Computer Science and Application 计算机科学与应用, 2022, 12(5), 1314-1330
Published Online May 2022 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2022.125131

Hans汉斯

基于原子间作用势的晶体结构去重算法及其应用

孔艳婷[1]，陈　品[1,2]，颜　辉[1,2]，陈志广[1,2*]

[1]中山大学计算机学院，广东 广州
[2]国家超级计算广州中心，广东 广州

收稿日期：2022年4月17日；录用日期：2022年5月16日；发布日期：2022年5月23日

摘　要

随着科学技术的发展，人们收集“大数据”的能力远远超过了分析它的能力。科学研究正经历从传统的实验观测、理论推演以及计算仿真向大数据研究转型，形成第四研究范式。材料数据信息学作为应用科学的重要分支之一，驱动着新材料的设计与发现。本文针对材料领域数据的冗余问题，提出一种基于原子间作用势的去重算法，可以有效地鉴别材料中的相同或相似结构，从而实现冗余数据的去重。实验结果表明，提出的算法与已有的材料结构去重算法相比，在准确性、鲁棒性以及计算效率各指标上均表现优异。进一步，本文利用开发的算法对材料领域三大知名数据库ICSD、CSD和COD数据库超过159万数据进行去重分析，有效地去除了101,643个相同和相似结构，并构建了开放共享的去冗余数据库，数据发布于：https://matgen.nscc-gz.cn/。

关键词

材料大数据，结构相似度，数据去重

Crystal Structure Deduplication Algorithm Based on Interatomic Potential and Its Applications

Yanting Kong[1], Pin Chen[1,2], Hui Yan[1,2], Zhiguang Chen[1,2*]

[1]School of Computer, Sun Yat-sen University, Guangzhou Guangdong
[2]National Supercomputer Center in Guangzhou, Guangzhou Guangdong

Received: Apr. 17th, 2022; accepted: May 16th, 2022; published: May 23rd, 2022

*通讯作者。

文章引用: 孔艳婷, 陈品, 颜辉, 陈志广. 基于原子间作用势的晶体结构去重算法及其应用[J]. 计算机科学与应用, 2022, 12(5): 1314-1330. DOI: 10.12677/csa.2022.125131

Abstract

With the development of science and technology, the ability of people to collect big data exceeds the ability to analyze it. Scientific research is undergoing a transformation from traditional experimental observation, theoretical deduction and simulation research to big data research, forming the fourth research paradigm. As one of the important branches of applied science, materials data informatics drives the design and discovery of new materials. Aiming at the problem of redundancy in the field of materials, this paper proposed a deduplication algorithm based on the interatomic potential, which can effectively identify the same and similar structures, thereby realizing the deduplication of redundant data. The experimental results show that this algorithm has excellent performance in accuracy, robustness and computational efficiency compared with the existing algorithms. Further, this paper used the proposed algorithm to deduplicate more than 1.59 million data in ICSD, CSD and COD databases, which effectively remove 101,643 same and similar structures, and built an open and shared de-redundancy database. The data was published at: https://matgen.nscc-gz.cn/.

Keywords

Material Big Data, Structure Similarity, Data Deduplication

1. 引言

1.1. 研究背景

随着数据的爆发性增长以及计算能力的不断提高，人们认识事物的思维方式和研究方法已经逐步从实验观测、理论推演、计算仿真发展为第四种研究新模式——“数据探索”[1]。新模式以数据驱动为主导，建立了数据之间的关系，带来了颠覆性的效果，也是引发社会变革的根本性原因之一。海量数据的到来带来了机遇的同时也在考验着现如今对大数据收集、分析、管理和访问等的处理能力。材料基因工程的提出和发展，促进了多学科交叉领域的发展，也强化了计算机应用范围，形成了材料领域学科的重要发展方向之一。而材料数据作为材料基因工程的三大支撑工具之一[2]，将计算机技术应用于材料领域，针对材料数据的分析和处理过程进行优化，将有利于推进材料领域的科研发展。然而由于实验仪器设备、实验操作人员的习惯以及实验观测环境等因素，针对同一材料会有不同的记录结果，从而导致大量冗余材料数据，因此有必要对冗余数据进行去重，减少其空间占用和计算的成本。

1.2. 研究现状

冗余数据的分析和处理一直都是一个重要的研究课题。一方面，通过冗余数据的分析和处理技术，一定程度上可以提高空间的利用率，另一方面，还可以降低传输及计算的成本。目前，计算机领域上冗余数据的缩减主要通过数据压缩、差分编码以及重复数据删除三种典型的技术[3]。三种冗余数据的缩减技术本质上都是通过检测冗余数据的方式并采取更短的指针实现数据的缩减。冗余数据的分析处理技术则可以分为两类：一是相同数据的检测技术，包括了完全文件检测技术、FSP 技术、CDC 技术以及

Slidingblock 技术等等[4]。第二类是相似数据的检测和编码技术，可以通过 shingle 技术、bloomfilter 技术或者模式匹配等技术计算指纹相似性[5]，再使用 delta 编码技术进行数据压缩或直接删除。然而计算机领域中的数据去重更注重于数据物理层面上的相同或相似，而材料科学领域的去重则更注重于数据逻辑层面上的相同或相似。因此，传统领域上的数据去重技术不完全匹配于结构数据的去重，在对结构材料数据进行去重时应当结合材料结构数据特点，对传统数据去重技术进行调整和完善。

冗余材料数据的去除通常包括以下步骤：数据特征提取、指纹计算、数据删除。在材料科学领域的应用上已有多种识别结构相似性的算法，主要从以下几个方面对其指纹进行计算：1) 原子基本信息的描述。包括径向分布信息[6]、光谱衍射信息[7]、键合信息[8]和力场及键合模式[9]等。2) 结构拓扑信息的描述[10]。3) 结构间点模式映射[11]。这些指纹计算方式在一定程度上都能表征相应晶体结构并计算出结构间的相似性。但这些算法依旧存在着准确率不高、鲁棒性低、计算效率低下等问题。

2. 基于原子间作用势的晶体结构去重算法设计

2.1. 相关知识

本文中所研究的材料对象为晶体结构，为更好地了解晶体结构的描述方式以及去重算法的原理，首先对涉及的相关知识做简单的介绍：

由于晶体结构具有周期性，一个晶体结构通常使用晶格及晶格内原子的位置进行描述。对晶格的描述通常由三个晶格矢量 A、B、C 以及中间角度 α、β、γ 给出，晶格内原子位置则使用笛卡尔坐标或者分数坐标描述。

晶格：将具有一定的原子群使用假想的线连接起来，构成的一个平行六面体的框架，即为晶格。如图 1(a)所示。

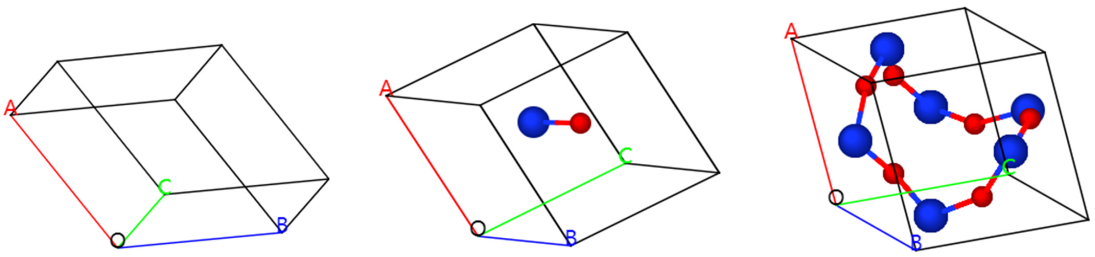
素晶胞：即基元，是晶体微观空间中最小重复单元。如图 1(b)所示。

空间群：空间群指的是晶体内部中全部对称要素的集合。

晶胞：是构成晶体的最小重复单元，包括了晶格、晶格内原子的位置以及空间群三部分信息，用于描述晶体内部的原子和粒子的分布情况，如图 1(c)所示。

超晶胞：对晶胞进行扩胞处理，即形成超胞，可认为是对晶胞的扩展。

范德华力：分子间作用力，存在于中性分子或原子间的弱碱性的电性吸引力。可以分为诱导力、色散力以及取向力。



(a) 晶格　(b) 素晶胞　(c) 晶胞

Figure 1. Schematicdiagram of lattice, primitive cell, unit cell
图 1. 晶格、素晶胞、晶胞示意图

2.2. 数据特征提取及指纹描述

数据压缩、差分编码以及重复数据删除这三种典型的缩减技术，除了数据压缩技术是以信息论为基

础，差分编码和重复数据删除均以数据块为单位分析数据的匹配情况，实现数据的缩减。材料结构数据的缩减与计算机领域的冗余数据缩减技术具有异曲同工之处。首先，它们都需要对数据进行划分，然后对数据块进行指纹计算，通过指纹判断数据的匹配程度，实现数据的缩减。然而，针对材料结构数据，由于实验仪器设备、实验操作人员的习惯以及实验观测环境等因素，会导致针对同一材料会有不同的记录结果，从而导致冗余数据。这一冗余数据在字符串层面是不同的，因此传统的方法失效，需要从数据的物理层面鉴定不同结构数据的相同以及相似程度。

本文提出了一种基于原子间范德华作用势(van der Waals, VDW)的晶体结构的去重算法，旨在提高晶体结构去重算法的适用性，降低计算成本，提高计算效率。算法通过对晶体结构信息进行提取和降维处理，基于原子间相互作用信息，构建晶体指纹，再利用加权余弦距离计算晶体结构间的相似性，最后采用聚类的方法对晶体结构进行去重。图 2 显示了算法的主要步骤：

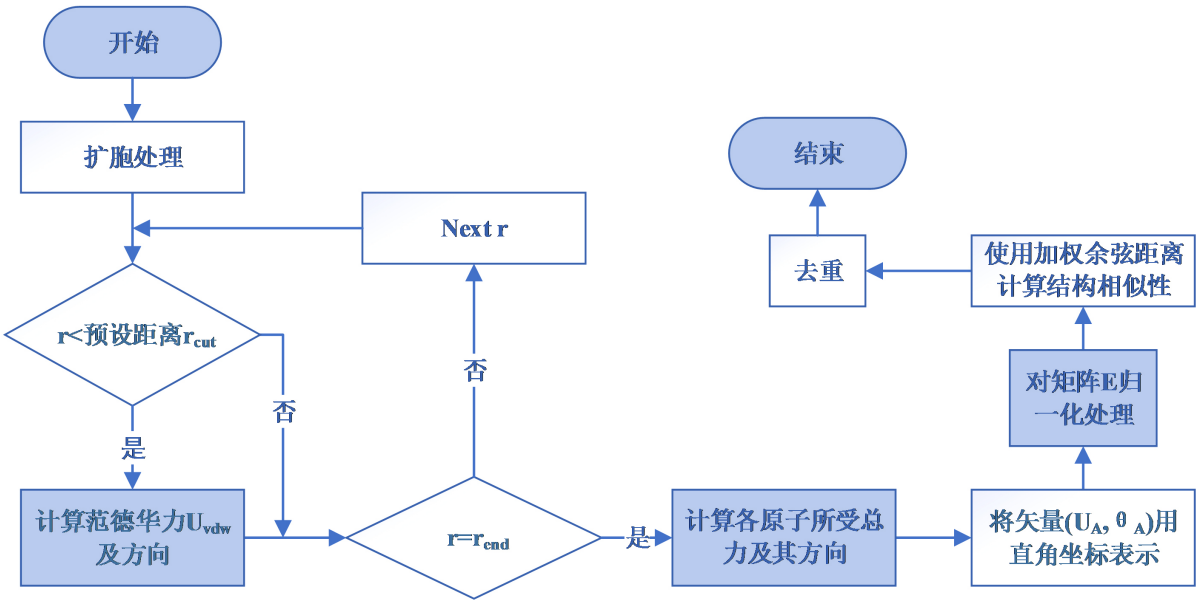


Figure 2. Flow chart of deduplication algorithm based on interatomic potential
图 2. 原子间作用势去重算法流程图

步骤一：对晶胞进行扩胞处理，形成一个 3 × 3 × 3 的超晶胞。

步骤二：对距离小于 r_{cut} 的原子对计算原子所受的范德华势 U_{vdw}。

步骤三：计算各原子受到力的总和，并用 (U_A,θ_A) 表示，其中 U_A 表示原子 A 所受原子间作用势的大小信息，θ_A 存储原子 A 所受原子间相互作用的方向信息。

步骤四：将矢量 (U_A,θ_A) 转换成直角坐标 (X_A,Y_A,Z_A) 的表示方式。

步骤五：使用三维矩阵存储各原子的受力信息 (X_A,Y_A,Z_A)，并进行归一化处理。

步骤六：利用加权余弦距离计算两个晶体结构的相似性。结果越接近 0，说明结构越相似；结果越接近 1，说明结构越不相似。

步骤七：通过聚类的方式对相似晶体去重，其中相似度大于 0.75 的结果记为两个晶体相似。

详细说明见下文。

晶体结构的相似性检测与文本、文件数据的相同或相似数据的检测不同。晶体结构使用文件的形式记录并存储，一个文件表示一个结构，简单地对结构文档内的数据进行相同或相似检测无法达到结构去

冗余的效果。由于实验中的噪音、误差，相同的结构可能会出现不相同的实验数据，对文档数据去重无法高效地做到相同晶体结构的去重。此外，由于材料数据的维度多、数据间关联性高，使用传统的相似度检测的方式容易将不相似的结构误判为相似，造成检测过度的情况。

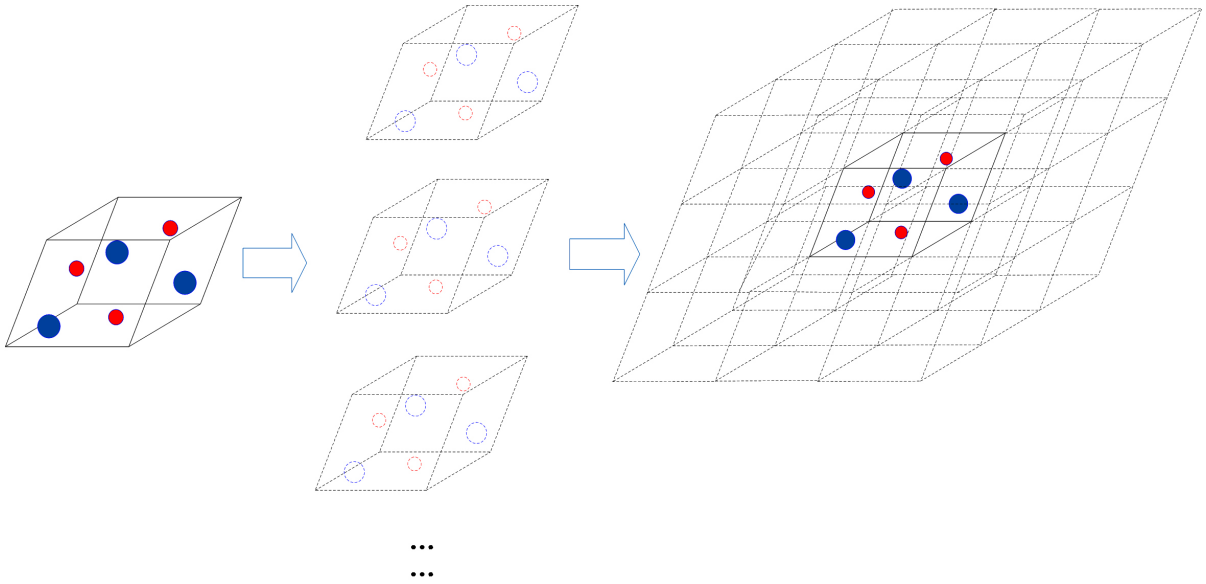


Figure 3. Schematic diagram of cell expansion
图 3. 扩胞处理示意图

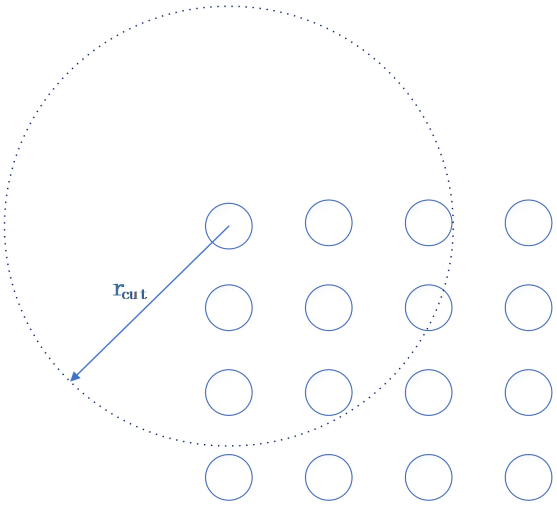


Figure 4. Scope diagram
图 4. 作用域示意图

针对材料结构数据维度多、结构关联性高的特殊性，本文提出的基于范德华势的晶体结构去重算法，对结构数据进行提取、降维，再对数据进行相似度计算、去重。下面对材料结构数据的提取和降维过程进行详细介绍。

该去重算法结合了原子类型、原子间类型、原子能、原子间相互作用力及其方向等因素，旨在提高算法的准确性、适用性和实用性。一个晶体结构由多个相同的晶胞组合而成，而结构文档数据是对晶胞

的描述文件。为了范德华势的计算更能准确地表示结构的真实情况，使提取的信息的具备唯一性和真实性。算法首先会根据结构的空间组信息、晶格信息以及原子的相关信息，对晶胞进行扩胞处理，如图 3 所示，形成一个 3 × 3 × 3 的超晶胞。

为了降低材料结构的计算成本，该算法还根据材料结构的特性，设置阈值，降低算法时间复杂度，提高计算速度。根据分子作用力与原子间的关系，当原子距离 $r \geq 10r_0 = r_{cut}$ 时，分子作用力几乎为零，其中 $r = r_0$ 处为斥力和吸引力平衡的地方。因此，算法规定仅对 $r < r_{cut}$ 的原子对计算其分子间作用力，如图 4 所示。

在计算原子间作用力时，使用具较高稳定性的 Lennard-Jones 6-12 方法进行计算，公式如下：

$$U_{vdw} = D_{AB}\left\{-2\left[\frac{d_{AB}}{r_{AB}}\right]^6 + \left[\frac{d_{AB}}{r_{AB}}\right]^{12}\right\}. \tag{1}$$

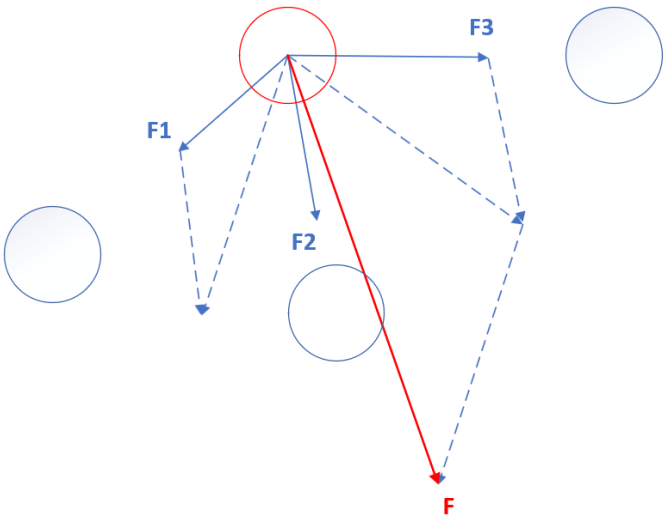


Figure 5. The sum of the forces
图 5. 力的总和

计算得到的范德华势的大小和方向，使用矢量 (U_{AB}, θ_{AB}) 进行描述，其中 A 为晶胞内的原子，B 为扩胞内的原子。$D_{AB} = \sqrt{D_A * D_B}$，$d_{AB} = \sqrt{d_A * d_B}$，$D_A$ 为原子 A 的范德华原子能，d_A 为原子 A 的范德华距离，r_{AB} 为原子 A 和 B 间的距离，$r_{cut} = \max_{AB}(10r_0)$，$U_{AB}$ 表示 A 原子与 B 原子范德华势的大小，θ_{AB} 表示范德华势的方向，A、B 原子则为两个不同的原子。

随后通过计算各原子的势能之和，记录晶胞内各原子受到的范德华势信息。根据矢量求和的方式，计算各个原子范德华势的和，如图 5 所示，并用 (U_A, θ_A) 表示。其中 U_A 表示原子 A 所受范德华势之和的大小，θ_A 表示原子 A 所受范德华势之和的方向，使用向量 (x_A, y_A, z_A) 表示。

为统一格式并便于后续计算，将 (U_A, θ_A) 的表达方式使用直角坐标描述，表示为 (X_A, Y_A, Z_A)，其中：

$$U_A = \sqrt{X_A^2 + Y_A^2 + Z_A^2}, \tag{2}$$

$$\frac{x_A}{X_A} = \frac{y_A}{Y_A} = \frac{z_A}{Z_A}. \tag{3}$$

然后将每个原子的信息 (X_i, Y_i, Z_i) 整合，并用三维矩阵 E' 进行描述。

到这里，已经将晶体结构的信息进行降维、提取。将描述晶体结构的文档信息提取至一个三维矩阵中，所得到的矩阵 E' 即可唯一代表一个晶体结构信息。算法解决了结构数据维度多、结构关联系高、无法直接通过对文本信息检测相同或相似数据的问题。成功将文档数据的检测对比转换为数学问题，算法仅仅需要分析矩阵信息，便可判断两个结构是否相似。

但是由于对结构的表示不统一，标准不一的问题，各个结构间无法直接进行比较或者计算他们的相似性，因此，还需要对矩阵 E' 进行“归一化”处理，得到矩阵 E，矩阵 E 即用于表征该晶体结构。

2.3. 指纹相似度计算及删除

对晶体结构信息进行提取、降维以后，需对晶体结构进行相似度计算。由于材料数据的影响因素比较复杂，各个因素的影响程度不一，同时考虑到各原子受到的范德华势的角度的影响，该算法使用加权余弦距离计算结构间的相似度。两个结构对应的原子 A、A' 两个原子的相似度计算公式为：

$$dist\left(A',A\right)=\frac{1}{2}w_{AA'}\left(1-\frac{F_A*F_{A'}}{\left\|F_A\right\|\left\|F_{A'}\right\|}\right), \tag{4}$$

$$w_{AA'}=\frac{\left\|F_A-F_{A'}\right\|}{\sum_N\left\|F_i-F_{i'}\right\|}. \tag{5}$$

类似，则两个结构的相似性计算公式为：

$$D_{\cos ine}=\sum_N\frac{1}{2}w_{aa'}\left(1-\frac{\sum_N E_a\left(n\right)*E_{a'}\left(n\right)}{\sqrt{\sum_N E_a^2\left(n\right)}\sqrt{\sum_N E_{a'}^2\left(n\right)}}\right), \tag{6}$$

$$w_{aa'}=\frac{\left\|E_a-E_{a'}\right\|}{\sum_N\left\|E_i-E_{i'}\right\|}. \tag{7}$$

距离 D 即为用于描述两个结构的相似程度的标准，当 D_{cosine} 的值越接近 0，则表示两个晶体结构的差异程度越高；当 D_{cosine} 的值越接近 1，则表示两个晶体结构越不相似。

最后，根据实验经验，本文对于相似度大于 0.75 的结果记为两个晶体相似，通过聚类的方式对相似晶体去重，保留其中一个最新发表的晶体结构。

3. 性能评估

为了说明基于 van der Waals (VDW)势的晶体结构去重算法的性能，本文选取了几个比较经典以及创新性较好的晶体结构去重算法进行复现，与本文提出的算法进行比较实验。所选取的算法如下：

1) 径向分布函数算法(Radial Distribution Function, RDF)：算法于 1998 年提出，是最早提出的也是最为经典晶体结构的去重算法，后来该算法也经过改善优化使得性能大幅度提高，本文复现实验参考 MAISE 软件[6]相似度计算方法，根据粒子在空间的分布机率计算结构建相似性。

2) 键特征矩阵算法(Bond Characterization Matrix, BCM) [8]：算法利用键表矩阵存储键信息，包括键矢量、键角、键类型、键长、键数目信息，用以表征整个晶体结构。该算法对晶体结构的键信息着重描述提取，以此表征整个结构。该算法详细地对某一因素的信息进行提取，具有一定的代表性。

3) 图论算法(Graph Theory, GT)：算法使用结构的拓扑信息表征整个晶体结构，每一种晶体结构对应了唯一一个或几个拓扑结构，通过对比其中的拓扑结构，判断两者是否为同一晶体结构。该算法于 2019 年提出[10]，是具备创新性的方法之一，选取该算法进行复现并对比具有代表性。

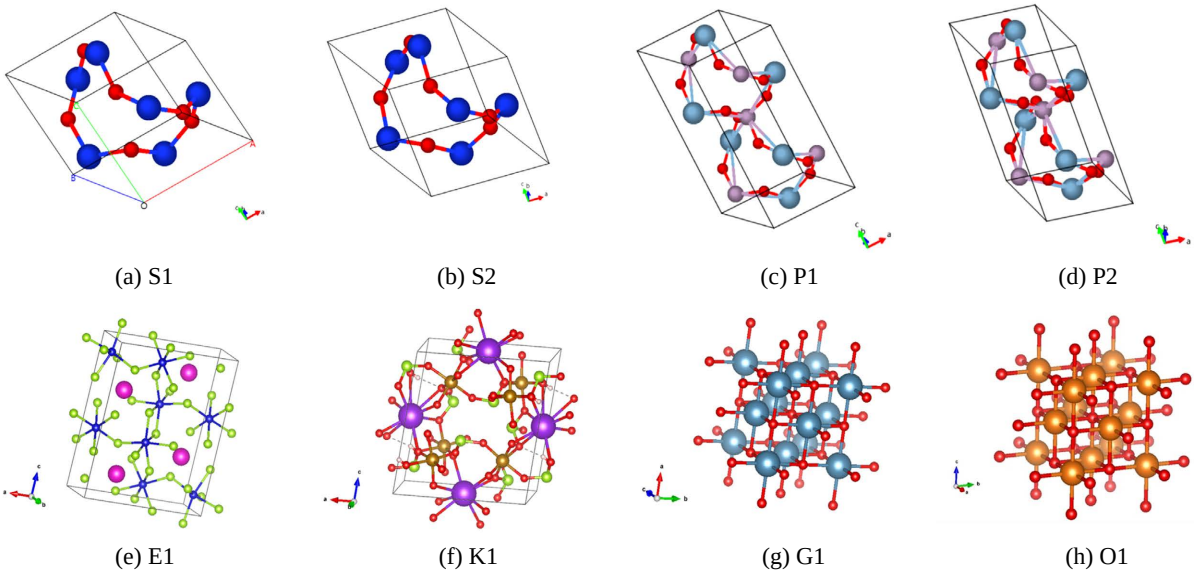
表 1 对各个算法进行了简单描述。为了说明各种算法的优缺点，实验中选取了部分最具代表性的晶

体结构进行对比实验。图 6 给出了部分晶体结构图，结构原子由不同颜色和大小表示。

实验中选取了 12 个晶体结构数据，8 种晶体类型进行对比实验，分别编号为 S1、S2、P1、P2、E1、K1、G1、O1。由于 GT 算法描述结构中原子间的连接情况，通过直接对比拓扑信息图判断结构是否相同，其相似度使用 0 和 1 表示，0 表示不相同，1 表示相同。而 RDF 算法、BCM 算法以及 VDW 算法给出 0 到 1 范围内连续的打分值，其结果越接近 1 表示该结构越相似。

Table 1. The comparison of algorithm
表 1. 算法对比

算法	表征方式	相似度/不相似度计算
RDF	径向分布信息：$RDF_{k,s1,s2}(R_n)$	点积：$\sum_{n}^{N_{bin}}\sum_{s1}^{N_{spc}}\sum_{s2}^{N_{spc}}\frac{RDF_{1,s1,s2}(R_n)RDF_{2,s1,s2}(R_n)}{n_1 n_2}$
BCM	键表征矩阵：$Q_l^{\delta_{AB}}$	欧氏距离：$\left[\sum_{\delta_{AB}}\sum_{l}\left(Q_l^{\delta_{AB},u}-Q_l^{\delta_{AB},v}\right)^2\right]^{\frac{1}{2}}$
GT	拓扑信息：$GT_A(R_n)$	异或操作：$GT_A(R_n)\oplus GT_B(R_n)$
VDW	VDW 势信息：$U_k(F_n)$	加权余弦距离：$\sum_{N}\frac{1}{2}w_{aa'}\left(1-\frac{\sum_N F_a(n)*F_{a'}(n)}{\sqrt{\sum_N F_a^2(n)}\sqrt{\sum_N F_{a'}^2(n)}}\right)$



(a) S1　(b) S2　(c) P1　(d) P2

(e) E1　(f) K1　(g) G1　(h) O1

Figure 6. Part of crystal structure
图 6. 部分晶体结构图

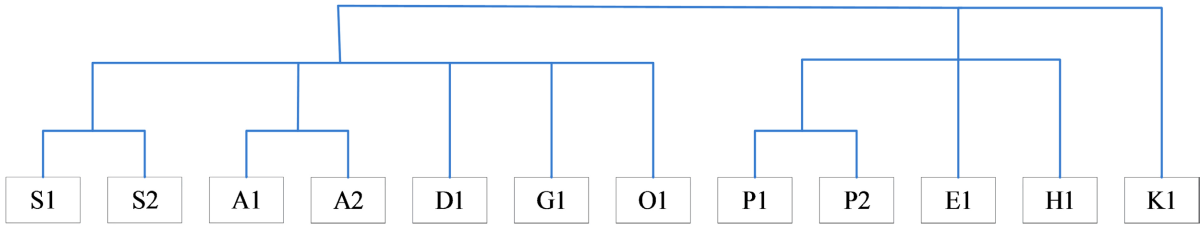


Figure 7. Result of deduplication
图 7. 去重结果

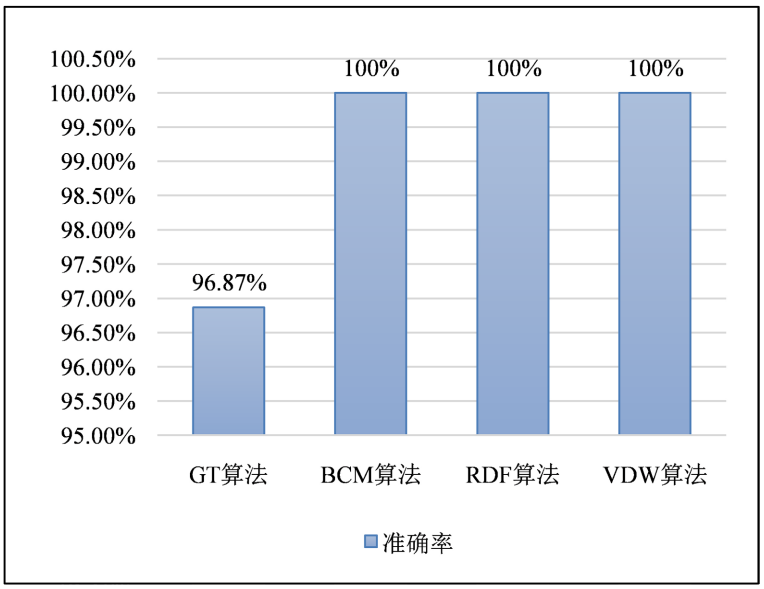


Figure 8. The accuracy of the algorithm
图 8. 各算法准确率

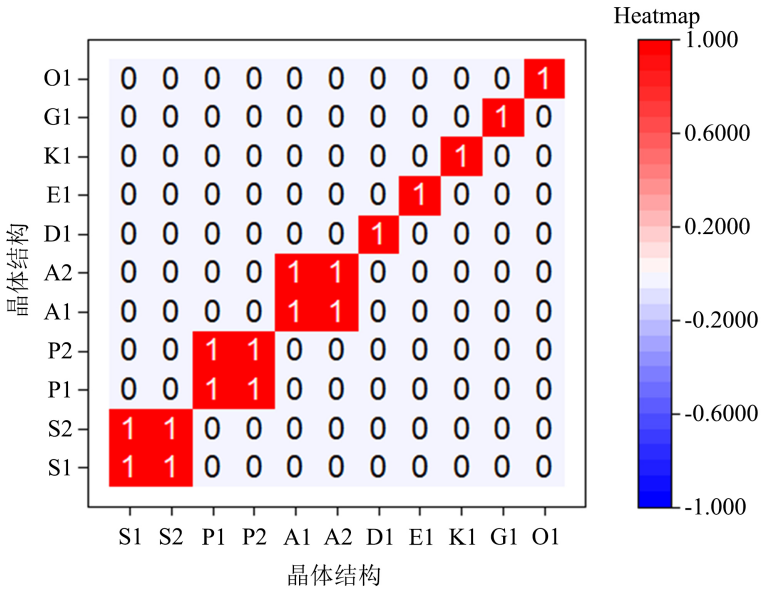


Figure 9. The result of the VDW algorithm
图 9. VDW 算法的计算结果

3.1. 准确性评估

准确性是评价一个算法性能的核心的指标。准确性通常用于描述实验中的某一实验指标或数据与真值的接近程度。本文选择了 12 个典型结构对算法的准确性进行评估，用于判断算法计算结果是否正确，能否应用于实际应用中。根据相似度的计算结果，VDW 算法将相似度大于 0.75 的结构对判定为结构相同。然后使用聚类的方式，可以对晶体结构进行去重，得出结果如图 7 所示。

分别使用 GT 算法、BCM 算法、RDF 算法以及 VDW 算法对数据进行相似性计算，并判断结构是否相似。经统计，可得各算法的去重准确率如图 8。其中 VDW 算法的计算结果可见图 9，“1”表示两个结果相同，“0”为不同。

分析统计结果，BCM 算法、RDF 算法以及 VDW 算法均能有效的识别结构是否相同，能有效的对冗余结构进行去重。而 GT 算法由于仅根据各原子见的连接情况进行判断，导致无法正确判断两个原子类型不同而原子连接情况相同的原子是否相同，例如编号为 G1 和 O1 的晶体结构。

3.2. 鲁棒性评估

鲁棒性通常用于描述一个算法能否在系统存在一定的不确定的扰动下，维持其性能不变的特性。一个算法是否具有鲁棒性，是算法能否应用于实际情况的关键，是算法的一个最重要的设计指标。为了探讨各算法的鲁棒性，本文首先从算法在不同结构上的计算结构进行了相似度计算，研究算法的通用性。图 10~13 分别给出了算法 GT、RDF、VDW 以及 BCM 的结构相似度结果热力图(具体数据可见附录)。热力图的横纵坐标表示结构编号，表格中 S_{ij} 的数据表示第 i 个结构与第 j 个结构的相似度，颜色越深表示结果约接近 1，两个结构越相似。同理，颜色越浅表示结果越接近 0，两个结构越不相似。

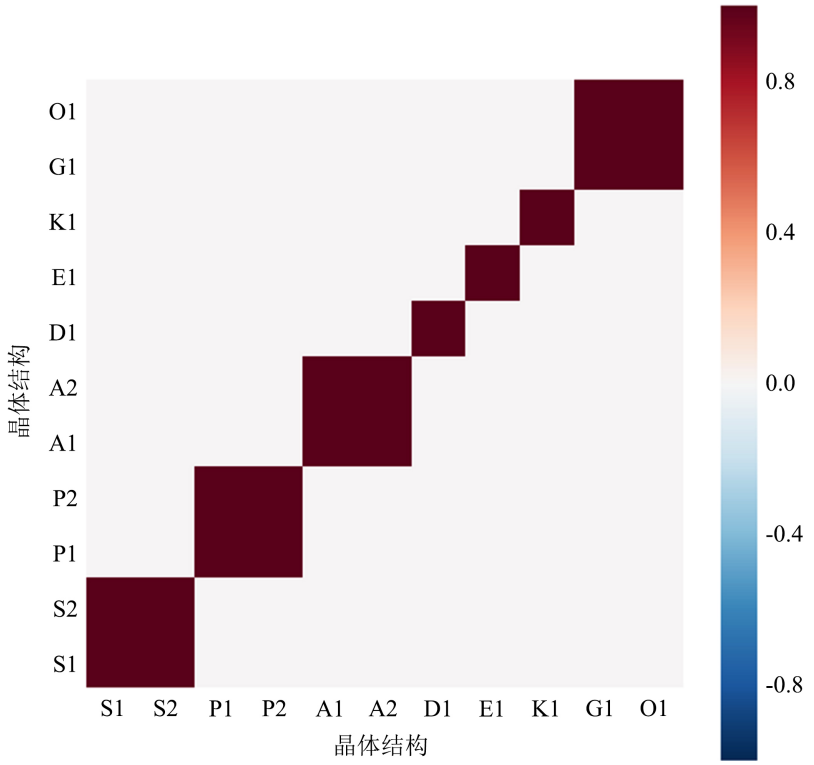


Figure 10. The accuracy of GT algorithm
图 10. GT 算法的相似度热力图

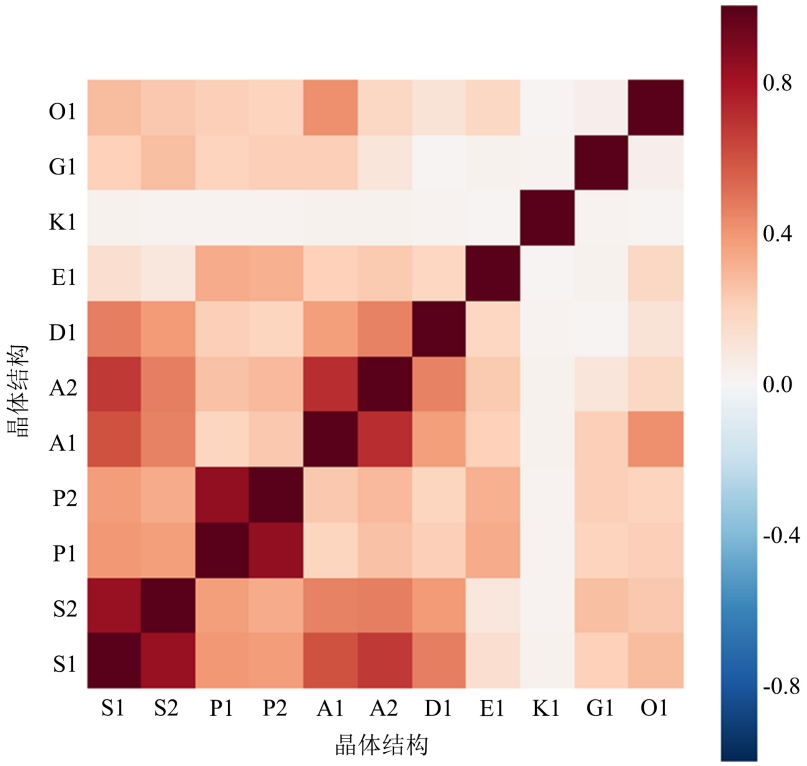


Figure 11. The accuracy of RDF algorithm
图 11. RDF 算法的相似度热力图

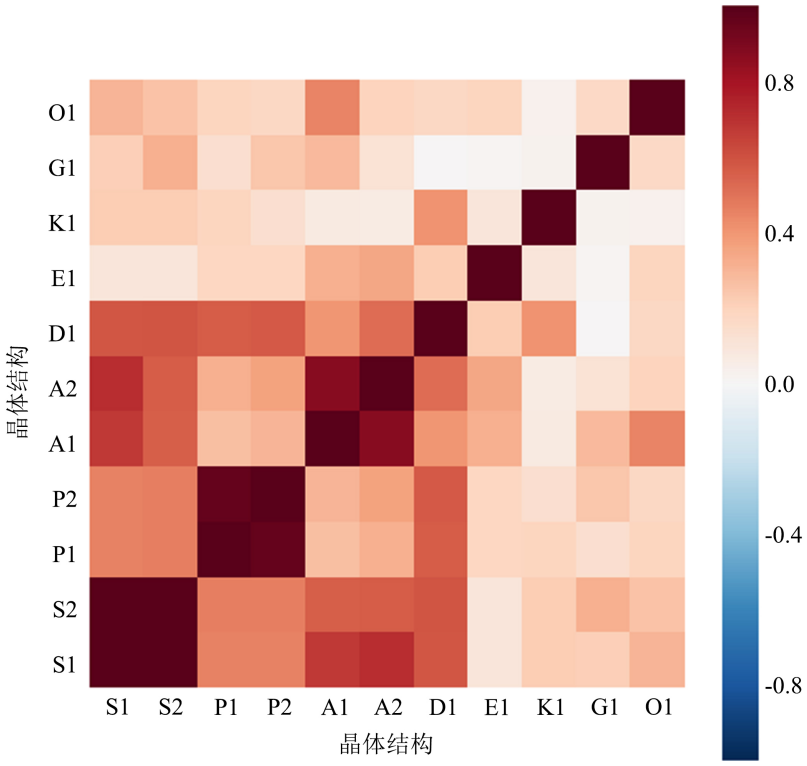


Figure 12. The accuracy of VDW algorithm
图 12. VDW 算法的相似度热力图

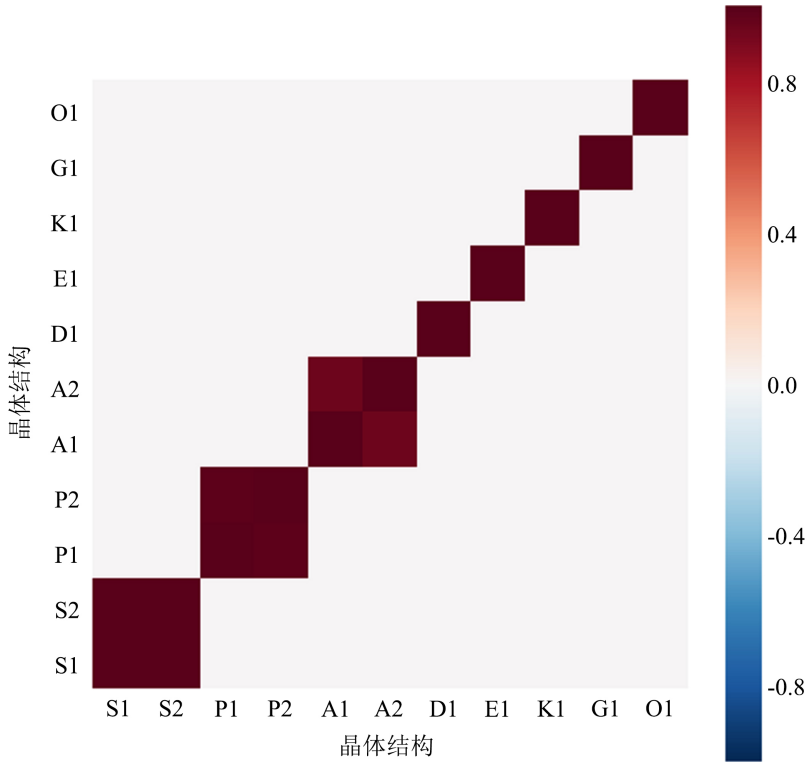


Figure 13. The accuracy of BCM algorithm
图 13. BCM 算法的相似度热力图

从图中可以看出，GT 算法仅有 0 和 1 这两个计算结果，它仅能用于判断任意两个结构的相同与不相同，而无法计算结构间的相似度。RDF 算法、BCM 算法以及 VDW 算法计算结果的域值为[0, 1]，能计算结构间的相似度。但是由于 BCM 算法的特殊性，该算法不能对不同原子类型和数量的结构进行相似度计算。因此，相比之下，RDF 算法以及 VDW 算法可以对任意两个晶体结构的相似度进行计算，从通用性的角度展现了算法的优势，具有更强的鲁棒性。

Table 2. The indicator of algorithm stability
表 2. 算法稳定性指标

	RDF 算法	VDW 算法
稳定性指标	0.9476	0.9654

进一步，本文从稳定性的角度研究算法的鲁棒性。使用如下计算公式(8)对 RDF 算法与 VDW 算法稳定性对比。其中，S_{alg} 为算法的稳定性数值，$Sim_{\alpha_1\beta_{i1}}$ 表示 α_1 结构与 β_{i1} 结构的相似度，α_1 与 α_2 为两个几近相同的结构。

$$S_{alg}=1-\sqrt{\frac{\sum_{i=1}^{N}\left(Sim_{\alpha_1\beta_i}-Sim_{\alpha_2\beta_i}\right)^2}{N}} \tag{8}$$

计算结果可见表 2，VDW 算法稳定性指标达到 0.9654，比 RDF 算法高出了 0.0178。实验表明，VDW 算法具有更好的稳定性。综合通用性和稳定性的研究结果，VDW 算法的鲁棒性性性能更好。

3.3. 计算效率评估

随着数据量的指数增长，人们对计算效率的要求也逐渐提高。为说明 VDW 算法的计算效率，本论文通过实验统计了各算法平均每个结构的处理时间，结果如图 14 所示：

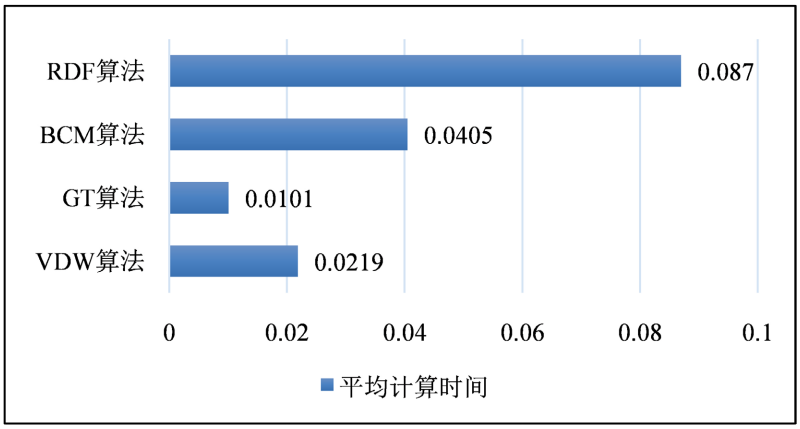


Figure 14. Average computation time of each algorithm
图 14. 各算法完成计算的平均时间

结果显示，GT 算法在计算效率上更具优势，用时更短，这一结果是符合预期的。产生这一现象的主要原因是由于 GT 算法的时间复杂度更低，为 $O(n^2)$。另外，虽然 VDW 算法、BCM 算法以及 RDF 算法的时间复杂度均为 $O(n^3)$，但是在计算效率上，VDW 算法仍然比其他两种算法快，分别快了 0.0185s 和 0.0651 s。这是因为 VDW 算法在提取特征数据时，根据物理知识，设置了阈值，减少了计算成本。BCM 则在计算结构相似度前对结构进行了分类和筛选，减少了不必要的计算。而 RDF 算法则是基于全局的，对结构内的任意原子进行了计算，无疑会导致计算效率的降低。尽管 VDW 算法和 BCM 算法分别通过设置阈值、筛选部分结构提高计算效率，但是 VDW 在计算效率上仍然占优势，体现了 VDW 算法具有较高的计算效率。

3.4. 应用效果

为了证明 VDW 算法的可行性，本文从 ICSD、CSD 和 COD 三个数据库分别收集了 20.8 万、97.3 万、41.3 万，共 159.4 万数据进行研究。

根据数据的特点，我们将数据分为有机晶体数据以及无机晶体结构两种晶体结构进行处理。由于数据中存在原子占位数小于 1.0 的结构情况较复杂，需要进行特殊处理，所以本文暂不考虑该类结构。

1) 有机晶体结构的处理

CSD 数据库是有机晶体结构的数据源。从 CSD 数据库中收集了 973,631 个晶体结构。在通过数据处理，得到了 316,380 个数据。处理过程如图 15 所示。

第一步，去除原子占位数小于 1.0 的晶体结构。对于可以提取其占位数的结构，删除原子占位数小于 1.0 的晶体结构数据。对于无法提取其占位数的结构，将其占位数默认为 1.0，然后去除含有金属元素的结构。至此，本文去除了 532,292 个晶体结构数据。

第二步，去除无序结构。若晶体结构中存在两个原子间距离 r_{AB} 与原子对应的范德华半径和 d_{A+B} 之比小于 0.5，即 $r_{AB}/d_{A+B} < 0.5$ 时，该结构即为 disorder 结构，需要时去除。经过这一步，本文去除了 42,153 个晶体结构数据。

第三步，对结构数据进行溶剂检测，去除其中的溶剂分子。这里去除了 50,695 个晶体结构数据。

第四步，调用第三章提出的 VDW 算法，对相同或者相似的晶体结构计算其相似度，进行去重，并根据数据的更新情况，选择最近发表的晶体保留，去除了 32,111 个晶体结构数据。

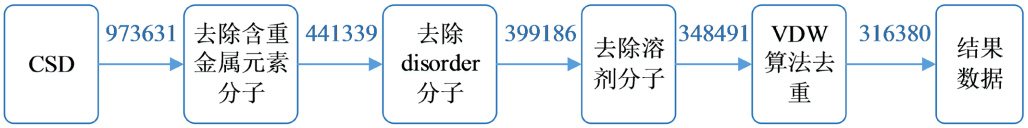


Figure 15. Processing of organic crystals
图 15. 有机晶体处理过程

2) 无机晶体结构的处理

无机晶体结构数据主要从 ICSD 数据库以及 COD 数据库中获取。本文从 ICSD 数据库中收集了 208,425 个晶体结构数据，从 COD 数据库中收集了 413,357 个晶体结构。经过数据的处理与合并后，最终获得了 92,791 个结构数据。具体处理过程如图 16 所示。

由于 ICSD 数据库和 COD 数据库的数据存储结构不同，本文在预处理的过程时，分别对两种数据进行处理。与有机晶体结构的处理过程相同，首先去除原子占位数小于 1.0 的晶体结构数据，此时 ICSD 数据库剩余 116,335 个晶体结构，COD 数据库剩余 281,400 个晶体结构。此外，由于 COD 数据库中存在部分有机晶体结构，本文将结构中存在 C-C 键和 C-H 键的结构认定为有机晶体结构，去除有机晶体，最终获得 45,988 个晶体结构。

由于无机晶体结构数据的复杂性较高，无机晶体结构数据的去重计算量过大，为减少计算量，本文分别对 ICSD 数据库以及 COD 数据库使用 VDW 算法去重处理后将数据合并，并再次去重，最终获得了 92,791 个晶体结构。

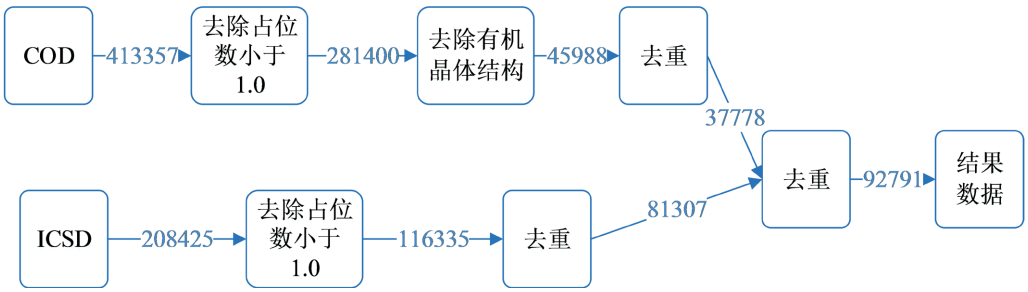


Figure 16. Processing of inorganic crystals
图 16. 无机晶体处理过程

实验证明，无论是针对有机晶体结构还是无机晶体结构，基于原子间相互作用势能的晶体结构去重算法都能在实际场景中有效应用，该算法具有实际应用效果。

4. 总结及展望

本文以材料数据为研究对象，针对传统材料结构相似数据去冗余算法效率低下、数据质量低等问题，提出了基于原子间作用势的晶体结构去重算法。实验结果表明该算法能有效识别相同以及相似晶体结构并去重。与已有算法比较，该算法在保证算法准确性的同时，其鲁棒性、计算效率均占有优势。通过对材料领域ICSD、COD以及CSD超过159万无机以及有机晶体结构数据进行相似度计算，有效去重了10.16

万个结构构建了去冗余的晶体结构数据库，相关数据发布于网站：https://matgen.nscc-gz.cn/。

基金项目

广东省重点领域研发计划(2019B010942001)；广东省引进创新创业团队项目(2016ZT06D211)。

参考文献

[1] Tolle, K.M., Tansley, D.S.W. and Hey, A.J.G. (2011) The Fourth Paradigm: Data-Intensive Scientific Discovery [Point of View]. *Proceedings of the IEEE*, **99**, 1334-1337. https://doi.org/10.1109/JPROC.2011.2155130

[2] Boubchir, M. and Aourag, H. (2021) Materials Genome Project: Mining the Ionic Conductivity in Oxide Perovskites. *Materials Science & Engineering*: *B*, **267**, Article ID: 114984. https://doi.org/10.1016/j.mseb.2020.114984

[3] Prajapati, P. and Shah, P. (2020) A Review on Secure Data Deduplication: Cloud Storage Security Issue. *Journal of King Saud University-Computer and Information Sciences*, https://doi.org/10.1016/j.jksuci.2020.10.021

[4] Xia, W., Zou, X., Jiang, H., Zhou, Y., Liu, C., Feng, D., Hua, Y., Hu, Y. and Zhang, Y. (2020) The Design of Fast Content-Defined Chunking for Data Deduplication Based Storage Systems. *IEEE Transactions on Parallel and Distributed Systems*, **31**, 2017-2031. https://doi.org/10.1109/TPDS.2020.2984632

[5] Vinod-Prasad, P. (2020) A Novel Mathematical Model for Similarity Search in Pattern Matching Algorithms. *Journal of Computer and Communications*, **8**, 94-99. https://doi.org/10.4236/jcc.2020.89008

[6] Weng, M., Wang, Z., Qian, G., Ye, Y., Chen, Z., Chen, X., Zheng, S. and Pan, F. (2019) Identify Crystal Structures by a New Paradigm Based on Graph Theory for Building Materials Big Data. *Science China Chemistry*, **62**, 982-986. https://doi.org/10.1007/s11426-019-9502-5

[7] Zhu, L., Amsler, M., Fuhrer, T., Schaefer, B., Faraji, S., Rostami, S., Ghasemi, S.A., Sadeghi, A., Grauzinyte, M., Wolverton, C. and Goedecker, S. (2016) A Fingerprint Based Metric for Measuring Similarities of Crystalline Structures. *The Journal of Chemical physics*, **144**, Article ID: 034203, https://doi.org/10.1063/1.4940026

[8] Rooymans, C.J.M., Rabenau, A. and Stanley Whittingham, M. (2019) Crystal Structure and Chemical Bonding in Inorganic Chemistry. *Journal of the Electrochemical Society*, **123**, 193C. https://doi.org/10.1149/1.2132930

[9] Samet, D. and Adem. T. (2021) FFCASP: A Massively Parallel Crystal Structure Prediction Algorithm. *Journal of Chemical Theory and Computation*, **17**, 2586-2598. https://doi.org/10.1021/acs.jctc.0c01197

[10] Alexandrov, E., Golov, A. and Shevchenko, A. (2018) Complex Approach to Analysis of Crystal Structures Based on a Unified Topological Model. *Acta Crystallographica Section A Foundations and Advances*, **74**, 153-154. https://doi.org/10.1107/S2053273318092975

[11] Therrien, F., Graf, P. and Stevanović, V. (2020) Matching Crystal Structures Atom-to-Atom. *The Journal of Chemical Physics*, **152**, Article ID: 074106. https://doi.org/10.1063/1.5131527

附录

Table A1. Results of RDF algorithm
表 A1. RDF 算法结果

	S1	S2	P1	P2	A1	A2	D1	E1	K1	G1	O1
S1	1.0000	0.0431	0.0139	0.1737	0.1143	0.1754	0.4160	0.1989	0.2177	0.2372	0.2782
S2		1.0000	0.0185	0.0286	0.0128	0.0945	0.2179	0.2164	0.1969	0.2730	0.2052
P1			1.0000	0.0129	0.0165	0.0269	0.0280	0.0167	0.0159	0.0217	0.0269
P2				1.0000	0.1822	0.2331	0.2059	0.3185	0.3317	0.0863	0.1391
A1					1.0000	0.4596	0.3705	0.1877	0.2162	0.3843	0.4656
A2						1.0000	0.7139	0.2866	0.2583	0.4660	0.6746
D1							1.0000	0.2385	0.1889	0.4606	0.6000
E1								1.0000	0.8473	0.3354	0.3789
K1									1.0000	0.3674	0.3945
G1										1.0000	0.8323
O1											1.0000

Table A2. Results of BCM algorithm
表 A2. BCM 算法结果

	S1	S2	P1	P2	A1	A2	D1	E1	K1	G1	O1
S1	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
S2		1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
P1			1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
P2				1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
A1					1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
A2						1.0000	0.9388	0.0000	0.0000	0.0000	0.0000
D1							1.0000	0.0000	0.0000	0.0000	0.0000
E1								1.0000	0.9900	0.0000	0.0000
K1									1.0000	0.0000	0.0000
G1										1.0000	0.9950
O1											1.0000

Table A3. Results of GT algorithm
表 A3. GT 算法结果

	S1	S2	P1	P2	A1	A2	D1	E1	K1	G1	O1
S1	1	1	0	0	0	0	0	0	0	0	0
S2		1	0	0	0	0	0	0	0	0	0
P1			1	0	0	0	0	0	0	0	0
P2				1	0	0	0	0	0	0	0
A1					1	0	0	0	0	0	0
A2						1	1	0	0	0	0
D1							1	0	0	0	0
E1								1	1	0	0
K1									1	0	0
G1										1	1
O1											1

Table A4. Results of VDW algorithm
表 A4. GT 算法结果

	S1	S2	P1	P2	A1	A2	D1	E1	K1	G1	O1
S1	1.0000	0.1698	0.0338	0.1936	0.1771	0.1964	0.4489	0.1743	0.1926	0.2574	0.2988
S2		1.0000	0.0258	0.0124	0.0076	0.1247	0.2873	0.2447	0.1379	0.3143	0.2174
P1			1.0000	0.0953	0.4115	0.0669	0.0733	0.1397	0.1926	0.2235	0.2258
P2				1.0000	0.2228	0.3477	0.3142	0.1834	0.1852	0.1005	0.0971
A1					1.0000	0.5169	0.4008	0.5704	0.5664	0.5930	0.5855
A2						1.0000	0.8749	0.3629	0.3174	0.5661	0.7160
D1							1.0000	0.2986	0.2731	0.5608	0.6743
E1								1.0000	0.9616	0.4634	0.4577
K1									1.0000	0.4635	0.4561
G1										1.0000	0.9949
O1											1.0000