_{XX'YY'}[k(X, X')\ell(Y, Y')] \tag{10}$$

$$+ \mathbb{E}_{XX'}[k(X, X')]\mathbb{E}_{YY'}[\ell(Y, Y')]$$

$$- 2\mathbb{E}_{XY}\left[\mathbb{E}_{X'}[k(X, X')]\mathbb{E}_{Y'}[\ell(Y, Y')]\right]$$

HSIC的这一定义得出一个正实数，当 X 和 Y 的相关性越强时，该数值越大。当 k 和 ℓ 是线性核时，HSIC 简化为 $\left|\mathrm{cov}[X, Y]\right|^2$，与基于方差的敏感性准则相比没有增加任何价值。对于非线性核，HSIC还可以检测 X 和 Y之间的非线性相关性。理想情况下， k 和 ℓ 是特征核，[17] 对于这些核，HSIC $= 0$ 当且仅当 X 和 Y是独立的（$\mathbb{P}_{XY} = \mathbb{P}_X\mathbb{P}_Y$）。在常见的核函数中，高斯核是特征核，而线性核和多项式核则不是。HSIC可以使用非特征核，但这样会失去捕获所有依赖关系的保证。[28]

通常无法显式获得 $\mathbb{P}_{xy}$，但是，如果我们对分布进行采样，就可以估计 HSIC。我们将一组样本定义为:

$$Z \coloneqq (\mathbf{X}, \mathbf{y}), \tag{11}$$

其中 $\mathbf{y} \in \mathbb{R}^{n \times d}$ 是一个包含 n 行均匀随机采样的参数向量{的矩阵$\mathbf{y}_1, \mathbf{y}_2, \ldots \mathbf{y}_n$}，并令 $\mathbf{y} \in \mathbb{R}^n = \{\mathrm{SSE}(\mathbf{y}_1), \mathrm{SSE}(\mathbf{y}_2) \ldots, \mathrm{SSE}(\mathbf{y}_n)\}$ 为相应的损失函数值集合。

为了测量索引为 ω的参数的敏感性，我们取其对应的包含 n 个样本值的列，$\mathbf{Y}_{\cdot\omega}$，以及损失值:

$$Z_\omega \coloneqq (\mathbf{X}_{\cdot\omega}, \mathbf{y}), \tag{12}$$

并且我们采用了 Song 等人[28]的HSIC 无偏 估计量：

$$\mathrm{HSIC}(\mathcal{F}_x, \mathcal{F}_y, Z_\omega) \approx \frac{1}{n(n-3)} \left[\rho_1 + \rho_2 - \rho_3\right], \tag{13}$$

其中：

$$\begin{aligned}
\rho_1 &\coloneqq \mathrm{tr}(\tilde{\mathbf{K}}\tilde{\mathbf{L}}) \\
\rho_2 &\coloneqq \frac{\mathbf{1}^\top \tilde{\mathbf{K}} \mathbf{1}\mathbf{1}^\top \tilde{\mathbf{L}} \mathbf{1}}{(n-1)(n-2)} \\
\rho_3 &\coloneqq \frac{2}{n-2} \mathbf{1}^\top \tilde{\mathbf{K}}\tilde{\mathbf{L}}\mathbf{1} \\
\tilde{\mathbf{K}}_{ij} &\coloneqq (1-\delta_{ij}) k(x_i, x_j) \quad \forall x_i, x_j \in X_{\cdot\omega} \\
\tilde{\mathbf{L}}_{ij} &\coloneqq (1-\delta_{ij}) \ell(y_i, y_j) \quad \forall y_i, y_j \in \mathbf{y}.
\end{aligned}$$

HSIC是一种便捷的指标，可用于对参数对输出的影响进行排序；不过，为了让解释更加容易，我们将参数的敏感性定义为归一化的 HSIC 值：[17]

$$s_\omega = \frac{\mathrm{HSIC}(\mathcal{F}_x, \mathcal{F}_y, Z_\omega)}{\sum_{i=1}^{d} \mathrm{HSIC}(\mathcal{F}_x, \mathcal{F}_y, Z_i)}. \tag{14}$$

按照构造，敏感性值有界（$0 \le s_\omega \le 1$）且 $\sum s_\omega = 1$。这使得敏感性相比 HSIC 成为更直观、更易于解读和理解的值。

有偏HSIC 估计量[28] 的优点是其构造方式保证结果为正。本文使用的无偏估计量虽然更准确，但对于敏感性较低的项目，有可能产生较小的负值。在敏感度分析中，结果的定性考量或许比定量考量更重要，因此我们在计算敏感性之前，直接将所有负的 HSIC 值截断为零（公式 14）。

如需更详细地了解 HSIC，建议读者参阅 Gretton 等人[25]关于其最初介绍的文献，以及 De Lozzo 和 Marrel [23]、Da Veiga [24]、Song

等人 [28] 和 Spagnol 等人 [17] 关于近期研究成果的文献。

请注意，也可以考虑更细粒度的敏感性方法，并令$\mathbf{y}$表示各个残差值（即 $y_i = b_{\mathrm{ref},i} - b(\mathbf{x},\ \mathbf{G}_i)$）。这样便可以确定特定训练集项的敏感性，从而为训练集设计提供更多信息。这代表了当前工作的扩展，本文不对此进行讨论，而是将其留作进一步研究的方向。

2.4 核函数

图1展示了本文所使用的核分布，并在适当情况下采用了不同的参数值。

我们为参数核选择了具有代表性的高斯核：

$$k(x, x') \coloneqq \exp\left[-\frac{(x - x')^2}{2\sigma^2}\right]. \tag{15}$$

该核捕捉了参数之间的相似性与不相似性，并允许分布呈现任意形状。为清晰起见，σ 由于其在公式15中的使用方式，与公式 7 中的同一符号无关。

我们测试了损失核的多个候选项。首先考虑 Spagnol 等人引入的非特征性阈值核 [17]：

$$\ell(y, y') \coloneqq h(y)h(y') \tag{16}$$

$$h(y) \coloneqq \begin{cases} 1 & \text{if } y < q_\alpha, \\ 0 & \text{otherwise.} \end{cases} \tag{17}$$

q_α 是数据的 α 分位数对应的值。构造上述核是为了让分布的权重集中在 $\mathbf{y}$ 的较小数值上，因为这些值代表了良好的损失函数值，也是我们最希望深入了解的。Spagnol

等人[17] 表明，阈值化对于识别极小值附近的敏感性至关重要，并可避免计算被非常大的函数值所淹没。

接下来，我们考虑阈值核的一种特征性连续近似，称之为合取高斯（CG）核：

$$\ell(y, y') \coloneqq \exp\left[-\frac{(y^2 + y'^2)}{2\gamma^2}\right], \tag{18}$$

其中，较小的 γ 参数值会使分布更加集中于最佳值。

作为比较，我们还测试了一种简单的非特征线性核：

$$\ell(y, y') \coloneqq yy', \tag{19}$$

以及公式15中介绍的高斯核。还可以使用其他核函数，但我们将留待后续工作再作探讨。

为稳定数值计算并处理数量级问题，我们会先对损失值进行变换和缩放，再应用损失核：

$$\tilde{y}(y) = \frac{\ln y - \min(\ln \mathbf{y})}{\max(\ln \mathbf{y}) - \min(\ln \mathbf{y})}. \tag{20}$$

3 软件

训练　数据是使用 Amsterdam Modeling Suite (AMS) 2022 计算的。

敏感性计算已使用 Python 编程语言实现，并集成到 ParAMS 2023,[14] 重新参数化工具包的开发版本中，该工具包与 AMS 捆绑提供。它还用于通过 GloMPO 优化管理软件处理力场优化。[4]

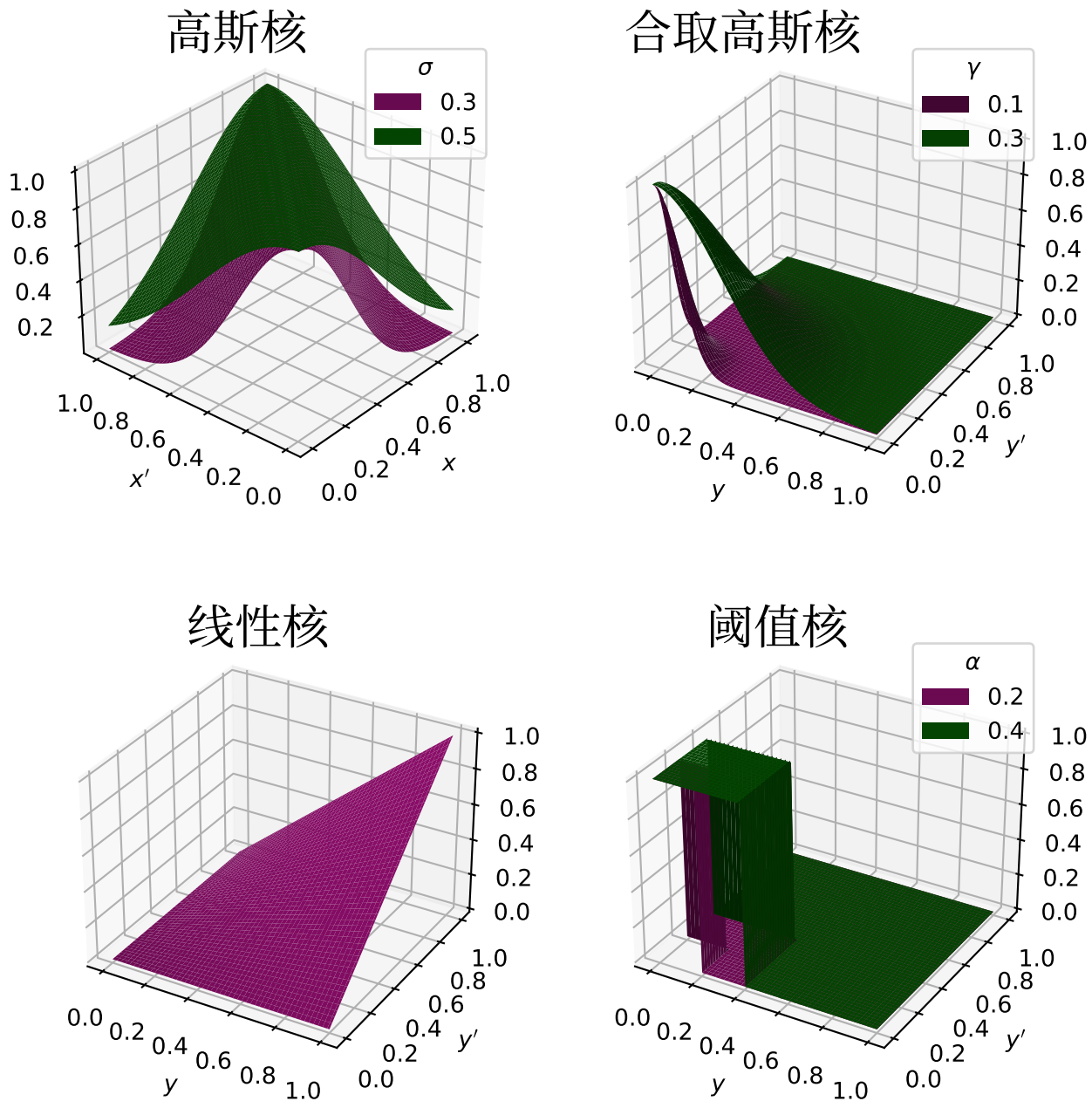


图1：所测试核函数的绘图。

4 参数化问题

本工作的主要目标是展示我们新的敏感性方法及其效用。因此，我们设计的训练集比通常创建生产级参数化所需的训练集更小。设计一个大型训练集本身就需要一篇完整的论文，也会严重偏离本文的重点。较小的训练集有助于我们使讨论更加集中和清晰。

无论如何，训练集设计始终是一个迭代过程：随着对新力场进行验证测试并发现原始训练集存在不足，会不断添加更多条目。[50] 这项工作可视为第一轮迭代，后续工作可以在此基础上继续推进。

4.1 初始值力场

我们的目标是创建一种新的 ReaxFF 力场，能够正确模拟 H_2S 在硫化锌上的吸附。硫化锌是一种吸收剂，因其优异的电子特性而受到了一些关注。[51–55] H_2S 是一种常见但有毒的气体，其在硫化锌上的吸附行为已针对气体检测[55] 和气体去除进行了研究。[53] 据我们所知，目前尚无专门设计的 Zn/S/H ReaxFF 力场来模拟这一行为。

表1详细说明了我们用于指代构成 ReaxFF 力场文件的各个参数块的命名法。[56] 关于参数块的讨论，请参阅第 2.1 节。

我们的初始力场融合了已发表力场中的参数值：

1.ATM:S、BND:S.S 和 ANG:S.S.S参数块均填入了 Islam 等人 Li/S 力场中的数值[57]；

2. ATM:H、BND:S.H 和 ANG:H.S.H 的数值取自 Müller 和 Hartke 的 C/H/O/S 力场[50]；

表1：ReaxFF 力场文件格式中用于指代不同参数块的别名。

别名	描述	参数数量
GEN	通用参数	41
ATM:W	W 类型原子的参数	每种原子类型 32 个
BND:W.X	W-X 键的参数	每个 W.X 对 16 个
OFD:W.X	原子对的非对角定义	每个 W.X 对 6 个
ANG:W.X.Y	由 W-X-Y 原子形成的角度参数	每个角度组 7 个
TOR:W.X.Y.Z	由 W-X-Y-Z 原子形成的扭转角参数	每个角度组 7 个
HBD：W.H.X.	氢键参数（原子 W 和 X）	每个W.X对4个

3. GEN、ATM:锌、BND:H.H、BND:锌.H、BND:锌.锌、BND:锌.S、ANG:锌.锌.S、ANG:锌.S.锌 ,

   ANG:Zn.S.S、ANG:S.Zn.S 和 HBD:S.H.S参数块填入了Raymand 等人在Zn/O/H力场中发表的参数[54]。该力场不包含任何硫化合物，因此使用相应的与氧相关的参数块。例如，将BND:Zn.O 参数填入初始力场的 BND:Zn.S 参数块中。

力场包含279个参数。根据训练集的内容和参数的描述，我们运用一些常见的直觉，从中选择53个参数进行初始优化。该选择是“贪心式”的，即当一个参数满足以下条件时，就会选择该参数进行优化：

1. 它可能会影响某个训练集项目的行为，

2. 训练集包含与该参数相关的项目，并且

3. 该参数适用于优化。

例如，元素的原子质量严格来说属于参数，但更改它们并不合适。同样，训练集中不存在 π- 键，因此不会选择其相关参数进行优化。最后，我们不会优化任何原子参数或通用块参数。通常认为，这些参数属于更“专家级”的参数，不适合用于初步阶段优化。[14]

表 2 列出了所选的 53 个用于优化的参数。我们使用了 ParAMS[14] 提供的默认参数范围，相关信息可见补充信息。

表 2：初始值的朴素参数选择。方程编号和参数名称沿用.[58]中使用的符号。

区块名称		Eqn.	描述	原子
ANG	-p_hb2	18	氢键/键级	H.S.H
	-p_hb3	18	氢键参数	H.S.H
	p_hb1	2	氢键能	H.S.H
	p_val1	13a	价角参数	H.S.H S.S.锌 S.锌.S S.锌.锌 锌.S.锌
	p_val2	13a	价角参数	H.S.H S.S.Zn S.Zn.S S.Zn.Zn Zn.S.Zn
	p_val4	13b	价角参数	H.S.H S.S.锌 S.锌.S S.锌.锌 锌.S.锌
	p_val7	13c	配位不足	H.S.H S.S.Zn S.Zn.S S.Zn.Zn 锌.S.Zn
	Theta_0,0	13g	180°-（平衡角度）	H.S.H S.S.Zn S.Zn.S S.Zn.Zn 锌.S.Zn
HBD	r_hb^0	18	氢键平衡距离	S.H.S
BND	D_e^sigma	6, 11a	σ键解离能	H.S S.S S.Zn Zn.Zn
	p_be1	6	键能参数	H.S S.S S.Zn Zn.Zn
	p_be2	6	键能参数	H.S S.S S.Zn Zn.Zn
	p_bo1	2	σ键级	H.S S.S S.Zn Zn.Zn
	p_bo2	2	σ键级	H.S S.S S.Zn Zn.Zn
	p_ovun1	11a	过度配位惩罚	H.S S.S S.Zn Zn.Zn

4.2 训练数据

训练集使用 BAND 和 PBEsol 及 DZ 或 DZP 基组计算得出，由 16 个任务组成，从中提取了 471 个独立训练点。

AMS BAND 使用 Hirshfeld、Voronoi 形变、Mulliken 和 CM5 方法计算电荷。我们已将 Hirshfeld 电荷纳入训练集。尽管这些电荷并非没有问题，[59] 我们仍认为它们是目前最佳的选择。由于我们在初始值参数选择中没有启用任何与电荷相关的参数（见第 4.1节），因此这些电荷的准确性并非特别重要；我们纳入训练集的少量电荷主要用于检验结果是否合理。

训练集由以下内容组成：

1.能量-体积扫描、电荷以及硫化锌的立方闪锌矿、立方岩盐和六方纤锌矿结构的生成焓；

2. H_2S 的优化几何构型、电荷、H-S键扫描和 H-S-H键角扫描；

3. 周期性闪锌矿表面的生成能、键长、电荷和键角；

4. 畸变闪锌矿表面的力；

5. 闪锌矿上吸附的 H_2S 的电荷、键长和吸附能。

训练集的详细信息见补充信息的 F节。

σ 使用 ParAMS 默认值。在初始训练集中，单个能量、键角、距离、电荷和扭转角以及力组均采用相同权重。完整的训练集、作业集合和初始参数接口请参阅补充信息。

5 结果与讨论

我们的工作流程如下：

基于朴素的参数选择（纯粹出于比较目的）对初始力场进行重新参数化；

进行敏感性计算，以确定参数排序（完全独立于上一步）；

再次进行重新参数化，使用敏感度分析确定的最不敏感和最敏感参数中的 10、20、33 和 43 个；

进行验证测试，使用初始参数化以及找到的一些最佳参数化结果。

我们的目标是证明：

1. HSIC 敏感性方法能够正确识别最敏感的参数；2. 在由高度敏感参数构成的低维空间内进行重新参数化，可以在更短时间内找到质量相近的力场；

3. 高维度的重新参数化存在过拟合的风险。

5.1 初始优化

使用协方差矩阵自适应进化策略（CMA-ES）[60] 执行了初始的十六次重新参数化，据报道该方法在 ReaxFF 重新参数化中表现良好。[4,13] 所有优化器均从第4.1节所述的初始力场值开始，并采用较宽的初始采样分布（$\sigma_0 = 0.5$）。根据 ParAMS 自动建议的边界，所有参数均缩放至零到一之间。

在以下情况下，优化器会停止：

1. 在最近的 2000 次函数评估中，其最低值没有改善，且探索到的值非常相似；或

2. 触发了任何默认的内部 CMA-ES 收敛准则。

这些优化器并行运行，累计使用 300000 次函数评估后，整个优化过程停止。这些条件都相当严格，为所有优化器提供了充足的时间来寻找尽可能好的最小值。

图 2 显示了这些重新参数化的优化器轨迹，并展示了原始损失 20 662 降低至 410 至 308 之间的结果。第 5.3 节将把这些结果与其他优化结果进行比较，并作更详细的讨论。每个优化器的详细结果可见补充信息的 A 节。

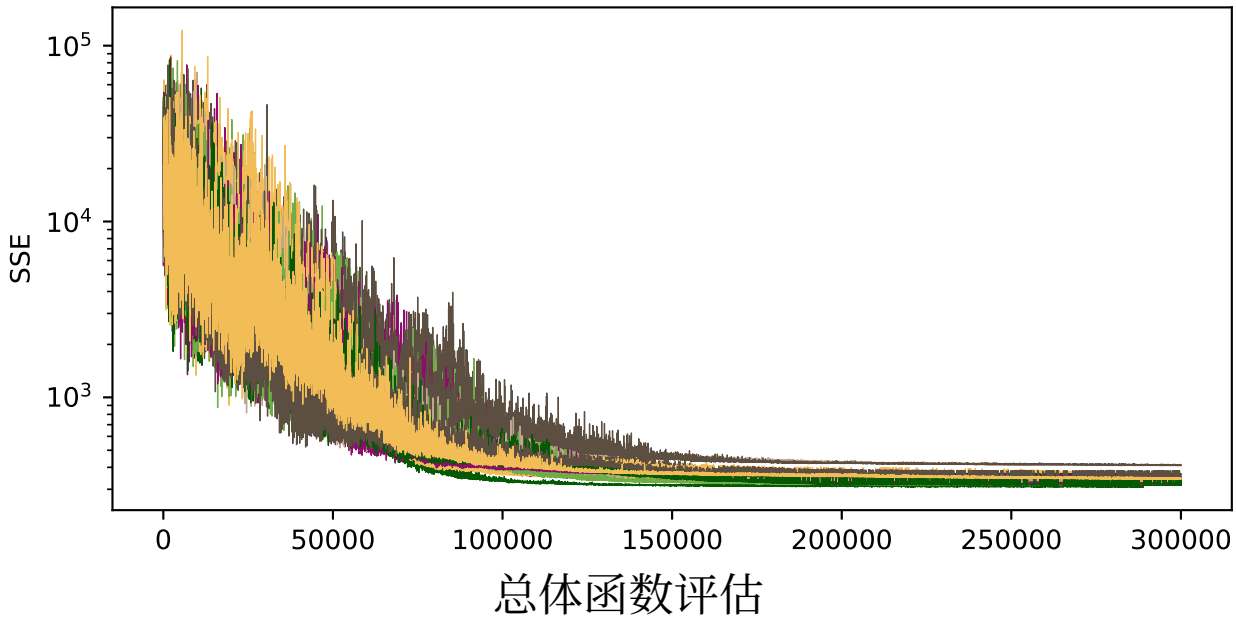


图2：采用朴素参数选择的初始力场的十六种 CMA-ES 重新参数化的优化器轨迹。

5.2 敏感度分析

为进行敏感度分析，从原始优化空间中收集了 10 000 个均匀随机生成的样本。在一个 64 核节点（AMD EPYC 7513 @ 2.6 GHz）上，采样过程总共耗时 142 分钟。该采样仅进行了一次，所有敏感性计算均使用这组相同的样本。

采样损失值的分布（以及转换后的值 $\tilde{y}$）如图3所示；最低值为2651。大约一半的样本实际上比初始损失值更差，而且这组样本显然不包含通过第5.1节中的原始优化找到的优良极小值。我们将证明，要准确估计参数敏感性，并不要求样本集中包含这类极小值。

我们没有在单次 HSIC 计算中一次性使用全部样本，而是采用自助法重复计算了 10 次。每次自助抽样均从原始的 10000 个样本集中随机抽取 2000 个子样本。每次计算使用较少的样本，可显著加快速度。

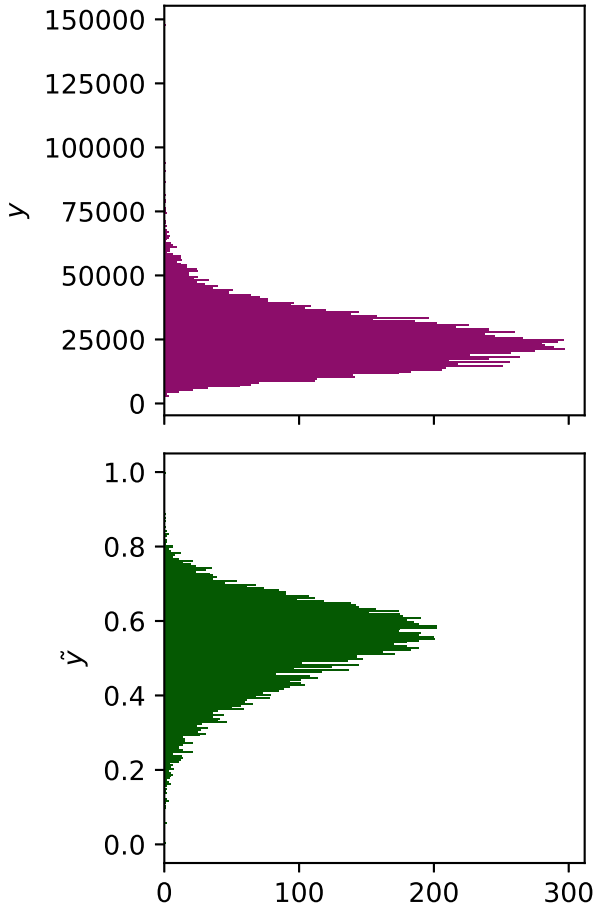


图 3：采样过程中收集并用于敏感性分析的损失值分布（上图），以及根据公式 20 缩放后的损失值（下图）。

计算时间，而结果的离散程度则能反映误差大小。除非另有说明，以下讨论的敏感性值均指 10 次重复的平均值。算法 1 给出了计算过程的伪代码；为简洁明了起见，我们未严格遵循符号规范。

算法 1 敏感性计算工作流。

要求： SSE，k，ℓ

随机生成 $\mathbf{y} \in \mathbb{R}^{10\,000\times 53}$

$\mathbf{y} \in \mathbb{R}^{10\,000} \leftarrow$ SSE(X)　　评估损失

$\tilde{\mathbf{y}} \leftarrow \tilde{y}(\mathbf{y})$　　。缩放损失值

针对 $i = 1, \ldots,$ **10 执行**　　。重复该计算

　$\mathbf{X}^{\mathrm{sub}} \leftarrow 2000$ 随机行 $\mathbf{X}$

　$\tilde{\mathbf{y}}^{\mathrm{sub}} \leftarrow$ 对应的 2000 个损失值

　$\tilde{\mathbf{L}} \leftarrow \ell(\tilde{\mathbf{y}}^{\mathrm{sub}})$

　对于 $\omega = 1$、.....、53 **执行**　　。每个参数

　　$\tilde{\mathbf{K}} \leftarrow k(\mathbf{y})$

　　$\mathbf{H}_{i\omega} \leftarrow$HSIC($\tilde{\mathbf{K}}, \tilde{\mathbf{L}}$)

$\mathbf{h} \leftarrow$均值$(\mathbf{H}_1, \ldots, \mathbf{H}_{10})$　　. 平均重复次数

$g_j \leftarrow \frac{h_j}{\sum_i h_i}$　　. 计算敏感性

以下核函数应用于损失值：CG 核（$\gamma = 0.3$）、高斯核（$\sigma = 0.3$）、线性核和阈值核（$\alpha = 0.4$）。在所有情况下，高斯核（$\sigma = 0.3$）均应用于参数值。由于所有值都缩放到 0 到 1 之间，$\sigma = 0.3$ 的值近似于均匀分布在这些值之间的所有样本的中位距离。多次报告表明，这是一个合适的值。[17,61]

执行敏感性计算所需的平均时间（与损失核无关，并包括重复次数）约为60 秒。优化需要数天才能完成，因此，将这些采样和敏感性步骤纳入重新参数化工作流程不会成为限制因素，并且正如下面将展示的那样，还会带来若干优势。

我们注意到，我们还使用包含 30000 个样本的样本池，并在每次计算中使用10000 个子样本，重复进行了计算。这些结果在.....之间产生了更为集中的分布

增加重复次数会耗费更多时间，但平均敏感度值基本相同。换句话说，这些重复次数不值得投入精力，因此下文不再展示。

敏感性计算的规模为 $\mathcal{O}(rdn^2)$，其中 r 是重复次数，d 是参数的数量， n是每次计算使用的子样本数量。增大矩阵的规模对计算时间的影响最大，因此能够通过使用小型子样本进行重复计算，获得稳健的参数排序，具有很大的优势。

图4比较了每项计算针对每个参数确定的敏感性值。尽管不同损失核之间存在很大差异，但对于最敏感的参数，其排序非常稳健。显著偏差仅出现在最不敏感的参数上，而这些参数的数值彼此都很接近（请注意对数坐标轴）。这表明我们采用的 HSIC 方法具有稳健性，并不过度依赖损失核或超参数的选择。

在区分敏感性方面最为困难的损失核是阈值核。我们认为，这是因为该核是不连续的，并且由于数值会被切换为零或一，因此为计算提供的信息非常少。我们对这一核的研究受到了 Spagnol等人[17]工作的启发。他们证明了这一核对于将计算聚焦于较小的损失值（接近所关注的极小值）并消除较大数量级差异的影响是必需的。正如第2.4节所讨论的，我们的实现总是先对损失值取对数，然后对结果进行缩放；这似乎解决了数量级问题。随机采样通常也不会包含非常好的极小值（如上文所述，在此情况下确实如此），因此阈值核的第二个优势就不那么重要了。其他核函数所提供的连续性似乎使参数的敏感性值更容易收敛。

我们还探讨了在补充信息的 B节中使用不同核参数的影响。只要选择合理的核参数值，使核值不会全部非常接近，这对敏感性结果的影响很小

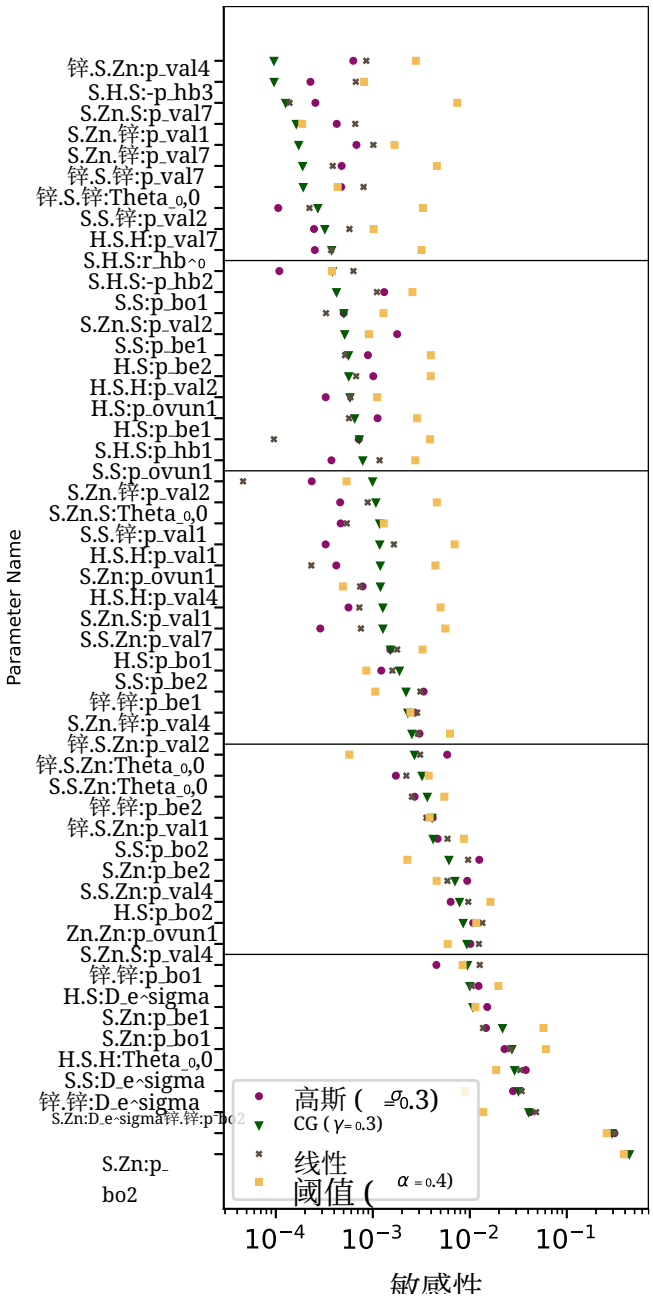


图4：使用不同核函数对损失值进行处理后，确定的各参数平均敏感度值（每种方法运行十次）。参数按照 CG 核的结果排序，该结果用于后续的重新参数化。黑线标出了最敏感的 10、20、33 和 43 个参数。

变为零或一。

鉴于使用各种损失核得到的排序相似，我们将继续仅讨论使用 CG 核（$\gamma = 0.3$）得到的结果。这一选择多少有些随意，但可以认为它具有最好的理论基础，因为它强调了低极小值的敏感性。

图 5 展示了按 a) 参数组和 b) 参数名称分组后得到的敏感性。例如，“S.Zn”指与硫-锌键相关的所有键参数，而 p_bo2 指所有 p_bo2 键参数，无论涉及哪些原子。p_bo2 用于计算 σ-键级，是一个指数项。D_e~sigma 是一个线性参数，与 σ-键级相乘，以确定键能对总势能的贡献（见公式 6）。

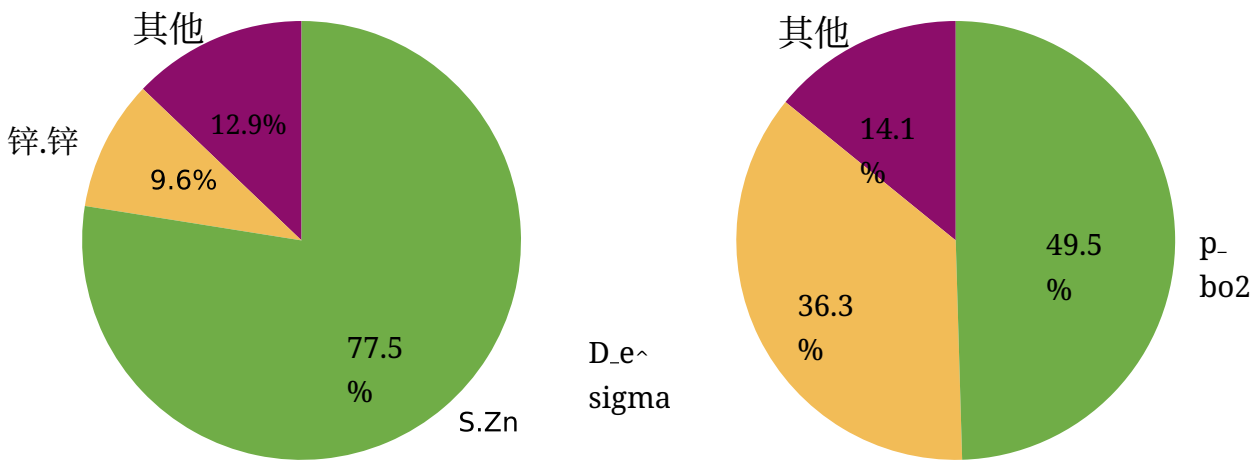


图 5：使用 CG 核进行 HSIC 计算得到的分组敏感性。按参数组（左）和参数名称（右）分组。

这些 σ-键参数的识别，尤其是硫-锌键和锌-锌键的参数，考虑到训练集的组成，是相当合理的。然而，敏感性竟然主要由这些参数决定，其程度可能令人惊讶。尽管如此，这些结果似乎符合直觉。

5.3 仅使用敏感参数重新参数化

基于上述确定的敏感度，我们使用以下参数进行了重新配置的重新参数化：最敏感的10、20、33和43个参数，以及最不敏感的10、20、33和43个参数。最敏感参数的分组如图4所示。我们做出这些选择，使得最敏感的10个参数与最不敏感的43个参数互为补集；两者合起来涵盖原来的53个参数。其他组合也遵循同样的原则。每次重新参数化都按照第5.1节所述的原始方法进行。

在更实际的情况下，我们不主张盲目地只激活最敏感的 n参数；我们在此这样做只是为了避免将人为决策引入结果。在实践中，应将敏感性分析作为指导，以更好地理解训练数据如何影响损失函数，并在选择参数时结合一些人的直觉。

为清晰起见，我们引入以下命名法来指代不同的优化：

O53 使用原始 53 个活跃参数进行重新参数化；**M**$\#$ 使用最敏感的 # 个参数进行重新参数化；**L**$\#$ 使用最不敏感的 # 个参数进行重新参数化。

图6展示了对于每种优化配置，十六个并行优化器中任一优化器所得到的当前最佳损失值（补充信息的A部分包含各个优化器的结果）。表3展示了最低损失值与初始损失值之差占初始损失值的比例。换句话说，这是优化过程中“移除”的初始损失所占的百分比。不出所料，O53产生了最低的损失，因为它具有最多的自由度。可以预期，活跃参数数量的任何减少都会使可实现的损失值变差。事实上，所有其他优化都找到了比原始优化更差的极小值。然而，使用最敏感参数与最不敏感参数之间的差异非常明显。

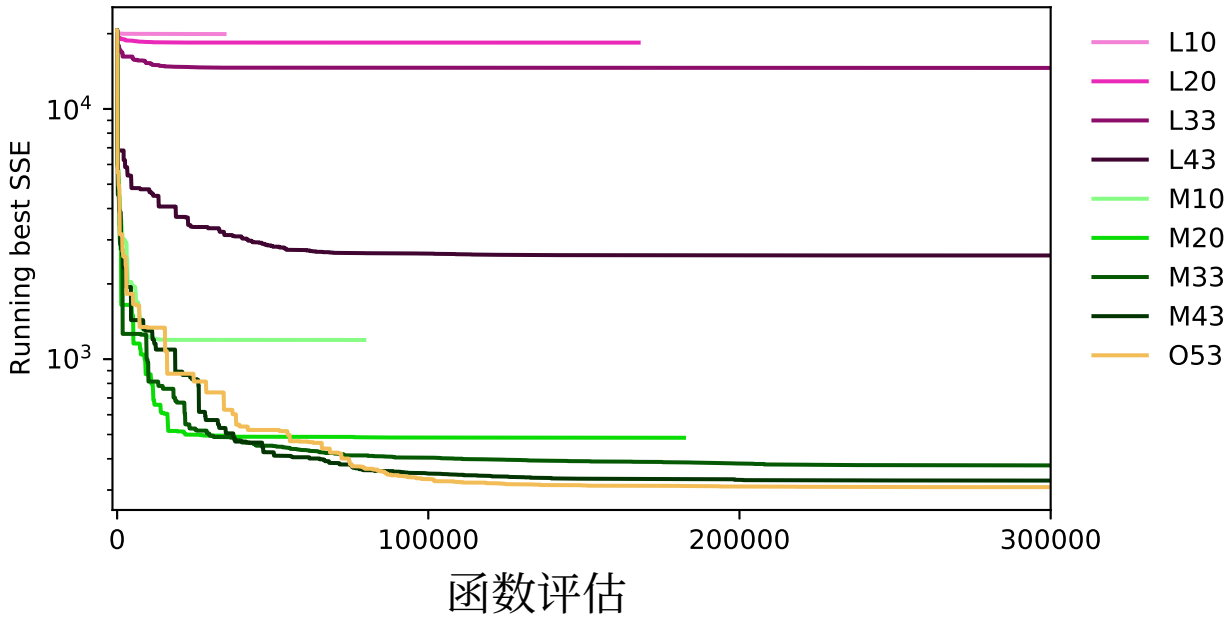


图6：各优化配置下，十六个并行优化器中任意一个所达到的运行中最佳值。

Ta表3：优化过程中损失初始值减少的比例 .

	最敏感（%）	最不敏感（%）
10	94.2	3.61
20	97.7	10.9
33	98.2	29.4
43	98.4	87.4
53	98.5	

在所有情况下，使用最敏感的参数都能消除原始误差的 $>93\%$。事实上，一些 M33 和 M43 优化器找到的极小值优于一些 O53 优化器找到的极小值。相比之下，使用数量相同的最不敏感参数进行优化，效果非常差，无法显著降低损失值。尤其值得注意的是，M10 找到的极小值优于其补集 L43。同样，L10 只能将损失降低 4%，而 M10 可降低 94%。

还需注意，M10 和 M20 能够在达到 300 000 次函数评估上限之前很久就收敛，所用时间约为 O53 的三分之一和三分之二。这节省了大量时间，因为 O53 耗时约两天才完成，远超运行敏感度分析以及 M10 或 M20 所需的时间。

这些结果表明，我们提出的 HSIC 敏感性方法能够快速识别用于重新参数化的最敏感参数，而且由此指导的参数选择能够在更短时间内得到较好的极小值。

5.4 力场比较与验证

为验证新力场的质量并展示过拟合的风险，我们使用初始值力场，以及 O53、M10 和 M20 在各自执行的 16 次优化中产生的误差最低的力场，进行了多项验证测试。

我们考虑：

1. 训练集项上的误差；

2. H_2S 在闪锌矿和纤锌矿晶片上的吸附分子动力学模拟；

3. 闪锌矿体相晶体的分子动力学模拟；

4. H_2S 分子在 (110)闪锌矿表面上的吸附能；以及 5. (110)闪锌矿表面和 $(11\bar{2}0)$ 纤锌矿的表面能。

我们将结果与参考值密度泛函理论计算结果以及文献值进行比较。对于涉及晶体表面的计算，我们使用闪锌矿的 (110)晶面和纤锌矿硫化锌的$(11\bar{2}0)$ 晶面，因为研究发现它们在能量上最为有利。[52,62] 板状结构始终由使用BAND通过 PBEsol/DZP 理论水平计算得到的预优化晶格参数构建。

5.4.1 训练集误差

我们首先通过将每个力场的总体损失值分解为各个独立贡献来开展分析。图 7 和图 8 比较了各训练集项中力场的预测值与参考值。

图 7 展示了可归为沿某一坐标进行势能面扫描的训练集项。除 H_2S 角度扫描外，原始力场对这些项的表现都极差，而在所有情况下，优化后的力场都产生了显著改善的结果。O53 力场在所有扫描中的准确度最高，但 M20 通常也很有竞争力，误差与其相近。唯一的例外是 H_2S 分子中的 H-S键扫描，M20 未能很好地复现该扫描。

图8 展示了其余的训练集项，其中我们将参考值与所选力场得出的预测值进行了比较。尽管键角和力得到了改进，但在大多数情况下仍存在显著误差，并且原始力误差的很大一部分仍保留在最终力场中。能量构成总体损失值的很大一部分，但实际上总体上看起来拟合得很好，大部分误差都来自单个训练集项；该项是畸变的和规则的(110)闪锌矿表面之间的能量差。我们还可以看到，在重新参数化过程中，电荷实际上并未得到改善。这是因为在我们的参数选择中，没有启用与电荷相关的参数。原子间距离是改进最明显的一类训练集项。这可以解释为，原本与硫相关的参数实际上是针对氧进行训练的，因此初始力场中的键长过短，无法容纳更大的硫原子，并且

不得不加长。

在大多数情况下，M10 的误差明显大于 M20。另一方面，M20 的误差通常与 O53 非常接近（而在键角方面则优于 O53）。与 O53 相比，M20 在力和能量方面的表现最差。不过，如上所述，能量误差主要集中在一个项目上，而且似乎没有任何力场能够很好地预测力。从这些观察结果来看，没有明显迹象表明 O53 是显著优于 M20 的力场。

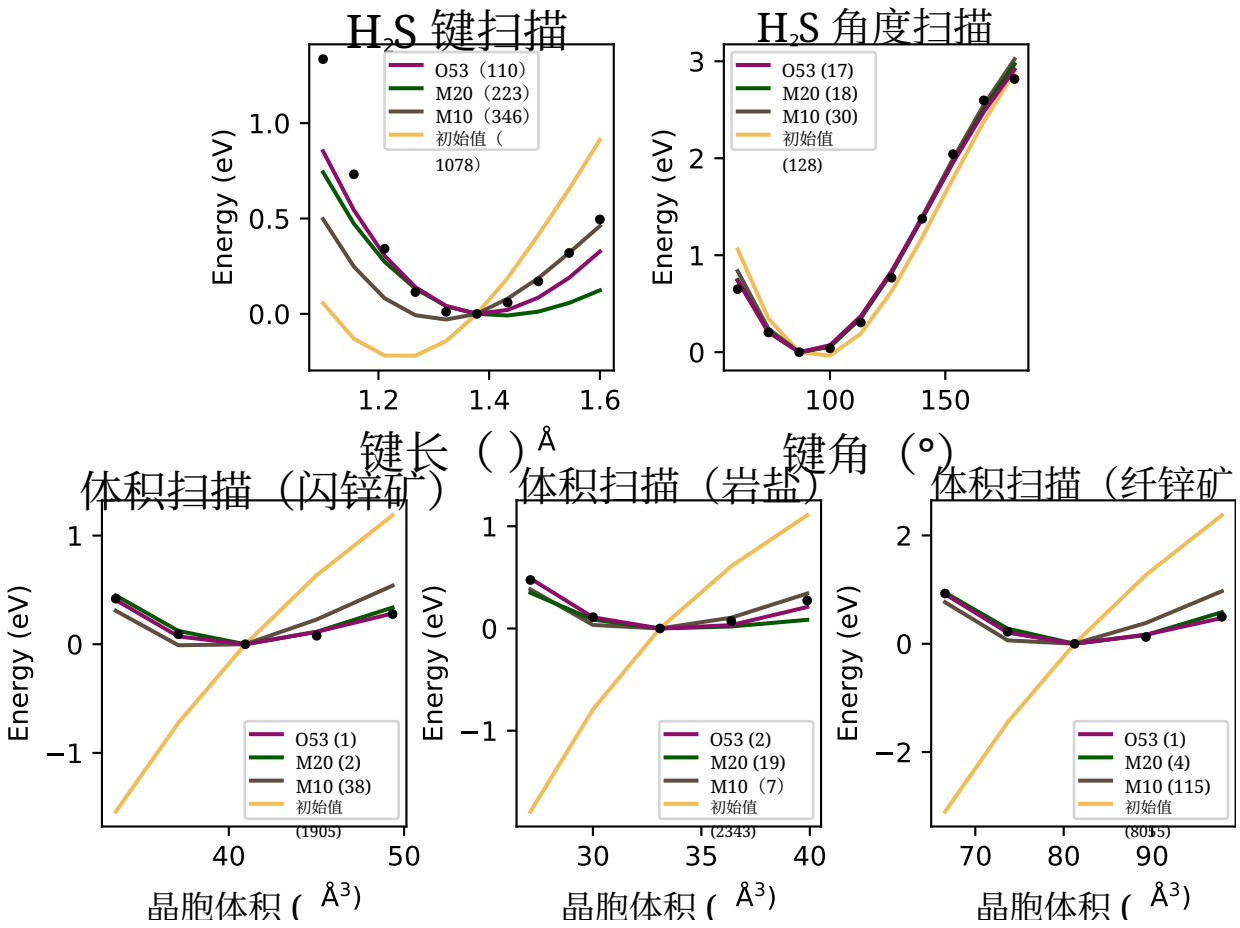


图 7：构成各种势能面扫描一部分的训练集项。参考值以黑点表示，各力场的损失值在图例中以括号标出。

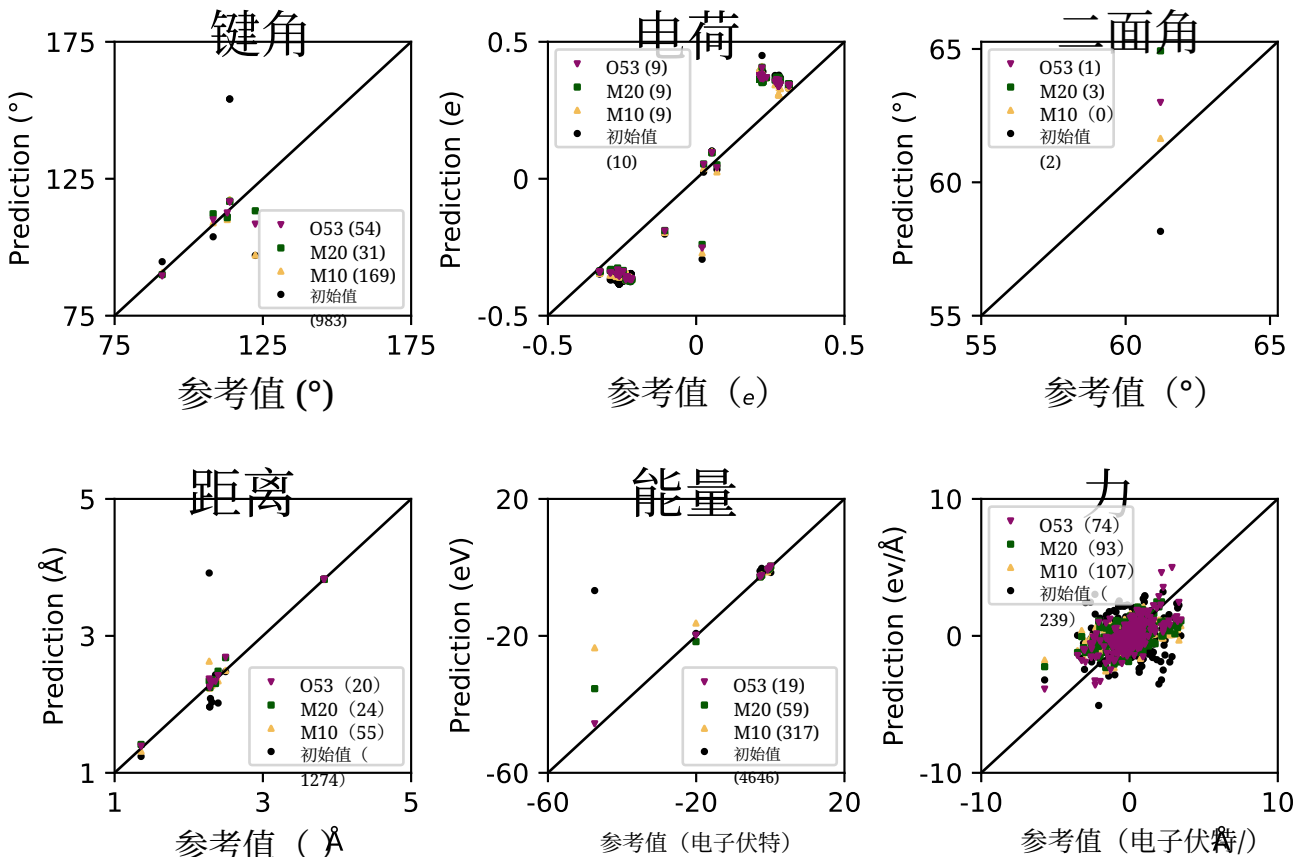


图8：按项目类型分组比较各训练集项和力场的参考值与预测值。图例中括号内列出了某组及力场的损失值。

5.4.2 分子动力学模拟

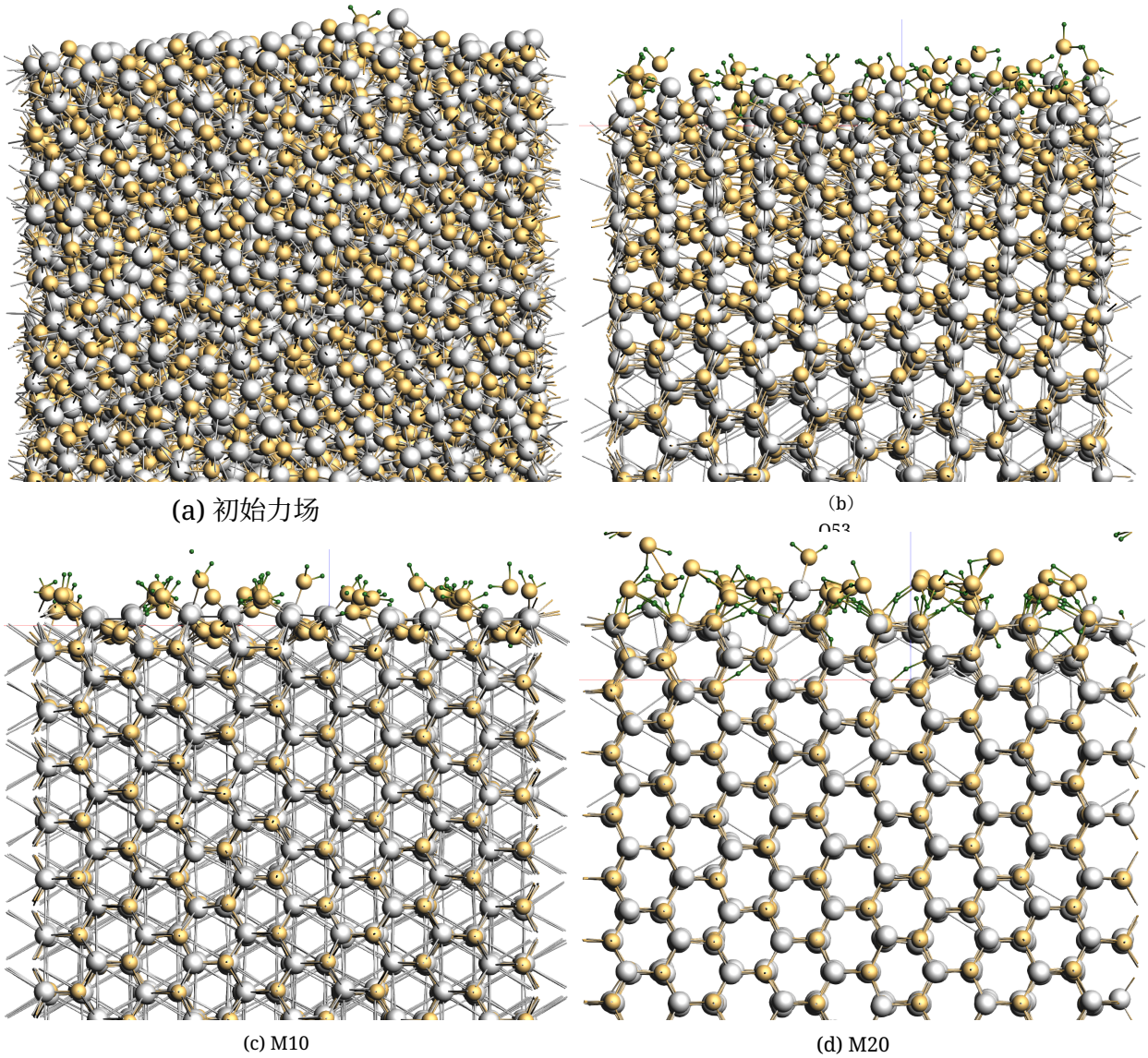
晶体表面吸附我们的第一项验证测试是对 H_2S分子在硫化锌板状结构上的吸附进行分子动力学模拟。我们参考文献来构建一个真实的模拟场景。Zhang等人[62] 通过计算和实验表明，粒径小于 ~7nm的立方硫化锌在室温下不稳定，很容易转变为六方多晶型。在实验中，Dloczik 等人[51] 成功制备了壁厚在10nm到 30nm之间的稳定硫化锌柱，其中检测到了立方相和纤锌矿相。Qi等人[55] 随后研究了 H_2S 在室温下硫化锌表面上的吸附。

利用这些来源，我们针对两种形貌创建了厚度为 ~10nm 的分子动力学模拟板状结构。温度维持在298K，100个 H_2S 分子随机分布在板状结构的上方和下方，模拟进行 300 000 个时间步，每步为 0.25 飞秒。

图 9 和图 10 分别展示了纤锌矿和闪锌矿模拟的最终表面几何构型。不出所料，初始力场表现不佳，因为板状结构立即发生塌陷。然而，M10 和 M20 模拟的表现更好。板状结构保持了其晶体形状，并将 H_2S 分子吸附到板状结构表面。

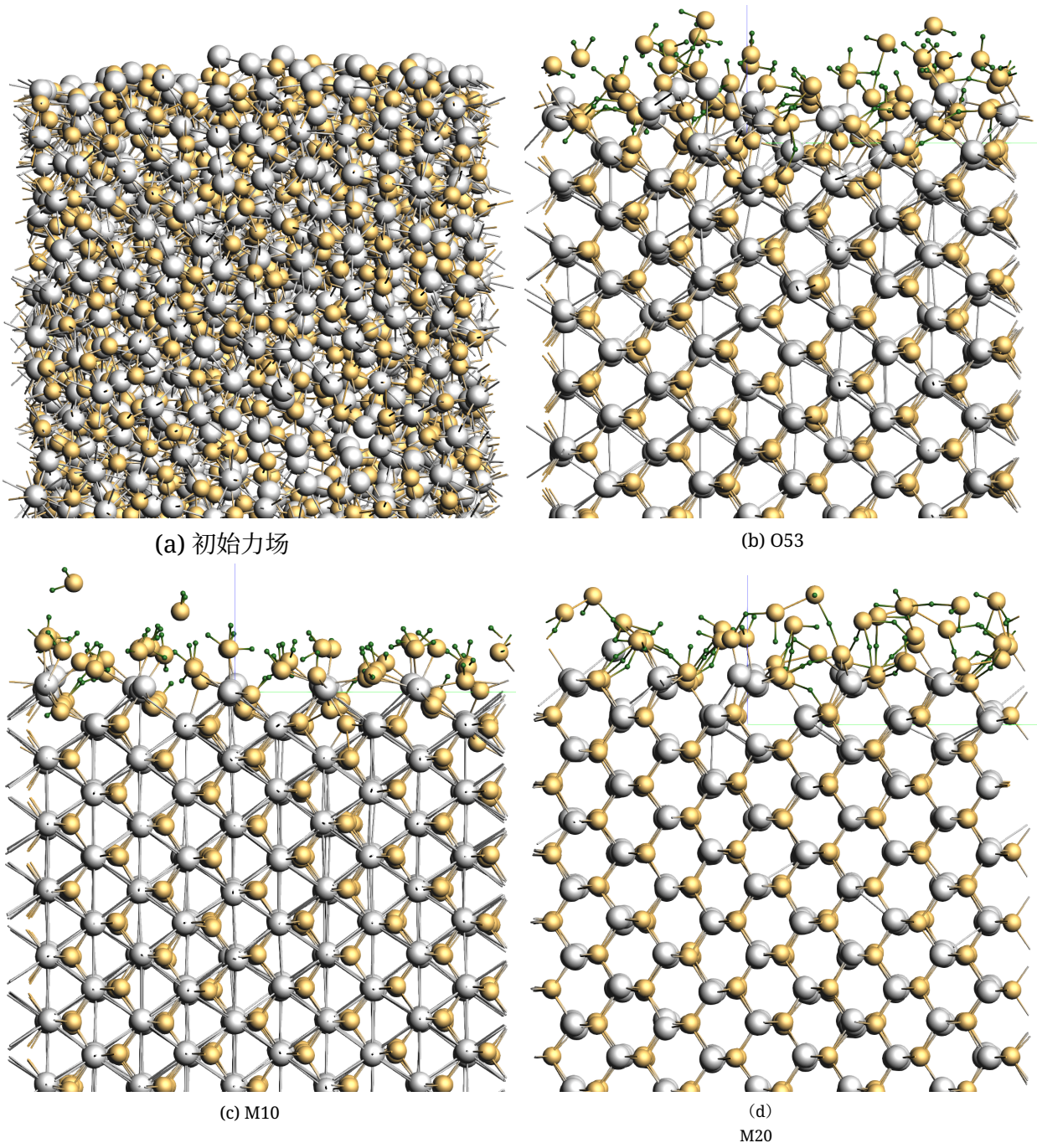
O53 模拟的表现并不相同。这些板状结构难以保持其表面。在纤锌矿情况下，表层几乎已经变为非晶态，尽管体相晶体保持得相对较好，但形变程度延伸至数层深处。闪锌矿表面更加有序，但其保持程度不如 M10 或 M20 那样稳固，而且似乎有一些 H_2S 分子被化学吸收到结构中。

如上文所述，文献表明，在这些条件下，这些板状结构应当是稳定的刚性固体，因此 O53 力场的表现令人意外。我们将在第 5.4.3 节讨论这种表现的原因。



(a) 初始力场 (b) O53 (c) M10 (d) M20

图 9：使用 a) 初始值、b) O53、c) M10 和 d) M20 力场时，100 个 H_2S 气体分子位于具有（$11\bar{2}0$）表面的 10 nm 厚纤锌矿晶片上方和下方，经过 75皮秒分子动力学模拟后最终帧中的一个表面。硫原子为黄色，锌为银色，氢为绿色。



(a) 初始力场

(b) O53

(c) M10

(d) M20

图10：在使用a)初始值、b) O53、c) M10 和d)M20力场，对闪锌矿晶体薄板上下方的100个 H_2S气体分子进行75皮秒分子动力学模拟的最终帧中显示的一个表面。硫原子为黄色，锌为银色，氢为绿色。

体相晶体 上述模拟定性地表明，O53力场较低的损失值并不能保证更好的性能。然而，我们希望定量地证明，M20 参数化能够产生更稳健的结果。

为此，我们对一个块体闪锌矿晶体（共 64 个原子）进行了第二次分子动力学模拟，在 500 K 下运行 10 ps，时间步长为 0.5 飞秒。我们进行了两次模拟：一次使用十六个 M20 优化器各自产生的最佳力场，另一次使用十六个 O53 优化器各自产生的最佳力场。在每次模拟中，力场均作为一个委员会使用。换言之，在每个时间步，将每个力场应用于相同的晶体几何构型。随后，对这些结果中的力取平均值，以更新下一时间步的原子位置。

图11显示了16种力场中力的标准差。我们并不关注特定时间、特定原子或特定方向上的结果，因为力的平均值始终为零。因此，直方图只显示所有标准差，不区分上述因素。16种 M20 力场的一致性明显高于 O53 力场。这表明，较低维度的优化稳健得多，生成的力场结果分布也更集中。较高维度的优化存在更多极小值，优化器可能会陷入其中。由于这些额外的极小值是由不敏感参数的存在造成的，因此它们更可能代表过拟合，而非由重要参数驱动的真正吸引子。有趣的是，尽管 M20 的16种力场之间的损失值分散程度为 98，而 O53 为33，但不能仅凭损失值的分散程度推断出这里所见的 M20 结果分布较为理想。

5.4.3 吸附和表面能量

我们更具定量性的验证测试如下：

1. 闪锌矿 (110)表面上 H_2S 分子的吸附能 (E_{ads}) 计算；以及

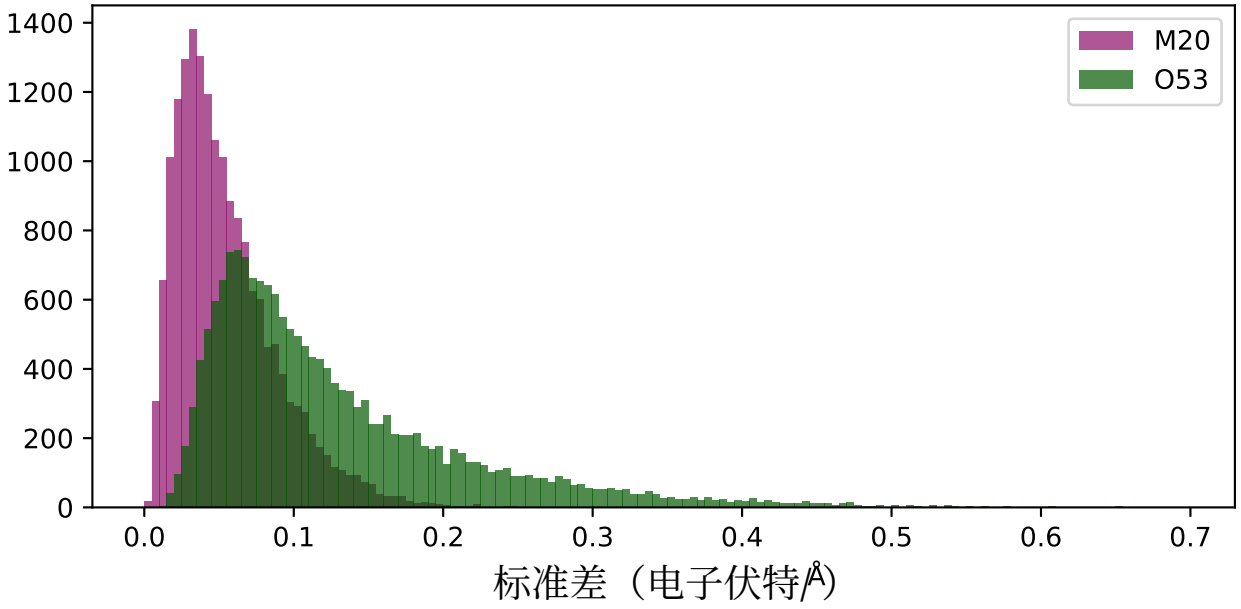


图11：使用十六个 M20和 O53力场进行委员会模拟时，闪锌矿晶体体相模拟中所有时间步、所有方向上所有原子所受力的标准差。

2. 闪锌矿 (110)表面和纤锌矿 $(11\bar{2}0)$ 表面的表面能 (E_{surf}) 计算。

这些计算的详细信息分别包含在补充信息的 C节和 D节中。

在本例中，我们使用O53 和 M20 集合中每个优化器得到的最佳力场来分析结果的分布。由于表现不佳，我们不考虑全部十六个M10结果。结果以力场相关 SSE 为函数展示于图12中。各个优化器的详细结果见补充信息中的 E节。在这些图中，我们还纳入了 DFT、初始力场和最佳M10力场的结果，以便进行比较。初始值和 M10结果未显示其对应的损失值，因为这些值过大，会使图表难以阅读。在接下来的讨论中，我们将继续使用“最佳”来指代损失值（SSE）最低的力场；这并不意味着该力场具有最准确的验证结果。

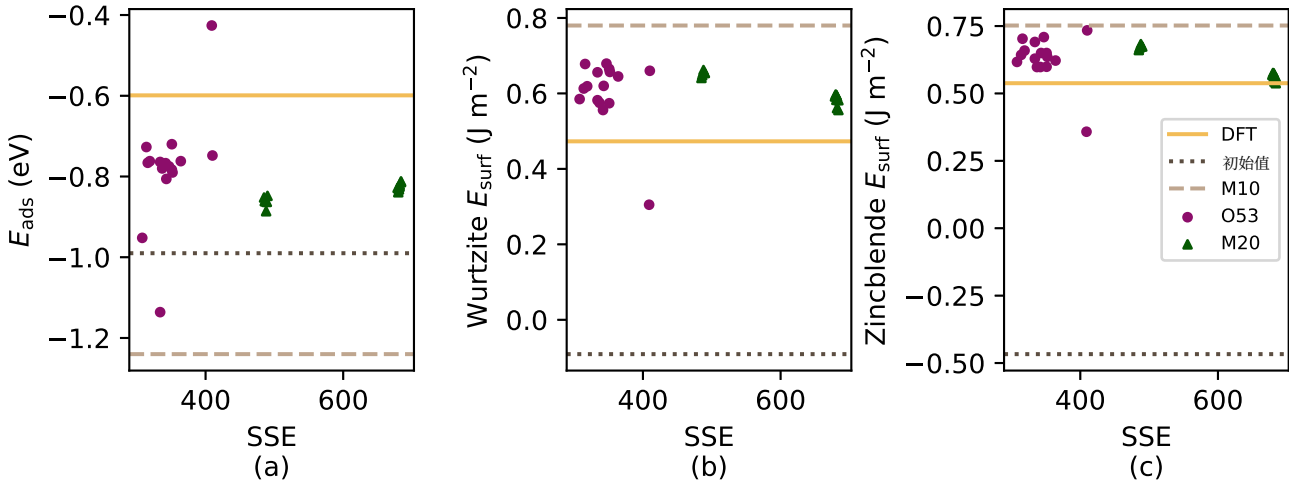
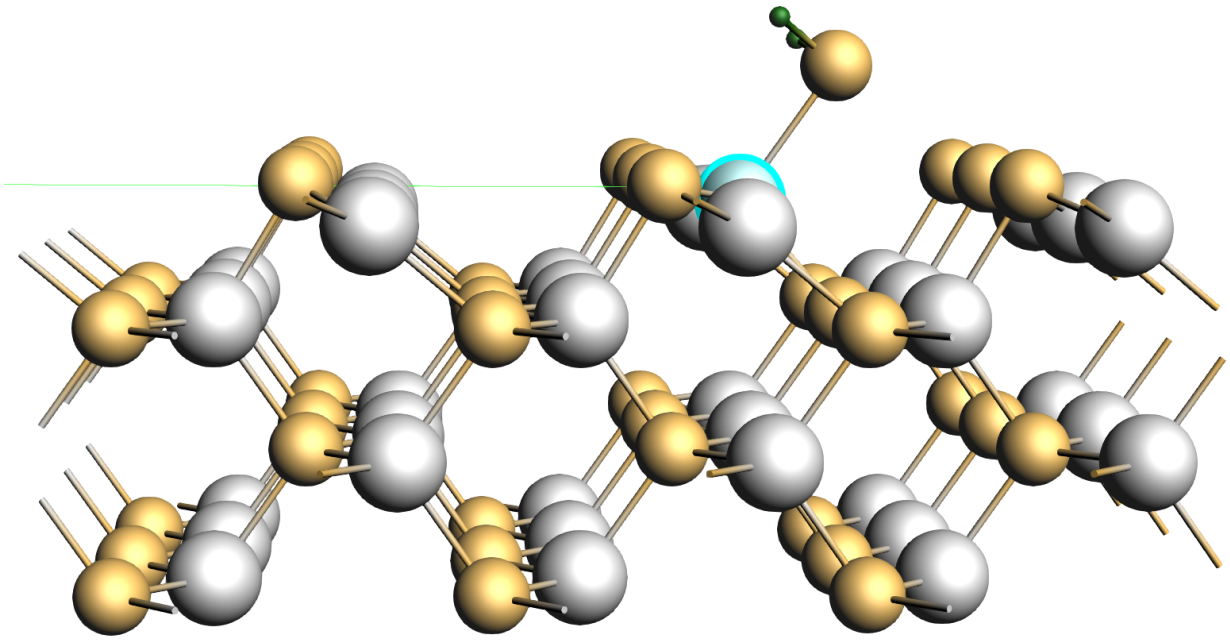


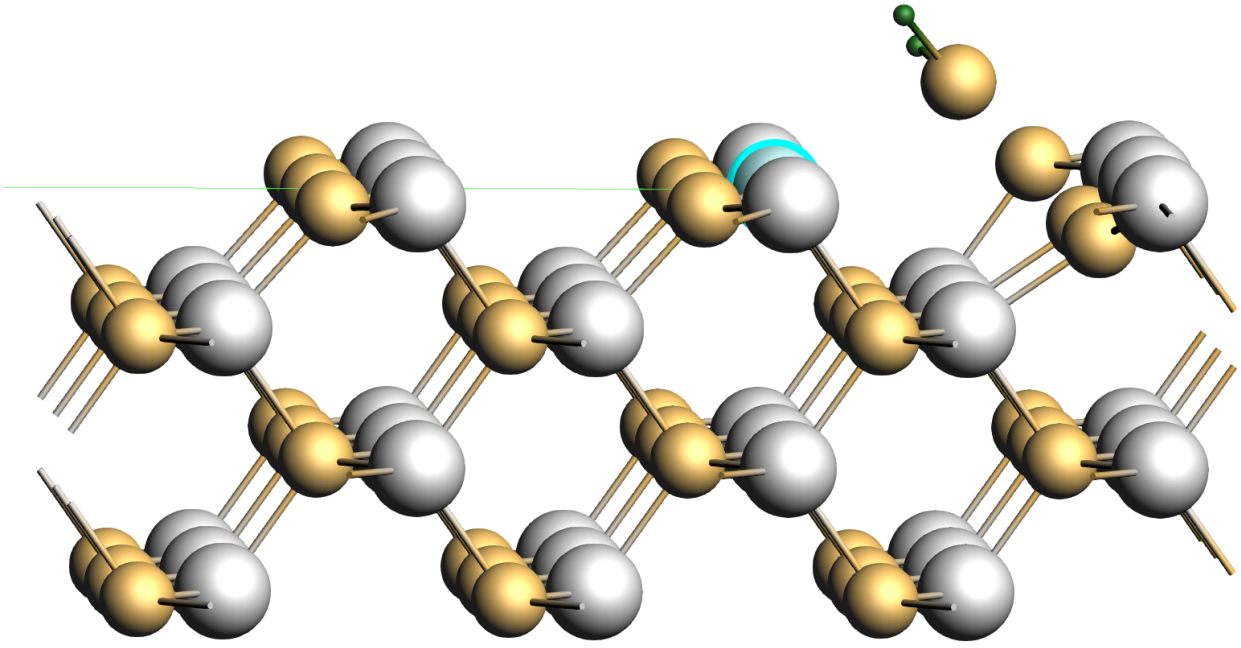
图12：对于在O53和M20重新参数化中运行的十六个优化器各自找到的最佳力场，a）H_2S吸附能，b）纤锌矿（$11\bar{2}0$）表面能，以及c）闪锌矿（110）表面能随SSE的变化。参考值密度泛函理论结果以水平黄色线显示。

吸附能 我们首先从吸附能结果开始评论。与我们的密度泛函理论参考值相比，所有力场在这方面都不太准确，不过最佳的M20 力场比O53 的结果更准确。有趣的是，初始力场的预测优于 M10，且仅略逊于最佳的O53 力场。然而，更广泛地说，所有 M20力场预测的吸附能都处于从 -0.89电子伏特到 -0.81电子伏特的狭窄范围内，这比 O53的结果更具可重复性，后者的范围为从 -1.14电子伏特到 -0.43电子伏特。

这种变异性可能是由若干次 O53 几何优化中出现的异常行为所致。图 13 展示了使用密度泛函理论和最佳 O53 力场进行吸附分子几何优化时的最终帧。H_2S 分子最初被放置在蓝色突出显示的锌原子附近，并使用密度泛函理论稳定在该位置。然而，使用 O53 力场时，H_2S 分子远离该原子，移向一排硫原子；反过来，这些硫原子也受到它的排斥。几何优化已收敛，这意味着这个带有变形表面的脱附位置在能量上比位于锌上的吸附位置更有利——但这是不正确的。第 5.4.2 节中观察到的奇怪表面效应可能与此有关。



(a)



(b)

图 13：使用 a) 密度泛函理论和 b) 最佳 O53 力场对吸附在 (110)闪锌矿表面上的 H_2S 分子进行几何优化的最终帧。该分子最初位于蓝色锌原子附近。

表面能 如果考察表面能结果，O53力场的分布始终大于 M20 结果的分布。平均而言，对于纤锌矿表面，O53 结果与M20 一样准确。对于闪锌矿表面，其中一个 M20 簇实际上相当准确。

总体而言，没有任何一种力场在所有验证测试中都表现良好，但它们都优于初始力场。如果希望获得更高的准确性，则需要更加关注训练集的组成；不过，这并非本文的重点。

与损失值的相关性 这些结果中最有趣的结论来自将其与相关力场的SSE进行比较。全部十六个 O53 优化器得到的 SSE值都紧密聚集在 308 到410 之间，比聚集在485到684 之间的 M20 优化器的 SSE 值低得多，且变化幅度也小得多。如果只考虑损失值，那么可能会预期 O53 力场所作出的预测会比M20 的预测精确得多；然而，我们发现事实并非如此。O53验证结果的变化幅度明显大于 M20 的验证结果，而且平均来看，似乎只略微更准确。

最重要的是，O53 验证结果的高度波动与 SSE 并不相关。换言之，对训练集项的预测更好，并不意味着对验证测试的预测更准确。我们认为，这些现象以及第 5.4.2 节中 O53 分子动力学模拟的异常行为，都可以归因于过拟合。有趣的是，M20 结果似乎与损失值呈反向相关，也就是说，损失越高，验证结果越准确。这可能表明 M20 也出现了一些过拟合现象；不过，由于所有优化器的 SSE 值都集中在仅有的两个数值附近，因此没有足够的数据来验证这一点。

过拟合是 ReaxFF 重新参数化期间的常见现象[9,13,50]，并且允许更改的参数越多，发生过拟合的可能性就越大。这些验证测试突出了

重新参数化时激活过多维度的风险在于，自由度越多，过拟合的可能性就越大。我们介绍使用过多参数时会出现的两种风险。

首先，必须强调参数对某个损失函数（该函数基于用户生成的训练集）的敏感性 与其在势能函数中的实际重要性 之间的区别。如果训练集中的条目不足以体现某个参数的重要性，那么该参数就会表现为不敏感。如果在重新参数化过程中启用了一个重要但不敏感的参数，优化器就会近乎随机地改变它。这可能导致得到损失值很低、但在实际运行中表现很差的力场；敏感参数已被正确设置，但其他重要参数却被错误地更改了。

第二个隐患是参数可能产生补偿效应。如果启用了许多参数，那么可能存在多组参数都能满足训练数据，但在模拟过程中势能表现不佳。使用过多参数会使优化器有机会通过调整不敏感参数来降低损失，而不是为少数合适的参数找出“正确”的值。

在我们的示例中，53 维的重新参数化似乎激活了过多的维度。尽管它们产生的总体损失值最低，但并未使每类项目的损失都达到最低，在验证测试中也并非总能做出最准确或最精确的预测。在分子动力学模拟中，我们观察到了一些异常行为；结合前面讨论的精度问题，我们认为这些是过拟合的证据。

相反，仅使用 10 个最敏感参数的重新参数化似乎没有足够的自由度。与 O53 和 M20 相比，这些力场产生了显著的训练集误差，并且在验证测试中表现不佳。

使用 20 个最敏感参数生成的力场质量良好，在某些测试中的表现优于 O53 力场，尽管训练集损失略差。验证测试的结果也始终比

O53 的结果更精确，且准确度相当。分子动力学模拟符合预期，而且生成这些力场所用的时间短于原始优化所需的时间。

6 展望

我们认为，有几个领域值得进一步研究。首先，此处使用的训练集可以作为起点，用于生成质量更高的 Zn/S/H 力场。这需要大幅扩展训练集，并涉及与训练集设计相关的所有众所周知且棘手的问题。[50]

其次，参数的敏感性很大程度上取决于允许其变化的范围。在本工作中，我们无需对此进行过多考虑，因为ParAMS软件包提供了建议范围。然而，无法确定这些范围是否始终适用。对适当范围进行形式化，或建立一种稳健的机制来确定这些范围，将大大有助于确保敏感性得到适当的确定。

第三，我们确定采样是敏感性分析流程中的限制步骤。研究如何从优化器轨迹中提取样本可能会很有意义。然而，由于HSIC 需要均匀分布的随机样本，因此需要谨慎地从轨迹中提取样本。一种可能的方法是Kennard-Stone算法，[63]但这种方法速度较慢且是串行的。尽管如此，“闭环”并允许用户从优化结果中提取敏感性似乎是一个很有吸引力的前景。

7 结论

作为一项介绍性工作，我们已证明，将HSIC 敏感性分析应用于 ReaxFF重新参数化能够成功识别最敏感参数。我们进一步表明，在优化过程中仅使用最敏感参数可以实现更快的收敛，并降低过拟合的可能性。即使仅进行定性分析，采用此类敏感度分析也能为用户深入了解.....的组成提供有价值的见解。

训练集。总体而言，我们认为，HSIC 敏感性分析是一种有用、稳健且易于使用的工具，有望大力助益于 ReaxFF 力场的重新参数化。

作者贡献

MG 设计并实施了敏感性测试，开展了测试工作，并撰写了本文。MH 设计了训练集，并指导了部分验证测试。TV 提供了监督、修改意见、指导和建议。

致谢

作者感谢弗兰德斯超级计算中心（VSC），该中心由根特大学、FWO 和弗兰德斯政府资助，使作者得以使用其计算资源。

本项目获得欧盟地平线 2020 研究与创新计划资助，资助协议编号为 814143。TV 和 MG 还获得根特大学研究委员会（ BOF ）的资助，资助编号为 01IT2322。

可获取的补充信息

以下内容包含在补充信息中：

si.pdf 完整优化结果、吸附和表面能的计算方法、完整的吸附和表面能结果、训练集组成；job_collection.yaml 中的训练集几何结构和作业计算设置；engine_collection.yaml 中针对输入几何结构的任务设置；training_set.yaml 中训练数据的参考值、权重和 si gma 值；

initial_parameters.yaml 中原始 ReaxFF 参数值及其范围（将 O53 参数设为活动参数）。

YAML 文件便于人类阅读，并由 ParAMS 使用。[14]此信息可通过互联网免费获取，网址为 http://pubs.acs.org。

参考文献

(1) van Duin, A. C. T.; Dasgupta, S.; Lorant, F.; Goddard, W. A. ReaxFF：用于烃的反应力场。《物理化学杂志 *A*》**2001**，*105*，9396–9409。

(2)Senftle, T. P.; Hong, S.; Islam, M. M.; Kylasa, S. B.; Zheng, Y.; Shin, Y. K.; J unker- meier,C.; Engel-Herbert, R.; Janik, M.J.; Aktulga, H. M.; Verstraelen, T.; Grama,A.; vanDuin,A. C. T. ReaxFF反应力场：发展、应用与未来方向。*npj*《计算材料》**2016**，*2*，15011。

(3) Komissarov, L.; Rüger, R.; Hellström, M.; Verstraelen, T. ParAMS：用于原子与分子模拟的参数优化。《化学信息与建模杂志》**2021**, *61*,3737–3743.

(4) Freitas Gustavo, M.; Verstraelen, T. GloMPO（全局管理并行优化）：一种用于高成本黑箱优化的工具及其在 ReaxFF 重新参数化中的应用。《化学信息学杂志》**2022**, *14*, 7.

(5) Chenoweth, K.; van Duin, A. C.; Goddard, W. A. 用于烃氧化分子动力学模拟的ReaxFF反应力场。《物理化学杂志 *A*》**2008**, *112*, 1040–1053.

(6) Barcaro, G.; Monti, S.; Sementa, L.; Carravetta, V. 用于硅纳米颗粒分子动力学模拟的反应力场（ReaxFF）参数化。《化学理论与计算杂志》**2017**, *13*, 3854–3861。

(7) Bae, G.-T.; Aikens, C. M. Au–S–C–H体系的改进 ReaxFF 力场参数。《物理化学杂志 *A*》**2013**，*117*，10438–10446。

(8) Liu, Y.; Hu, J.; Hou, H.; Wang, B. 用于全氟酮热分解分子动力学研究的 ReaxFF反应力场开发与应用。《化学物理》**2020**, *538*, 110888.

(9) Hubin, P. O.; Jacquemin, D.; Leherte, L.; Vercauteren, D. P. 脯氨酸催化的羟醛反应 ReaxFF反应力场参数化。《计算化学杂志》**2016**, *37*, 2564–2572.

(10) Hu, X.; Schuster, J.; Schulz, S. E. 用于模拟铜原子层沉积的 ReaxFF反应力场多参数并行优化。《物理化学杂志 *C*》**2017**, *121*, 28077–28089.

(11) LaBrosse, M. R.; Johnson, J.K.; van Duin, A. C. 钴可转移反应力场的开发。《物理化学杂志 *A*》**2010**, *114*, 5855–5861.

(12) Larsson, H. R.; van Duin, A. C.; Hartke, B. SiOH 反应力场 ReaxFF 参数的全局优化。《计算化学杂志》**2013**，*34*，2178–2189。

(13) Shchygol, G.; Yakovlev, A.; Trnka, T.;van Duin, A.C.;Verstraelen,T. ReaxFF 参数优化与蒙特卡罗和进化算法：指南与见解。《化学理论与计算杂志》**2019**，*15*，6799–6812。

(14) Software for Chemistry and Materials，AMS 的参数化工具 (2022.204.r105828)。2022；https://www.scm.com/doc/params [访问日期：6 月2022]。

(15) Saltelli, A.; Aleksankina, K.; Becker, W.; Fennell, P.; Ferretti, F.; Holst, N.; Li, S.;

Wu, Q. 为什么如此多已发表的敏感性分析是错误的：敏感度分析实践的系统综述。环境建模与软件 **2019**，*114*，29–39。

(16) Baroni, G.; Francke, T. 一种结合基于方差和基于分布的全局敏感性分析的有效策略。环境建模与软件 **2020**，*134*，104851。

(17) Spagnol, A.; Le Riche, R.; Da Veiga, S. 用于优化和变量选择的全局敏感性分析。*SIAM/ASA* 不确定性量化期刊 **2019**，*7*，417–443。

(18) Borgonovo, E. 一种新的不确定性重要性度量。可靠性工程与系统安全 **2007**，*92*，771–784。

(19) Saltelli, A.; Ratto, M.; Andres, T.; Campolongo, F.; Cariboni, J.; Gatelli, D.; Saisana, M.; Tarantola, S. 全局敏感性分析：入门；John Wiley & Sons, Ltd，2008。

(20) Herman, J. D.; Kollat, J. B.; Reed, P. M.; Wagener, T. 技术简报：Morris 方法可有效降低分布式流域模型进行全局敏感性分析的计算需求。水文与地球系统科学 **2013**，*17*，2893–2903。

(21)Nossent, J.; Elsen, P.; Bauwens, W. 复杂环境模型的Sobol' 敏感度分析。 环境建模与软件 **2011**，*26*，1515–1525。

(22) Campolongo, F.; Saltelli, A.; Cariboni, J. 从筛选到定量敏感度分析：一种统一方法。计算机物理通讯**2011**，*182*，978–988。

(23) De Lozzo, M.; Marrel, A. 在敏感度分析和筛选中使用依赖关系度量的新改进。统计计算与模拟杂志**2016**, *86*,3038–3058.

(24) Da Veiga,S. 基于依赖关系度量的全局敏感度分析。统计计算与模拟杂志 **2015**, *85*, 1283–1305.

(25) Gretton, A.; Bousquet, O.; Smola, A.; Schölkopf, B. 载于算法学习理论；Jain, S., Simon, H. U., Tomita, E. 编；Springer Berlin Heidelberg，2005；第63–77页。

(26) Yamada, M.; Tang, J.; Lugo-Martinez, J.; Hodzic, E.; Shrestha, R.; Saha, A.; Ouyang, H.; Yin, D.; Mamitsuka, H.; Sahinalp, C.; Radivojac, P.; Menczer, F.; Chang, Y. 面向大型生物数据的超高维非线性特征选择。*IEEE*知识与数据工程汇刊 **2018**, *30*, 1352–1365.

(27) Yamada, M.;Jitkrittum, W.; Sigal,L.;Xing,E. P.; Sugiyama,M.基于逐特征核化套索的高维特征选择。神经计算 **2014**，*26*，185– 207。

(28) Song,L.; Smola, A.;Gretton,A.;Bedo, J.; Borgwardt, K. 通过依赖性最大化进行特征选择。《机器学习研究杂志》 **2012**，*13*，1393–1434。

(29) Climente-González, H.; Azencott, C.-A.; Kaski, S.; Yamada, M. 分块HSIC套索：用于超高维数据的无模型生物标志物检测。生物信息学 **2019**,*35*, i427–i435.

(30) Feng, L.; Di, T.; Zhang, Y. 用于故障监测的基于HSIC的核独立成分分析。化学计量学与智能实验室系统 **2018**, *178*, 47–55。

(31) Qian, J.; Lu, M.; Tian, F.; Liu, R. 医疗电子鼻用于伤口感染检测的传感器阵列优化研究。*IEEE*电路与系统汇刊第二辑：简报 **2022**，*69*，1867–1871。

(32) Li, Y.; Pogodin, R.; Sutherland, D. J.; Gretton, A. 收录于神经信息处理系统进展；Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P. S., Vaughan, J. W, 编；Curran Associates, Inc., 2021; 第34卷；第15543–15556页。

(33) Wang, T.; Li, W. 基于希尔伯特–施密特独立性准则的核学习与优化。国际机器学习与控制论杂志**2018**，*9*，1707–1717。

(34) Kaymak, M. C.; Rahnamoun, A.; O'Hearn, K. A.; van Duin, A. C. T.; Merz, K. M. J.; Aktulga, H. M. JAX–ReaxFF：一种基于梯度的框架，用于快速优化反应力场。《化学理论与计算杂志》**2022**，*18*，5181– 5194。

(35) Pahari, P.; Chaturvedi, S. 使用遗传算法确定反应力场的最佳拟合势能参数。《分子建模杂志》 **2012**，*18*，1049– 1061。

(36) Trnka, T.; Tvaroška, I.; Koča, J. 面向酶促反应能量学的 ReaxFF反应力场自动训练。《化学理论与计算杂志》**2018**，*14*，291–302。

(37) Jaramillo-Botero, A.; Naserifar, S.; Goddard, W. A. 常规项多目标力场优化框架及其在碳化硅反应力场中的应用。《化学理论与计算杂志》**2014**，*10*，1426–1439。

(38) Boothroyd, S.; Wang, L.-P.; Mobley, D. L.;Chodera, J.D.; Shirts,M.R.开放式力场评估器：一个用于从分子模拟中估算物理性质的自动化、高效且可扩展的框架。《化学理论与计算杂志》**2022**，*18*，3566–3576。

(39) Örn Jónsson, E.; Rasti, S.; Galynska, M.; Meyer, J.; Jónsson, H. 基于单中心多极展开的适用于柔性 H_2O 分子的可迁移势能函数。《化学理论与计算杂志》**2022**，*18*，7528–7543。

(40) Mondal, A.; Young, J. M.; Barckholtz, T. A.; Kiss, G.; Koziol, L.; Panagiotopoulos, A. Z. 遗传算法驱动的熔融碱金属力场参数化

碳酸盐和氢氧化物盐。《化学理论与计算杂志》 **2020**，*16*，5736–5746。

(41) Verstraelen, T.; Bultinck, P.; van Speybroeck, V.; Ayers, P. W.; van Neck, D.; Waro- quier, M. 电荷平衡模型中参数的重要性。《化学理论与计算杂志》**2011**，*7*，1750–1764。

(42) He, Z.; Xiong, X.; Yang, B.; Li, H. 使用先进的基于代理的非线性模型表示方法对高速列车头型进行气动优化。优化与工程**2022**，*23*，59–84。

(43) Wang, T.; Wang, C.; Xu, Z.; Cui, C.; Wang, X.; Demitrack, Z.; Dai, Z.; Bagtzoglou, A.; Stuber, M. D.; Li, B. 利用非理想非均匀混合模型和高保真传感实现对给水与废水处理系统的精确控制。化学工程杂志**2022**，*430*，132819。

(44) Zhang,T.;Di, X.;Chen,G.; Zhu, L. 单层蓝磷烯的COMPASS力场参数化。纳米技术 **2020**，*31*，145702。

(45) Gopalakrishnan, S.; Dash, S.; Maranas, C. K-FIT：一种使用稳态通量组学数据的加速动力学参数化算法。代谢工程**2020**，*61*，197–205。

(46)Mohammed,I.; Gnedin,N. Y.宇宙剪切层析中的重子效应：主成分分析参数化与极端重子模型的重要性。天体物理学杂志 **2018**，*863*，173。

(47) Quinn, K. N.; Wilber, H.;Townsend,A.;Sethna, J.P.切比雪夫逼近与模型预测的全局几何。物理评论快报 **2019**，*122*，158302。

(48) Mortier, W. J.; Ghosh, S. K.; Shankar, S. 分子中原子电荷的电负性均衡法

计算。《美国化学会志》**1986**，*108*，4315–4320。

(49) Verstraelen, T.; Ayers, P. W.; van Speybroeck, V.; Waroquier, M. ACKS2：原子凝聚的 Kohn–Sham 密度泛函理论二阶近似。《化学物理学杂志》**2013**，*138*，074108。

(50) Müller, J.; Hartke, B. 基于多参考从头算数据拟合的用于二硫键机械化学的ReaxFF反应力场。《化学理论与计算杂志》**2016**,*12*, 3913–3925.

(51) Dloczik, L.; Engelhardt, R.; Ernst, K.; Fiechter, S.; Sieber, I.; Könenkamp, R. 通过单晶氧化锌柱的化学转化制备硫化锌六边形纳米管。《应用物理快报》**2001**，*78*，3687。

(52) Hamad, S.;Cristol,S.; Catlow, C. R. A. 硫化锌的表面结构和晶体形貌：计算研究。《物理化学杂志 *B*》 **2002**，*106*，11002–11008。

(53) Li,S.;Wei,X.;Zhu, S.;Gui, Y. 二氧化硫、硫化氢吸附于铁–硫化锌表面的第一性原理分析。《传感器与执行器*A*辑：物理》 **2021**，*329*，112827。

(54) Raymand, D.; vanDuin, A. C.; Spangberg, D.; Goddard, W.A.;Hermansson, K. 通过分子动力学模拟研究台阶状氧化锌表面上的水吸附。《表面科学》**2010**，*604*，741–752。

(55)Qi,G.;Zhang,L.;Yuan, Z.氧化锌纳米棒由几纳米厚的硫化锌薄层修饰后的硫化氢气体传感性能改进。《物理化学化学物理》 **2014**,*16*, 13434–13439。

(56) Software for Chemistry and Materials，力场格式规范。2022；https: // tinyurl.com/5aerwd38 [访问日期：6月 2022]。

(57) Islam, M. M.; Ostadhossein, A.; Borodin, O.; Yeates, A. T.; Tipton, W. W.; Hennig, R. G.; Kumar, N.; van Duin, A. C. T. 针对锂化硫正极材料的 ReaxFF 分子动力学模拟。《物理化学化学物理》**2015**，*17*，3383–3393。

(58) Software for Chemistry and Materials，ReaxFF：SCM 开发者文档。2022；https://tinyurl.com/39vkburf [访问日期：6月 2022]。

(59) Verstraelen, T.; Vandenbrande, S.; Heidar-Zadeh, F.; Vanduyfhuys, L.;van Spey- broeck, V.;Waroquier, M.;Ayers, P. W. 最小基组迭代分区法：用于力场开发的分子中的原子。《化学理论与计算杂志》**2016**，*12*，3894–3912。

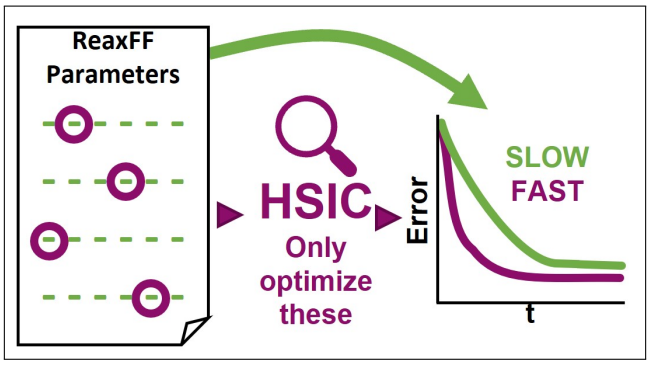
(60) Hansen,N.;Ostermeier, A.进化策略中的完全去随机化自适应。《进化计算》**2001**，*9*，159–195。

(61) Gretton, A.; Borgwardt, K. M.; Rasch, M. J.; Schölkopf, B.; Smola, A. 核双样本检验。《机器学习研究杂志》**2012**，*13*，723–773。

(62) Zhang, H.; Huang, F.; Gilbert,B.; Banfield, J. F. 硫化锌纳米颗粒相稳定性的分子动力学模拟、热力学分析和实验研究。《物理化学杂志 B》《物理化学杂志 *B*》**2003**，*107*，13051–13060。

(63) Kennard, R. W.; Stone, L. A. 计算机辅助实验设计。《技术计量学》1969，11，137–148。

Graphical TOC Entry



arXiv:2304.05046v1 [physics.chem-ph] 11 Apr 2023

补充信息

使用希尔伯特–施密特独立性准则进行 ReaxFF 重新参数化的敏感度分析

迈克尔·弗雷塔斯·古斯塔沃[1,2],ORCID:0000-0002-1832-8413马蒂·赫尔斯特伦¨[2],ORCID:0000-0003-3053-5658图恩 Verstraelen[1,*],
ORCID:0000-0001-9288-5608

[1] 根特大学分子建模中心，比利时根特[2] Software for Chemistry and MaterialsBV，荷兰阿姆斯特丹

[*] 通讯作者： toon.verstraelen@ugent.be

2023年4月12日

目录

A 完整优化结果 1 B 核参数的敏感性结果 2 C 吸附能计算 4 D 表面能计算 5 E 表面能和吸附能的完整结果 6 F 训练集的组成 7

A 完整优化结果

表 S1：最低 SSE l　　每组重新参数化由每个并行优化器找到的最低损失值　　.

优化器	O53	M10	M20	M33	M43	L10	L20	L33	L43
1	319	1871	680	486	380	19 918	18 414	14 588	2650
2	308	1876	680	497	387	19 918	18 431	14 588	2622
3	351	1869	682	419	392	19 918	18 431	14 589	2703
4	334	2428	488	415	327	19 918	18 414	14 588	2607
5	409	1873	490	417	343	19 918	18 413	14 596	2650
6	343	1873	684	479	489	19 918	18 414	14 588	2621
7	352	1879	681	491	355	19 918	18 415	14 589	2637
8	364	1875	682	471	412	19 918	18 414	14 588	2901
9	410	1869	684	478	336	19 918	18 413	14 587	2747
10	334	1195	486	484	351	19 918	18 432	14 642	2597
11	337	1874	679	424	390	19 918	18 430	14 589	2616
12	314	1871	487	530	397	19 918	18 413	14 589	2617
13	351	1875	681	378	389	19 918	18 415	14 591	3097
14	316	1870	489	577	356	19 918	18 413	14 589	2617
15	347	1888	681	376	330	19 918	18 412	14 588	2606
16	342	1194	485	424	388	19 918	18 412	14 588	2616
Min.	308	1194	485	376	327	19 918	18 412	14 587	2597
Max.	410	2428	684	577	489	19 918	18 432	14 642	3097
标准差	33	339	98	62	48	0	8	17	162

B 核参数函数下的敏感性结果

图 S1 展示了使用 CG 核时，形状参数取不同值 γ 所得到的敏感性结果。ReaxFF参数按照所有核函数的平均敏感性值排序，因此不会偏向任何一个核函数。所有核函数都能识别出相似的最敏感参数排序；然而， γ 值过低或过高时，核值的取值范围都不够宽，导致该方法无法准确解析敏感性。对于 $\gamma = 0.1$ ，十次重复得到的结果分散程度很大，中等敏感参数的排序在不同重复之间变化显著。这是因为几乎所有核值都非常接近零，因此难以区分。这使得图中显示的平均值与使用较大的 γ 值计算出的结果大相径庭。 $\gamma = 0.7$也存在同样的问题，但此时核值非常接近 1。例如可以看到，根据其他所有核函数均为中等敏感度的参数 S.Zn:p ovun1，其敏感性为零。这主要是我们将负 HSIC 值截断为零时进行阈值化所造成的假象，但其出现的原因是使用较大的该参数值时，原始 HSIC 值极小。 $\gamma = 0.3$ 和 $\gamma = 0.5$ 的适中取值会产生相似的结果。这表明排序具有稳健性；只要核参数值合理，用户就不必过分纠结于如何选择这些值。

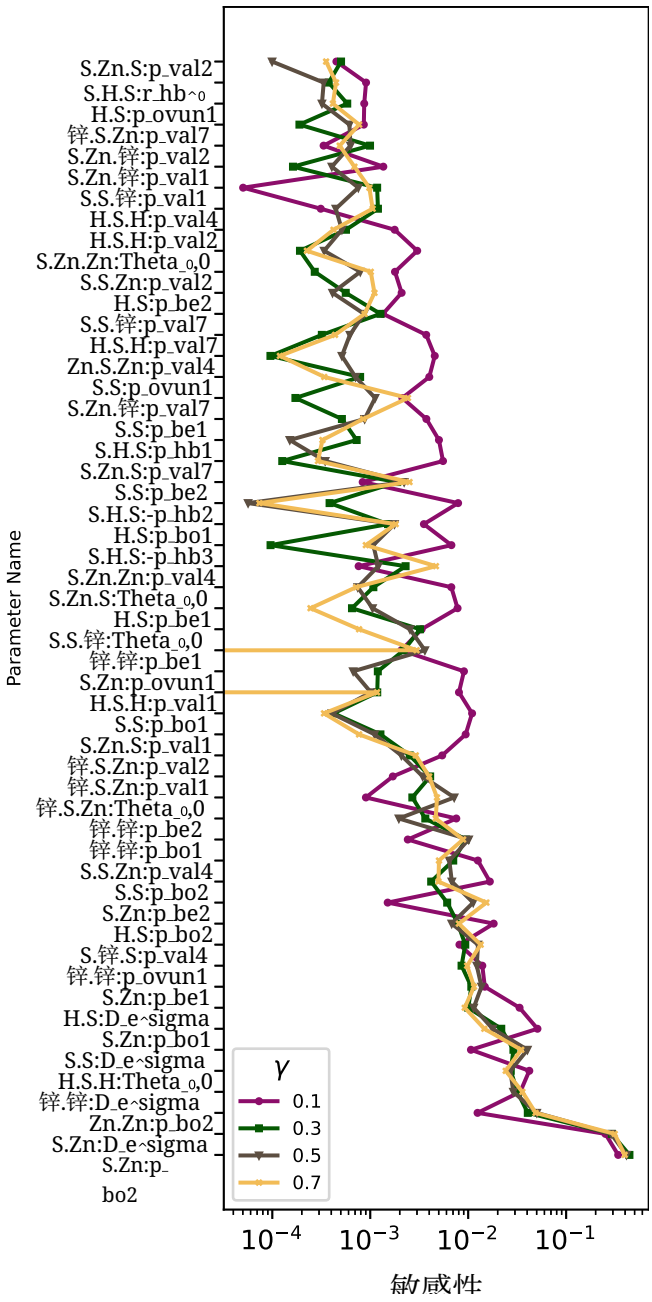


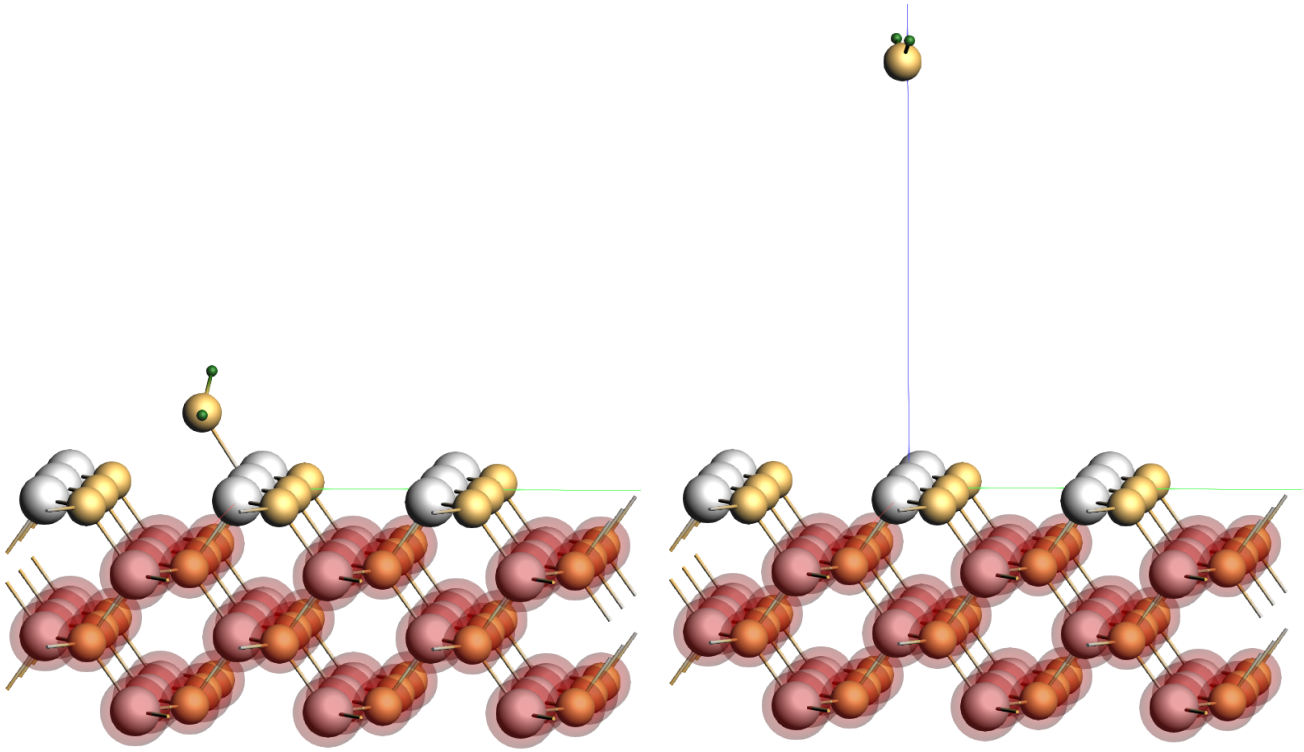
图 S1：使用 CG 核的不同γ 值确定的每个参数的平均敏感度值（每个值基于十次运行）。参数按照跨核函数的平均值排序。

C 吸附能的计算

对图 S2 中的体系进行了几何优化。在这些优化中，红色圈出的原子固定不动。吸附能通过胞内法计算，即两个体系之间形成能的能量差：

$$\Delta E_{\text{ads}} = E_{\text{slab+adsorbed molecule}} - E_{\text{slab+far away molecule}}. \tag{S1}$$

表面采用预先优化的晶格参数构建。



(a) (b)

图 S2：体　　系模型：(a) H_2S 分子位于 (110) 闪锌矿表面上；(b) 位于该表面上方。

D 表面能的计算

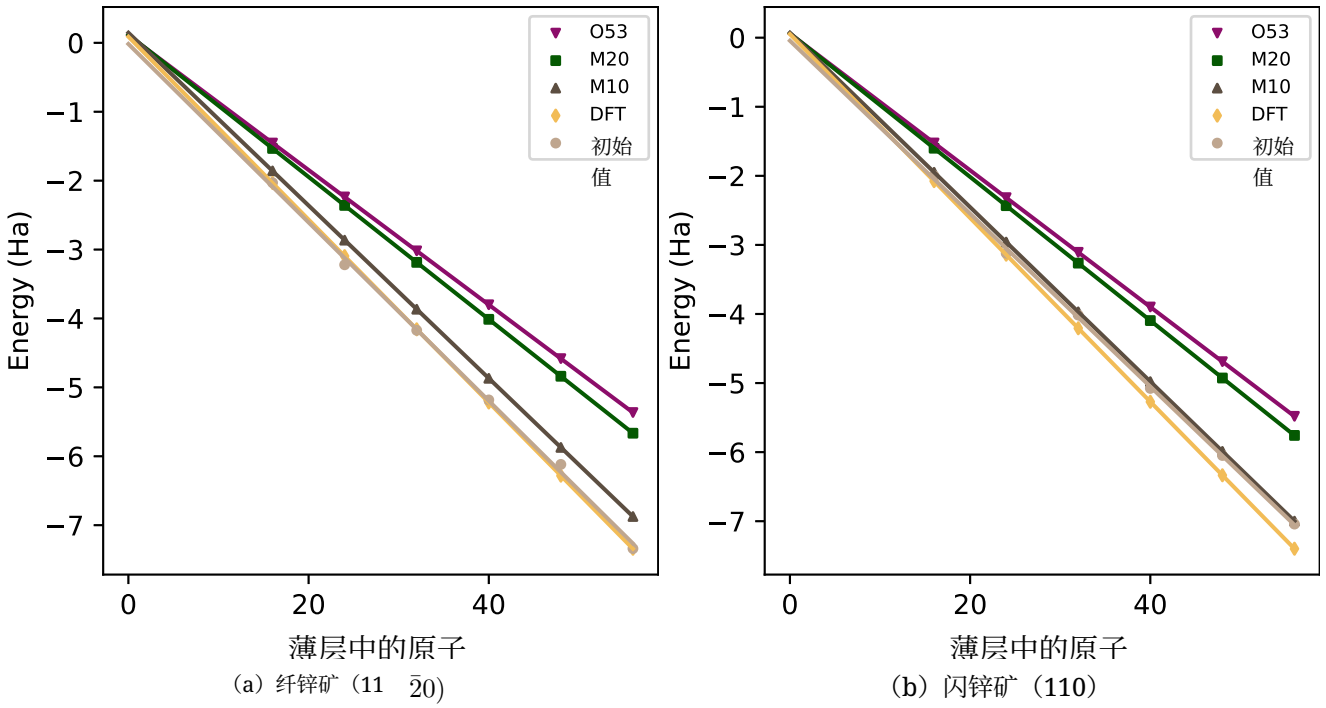
对厚度各异的（110）闪锌矿和（$11\bar{2}0$）纤锌矿板状结构进行了几何优化。板状结构采用预先优化的晶格参数构建。能量根据板状结构中的原子数进行了线性回归（见图 S3）。表面能根据这些回归的截距，使用以下公式计算：

$$E_{\text{surf}} = \frac{c}{2A}, \tag{S2}$$

其中，c为截距，A为薄层表面积，分别按 (110) 闪锌矿和 ($11\bar{2}0$) 纤锌矿计算为 $20.7\ ^{\circ}\text{A}^2$ 和 $40.9\ ^{\circ}\text{A}^2$。表 S2 列出了拟合计算结果。

表 S2：不同厚度板状结构的能量（E）线性回归（$E = mn + c$），用于确定纤锌矿晶体的（$11\bar{2}0$）面和闪锌矿晶体的（110）面表面能，其中 n 是板状结构中的原子数。

	纤锌矿		闪锌矿	
	m	c	m	c
DFT	−0.133	0.089	−0.133	0.051
初始值	−0.129	−0.017	−0.125	−0.044
O53	−0.098	0.110	−0.099	0.059
M20	−0.103	0.120	−0.104	0.063
M10	−0.125	0.146	−0.126	0.071



（a）纤锌矿（$11\bar{2}0$）　（b）闪锌矿（110）

图 S3：通过线性回归计算表面能。

E 完整表面和吸附能结果

表 S3：十六个O53 和M20 优化器各自找到的最佳力场的 SSE、吸附能和表面能结果。损失值最低的行/优化器以粗体显示。

	O53				M20			
			E_{surf} (J m^{-2})				E_{surf} (J m^{-2})	
Opt.*	SSE†	E_{ads} (电子伏特)	W. (11$\bar{2}$0) ‡	闪锌矿（110）¶	SSE†	E_{ads} (电子伏特)	W. (11$\bar{2}$0) ‡	闪锌矿（110）¶
1	319	-0.762	0.619	0.659	680	-0.839	0.597	0.576
2	**308**	**-0.952**	**0.585**	**0.617**	680	0.593	0.571	-0.828
3	351	-0.783	0.665	0.650	682	-0.820	0.561	0.544
4	334	-0.764	0.656	0.691	488	-0.886	0.662	0.683
5	409	-0.426	0.305	0.358	490	-0.848	0.654	0.674
6	343	-0.806	0.620	0.650	684	-0.812	0.556	0.538
7	352	-0.790	0.657	0.636	681	-0.832	0.596	0.575
8	364	-0.762	0.645	0.622	682	-0.824	0.591	0.568
9	410	-0.748	0.660	0.734	684	-0.814	0.582	0.560
10	334	-1.136	0.582	0.629	486	-0.861	0.648	0.670
11	337	-0.780	0.575	0.598	679	-0.828	0.592	0.569
12	314	-0.727	0.613	0.643	487	-0.861	0.654	0.673
13	351	-0.720	0.574	0.599	681	-0.827	0.595	0.572
14	316	-0.766	0.678	0.703	489	-0.863	0.656	0.676
15	347	-0.775	0.679	0.709	681	-0.827	0.583	0.561
16	342	0.556	0.598	-0.767	**485**	**-0.852**	**0.641**	**0.661**
Ave.	346	0.604	0.631	-0.779	609	0.610	0.605	-0.839
Min.	308	0.305	0.358	-1.136	485	0.556	0.538	-0.886
Max.	410	0.679	0.734	-0.426	684	0.662	0.683	-0.812
Std.	29	0.090	0.084	0.139	97	0.036	0.056	0.021

*优化器编号
† 平方误差和（见公式 7）
‡ 纤锌矿（11$\bar{2}$0）表面
¶闪锌矿（110）表面

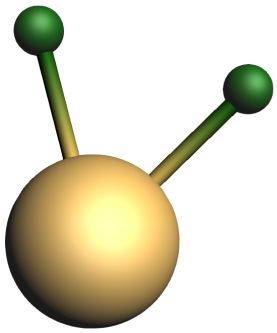
F 训练集的组成

本附录介绍了提取训练数据的所有作业（按字母顺序排列）。每项作业均提供以下信息：

- 用于引用条目的唯一作业编号；
- 系统初始几何构型的图像；
- 作业运行说明；
- 从作业中提取的训练数据详情。

参考数据的 DFT 计算使用 BAND 和 PBEsol 泛函，并采用 DZ 或 DZP 基组。对于体积扫描，k 空间质量设为“良好”，相当于每 $^{\circ}A^3$ 约有 27 个 k 点。

作业编号： 键角扫描 硫化氢 PBEsol



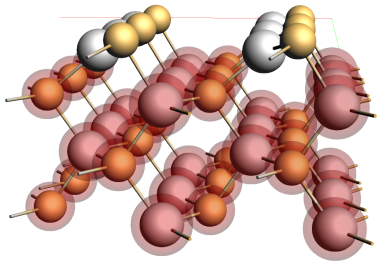
工作描述

- 10点角度扫描，范围从60°到180°。
- 每个角度最多进行 30 点几何优化

提取的训练数据

- 10 种几何构型各自的能量

作业编号： band110



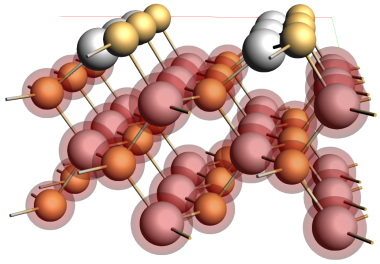
作业描述

- 最多进行 30 步几何优化
- 以红色光晕标记的原子为固定原子

提取的训练数据

- band 110 与 band distorted clean 110 之间的能量差
- 硫2化氢吸附能：（band h2s 110 - band 110 - band h2s）

作业编号： band 110 noconstraints



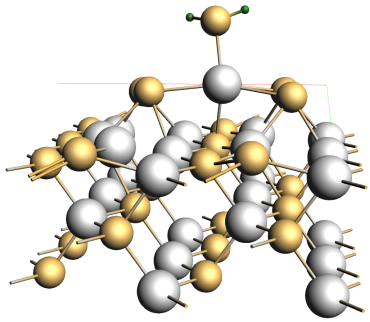
作业描述

- 最多进行 30 点几何优化
- 未固定任何原子位置

提取的训练数据

- 原子电荷（共 48 个）
- 5 薄片内部和表面的键角
- 1S-Zn-S-Zn 二面角
- 9薄片内部和表面的原子间距离
- band 110 noconstraints 与纤锌矿 sp 之间的能量差

作业编号：畸变吸附带 110



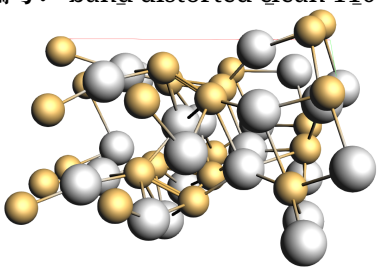
作业说明

- 单点计算
- 通过对分子动力学快照进行密度泛函理论几何优化得到的几何构型

提取的训练数据

- 原子力（共 153 个）
- 畸变吸附带 110 与 band h2s 110 之间的能量差

作业编号： band distorted clean 110



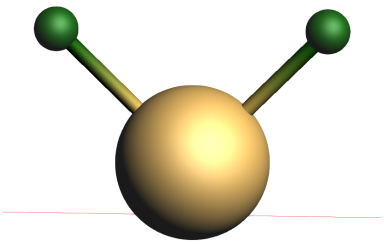
作业描述

- 单点计算
- 由分子动力学快照经密度泛函理论几何优化获得的几何构型

提取的训练数据

- 原子力（共 144 个）
- band distorted clean 110 与 band 110 之间的能量差

作业编号： bandh2s



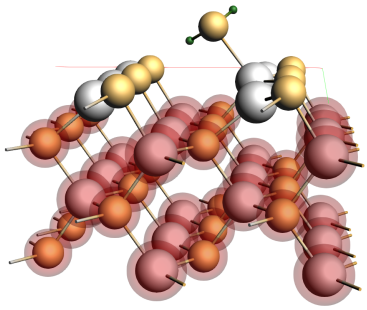
作业描述

- 最多 30 步的几何优化

提取的训练数据

- 原子电荷（共 3 个）
- 键角
- H-S 键长
- H_2S 吸附能：(band h2s 110 - band 110 - band h2s)

作业编号： band h2s 110



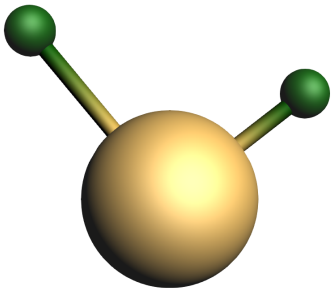
职位描述

- 最多 30 个点几何优化
- 以红色光晕标出的原子是固定的

提取的训练数据

- 原子电荷（共 51 个）
- S-锌 吸附距离
- H-S 吸附距离
- H_2S 吸附能：(band h2s 110 - band 110 - band h2s)
- 畸变吸附带 110 与 band h2s 110 之间的能量差

作业编号： 键长扫描 硫化氢 PBEsol



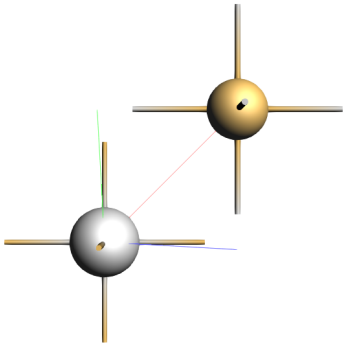
职位描述

- 10 从 1.1 °A 到 1.6 °A 的点键长扫描
- 每个键长最多进行 30 点几何优化

提取的训练数据

- 10种几何构型各自的能量

作业编号： 岩盐



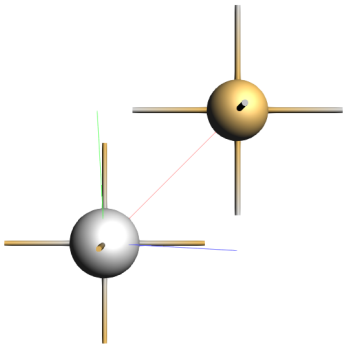
任务描述

- 5 晶胞体积扫描，缩放范围为 85% 至 125 %
- 对每个几何构型进行单点能计算

提取的训练数据

- 5 种几何构型各自的能量

作业编号：岩盐 sp_



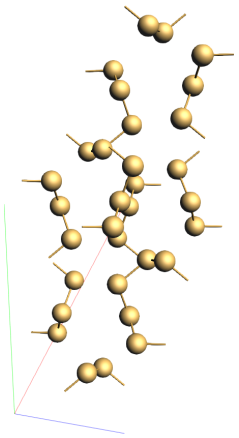
作业描述

- 单点计算

提取的训练数据

- 原子电荷（共 2 个）
- 能量差（岩盐 sp - 0.03125硫_- 0.5锌）
- 岩盐 sp 与闪锌矿 sp 之间的能量差 _ _

作业编号：硫



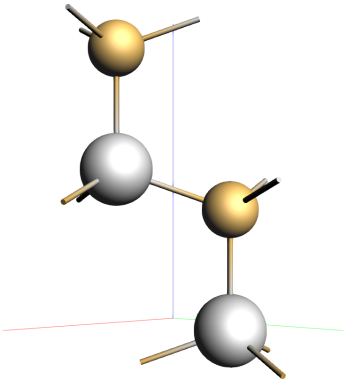
职位描述

- 单点计算

提取的训练数据

- 能量差（岩盐 sp - 0.03125硫_- 0.5锌）
- 能量差（闪锌矿 sp - 0.03125硫_- 0.5锌）
- 能量差 (0.5纤锌矿 sp - 0.03125硫 - 0.5锌)

作业编号： 纤锌矿



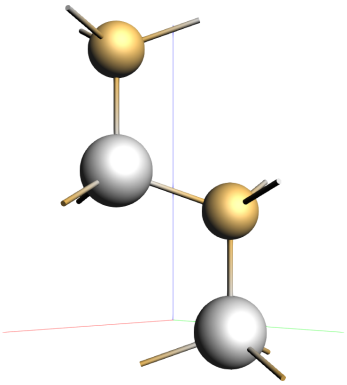
职位描述

- 5 晶胞体积扫描，缩放范围为 85% 至 125 %
- 每个几何构型的单点能计算

提取的训练数据

- 5种几何构型各自的能量

作业编号： 纤锌矿 sp



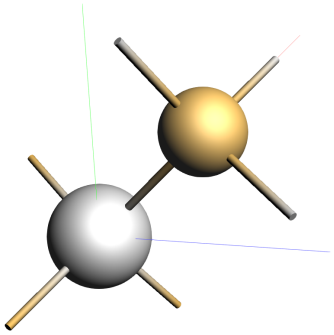
职位描述

- 单点计算

提取的训练数据

- 原子电荷（共 4 个）
- band 110 noconstraints 与纤锌矿 sp 之间的能量差
- 闪锌矿 sp 与纤锌矿 sp 之间的能量差
- 能量差（0.5纤锌矿自旋极化 - 0.03125硫 - 0.5锌）

作业编号： 闪锌矿



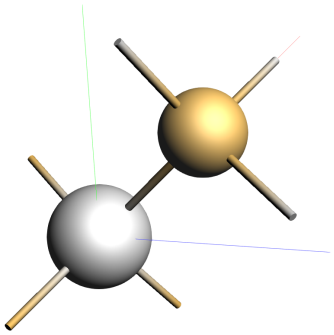
职位描述

- 5 晶胞体积扫描，缩放范围为 85% 至 125%
- 每个几何构型的单点能计算

提取的训练数据

- 5种几何构型各自的能量

作业编号： 闪锌矿 sp_



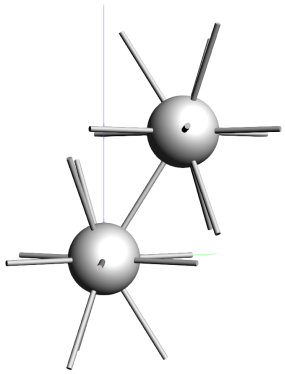
职位描述

- 单点计算

提取的训练数据

- 原子电荷（共 2 个）
- 岩盐 sp 与闪锌矿 sp 之间的能量差 _ _
- 闪锌矿 sp 与纤锌矿 sp 之间的能量差 _ _
- 能量差（闪锌矿 sp - 0.03125硫_- 0.5锌）

作业编号: 锌



职位描述

- 单点计算

提取的训练数据

- 原子电荷（共 2 个）
- 能量差（0.5纤锌矿 sp - 0.03125硫 - 0.5锌）
- 能量差（闪锌矿 sp - 0.03125硫 - 0.5锌）
- 能量差（岩盐 sp - 0.03125硫 - 0.5锌）