单纯复合体特征工程的理论、方法与实现

摘要

功能材料的理性设计是贯穿现代物理、化学与材料科学的核心挑战，其成功深度依赖于能够精确捕捉复杂"结构-性质"关系的描述符体系。传统描述符虽被广泛应用，但在编码晶体材料内蕴的精细几何、拓扑及量子力学信息方面往往表现出局限性。为突破此瓶颈，本研究构建了一套全新的、基于**单纯同调(Simplicial Homology)**思想的晶体材料特征工程学理论框架。该框架通过一个从0-单纯形（原子）到高阶构型（化学键、三体相互作用乃至全局性质）的层次化体系，对晶体结构进行系统性的数学表征。

本工作的核心创新在于，将**李代数(Lie Algebra)**、**辛几何(Symplectic Geometry)**与**商代数(Quotient Algebra)**等现代数学形式，同量子化学与凝聚态物理学的基本原理进行深度融合，旨在构建一套具有更高维度信息、更深刻物理内涵的特征体系。此框架的一个基石性原则是**物理不变性(Physical Invariance)**的严格执行：所有描述局域原子环境的特征，均被构造为在平移、旋转及对同种原子的置换操作下保持不变，从而保证其物理意义的普适性与数学表达的稳健性。

由此产生的特征体系，通过量化诸如局域对称性破缺、几何各向异性、电子-结构耦合以及拓扑连通性等关键概念，超越了传统描述符的表达能力，为材料提供了一个多尺度、高信息密度且物理内涵清晰的表示。本工作旨在为先进机器学习模型提供更具洞察力的输入，从而为实现材料的**逆向设计(Inverse Design)**与**物理规律的自动发现(Automated Discovery of Physical Laws)**这两大前沿目标奠定坚实的理论基础。

第一章：0-单纯形特征 - 原子中心环境

0-单纯形(0-simplex)代表单个原子，是构成晶体的基本单元。此层级的特征旨在全面描述原子的内禀性质及其局部配位环境的几何、拓扑与电子结构。本章将系统阐述一个描述原子中心环境的统一特征集，此集合从四个层面——**基础物理化学**、**量子化学**、**经典代数几何**与**融合量子-代数**——对原子进行全方位刻画，为构建多尺度材料基因组提供基石。

1.1 理论框架与特征体系

在单纯形理论的层次化框架中，0-单纯形（原子）构成了晶体结构的最基本单元。对单个原子及其局部环境进行全面而深刻的表征，是构建任何可靠的、自下而上的多尺度材料模型的逻辑起点。本章致力于为0-单纯形建立一个完备的特征体系，旨在超越传统描述符在信息维度和物理精度上的局限性。

传统的原子环境描述符，例如配位数(coordination number)，虽然在直观上具有解释力，但其固有的离散性和对几何细节的忽略，导致了严重的信息损失。例如，一个配位数为6的原子环境，既可能对应一个完美的正八面体（点群 O_h），也可能是一个经历了显著Jahn-Teller畸变的拉长或压扁八面体（点群 D_{4h}）。仅用整数"6"来描述这两种在物理性质（如d轨道能级分裂）上截然不同的构型，是远远不够的。此类描述符的非连续性使其难以捕捉结构畸变的微小、连续变化，从而限制了其在梯度优化算法和高精度机器学习模型中的应用潜力。

为克服此局限，本章的核心理论诉求是：必须构建一套能够将原子内禀属性及其局部环境的对称性、几何形状、电子结构拓扑以及化学-结构耦合效应，精确地量化为**连续、可微且在刚体运动（平移、旋转）下不变的**标量或张量不变量。这些特征不仅为机器学习模型提供了更丰富的输入，更重要的是，它们本身即构成了对局部物理化学状态的深度编码。

为实现此目标，本章将从以下四个相互关联但侧重点各异的层面，系统性地构建0-单纯形的特征集：**基础物理化学**（原子的内禀身份）、**量子化学**（电子结构的深度刻画）、**经典代数几何**（局部环境的形状与对称性），以及最终的**融合量子-代数**（结构与电子效应的耦合）。这一多视角、多层次的特征体系，共同为构建完整的材料基因组提供了坚实的基石。

1.2 基础物理化学特征

此部分特征旨在描述原子的内禀属性，这些属性主要通过标准化学信息库（如 `mendeleev` 、 `pymatgen` ）直接获取，或通过基础晶体化学分析得出。它们是定义原子身份及其在晶格中基本角色的基石。

1.2.1 元素周期表属性

这些特征是元素的固有属性，直接从化学信息库中获取，无需复杂计算。

- **原子序数 (`atomic_number`)**: Z，元素的唯一标识符，决定其在周期表中的位置和化学行为的基础。
- **电负性 (`electronegativity`)**: 鲍林(Pauling)电负性 χ，衡量原子在化学键中吸引电子能力的相对标度，是判断键极性的核心参数。
- **第一电离能 (`ionization_energy`)**: 从气态中性原子中移去一个电子至无穷远处所需的最低能量，直接反映原子失去电子的难易程度。
- **电子亲和能 (`electron_affinity`)**: 气态中性原子获得一个电子形成负离子时所释放的能量，反映原子获得电子的能力。
- **价电子数 (`valence_electrons`)**: 位于原子最外层、参与形成化学键的电子数目，是决定元素化合价和成键模式的关键。
- **离子半径 (`ionic_radius`)**: 在特定氧化态和配位数下，离子在离子晶体中所表现出的有效半径。
- **共价半径 (`covalent_radius`)**: 在共价键中，同种原子核间距的一半，反映原子在形成共价键时的尺寸。

1.2.2 局部拓扑与价态特征

1.2.2.1 配位数 (`coordination_number`)

为规避传统基于距离截断法所引入的经验性和主观性，我们采用基于**沃罗诺伊镶嵌 (Voronoi Tessellation)** 的严格几何方法来定义配位数。对于晶格中的中心原子 A，其配位数 CN_A 被无歧义地定义为在沃罗诺伊分解中与其共享一个有效界面（沃罗诺伊多面体的面）的邻居原子的总数：

$$\mathrm{CN}_A = |\{j \in \text{Neighbors} : V_A \cap V_j \neq \emptyset \text{ and has area} > 0\}|$$

其中 V_A 和 V_j 分别是原子A和j的沃罗诺伊多面体。此定义确保了配位数的确定性与客观性，不受人为参数选择的影响。

1.2.2.2 键价和 (`bond_valence_sum`)

原子的形式氧化态通过**键价和理论 (Bond Valence Sum, BVS)** 进行估算。该理论假定，一个离子 i 的形式价态 V_i 可以通过与之成键的所有邻居离子 j 对其贡献的键价 s_{ij} 的总和来近似：

$$V_i = \sum_j s_{ij}$$

其中，单根键的键价 s_{ij} 与键长 d_{ij} 之间存在一个被广泛验证的经验性指数关系：

$$s_{ij} = \exp\left(\frac{R_0 - d_{ij}}{B}\right)$$

在此表达式中，R_0 是一个依赖于成键原子对 (A-B) 类型的经验参数，代表该键对的理想键长，而 B 通常取为一个普适常数，约为 0.37 Å。该特征为原子提供了一个化学上合理、物理上自洽的氧化态估计。

1.3 量子化学核心特征

此部分特征直接从**第一性原理密度泛函理论 (DFT)** 计算的输出中提取，旨在提供对原子局部电子结构、成键环境以及电磁性质的深度洞察。这些特征是连接原子核排布与材料宏观物理性质的关键桥梁。

1.3.1 原子电荷与静电势

1.3.1.1 Bader电荷 (`bader_charge`)

Bader电荷是一种基于**"分子中原子"的量子理论 (Quantum Theory of Atoms in Molecules, QTAIM)** 的原子电荷划分方案。该理论通过分析总电子密度场 $\rho(\mathbf{r})$ 的拓扑性质来定义原子。具体而言，它在原子间寻找电子密度的"谷底"，即梯度为零的**零通量面 (zero-flux surface)**。这些曲面将总电子密度明确地划分为若干个互不交叠的原子盆 (atomic basins)。原子的Bader电荷 q_{Bader} 定义为该原子的原子核电荷 Z 与其原子盆 Ω 内积分电子密度之差：

$$q_{\text{Bader}} = Z - \int_{\Omega} \rho(\mathbf{r}) d^3\mathbf{r}$$

相较于其他电荷分配方案（如Mulliken电荷），Bader电荷被认为是一种物理意义更明确、对计算基组依赖性更小的电荷描述符。

1.3.1.2 原子核处静电势 (`electrostatic_potential_at_nucleus`)

该特征是原子核位置 $\mathbf{R}_A$ 的总静电势（哈特里势） $\Phi(\mathbf{R}_A)$。此静电势由体系中所有原子核的正电荷与所有电子的负电荷密度通过泊松方程自洽决定。它精确地反映了该原子核所感受到的、来自体系其余所有部分的平均静电场环境，是衡量局部极化和化学反应活性的一个关键指标。

1.3.2 原子核处电子密度与局域化函数

1.3.2.1 原子核处电子密度 (`electron_density_at_nucleus`)

即在原子核位置 $\mathbf{R}_A$ 的总电子密度值 $\rho(\mathbf{R}_A)$。根据量子力学，该值与原子核的s轨道电子波函数在核位置的概率密度密切相关（尤其是对于较重元素），是衡量原子核附近电子聚集程度的最直接指标。它在分析超精细相互作用（如穆斯堡尔谱中的同质异能移）等现象时至关重要。

1.3.2.2 原子核处电子局域化函数 (`elf_at_nucleus`)

电子局域化函数 (Electron Localization Function, ELF) 是一个用于揭示化学体系中电子在空间中局域化程度的、值域为 $[0, 1]$ 的标量场。ELF的核心思想是通过比较在体系中任意一点 $\mathbf{r}$ 的同自旋电子对的动能密度，与具有相同密度的均匀电子气（一种理想化的自由电子模型）的动能密度，来判断该点电子的成对概率，从而衡量其局域性。

- **高ELF值 (接近1)**: 意味着该处的电子更倾向于成对出现，而不是自由运动。这通常对应于化学上直观的区域：共价键的中心、非键孤对电子或原子的内壳层电子。
- **中等ELF值 (接近0.5)**: 意味着该处电子的行为与均匀电子气类似，表现出典型的离域金属电子行为。
  原子核处的ELF值 $\mathrm{ELF}(\mathbf{R}_A)$ 尤其能反映原子内层电子的局域化状态和壳层结构。

1.3.3 局域磁矩 (`local_magnetic_moment`)

在考虑电子自旋的自旋极化计算中，每个原子位点 i 的局域磁矩 μ_i 被定义为在该原子周围特定空间区域 Ω_i 内，自旋向上电子密度 $\rho_\uparrow(\mathbf{r})$ 与自旋向下电子密度 $\rho_\downarrow(\mathbf{r})$ 之差的总积分：

$$\mu_i = \int_{\Omega_i} \left[\rho_\uparrow(\mathbf{r}) - \rho_\downarrow(\mathbf{r})\right] d^3\mathbf{r}$$

其中，积分区域 Ω_i 可以有多种定义方式，例如Bader原子盆或Muffin-tin球。该特征是描述材料磁有序（铁磁、反铁磁等）性质的根本。

1.3.4 局域与投影态密度(DOS)特征

1.3.4.1 费米能级处局域态密度 (`ldos_at_fermi`)

局域态密度 (Local Density of States, LDOS) $N_i(E)$ 描述了能量为 E 的电子态在原子 i 周围空间内的分布权重。费米能级 E_F 处的LDOS值 $N_i(E_F)$ 是判断材料导电性的关键指标：对于金属，$N_i(E_F) > 0$；对于半导体或绝缘体，$N_i(E_F) \approx 0$。

1.3.4.2 投影态密度(PDOS)的矩

投影态密度 (Projected Density of States, PDOS) $N_{il}(E)$ 进一步将LDOS根据角动量量子数 l（$l = s, p, d, f, \ldots$）分解到不同的原子轨道上。通过计算PDOS的能量矩，可以定量描述特定原子上特定轨道电子能带的形状、中心和宽度。

- **PDOS带中心 (`pdos_band_center_l`)**: 定义为被占据态PDOS的第一矩，即能量关于态密度的加权平均值。它表征了原子 i 的 l 轨道电子占据态的平均能量。

  $$E_{il,\text{center}} = \frac{\int_{-\infty}^{E_F} E \cdot N_{il}(E) dE}{\int_{-\infty}^{E_F} N_{il}(E) dE}$$

  带中心的位置与化学成键的强度和类型密切相关。
- **PDOS带宽 (`pdos_band_width_l`)**: 定义为被占据态PDOS的二阶中心矩的平方根（即标准差），表征了 l 轨道电子能量的分布宽度。

  $$W_{il,\text{width}} = \sqrt{\frac{\int_{-\infty}^{E_F} (E - E_{il,\text{center}})^2 \cdot N_{il}(E) dE}{\int_{-\infty}^{E_F} N_{il}(E) dE}}$$

  根据紧束缚理论，带宽与轨道间的交叠积分（hopping integral）成正比。因此，带宽越大，通常意味着电子离域性越强或轨道杂化越显著。

1.4 经典代数几何特征

这部分特征旨在将局部原子环境的几何信息，通过抽象的代数与几何框架进行编码。其核心目标是超越传统的几何描述符（如键长、键角），发展出能够捕捉局部环境对称性、形状和各向异性等高阶信息的、连续且具有不变性的数学构造。

1.4.1 键长畸变指数 (`bond_length_distortion_index`)

此特征旨在量化中心原子与其近邻配位原子之间键长的离散程度，是衡量配位多面体几何畸变的最直接指标之一。设中心原子与其第 i 个邻居的键长为 d_i ($i = 1, \ldots, N$)，则平均键长为 $\bar{d} = \frac{1}{N}\sum_{i=1}^{N} d_i$。键长畸变指数 σ_d 定义为这组键长的标准差：

$$\sigma_d = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(d_i - \bar{d})^2}$$

在一个完美的对称环境中（例如，理想BCC晶格中的原子），所有键长相等，因此 $\sigma_d = 0$。该值越大，表示配位多面体的键长畸变越严重，这通常与Jahn-Teller效应或晶格应变相关。

1.4.2 李代数不对称性 (`lie_algebraic_asymmetry`)

1.4.2.1 物理动机

原子配位环境偏离完美对称性的程度，是决定其局部电子性质（如晶体场分裂、d轨道能级简并解除、Jahn-Teller畸变）的关键物理因素。我们寻求一个能够整体捕捉这种不对称性的标量描述符。一个理想的描述符应当满足：对于任何具有反演对称中心的完美对称环境（例如，理想BCC晶格的中心原子），其值为零；并且该值应随着对称性破缺程度的增加而单调递增。

1.4.2.2 数学构造与理论基础

设中心原子位于坐标原点 $\mathbf{p}_0 = \mathbf{0}$。其局部环境由一组从中心原子指向其 N 个最近邻原子的**相对位置向量** $\{\mathbf{v}_i \in \mathbb{R}^3\}_{i=1}^{N}$ 完全定义。我们首先定义一个**局部结构不对称向量** $\mathbf{V}_{\text{struct}}$，作为这些相对位置向量的矢量和：

$$\mathbf{V}_{\text{struct}} = \sum_{i=1}^{N} \mathbf{v}_i$$

在任何具有反演对称性的点群中，对于每一个邻居向量 $\mathbf{v}_i$，都必然存在一个与之对应的邻居向量 $\mathbf{v}_j = -\mathbf{v}_i$。因此，在这些高度对称的环境中，$\mathbf{V}_{\text{struct}}$ 严格为零。当对称性被破坏时（例如，由于缺陷、应变或热振动），向量和将不再为零，其大小 $\|\mathbf{V}_{\text{struct}}\|$ 直观地量化了这种不对称的程度。

为了赋予该描述符更深刻的代数内涵并确保其旋转不变性，我们引入**李代数** `so(3)` ——三维空间中所有无穷小旋转所构成的代数。任何三维向量 $\mathbf{u} = (u_x, u_y, u_z)$ 都可以通过伴随表示映射到 `so(3)` 中的一个元素，即一个3×3的反对称矩阵 $M_{\mathbf{u}}$：

$$M_{\mathbf{u}} = \begin{pmatrix} 0 & -u_z & u_y \\ u_z & 0 & -u_x \\ -u_y & u_x & 0 \end{pmatrix}$$

李代数的一个核心概念是**卡西米尔不变量(Casimir Invariant)**，它是一个与代数中所有元素都对易的算子。对于半单李代数 `so(3)` ，其二阶Casimir不变量 C_2 作用在任意一个代数元素 M 上的值为 $C_2(M) = -\frac{1}{2}\mathrm{Tr}(M^2)$。将我们的物理量 $\mathbf{V}_{\text{struct}}$ 代入，我们首先计算矩阵 $M_{\mathbf{V}_{\text{struct}}}$ 的平方：

$$M^2_{\mathbf{V}_{\text{struct}}} = \begin{pmatrix} -(v_y^2+v_z^2) & v_x v_y & v_x v_z \\ v_x v_y & -(v_x^2+v_z^2) & v_y v_z \\ v_x v_z & v_y v_z & -(v_x^2+v_y^2) \end{pmatrix}$$

其中 v_x, v_y, v_z 是 $\mathbf{V}_{\text{struct}}$ 的分量。该平方矩阵的迹（Trace）为：

$$\text{Tr}(M^2_{\mathbf{V}_{\text{struct}}}) = -(v_y^2+v_z^2) - (v_x^2+v_z^2) - (v_x^2+v_y^2) = -2(v_x^2+v_y^2+v_z^2) = -2\left\|\mathbf{V}_{\text{struct}}\right\|^2$$

最后，根据Casimir不变量的定义，我们得到一个至关重要的恒等式：

$$C_2(M_{\mathbf{V}_{\text{struct}}}) = -\frac{1}{2}\text{Tr}(M^2_{\mathbf{V}_{\text{struct}}}) = -\frac{1}{2}\left(-2\left\|\mathbf{V}_{\text{struct}}\right\|^2\right) = \left\|\mathbf{V}_{\text{struct}}\right\|^2$$

这个结果表明，局部结构不对称向量的**平方欧几里得范数**，在数值上严格等于其在 `so(3)` 李代数伴随表示下的Casimir不变量。由于Casimir不变量在所有李群变换（即旋转）下保持不变，这就为我们的特征提供了坚实的旋转不变性保证。因此，我们将 `lie_algebraic_asymmetry` 特征严格定义为：

$$\mathcal{L}_{\text{asymm}} = \left\|\mathbf{V}_{\text{struct}}\right\|^2 = \left\|\sum_{i=1}^{N} \mathbf{v}_i\right\|^2$$

1.4.2.3 计算注解

在实践中，通过 `pymatgen` 或 `ASE` 等工具确定中心原子的 N 个近邻及其相对位置向量 `v_i` 后，仅需使用 `numpy.sum` 对这些向量求和，然后用 `numpy.linalg.norm` 计算其范数的平方即可，计算过程极为高效。

1.4.3 局部环境各向异性 (`local_environment_anisotropy`)

1.4.3.1 物理动机

除了是否存在由非零 $\mathbf{V}_{\text{struct}}$ 所描述的"极性"不对称之外，配位多面体的整体"形状"也至关重要。一个配位环境可以是各向同性的（球形），也可以是在某个平面上延展的（扁球形），或是在某个轴向拉长的（长球形）。这种形状的各向异性决定了轨道交叠的方向性、电荷传输的优势路径以及对外部应力的响应模式。该特征旨在定量地捕捉这种形状上的各向异性。

1.4.3.2 数学构造与理论基础

我们采用二阶矩张量，在此语境下称为**结构张量 (Structure Tensor)** $\mathbf{T}$，来捕捉邻近原子群的空间分布特征。它在形式上完全类似于刚体转动理论中的惯性张量，描述了邻居原子"质量云"的分布形状和主轴方向。其定义为所有邻居相对位置向量 $\mathbf{v}_i$ 的**外积 (Outer Product)** 之和：

$$\mathbf{T} = \sum_{i=1}^{N} \mathbf{v}_i \otimes \mathbf{v}_i^T$$

其中，$\otimes$ 代表外积。对于任意一个邻居向量 $\mathbf{v}_i = (v_{ix}, v_{iy}, v_{iz})^T$，其自身的外积 $\mathbf{v}_i \otimes \mathbf{v}_i^T$ 是一个3x3的对称矩阵：

$$\mathbf{v}_i \otimes \mathbf{v}_i^T = \begin{pmatrix} v_{ix}^2 & v_{ix}v_{iy} & v_{ix}v_{iz} \\ v_{iy}v_{ix} & v_{iy}^2 & v_{iy}v_{iz} \\ v_{iz}v_{ix} & v_{iz}v_{iy} & v_{iz}^2 \end{pmatrix}$$

结构张量 $\mathbf{T}$ 是将所有 N 个邻居的这种方向信息矩阵进行线性叠加，从而得到一个描述整个配位环境平均"形状"和"取向"的综合性对称矩阵。

作为一个实对称矩阵，$\mathbf{T}$ 具有三个实数特征值，$\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq 0$，它们分别代表了邻居原子分布在三个相互正交的主轴方向上的方差。

- 如果 $\lambda_1 \approx \lambda_2 \approx \lambda_3$，则环境是**各向同性**的（球形）。
- 如果 $\lambda_1 \gg \lambda_2 \approx \lambda_3$，则环境是**轴向的**（长球形）。
- 如果 $\lambda_1 \approx \lambda_2 \gg \lambda_3$，则环境是**平面**的（扁球形）。

为了构造一个单一的、归一化的标量来描述这种各向异性，我们采用**相对形状各向异性**的定义，它量化了特征值谱的离散程度：

$$\mathcal{A}_{\text{aniso}} = \frac{(\lambda_1 - \lambda_2)^2 + (\lambda_2 - \lambda_3)^2 + (\lambda_3 - \lambda_1)^2}{2(\lambda_1 + \lambda_2 + \lambda_3)^2} = \frac{\sum_i \lambda_i^2 - \sum_{i<j} \lambda_i \lambda_j}{(\sum_i \lambda_i)^2}$$

该表达式的一个重要优点是，它可以完全通过矩阵的迹不变量来计算，从而在实践中**完全避免**了耗时的特征值求解过程。利用矩阵理论中的恒等式：

- $\sum_i \lambda_i = \text{Tr}(\mathbf{T})$
- $\sum_i \lambda_i^2 = \text{Tr}(\mathbf{T}^2)$
- $\sum_{i<j} \lambda_i \lambda_j = \frac{1}{2}[(\text{Tr}(\mathbf{T}))^2 - \text{Tr}(\mathbf{T}^2)]$

代入上式并化简，我们得到最终的、计算上极为高效的公式：

$$\mathcal{A}_{\text{aniso}} = \frac{3}{2} \frac{\text{Tr}(\mathbf{T}^2)}{(\text{Tr}(\mathbf{T}))^2} - \frac{1}{2}$$

该特征是无量纲且有界的，$0 \leq \mathcal{A}_{\text{aniso}} \leq 1$。

- $\mathcal{A}_{\text{aniso}} = 0$ 对应完美各向同性环境 ($\lambda_1 = \lambda_2 = \lambda_3$)。
- $\mathcal{A}_{\text{aniso}} = 1$ 对应完美线性（一维）环境 ($\lambda_1 > 0, \lambda_2 = \lambda_3 = 0$)。

1.4.3.3 计算注解

在 `numpy` 中，可以先构建一个 $(N, 3)$ 的邻居向量矩阵 `V`，则结构张量 `T` 可以通过矩阵乘法 `V.T @ V` 高效计算。`T` 的迹 `Tr(T)` 可用 `numpy.trace(T)` 获得，而 `T` 的平方的迹 `Tr(T^2)` 可以通过 `numpy.trace(T @ T)` 算出。最终将这些值代入公式即可，完全无需调用特征值求解器。

1.4.4 商空间不变度量（`invariant_quotient_metric`）

1.4.4.1 物理动机

在真实的晶体中，观察到的任何局部几何畸变，都是**内禀化学驱动力**（原子自身的化学成键偏好，如sp³杂化倾向于109.5°角）和**外部环境约束**（如晶格应力、长程库仑相互作用）共同作用的复杂结果。例如，一个高电负性的F原子倾向于形成强方向性的共价键，其环境天然地具有一定程度的不对称性；而一个低电负性的Cs阳离子倾向于形成无方向性的离子键，其环境天然地趋向于高度对称。该特征旨在通过**定义一个将几何畸变用原子内禀化学性质进行归一化的新度量**，来解耦这两种效应，从而衡量出"超额"的、主要由外部环境主导的畸变。

1.4.4.2 数学构造

该特征在"商代数"的抽象框架内被概念化，即通过取两个具有明确物理意义的量的比值来构造一个新的、具有更深刻洞察力的度量。

1. **分子：无量纲化的几何不对称性 L'**。为了使几何不对称性的大小能够在不同的化学体系和不同的配位数之间进行有意义的比较，我们首先将其进行无量纲化。此过程涉及以下参数：
   - N: 中心原子的配位数。
   - $\mathbf{v}_i$: 从中心原子指向第 i 个近邻原子的相对位置向量。
   - $\bar{d}$: 中心原子到其所有近邻的平均键长（距离），$\bar{d} = \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{v}_i\|$。

   我们将无量纲的几何不对称性 L' 定义为，由平均键长归一化后的不对称向量的范数平方：

   $$L' = \left\| \frac{1}{N} \sum_{i=1}^{N} \frac{\mathbf{v}_i}{\bar{d}} \right\|^2$$

   此处的双重归一化（除以N和$\bar{d}$）确保了 L' 是一个与尺寸和邻居数量无关的、纯粹描述形状的量。
2. **分母：内禀化学驱动力代理**。我们选择原子的**鲍林电负性** χ 作为其内禀成键方向性偏好的标量代理。χ 越大，原子形成定向共价键的趋势越强，因此其对一定程度的几何不对称性的"容忍度"或"期望值"就越高。

我们将 `invariant_quotient_metric` $\mathcal{Q}_{\text{metric}}$ 定义为无量纲几何不对称性 L' 与电负性平方 χ^2 的比值：

$$\mathcal{Q}_{\text{metric}} = \frac{L'}{\chi^2}$$

分母中使用 χ^2 是基于物理化学的考量：在许多键合模型中（如Pauling关于离子性的公式），能量项与电负性之差的平方成正比。此处的平方旨在使分母在概念上与分子的"能量"量纲（向量范数平方）相匹配，从而使整个商度量在量纲上更加和谐。

1.4.4.3 解释与意义

该度量直观地代表了**单位化学驱动力所导致的几何不对称性**。一个高值的 $\mathcal{Q}_{\text{metric}}$ 强烈地暗示，该原子所处的环境畸变程度，远超其自身化学性质所能解释的范畴。这为我们提供了一个强有力的指标来识别那些承受着显著外部应力、长程相互作用或处于亚稳态的原子位点。在材料的逆向设计中，该特征可作为一个关键的筛选判据，用于寻找低应变、高稳定性的结构。

1.5 融合量子-代数特征

1.5.1 物理动机

前两节分别从量子化学和经典代数几何的角度对原子局部环境进行了刻画。然而，这两种描述在物理上并非独立，而是深度耦合的：原子的几何排布决定了电子波函数的形态，而电子的重新分布（即化学键的形成）又反作用于原子，使其弛豫到新的平衡位置。为了捕捉这种至关重要的**电子-结构耦合效应**，本节引入了一类创新的融合特征。其核心思想是，将由原子核位置决定的经典几何/代数不变量（"形"）与从DFT计算中得到的、描述电子行为的量子化学标量（"神"）进行数学上的融合（如乘积、加权等）。由此产生的复合描述符，旨在量化那些只有在同时考虑几何与电子结构时才会显现的高阶物理化学现象。

1.5.2 数学构造与物理解释

这些特征是通过将1.3节（量子化学核心特征）的标量与1.4节（经典代数几何特征）的标量进行元素级乘积而系统性地构建的。

1. **电荷加权李代数不对称性 (`charge_weighted_lie_asymmetry`)**
   此特征用原子的Bader电荷 q_{Bader} 来加权李代数不对称性，旨在量化一个**带电体系的几何不对称性**。

   $$\mathcal{L}_{\text{quantum}} = q_{\text{Bader}} \times \mathcal{L}_{\text{asymm}}$$

   物理图像: 一个带有净正电荷或负电荷的原子，如果其配位环境同时又是几何不对称的（即 $\mathcal{L}_{\text{asymm}} > 0$），那么该原子的正电荷中心（原子核）将与其周围电子云的负电荷中心不再重合。这种电荷中心与几何中心的分离，在物理上就构成了一个**局域电偶极矩**。本特征的数值正比于该诱导电偶极矩的大小，因此它直接量化了由于几何畸变导致的局域极化强度。

2. **电子态加权各向异性 (`electron_weighted_anisotropy`)**
   此特征用原子核位置的电子密度 $\rho(\mathbf{R}_A)$ 和电子局域化函数 $\mathrm{ELF}(\mathbf{R}_A)$ 来共同加权几何各向异性。

   $$\mathcal{A}_{\text{quantum}} = (\rho(\mathbf{R}_A) \cdot \mathrm{ELF}(\mathbf{R}_A)) \times \mathcal{A}_{\text{aniso}}$$

   物理图像: 纯粹的几何各向异性 $\mathcal{A}_{\text{aniso}}$ 只描述了原子核的空间排布形状。然而，从化学成键的角度看，这种几何形状只有在电子"实际参与"的区域才显得尤为重要。例如，在共价键或孤对电子所在的区域，电子密度高（ρ大）且高度局域化（ELF大），此时几何上的任何各向异性都将强烈影响轨道交叠和化学键的方向性。相反，在电子密度极低的区域，即使几何上存在各向异性，其化学意义也微乎其微。因此，加权因子 $(\rho \cdot \mathrm{ELF})$ 的作用就是**放大这种由成键电子或孤对电子主导的、具有化学活性的几何各向异性**。
3. **电荷调制商空间度量 (`charge_modulated_quotient_metric`)**
   此特征用一个与Bader电荷相关的函数来调制不变商空间度量，旨在探索电荷状态对原子感知"超额"畸变的非线性响应。

   $$\mathcal{Q}_{\text{quantum}} = f(q_{\text{Bader}}) \times \mathcal{Q}_{\text{metric}}$$

   其中 $f(q)$ 是一个用于探索非线性电荷效应的可选加权函数。其具体形式可根据研究目标进行设计，例如可选用指数函数 $f(q) = \exp(q)$ 来描述电荷对畸变的指数级敏感度，或选用多项式函数 $f(q) = (1 + |q|)^n$ 来建模更高阶的电荷效应。
   物理图像: 该特征旨在回答一个更深层次的问题：一个带电的离子（$q_{\text{Bader}} \neq 0$）与一个中性原子相比，它对由外部应力所引起的"超额"几何畸变（由 $\mathcal{Q}_{\text{metric}}$ 度量）是否更敏感或更不敏感？这种敏感性的依赖关系是否是线性的？通过引入可调的函数 $f(q)$, 该特征为机器学习模型提供了一个能够发掘这种复杂的、非线性的**电荷-应变耦合关系**的自由度。

这些融合特征是连接微观电子结构与局部几何构型的关键桥梁，有望在机器学习模型中捕捉到比单一来源特征更复杂、更深刻的构效关系。

第二章：1-单纯形特征 - 化学键环境

1-单纯形(1-simplex)代表连接两个原子（0-单纯形）的一条边，在化学中，其自然对应物是**化学键**。本章旨在为化学键这一核心概念建立一套完备的数学描述符。这些特征不仅要捕捉键本身的几何与量子化学属性，还必须作为桥梁，关联其所连接的两个原子的0-单纯形特征，从而构建一个真正具有层次化、关联性的特征体系。

2.1 几何与拓扑特征

2.1.1 键长 (`bond_length`)

键长是描述两个成键原子A和B之间相互作用强度的最直接几何量度。设原子A和B的核坐标分别为 $\mathbf{r}_A$ 和 $\mathbf{r}_B$，则键长 d_{AB} 被严格定义为它们之间的欧几里得距离：

$$d_{AB} = \|\mathbf{r}_A - \mathbf{r}_B\| = \sqrt{(r_{A,x} - r_{B,x})^2 + (r_{A,y} - r_{B,y})^2 + (r_{A,z} - r_{B,z})^2}$$

根据Morse势等经典的势能面模型，键长与键的强度和稳定性存在直接的、通常是指数的反比关系。在晶格动力学中，键长偏离其平衡值的程度直接量化了局部的结构应变，是计算弹性常数和声子频率的基础输入。

2.1.2 沃罗诺伊分析相关特征

2.1.2.1 理论背景

为了获得一个无参数、无偏见的原子配位环境描述，我们再次采用基于**沃罗诺伊镶嵌**的几何方法。沃罗诺伊分解为空间中的点集提供了一个唯一且完备的划分，从而为定义配位数、邻居关系和几何重要性提供了客观的理论基础。对于晶体中的任意原子 k，其沃罗诺伊多面体 V_k 定义为空间中所有到原子 k 的距离不大于到任何其他原子距离的点的集合。两个相邻原子 A 和 B 的沃罗诺伊多面体共享一个公共面 F_{AB}。

2.1.2.2 沃罗诺伊界面面积与立体角 (`voronoi_interface_area` , `voronoi_solid_angle`)

- **界面面积**: 共享界面 F_{AB} 的面积 A_{AB}，直接量化了原子A和B之间的"接触"程度。
- **立体角**: 对于化学键A-B，沃罗诺伊立体角 Ω_{AB} 量化了原子B在原子A的配位环境中占据的"空间份额"。它定义为从原子A的核位置观察共享沃罗诺伊界面 F_{AB} 所张开的立体角。

物理意义：这两个量共同描述了邻居原子B相对于中心原子A的几何重要性。一个具有较大界面面积和立体角的邻居，通常意味着它对中心原子的局部几何环境和电子结构具有更显著的影响。在所有邻居中，这些值的相对大小可以用来区分强相互作用和弱相互作用。

2.1.2.3 键两端原子的配位数 (`site1_coord_num` , `site2_coord_num`)

基于沃罗诺伊分解，我们已在1.2.2.1节中将原子的配位数严格定义为与其共享沃罗诺伊界面的邻居原子数量。对于一个化学键A-B， `site1_coord_num` 和 `site2_coord_num` 分别表示成键原子A和B各自的沃罗诺伊配位数。这两个数值为理解键的局部拓扑环境（例如，一个键是连接两个高配位数中心，还是连接一个高配位数中心和一个低配位数终端）提供了关键信息。

2.1.3 晶格对齐度 (`bond_lattice_alignment`)

2.1.3.1 理论动机

在晶体材料中，化学键相对于晶格主轴的取向直接影响材料的各向异性物理性质，例如方向依赖的弹性模量、热膨胀系数和电导率。为定量描述这种空间取向关系，我们引入了键的晶格对齐度特征。

2.1.3.2 数学定义

设化学键A-B的单位向量为 $\hat{\mathbf{v}}_{AB} = (\mathbf{r}_B - \mathbf{r}_A)/d_{AB}$，晶格基矢为 $\mathbf{a}, \mathbf{b}, \mathbf{c}$，其单位向量为 $\hat{\mathbf{a}}, \hat{\mathbf{b}}, \hat{\mathbf{c}}$。键向量与三个晶格基矢的对齐度被定义为一个三维向量：

$$\mathbf{A}_{\text{align}} = \left(|\hat{\mathbf{v}}_{AB} \cdot \hat{\mathbf{a}}|, |\hat{\mathbf{v}}_{AB} \cdot \hat{\mathbf{b}}|, |\hat{\mathbf{v}}_{AB} \cdot \hat{\mathbf{c}}| \right)$$

每个分量都是键向量与对应晶格基矢之间夹角余弦的绝对值，取值范围为 $[0, 1]$。

- **分量值为1**: 键向量与该晶格基矢完全平行。
- **分量值为0**: 键向量与该晶格基矢完全垂直。

物理解释：这个三维向量 $\mathbf{A}_{\text{align}}$ 完整地编码了化学键在晶格坐标系中的取向信息。例如，一个主要沿着a轴方向的键，其对齐度向量会接近 $(1, 0, 0)$。这个特征对于分析结构各向异性与材料性能之间的关系至关重要。

2.2 原子属性派生特征

2.2.1 理论框架：从0-单纯形到1-单纯形

此类特征是实现本工作核心的层次化特征体系的关键，完美体现了单纯同调学的基本思想：一个高阶单纯形的性质，应当能够通过其边界上低阶单纯形的性质来表达或派生。具体到1-单纯形（化学键A-B），其特征应通过其两个端点0-单纯形（原子A和B）的已知特征来构造，从而在不同特征层次间建立起深刻的内在联系和物理关联。

2.2.2 数学构造原理：差异与平均算子

设 $\mathcal{F}^{(0)}(A)$ 和 $\mathcal{F}^{(0)}(B)$ 分别为成键原子A和B的任意一个已定义的、标量型的0-单纯形特征（例如，电负性、Bader电荷等）。我们定义以下两类普适的派生算子，用于从原子特征生成键特征：

1. **差异算子 (Difference Operator)** Δ:

$$\Delta\mathcal{F}^{(0)} = \left| \mathcal{F}^{(0)}(A) - \mathcal{F}^{(0)}(B) \right|$$

   此算子旨在量化成键原子在特定属性上的**异质性 (heterogeneity)** 或不匹配程度。根据鲍林的经典理论，原子间电负性的差异是决定化学键极性和离子性的核心因素。因此，差异算子在描述键的极

性、电荷转移以及由于不匹配引起的局部应变等特征中，具有根本性的重要性。

2. **平均算子 (Averaging Operator)** $\overline{\mathcal{F}}$:

$$\overline{\mathcal{F}}^{(0)} = \frac{\mathcal{F}^{(0)}(A) + \mathcal{F}^{(0)}(B)}{2}$$

此算子描述了化学键所在区域的**平均化学环境**，反映了成键原子双方的协同效应。在分子轨道理论中，成键轨道的能量往往与参与成键的原子轨道能量的平均值相关。因此，平均算子对于描述键的整体能量标度、平均电子密度等性质至关重要。

2.2.3 具体派生特征列表

基于上述理论框架，我们系统性地将差异和平均算子应用于第一章中定义的各类0-单纯形特征，从而生成以下一系列具有明确物理意义的1-单纯形特征：

- **电负性差 (`delta_electronegativity`)**: $\Delta\chi = |\chi_A - \chi_B|$，直接关联键的离子性和极性。
- **离子半径差 (`delta_ionic_radius`)**: $\Delta r_{\text{ion}} = |r_{\text{ion},A} - r_{\text{ion},B}|$，量化成键原子尺寸不匹配度，与局部几何应变相关。
- **共价半径和 (`sum_covalent_radii`)**: $\Sigma r_{\text{cov}} = r_{\text{cov},A} + r_{\text{cov},B}$。根据Schomaker-Stevenson规则，此和值是理想共价键长的理论参考，可用于定义键长应变。
- **平均Bader电荷 (`avg_bader_charge`)**: $\bar{q}_{\text{Bader}} = (q_{\text{Bader},A} + q_{\text{Bader},B})/2$，量化键区域的平均电荷积累或亏损。
- **Bader电荷差 (`delta_bader_charge`)**: $\Delta q_{\text{Bader}} = |q_{\text{Bader},A} - q_{\text{Bader},B}|$，直接度量键上的净电荷转移量，是键极性的第一性原理描述符。
- **局部环境各向异性差 (`delta_local_anisotropy`)**: $\Delta\mathcal{A} = |\mathcal{A}_{\text{aniso},A} - \mathcal{A}_{\text{aniso},B}|$，描述成键原子两端的局部几何环境形状的差异程度。
- **平均商空间度量 (`avg_quotient_metric`)**: $\overline{\mathcal{Q}} = (\mathcal{Q}_{\text{metric},A} + \mathcal{Q}_{\text{metric},B})/2$，量化化学键所在区域的平均"超额"应变状态。
- **李代数不对称性差 (`delta_lie_asymmetry`)**: $\Delta\mathcal{L} = |\mathcal{L}_{\text{asymm},A} - \mathcal{L}_{\text{asymm},B}|$，描述成键原子两端局部不对称性的差异，可能与键的扭转或弯曲应力相关。

2.3 量子化学成键特征

2.3.1 理论基础：键临界点(BCP)与键中点(BMP)近似

化学键的本质在于电子在原子间的重新分布与共享。为了直接探测量子力学层面的成键信息，本节特征直接从第一性原理计算获得的电子结构数据中提取。

根据Bader的"分子中原子的量子理论"(QTAIM)，化学键的电子结构性质，最适宜在其**键临界点 (Bond Critical Point, BCP)** 处进行分析。BCP被严格定义为在三维电子密度场 $\rho(\mathbf{r})$ 中，一阶导数（梯度）为

零，且Hessian矩阵恰好具有一个正特征值和两个负特征值的点（即特征值符号为 $(+1, -1, -1)$ 的鞍点）。BCP是连接两个原子核的电子密度"山脊"上的鞍点。

然而，在三维标量场中精确地定位BCP需要迭代式的数值优化，这在高通量计算中可能成为效率瓶颈。因此，我们采用一个被广泛接受且计算上极为高效的近似：使用**键中点 (Bond Midpoint, BMP)** $\mathbf{r}_{\text{mid}} = (\mathbf{r}_A + \mathbf{r}_B)/2$ 作为BCP的替代。该近似的理论合理性基于：

1. **对称性论证**：对于任何同核键（如Si-Si），由于体系具有反演对称性，BCP与BMP严格重合。
2. **微扰分析**：对于异核键，BCP相对于BMP的偏移量与成键原子的电负性差成正比。对于绝大多数化学键，该偏移量远小于键长，因此BMP仍然位于电子密度鞍点的核心区域。
3. **数值验证**：已有研究表明，在BMP处和在精确BCP处计算的电子密度、ELF等标量值的差异通常小于5%，这一精度足以满足大规模材料筛选和机器学习建模的需求。

2.3.2 具体量子化学特征

2.3.2.1 键中点电子密度 (`bond_midpoint_density`)

该特征定义为在键中点处的总电子密度值：

$$\rho_{\text{mid}} = \rho(\mathbf{r}_{\text{mid}})$$

物理意义: 根据QTAIM理论，在BCP处的电子密度 ρ_{BCP} 与化学家直觉中的**键级 (bond order)** 存在强烈的、通常是幂律的正相关关系。高密度的 ρ_{mid} 值表示在该区域有显著的电子积累，对应于更强的共价相互作用（如双键、三键）；而低密度值则对应于较弱的相互作用，如离子键或范德华力。

2.3.2.2 键中点电子局域化函数 (`bond_midpoint_elf`)

该特征定义为在键中点处的电子局域化函数值：

$$\text{ELF}_{\text{mid}} = \text{ELF}(\mathbf{r}_{\text{mid}})$$

物理解释: ELF是一个描述电子局域化程度的无量纲标量。在键中点：

- $\text{ELF}_{\text{mid}} \approx 1$: 对应于高度局域化的电子对，是典型的**共价键**特征。
- $\text{ELF}_{\text{mid}} \approx 0.5$: 表示该处的电子行为类似于均匀电子气，是**金属键**的特征。
- $\text{ELF}_{\text{mid}} \to 0$: 表示该处电子密度极低，通常对应于**离子键**或原子间的空隙。

2.3.2.3 键中点电子密度拉普拉斯 (`laplacian_of_density_at_midpoint`)

该特征定义为在键中点处电子密度的拉普拉斯算符的值：

$$\nabla^2\rho_{\text{mid}} = \nabla^2\rho(\mathbf{r}_{\text{mid}}) = \left(\frac{\partial^2\rho}{\partial x^2} + \frac{\partial^2\rho}{\partial y^2} + \frac{\partial^2\rho}{\partial z^2}\right)\Bigg|_{\mathbf{r}_{\text{mid}}}$$

QTAIM化学键分类原理: 电子密度拉普拉斯的符号是QTAIM理论中用以区分不同类型化学相互作用的核心判据：

1. **共享相互作用 (共价键)**: $\nabla^2\rho < 0$。负值意味着电子密度在BCP（或BMP）区域是**局域集中**的，电荷被吸引并汇聚到键区，形成共享电子对。
2. **闭壳层相互作用 (离子键、范德华键)**: $\nabla^2\rho > 0$。正值意味着电子密度在BCP（或BMP）区域是**局域排空**的，电荷主要被束缚在各自的原子盆(atomic basin)内部，原子间相互作用主要通过静电力或色散力。

因此，该特征的符号为我们提供了一个基于第一性原理的、对化学键类型进行自动分类的强大工具。

2.4 融合量子-几何特征

2.4.1 理论动机与创新性

本类特征代表了材料特征工程的前沿探索，旨在捕获几何构型与量子效应之间在化学键尺度上的非线性耦合关系。传统方法往往将几何信息（如键长、键角）与电子结构信息（如电荷、态密度）作为独立的描述符输入模型，这在本质上忽略了它们在真实材料中通过化学键这一媒介所产生的强耦合特性。我们通过构造几何不变量与量子化学标量的乘积或更复杂的函数形式，创建能够同时反映结构与电子相互作用的、信息密度更高的复合特征。

2.4.2 量子加权李括号 (`quantum_weighted_lie_bracket`)

2.4.2.1 物理图像与理论基础

在复杂的晶体环境中，一个化学键两端的原子，其各自的局部结构不对称性可能表现出不同的"方向"和"大小"。当这些不对称性与键上的电荷转移同时存在时，将产生复杂的电荷-结构耦合效应，这在理解铁电性、压电性和许多结构相变的微观机制中至关重要。

2.4.2.2 数学构造

设键A-B两端原子的局部结构不对称矢量分别为 $\mathbf{V}_{\text{struct}}(A)$ 和 $\mathbf{V}_{\text{struct}}(B)$（定义见1.4.2节），Bader电荷分别为 q_A 和 q_B。量子加权李括号被定义为：

$$\mathcal{L}_{\text{bracket}} = |q_A - q_B| \cdot \|\mathbf{V}_{\text{struct}}(A) \times \mathbf{V}_{\text{struct}}(B)\|$$

物理解释:

- **叉乘项**: $\|\mathbf{V}_A \times \mathbf{V}_B\| = \|\mathbf{V}_A\| \, \|\mathbf{V}_B\| \, |\sin\phi|$，其中 ϕ 是两个不对称矢量间的夹角。在几何上，此项等于由两个不对称矢量张成的平行四边形的面积，它量化了键两端局部不对称性的"扭转程度"或"几何不兼容性"。
- **李代数关联**: 在李代数 $\mathfrak{so}(3)$ 中，三维向量的叉乘运算正是其**李括号 (Lie Bracket)** 运算：$[\mathbf{u}, \mathbf{v}] = \mathbf{u} \times \mathbf{v}$。李括号的非零性是衡量两个代数元素（在此对应两个无穷小旋转）非对易性的标

志。因此，该叉乘项的物理内涵可以被理解为，由A、B两端局部环境所定义的对称性破缺的"方向"是相互"别扭"的，无法通过一个共同的对称操作来消除。

- **电荷权重**: 差分Bader电荷 $\left|q_A - q_B\right|$ 代表了键A-B上的净电荷转移量。它作为权重因子，体现了电荷转移对这种几何扭转的调制或放大作用。一个几何上高度扭转的键，如果同时伴随着显著的电荷转移，那么它对体系的极化和应力贡献将尤为突出。

2.4.3 电子密度加权距离（`density_weighted_distance`）

2.4.3.1 理论动机

在多原子体系中，并非所有化学键对材料整体性质的贡献都是等价的。共价性更强的键通常具有更高的电子云重叠、更强的方向性和更大的力常数，因此在决定材料的机械、电子和光学性质方面起着更为重要的主导作用。传统的几何键长 d_{AB} 无法区分这种化学上的差异，需要引入电子结构信息对其进行权重修正。

2.4.3.2 数学定义与物理解释

电子密度加权距离被定义为几何键长与键中点电子密度的简单乘积：

$$d_{\text{weighted}} = \rho(\mathbf{r}_{\text{mid}}) \times d_{AB}$$

物理解释:

- **权重系数的物理意义**: 根据密度泛函理论的Hohenberg-Kohn定理，电子密度 $\rho(\mathbf{r})$ 是体系基态性质的基本变量。在键中点，$\rho(\mathbf{r}_{\text{mid}})$ 的大小直接反映了该化学键的"电子活跃度"或"共价性强度"。
  - 对于强共价键，$\rho(\mathbf{r}_{\text{mid}})$ 值较大，使得 d_{weighted} 被放大，从而在特征空间中突显其重要性。
  - 对于弱的离子键或范德华相互作用，$\rho(\mathbf{r}_{\text{mid}})$ 值很小，使得 d_{weighted} 被衰减，反映了其较小的贡献。
- **量纲分析**: $[d_{\text{weighted}}] = [\rho] \times [d] = (\text{electrons}/\text{Å}^3) \times \text{Å} = \text{electrons}/\text{Å}^2$。该量纲可以被理解为一种"面电荷密度"或沿键方向积分的"线电荷密度"，它不再是一个纯粹的几何量，而是一个携带了电子结构信息的、具有化学敏感性的长度度量。

通过这种方式， `density_weighted_distance` 实现了从纯几何描述符向"化学敏感"描述符的深刻转变，为机器学习模型区分不同化学键的贡献提供了更丰富的物理信息。

第三章：2-单纯形特征 - 三体相互作用

3.1 理论背景与物理动机

在单纯同调的几何层次中，2-单纯形是由三个顶点（0-单纯形）及其连接的边（1-单纯形）所构成的最简单的面（三角形）。在晶体化学的语境下，其自然对应物是由三个原子A、B、C构成的**三角形构型**。这一构型是描述体系中超越简单两体成对作用的、最基础的**三体相互作用**的载体。

三体相互作用在决定材料的许多关键性质方面扮演着核心角色。任何依赖于原子间特定角度关系的物理化学现象，其本质都源于三体相互作用。这些现象包括：

1. **键角弯曲势能**: 在分子动力学力场中，键角偏离其平衡位置产生的恢复力是描述振动谱和结构柔软性的关键能量项。
2. **原子轨道杂化**: VSEPR理论和分子轨道理论明确指出，原子的sp³、sp²、sp等杂化类型直接决定了其周围化学键的理想空间取向（如109.5°, 120°, 180°）。
3. **立体化学与配位几何**: 配位多面体（如四面体、八面体）的稳定性和形状，是由一系列共享中心原子的三体构型共同决定的。
4. **结构相变**: 许多由畸变驱动的相变（如Peierls畸变）都涉及到键角的协同变化。

因此，对2-单纯形进行精确的数学表征，是从成对相互作用模型迈向更真实的、能反映角效应的多体相互作用模型的关键一步。

3.2 几何与拓扑特征

3.2.1 键角（`bond_angle`）

对于以原子B为中心顶点的三原子构型A-B-C，键角 θ_{ABC} 定义为从B点出发指向A和C的两个键向量 $\mathbf{v}_{BA} = \mathbf{r}_A - \mathbf{r}_B$ 和 $\mathbf{v}_{BC} = \mathbf{r}_C - \mathbf{r}_B$ 之间的夹角。其数学表达式通过向量的点积定义得到：

$$\theta_{\mathrm{ABC}} = \arccos\left(\frac{\mathbf{v}_{BA} \cdot \mathbf{v}_{BC}}{\|\mathbf{v}_{BA}\| \, \|\mathbf{v}_{BC}\|} \right)$$

物理化学意义: 键角是原子杂化状态和局部电子构型的最直接几何反映。测量的键角与理论化学预测的理想值（如sp³的109.5°）的偏差，直接量化了由晶格应力、非键相互作用或电子效应（如孤对电子的排斥）引起的**角应变 (angular strain)**，这是评估结构稳定性的一个重要指标。

3.2.2 三角形几何描述符

3.2.2.1 三角形面积 (`triangle_area`)

由A, B, C三点构成的三角形面积，提供了量化这三个原子偏离共线程度的直接度量。利用向量叉积的几何意义（其模等于由两向量张成的平行四边形的面积），三角形面积的计算公式为：

$$\mathcal{A}_{\triangle} = \frac{1}{2} \|\mathbf{v}_{BA} \times \mathbf{v}_{BC}\| = \frac{1}{2} \|\mathbf{v}_{AB} \times \mathbf{v}_{AC}\|$$

几何与物理解释:

- $\mathcal{A}_{\triangle} \to 0$: 意味着三点接近共线（键角接近0°或180°）。这个极限情况在识别一维链状结构或线性分子时非常有用。
- $\mathcal{A}_{\triangle} > 0$: 形成一个真正的三角形构型。面积的大小结合键长信息，可以进一步用于分析三角形的形状（例如，是等边、等腰还是不等边）。

此特征在识别和区分材料的结构基元（motif）和维度特征（如1D链、2D层、3D网络）中具有重要作用。

3.2.2.2 三角形周长 (`triangle_perimeter`)

周长是描述三原子团簇整体尺寸的最简单特征：

$$P_{\triangle} = d_{AB} + d_{BC} + d_{CA} = \|\mathbf{r}_A - \mathbf{r}_B\| + \|\mathbf{r}_B - \mathbf{r}_C\| + \|\mathbf{r}_C - \mathbf{r}_A\|$$

尺寸效应的物理意义: 周长作为三原子体系的特征长度，其变化直接反映了局部结构的膨胀或收缩。这与多种物理现象密切相关，例如：

- **热膨胀**: 温度升高导致平均键长增加，周长相应增大。
- **压力效应**: 外部施加的静水压力导致结构压缩，周长减小。
- **化学取代/掺杂**: 将体系中的原子替换为尺寸不同（离子半径不同）的原子，会直接导致包含该原子的所有三体构型的周长发生系统性变化。

3.3 层次化派生特征

3.3.1 理论框架：从边界到整体

再次遵循单纯同调的层次化设计原理，一个2-单纯形（三角形A-B-C）的性质，应当可以由其边界元素的性质来系统性地派生。其边界由三个0-单纯形（顶点A, B, C）和三条1-单纯形（边AB, BC, CA）构成。我们通过对这些低阶单纯形的已知特征进行统计分析（如计算平均值、标准差、极差等），来构造2-单纯形的派生特征。这种方法确保了特征体系在不同几何尺度上的数学一致性和物理可解释性。

3.3.2 基于原子属性的统计特征

对于构成三角形的三个原子A, B, C的任意一个0-单纯形特征 $\mathcal{F}^{(0)}$（如电负性、Bader电荷等），我们可以定义其在三原子体系中的统计描述符：

- **平均值 (`avg_...`)**: 描述三原子区域的平均化学/物理环境。

$$\overline{\mathcal{F}}_{\triangle}^{(0)} = \frac{1}{3}\left(\mathcal{F}_{A}^{(0)} + \mathcal{F}_{B}^{(0)} + \mathcal{F}_{C}^{(0)}\right)$$

  具体示例: `avg_atomic_bader_charge`， `avg_electronegativity`.
- **极差 (`range_...`)**: 描述该属性在三原子间的不均匀性或分散程度。

$$\Delta\mathcal{F}_{\triangle}^{(0)} = \max(\mathcal{F}_{A}^{(0)}, \mathcal{F}_{B}^{(0)}, \mathcal{F}_{C}^{(0)}) - \min(\mathcal{F}_{A}^{(0)}, \mathcal{F}_{B}^{(0)}, \mathcal{F}_{C}^{(0)})$$

  具体示例: `range_atomic_bader_charge`， `range_electronegativity`.

3.3.3 基于键属性的统计特征

同理，对于构成三角形的三条化学键AB, BC, CA的任意一个1-单纯形特征 $\mathcal{F}^{(1)}$（如键长、键中点电子密度等），我们也可以定义其统计描述符：

- **平均值 (`avg_...`)**: 描述三体作用范围内的平均成键性质。

$$\overline{\mathcal{F}}_{\triangle}^{(1)} = \frac{1}{3}\left(\mathcal{F}_{AB}^{(1)} + \mathcal{F}_{BC}^{(1)} + \mathcal{F}_{CA}^{(1)}\right)$$

  具体示例: `avg_bond_length`， `avg_bond_midpoint_density`.
- **极差 (`range_...`)**: 描述成键性质的差异性，是衡量三体构型几何或化学畸变的关键指标。

$$\Delta\mathcal{F}_{\triangle}^{(1)} = \max(\mathcal{F}_{AB}^{(1)}, \mathcal{F}_{BC}^{(1)}, \mathcal{F}_{CA}^{(1)}) - \min(\mathcal{F}_{AB}^{(1)}, \mathcal{F}_{BC}^{(1)}, \mathcal{F}_{CA}^{(1)})$$

  具体示例: `bond_length_range`， `bond_midpoint_density_range`.

`bond_length_range` 的物理应用：Jahn-Teller畸变的量化

键长极差特征 Δd_{range} 对于识别和量化由**Jahn-Teller效应**引起的几何畸变特别有效。在某些高对称性的配位环境（如八面体）中，如果中心离子的d轨道存在电子简并，体系会自发地发生几何畸变以破除简并、降低总能量。这种畸变通常表现为原本等价的几根键（例如，八面体的轴向键和赤道键）的键长发生分裂，导致 Δd_{range} 从零变为一个显著的非零值。因此，通过计算包含中心离子的所有2-单纯形的 Δd_{range}，并对其进行统计分析（如取最大值或平均值），就可以构造一个量化整个配位多面体Jahn-Teller畸变程度的强大描述符。

3.4 融合辛几何特征

3.4.1 理论背景与创新动机：将相空间动力学引入晶体化学

本节旨在引入一个具有突破性的理论创新：将**辛几何 (Symplectic Geometry)** 的核心思想首次引入晶体结构特征的构造中。辛几何是哈密顿力学和经典相空间动力学的标准数学语言，其核心是定义了一个在时间演化中守恒的、被称为**辛2-形式 (symplectic 2-form)** 的数学结构，它在相空间中定义了守恒的"面积元"。

3.4.1.1 辛几何的物理类比

在经典哈密顿力学中，一个系统的状态由其在$2n$维**相空间**中的一个点来描述，该相空间由n个广义坐标$\{q_i\}$和n个共轭动量$\{p_i\}$张成。辛2-形式ω在这个空间中定义为$\omega = \sum_{i=1}^{n} dq_i \wedge dp_i$。

我们在此提出一个深刻的物理类比，以构建一个局部的、静态的"广义相空间"来描述三体相互作用：

- **"广义坐标"** q: 由体系的**几何自由度**来扮演，例如，从中心原子指向邻居的键向量。
- **"共轭动量"** p: 由体系的**化学或电子自由度**来扮演，例如，由电荷转移或电负性差异驱动的"化学势梯度"。

这种类比的物理基础在于：在任何真实的物理过程（如结构弛豫、声子振动、相变）中，几何结构的变化（坐标q的变化）与电子云的重新分布（动量p的变化）总是**共轭**发生、相互耦合的。因此，通过在静态的晶体"快照"中构造一个辛几何形式的特征，我们有望捕捉到这种潜在的、动态的**几何-电子耦合倾向**。

3.4.2 局部相空间通量 (`local_phase_space_flux`)

3.4.2.1 数学构造与推导

我们考虑以原子B为顶点的三原子系统A-B-C。我们在此局部环境中定义一对共轭的矢量：

1. **几何向量 (Coordinate-like Vector)** $\mathbf{G}$：代表空间坐标自由度，我们选择其中一个键向量来扮演此角色。

$$\mathbf{G} = \mathbf{v}_{BA} = \mathbf{r}_A - \mathbf{r}_B$$

2. **化学向量 (Momentum-like Vector)** $\mathbf{C}$：代表共轭的"动量"自由度。我们将其构造为由另一个键B-C上的电荷转移所驱动的、有方向的"化学动量"。

$$\mathbf{C} = (q_C - q_B) \cdot \frac{\mathbf{v}_{BC}}{\|\mathbf{v}_{BC}\|}$$

   该向量的**物理意义**是：其方向由键B-C的方向给定，其大小$|q_C - q_B|$正是键B-C上净的Bader电荷转移量。它代表了沿着键B-C方向的"电荷流"的强度。

现在，我们定义**局部相空间通量** $\mathcal{S}_{\text{flux}}$ 为这两个共轭矢量所张成的"相空间面积元"，在三维空间中，这自然地通过两个向量的**叉积**的模来量化：

$$\mathcal{S}_{\text{flux}} = \|\mathbf{G} \times \mathbf{C}\| = \left\| \mathbf{v}_{BA} \times \left((q_C - q_B) \frac{\mathbf{v}_{BC}}{\|\mathbf{v}_{BC}\|} \right) \right\|$$

利用向量叉积的齐次性（$\|\mathbf{u} \times k\mathbf{v}\| = |k| \|\mathbf{u} \times \mathbf{v}\|$），我们可以将标量因子 $|q_C - q_B|$ 提出：

$$\mathcal{S}_{\text{flux}} = |q_C - q_B| \cdot \left\| \mathbf{v}_{BA} \times \frac{\mathbf{v}_{BC}}{\|\mathbf{v}_{BC}\|} \right\|$$

再利用叉积的几何定义（$\|\mathbf{u} \times \mathbf{v}\| = \|\mathbf{u}\| \|\mathbf{v}\| |\sin\theta|$，其中 θ 是向量夹角），上式中的叉积项可以被展开：

$$\left\| \mathbf{v}_{BA} \times \frac{\mathbf{v}_{BC}}{\|\mathbf{v}_{BC}\|} \right\| = \|\mathbf{v}_{BA}\| \cdot \left\| \frac{\mathbf{v}_{BC}}{\|\mathbf{v}_{BC}\|} \right\| \cdot |\sin\theta| = d_{BA} \cdot 1 \cdot |\sin\theta|$$

其中 $d_{BA} = \|\mathbf{v}_{BA}\|$ 是键B-A的长度，θ 正是A-B-C键角。

将此结果代回，我们得到了一个极为简洁、优美且物理意义清晰的最终表达式：

$$\mathcal{S}_{\text{flux}} = |q_C - q_B| \cdot d_{BA} \cdot |\sin\theta|$$

3.4.2.2 物理解释与应用

最终表达式的物理分析：局部相空间通量是三个基本物理量的乘积：

1. **电荷转移强度** $|q_C - q_B|$: 量化了键B-C上的电荷不平衡程度。
2. **几何力臂** d_{BA}: 提供了该相互作用的空间尺度。
3. **几何耦合因子** $|\sin\theta|$: 描述了两个键向量的几何耦合效率。

物理图像: 此特征可以被直观地理解为：在原子B处，沿着B-C方向的"电荷流"或"电荷动量"，对B-A这条"杠杆"所产生的**扭矩的模**。当两个键共线时（$\theta \to 0^\circ$ 或 180°），$\sin\theta \to 0$，扭矩为零，符合物理直觉。当两个键垂直时（$\theta = 90^\circ$），$\sin\theta = 1$，扭矩效应最强。

量纲与应用:

- **量纲分析**: $[\mathcal{S}_{\text{flux}}] = [\text{charge}] \times [\text{length}] = e \cdot \text{Å}$。该量纲与**电偶极矩**的量纲完全相同。因此，该特征可以被看作是一种广义的、由三体相互作用诱导的"电荷-长度矩"。
- **应用前景**: 这种量化几何-电子耦合的特征，有望在描述需要多体相互作用和电荷-晶格耦合的复杂物理现象中发挥关键作用，例如：
  - **铁电性与压电性**: 高的平均通量值可能指示材料具有强的局域极化倾向和机电耦合效应。
  - **Peierls畸变**: 在一维链状材料中，该特征有望捕捉到电荷密度波与晶格畸变之间的耦合强度。
  - **声子非谐性**: 通量的变化可能与声子谱中的非谐性效应相关。

此特征代表了将来自不同物理领域的思想（哈密顿动力学）与晶体化学问题相结合的范式创新，为从静态结构中挖掘潜在的动力学和响应性质提供了全新的数学工具。

第四章：全局特征 - 晶胞的宏观性质

4.1 理论框架：从局部到全局的涌现

在单纯同调的层次化理论框架中，全局特征居于最高层级，旨在描述整个晶胞作为一个统一体系的宏观本征属性。这些特征不再局限于个别的原子或化学键，而是反映了所有微观组分（0-、1-、2-单纯形等）相互作用后所**涌现 (emergent)**出的集体行为。

从物理学的角度看，任何材料的宏观可观测性质（如密度、对称性、能带隙、弹性模量等），都是其内部无数微观相互作用在热力学和统计力学规律支配下的平均化或协同表现。因此，全局特征是连接微观作用机理与宏观材料功能的关键桥梁。在数学上，这意味着一个全局特征 $\mathcal{F}_{\text{global}}$ 应当是所有局部特征集合 $\{\mathcal{F}_{\text{local}}\}$ 的一个泛函：

$$\mathcal{F}_{\text{global}} = \mathcal{G}[\{\mathcal{F}_{\text{local},i}\}]$$

其中 $\mathcal{G}$ 代表一个复杂的、可能非线性的映射关系，例如平均、求和、取极值或更复杂的统计矩运算。本章构建的全局特征，正是对这种映射关系在不同物理维度上的具体实现。

4.2 基础几何与拓扑特征

这类特征直接源于对晶体周期性结构的几何分析和原子排布的统计，构成了材料最基本的"宏观身份"。

4.2.1 结构紧密度与尺寸

- **晶胞体积 (`volume`)**: $V_{\text{cell}} = |\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c})|$，晶格基矢 $\mathbf{a}, \mathbf{b}, \mathbf{c}$ 构成的平行六面体体积。
- **化学式单元数 (`num_formula_units`)**: 晶胞中包含的化学式单元的数量 N_{FU}。
- **每化学式单元体积 (`volume_per_formula_unit`)**: $V_{\text{FU}} = V_{\text{cell}} / N_{\text{FU}}$。这是一个归一化的体积度量，消除了因晶胞选取不同（如原始胞或超胞）导致的体积差异，是比较不同材料结构紧密度的公平指标。
- **原子堆积因子 (`packing_fraction`)**:

  $$\eta = \frac{V_{\text{atoms}}}{V_{\text{cell}}} = \frac{\sum_{i=1}^{N_{\text{atoms}}} \frac{4}{3}\pi r_i^3}{V_{\text{cell}}}$$

  其中 r_i 是第 i 个原子的离子或共价半径。该特征衡量原子在晶胞中的空间填充效率，直接反映结构的致密程度。

4.2.2 结构各向异性

- **晶格各向异性比 (`lattice_anisotropy`)**:

$$R_{\text{aniso}} = \frac{\max(a, b, c)}{\min(a, b, c)}$$

其中 a, b, c 是晶格常数。该比值量化了晶胞形状的各向异性程度。$R_{\text{aniso}} \approx 1$ 对应各向同性较强的立方或六方晶系，而 $R_{\text{aniso}} \gg 1$ 则表明晶胞在某个方向上被显著拉伸或压缩，对应于正交、单斜等低对称性晶系。

4.2.3 体相结构各向异性指数 (`bulk_anisotropy_index`)

4.2.3.1 物理动机与理论背景

当面临计算资源有限且需与高通量实验快速迭代的场景时，通过对每个可能的晶面进行高成本的DFT计算来获得物理加权平均的表面性质特征（如表面能、表面对称性破缺）变得不切实际。为此，我们提出一个全新的、计算成本极低的"代理"特征，其核心思想是：**一个材料会形成何种能量各异的表面，其物理"基因"已经蕴含在其体相晶格的几何各向异性之中。**

各向同性的晶体（如FCC金属），其原子在空间中均匀分布，不同方向的表面性质差异不大。相反，各向异性晶体（如层状材料）的成键在不同方向上存在巨大差异，其表面能和表面性质也必然随晶面方向而显著变化。因此，体相的**结构各向异性程度**，可以作为**表面性质各向异性程度**的一个高效、廉价的代理描述符。该特征旨在通过分析体相晶格中原子局部环境的平均"形状"，来定量捕捉这种内在的结构各向异性。

4.2.3.2 数学构造

该特征的构造过程分为三步：构建局部环境张量、全局平均化、以及计算最终的各向异性指数。

1. **构建局部环境张量**：对于晶格中的每一个原子，我们考察其周围的近邻原子。类似于1.4.3节中定义的局部环境结构张量，我们将从中心原子指向其近邻原子的相对向量 $\mathbf{v}_i$ 构建一个二阶**局部结构张量** $\mathbf{M}_{\text{local}}$:

$$\mathbf{M}_{\text{local}} = \sum_{i}^{\text{neighbors}} \mathbf{v}_i \otimes \mathbf{v}_i^T = \sum_{i} \begin{pmatrix} v_{ix}^2 & v_{ix}v_{iy} & v_{ix}v_{iz} \\ v_{iy}v_{ix} & v_{iy}^2 & v_{iy}v_{iz} \\ v_{iz}v_{ix} & v_{iz}v_{iy} & v_{iz}^2 \end{pmatrix}$$

该张量描述了单个原子周围邻居的几何分布形状。
2. **全局平均张量**：为了得到一个描述整个材料的宏观特征，我们将晶格中所有**不等价原子**的局部结构张量进行平均，得到一个**全局平均结构张量** $\bar{\mathbf{M}}$。

$$\bar{\mathbf{M}} = \frac{1}{N_{\text{unique}}} \sum_{j=1}^{N_{\text{unique}}} \left(\frac{1}{N_{j,\text{neigh}}} \sum_{k=1}^{N_{j,\text{neigh}}} \mathbf{v}_{jk} \otimes \mathbf{v}_{jk}^T \right)$$

其中 N_{unique} 是不等价原子的数量，第二重求和是对第 j 个不等价原子的 $N_{j,\text{neigh}}$ 个邻居进行的。这一过程提取了原子成键环境的平均几何形状信息。

3. **计算各向异性指数**：对对称的全局平均张量 $\bar{\mathbf{M}}$ 进行对角化，得到三个实数本征值 $\lambda_1, \lambda_2, \lambda_3$。这三个值描述了原子平均成键环境在空间三个相互正交的主轴上的"延展"程度。
   - 如果 $\lambda_1 \approx \lambda_2 \approx \lambda_3$，说明平均环境是球对称的，材料在结构上是**各向同性**的。
   - 如果三个值差异很大，说明平均环境是椭球状的，材料在结构上是**各向异性**的。

   我们采用这三个本征值的**归一化标准差**作为最终的各向异性指数 η：

$$\eta = \frac{\sqrt{\frac{1}{3}\sum_{i=1}^{3}(\lambda_i - \bar{\lambda})^2}}{\bar{\lambda}} = \frac{\text{StdDev}(\lambda_1, \lambda_2, \lambda_3)}{\text{Mean}(\lambda_1, \lambda_2, \lambda_3)}$$

其中 $\bar{\lambda}$ 是三个本征值的平均值。该指数是一个无量纲的标量，其值为0表示完全各向同性，值越大则各向异性越强。此计算完全基于晶体几何，无需任何DFT计算，因此极为高效。

4.2.3.3 数值实现与物理意义的讨论：为何本征值必须是实数？

在数值计算的实践中，可能会观察到求解出的本征值 λ_i 带有微小的、可忽略不计的虚部。在此，我们必须严格依据物理和数学原理，对这一现象进行澄清，并说明为何只保留实部是唯一正确的处理方式。

1. 数学保证：实对称矩阵的本征值为实数

我们的出发点是全局平均结构张量 $\bar{\mathbf{M}}$。根据其构造方式（一系列实向量外积之和的平均），$\bar{\mathbf{M}}$ 在数学上被严格保证为一个**实对称矩阵** (real symmetric matrix)。线性代数中的谱定理 (Spectral Theorem) 明确指出：**任何实对称矩阵的所有本征值都必然是实数。** 这意味着，从理论上讲，虚部是绝不可能出现的。

2. 虚部的来源：数值噪声 (Numerical Noise)

计算中实际出现的微小虚部，其来源并非物理效应，而是计算机**浮点运算的精度限制**。由于计算机无法以无限精度存储数字，重复的加法和乘法运算会引入微小的舍入误差。这些误差可能导致一个理论上完美的对称矩阵，在数值上变得极其微小地不对称（例如，矩阵元 M_{ij} 与 M_{ji} 的差异在 10^{-18} 量级）。

当一个通用的本征值求解算法（如 `numpy.linalg.eigvals` ）处理这样一个数值上"几乎"对称但非严格对称的矩阵时，它会返回数学上最通解——复数。然而，由于矩阵的非对称性极小，解的虚部也同样会极其微小（例如 $10^{-18}j$）。

3. 物理意义的阐释

- **实部的物理意义**: 张量 $\bar{\mathbf{M}}$ 的本征值的**实部**，直接对应于原子平均配位环境在三个相互正交的主轴方向上的空间分布方差。它是一个可测量的、经典的几何量，描述了原子云的"形状"是球形的、长球形的（沿某轴拉伸）还是扁球形的（在某平面延展）。**这是该特征的核心物理内涵。**
- **虚部的物理意义**: 在此经典几何的语境下，本征值的**虚部没有任何物理或化学意义**。它不对应任何可观测的物理量，而纯粹是数值计算过程中产生的噪声伪影 (artifact)。
  （注：在量子力学等其他物理领域，虚数扮演着核心角色，例如描述波函数的相位或体系的耗散。但对于当前所讨论的、描述原子核位置分布的经典结构张量，情况并非如此。）

4. 结论与正确实现

基于以上分析，我们必须将计算中出现的任何虚部视为非物理的数值噪声，并予以舍弃。在编程实践中，最严谨的做法不是简单地取实部，而是使用一个**预设了输入为对称/厄米矩阵的、并保证返回实数本征值的求解器**，例如 `numpy.linalg.eigvalsh` 。这正是我们在代码实现中所遵循的原则，它从算法层面就根除了产生非物理虚部的可能性，确保了最终特征的物理真实性。

4.3 DFT计算的电子基态属性

这类特征源于对整个晶胞进行第一性原理计算后得到的电子结构信息，描述了材料作为多电子体系的量子力学基态性质。

4.3.1 能量相关特征

- **总能量 (`total_energy`)**: 体系的DFT基态总能量 E_{total}。
- **原子平均总能量 (`total_energy_per_atom`)**:

$$E_{\text{per atom}} = \frac{E_{\text{total}}}{N_{\text{atoms}}}$$

  其中 N_{atoms} 是晶胞中的原子总数。这是一个尺寸归一化的能量，是在相同的计算参数下，比较不同结构热力学稳定性的核心指标。
- **费米能级 (`fermi_level`)**: 费米能级 E_F 是占据态的最高能量，在绝对零度下严格定义为体系总能量对电子数 N 的偏导数，即电子的化学势：

$$E_F = \mu(T = 0\text{K}) = \left(\frac{\partial E_{\text{total}}}{\partial N} \right)_{T=0}$$

  它是决定材料所有电子学行为（导电性、功函数、催化活性等）的关键基准能量。

4.3.2 带隙特征

- **能带隙 (`band_gap`)**: 导带底 (Conduction Band Minimum, CBM) 与价带顶 (Valence Band Maximum, VBM) 之间的能量差, $E_g = E_{\text{CBM}} - E_{\text{VBM}}$。它是区分金属 ($E_g \leq 0$)、半导体和绝缘体 ($E_g > 0$) 的根本判据。
- **直接/间接带隙 (`is_direct_gap`)**: 判断VBM和CBM在倒易空间中是否位于同一点（k点）的布尔特征。这决定了材料的光吸收和发射效率。

4.3.3 静电势统计特征

- **平均静电势 (`electrostatic_potential_mean`)**: 整个晶胞内哈特里静电势 $\Phi(\mathbf{r})$ 的体积平均值：

$$\bar{\Phi} = \frac{1}{V_{\text{cell}}} \int_{V_{\text{cell}}} \Phi(\mathbf{r})\, d^3\mathbf{r}$$

- **静电势方差 (`electrostatic_potential_variance`)**: 静电势在晶胞内的起伏程度，量化了内部电场的不均匀性：

$$\sigma_\Phi^2 = \left\langle (\Phi(\mathbf{r}) - \bar{\Phi})^2 \right\rangle = \frac{1}{V_{\text{cell}}} \int_{V_{\text{cell}}} [\Phi(\mathbf{r}) - \bar{\Phi}]^2\, d^3\mathbf{r}$$

  一个高的方差值意味着体系内部存在强烈的、变化剧烈的内建电场。

4.4 全局高阶代数特征

4.4.1 理论动机

本类特征基于李代数、辛几何、商代数思想，从不变物理张量与全局拓扑中提取创新特征。这些特征旨在捕获传统描述符难以表达的复杂多体相互作用和对称性破缺效应。

4.4.2 商代数拓扑特征

4.4.2.1 晶体结构哈希 (`structure_hash`)

基于Weisfeiler-Lehman图同构算法生成的、对晶体拓扑结构唯一的字符串标识符。其数学构造可概括如下：

1. 将晶体结构抽象为一个图，其中原子为节点，化学键为边。
2. 每个节点的初始标签（哈希值）为其原子类型（如原子序数）。
3. 进行多轮迭代更新。在第 $k+1$ 轮，节点 v 的新哈希值 $h_v^{(k+1)}$ 由其当前的哈希值 $h_v^{(k)}$ 与其所有邻居节点 $u \in N(v)$ 在第 k 轮的哈希值构成的多重集 $\{h_u^{(k)}\}$ 共同决定：

$$h_v^{(k+1)} = \text{hash}(h_v^{(k)}, \text{multiset}\{h_u^{(k)} : u \in N(v)\})$$

4. 经过足够轮数的迭代直至所有节点的哈希值收敛，最终的全局结构哈希值由所有节点最终哈希值的有序集合生成。

拓扑意义：此哈希值是晶格拓扑连接性的"指纹"。两个结构若拥有相同的哈希值，则意味着它们在拓扑上是同构的。

4.4.2.2 威科夫位置熵 (`wyckoff_position_entropy`)

$$S_{\text{Wyckoff}} = -\sum_{i} p_i \log p_i$$

其中 p_i 是晶胞中原子占据第 i 种威科夫(Wyckoff)位置的概率，即占据该位置的原子数与晶胞总原子数的比值。

对称性意义：衡量结构中对称不等价位点分布的复杂度。高熵值表明多种威科夫位置共存，对应于更复杂的晶体结构。

4.4.2.3 晶格连通性维度 (`lattice_connectivity_dimension`)

$$D_{\text{conn}} = \frac{N_{\text{inter-cell}}}{N_{\text{total bonds}}}$$

其中 $N_{\text{inter-cell}}$ 是跨越晶胞边界的化学键数，$N_{\text{total bonds}}$ 是晶胞内化学键的总数。

拓扑意义：衡量晶格在拓扑上的连通程度。$D_{\text{conn}} \to 0$ 对应于孤立的分子晶体，而较大的值则反映了高度连通的三维网络结构。

4.4.3 李代数对称性特征

4.4.3.1 全局不对称性范数 (`global_asymmetry_norm`)

$$\mathcal{L}_{\text{global}} = \left\| \sum_{i=1}^{N_{\text{atoms}}} \mathbf{V}_{\text{struct},i} \right\|$$

其中 $\mathbf{V}_{\text{struct},i}$ 是第 i 个原子的局部不对称矢量（见1.4.2节）。

对称性意义：这是对晶胞中所有局部不对称矢量（或等效的局部电偶极矩）的矢量和。其非零值指示了整个晶胞存在一个净的极化方向，是空间反演对称性破缺的直接宏观体现。

4.4.3.2 力协方差张量不变量

对于一个未经几何优化的结构，原子会受到残余的Hellmann-Feynman力。设 $\mathbf{f}_i$ 为第 i 个原子所受的力，$\mathbf{F}$ 是一个由所有力向量构成的 $N \times 3$ 矩阵。力协方差矩阵定义为 $\mathbf{C}_F = \frac{1}{N}\mathbf{F}^T\mathbf{F}$。其不变量可作为特征：

第一不变量 (`force_covariance_invariant_1`):

$$I_1 = \mathrm{Tr}(\mathbf{C}_F) = \frac{1}{N}\mathrm{Tr}(\mathbf{F}^T\mathbf{F}) = \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{f}_i\|^2$$

第二不变量 (`force_covariance_invariant_2`):

$$I_2 = \det(\mathbf{C}_F) = \det\left(\frac{\mathbf{F}^T\mathbf{F}}{N}\right)$$

物理意义: I_1 是均方力的大小，量化了结构偏离其力学平衡态（完全弛豫状态）的总体程度。I_2 与协方差矩阵的特征值之积相关，量化了原子间残余应力场的方向各向异性。

4.4.3.3 总扭转应力 (`total_torsional_stress`)

$$\mathcal{T}_{\text{total}} = \sum_{i=1}^{N_{\text{atoms}}} \|\mathbf{V}_{\text{struct},i} \times \mathbf{f}_i\|$$

物理意义: 该特征对每个原子上由局部环境不对称性 $\mathbf{V}_{\text{struct},i}$ 和作用力 $\mathbf{f}_i$ 产生的"扭矩"进行求和。它描述了局部环境不对称性与原子受力的耦合，反映体系内部的非中心力和扭转应力的大小。

4.4.4 辛几何电子应力特征

4.4.4.1 辛电子应力张量不变量

借鉴辛几何中坐标与动量耦合的思想，我们构造一个**辛电子应力张量** $\mathbf{S}_e = \langle \mathbf{E}(\mathbf{r}) \otimes \nabla\rho(\mathbf{r}) \rangle$，其中 $\mathbf{E}(\mathbf{r}) = -\nabla\Phi(\mathbf{r})$ 是内建电场，$\nabla\rho(\mathbf{r})$ 是电子密度梯度，$\langle\cdot\rangle$ 表示晶胞体积平均。

第一不变量 (`symplectic_stress_invariant_1`):

$$J_1 = \mathrm{Tr}(\mathbf{S}_e) = \mathrm{Tr}\left(\langle \mathbf{E}(\mathbf{r}) \otimes \nabla\rho(\mathbf{r}) \rangle\right) = \langle \mathbf{E}(\mathbf{r}) \cdot \nabla\rho(\mathbf{r}) \rangle$$

第二不变量 (`symplectic_stress_invariant_2`):

$$J_2 = \det(\mathbf{S}_e) = \det\left(\langle \mathbf{E}(\mathbf{r}) \otimes \nabla\rho(\mathbf{r}) \rangle\right)$$

物理意义: J_1 量化了内部电场与电子密度梯度的平均对齐程度，反映了电场驱动电子重新分布的趋势。J_2 则量化了这种电子应力耦合的三维体积效应。

4.4.4.2 总梯度范数 (`total_gradient_norm`)

$$G_{\text{total}} = \left\langle |\nabla\rho(\mathbf{r})|^2 \right\rangle = \frac{1}{V_{\text{cell}}} \int_{V_{\text{cell}}} |\nabla\rho(\mathbf{r})|^2 \, d^3\mathbf{r}$$

物理意义：平均电子密度梯度的平方，量化了电子云的整体不均匀性。在广义梯度近似(GGA)泛函中，此项直接关联到交换相关能的修正，因此也反映了体系共价成键性的强弱。

特征总结

本理论框架构建了一个跨越四个几何层次（0-、1-、2-单纯形及全局）的、多尺度、高维度的特征体系。

- **第1章 (0-单纯形)**: 描述原子中心环境，包含基础、量子、几何及融合特征。
- **第2章 (1-单纯形)**: 描述化学键环境，包含几何、派生、量子及融合特征。
- **第3章 (2-单纯形)**: 描述三体相互作用，包含几何、派生及融合辛几何特征。
- **第4章 (全局)**: 描述晶胞宏观性质，包含几何、电子基态及高阶代数特征。

该特征体系系统性地将晶体材料的结构与物性信息编码为数学上严谨、物理上深刻的描述符，为数据驱动的材料科学研究提供了坚实的理论与计算基础。