Computer Science and Application 计算机科学与应用, 2022, 12(4), 866-878
Published Online April 2022 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2022.124088

Hans汉斯

CSPTH：基于天河二号的晶体结构预测软件框架

刘瑾瑜[1]，陈　品[1,2]，卢宇彤[1,2*]

[1]中山大学，计算机学院，广东 广州
[2]中山大学国家超级计算中心，广东 广州

收稿日期：2022年3月6日；录用日期：2022年4月5日；发布日期：2022年4月12日

摘　要

晶体结构是深入理解材料的物理及化学性质的重要信息，发展可以从理论上预测晶体结构的方法具有重要意义。通过高性能计算集群甚至利用超级计算机来加速晶体结构预测已逐步成为趋势。本文中，我们基于天河二号超级计算机开发了一套开源的晶体结构预测软件框架，命名为CSPTH。在算法层面，我们基于当前效率最高的遗传算法进行晶体结构预测，并采用了多种技术提升结构预测效率，包括并行化生成种群结构；引入空间群限定，减少自由度搜索，提升结构多样性；引入晶体指纹进行相似性算法，排除相似结构干扰，避免“基因漂变”的问题。特别地，我们针对晶体结构预测算法的应用特点以及天河二号的系统环境，从任务以及数据管理两个方面做了优化。在任务管理上，我们设计了多层任务调度管理模块，根据计算任务的规模大小的分发细粒度作业(节点内)以及粗粒度作业(跨节点)，提升计算资源的高效使用；在数据管理上，我们将每个计算任务的数据都临时储存于计算节点的RAMDISK，提取有效信息后再存储于MongoDB数据库，避免大量小文件存储于公共存储。CSPTH已在15种已知一元、二元以及三元体系上进行了结构预测，实验结果表明CSPTH能根据给定的组分及外部压力条件下全部预测出相应的稳定结构。

关键词

晶体结构预测，遗传算法，高性能计算

CSPTH: A Crystal Structure Prediction Framework on Tianhe-2 Supercomputer

Jinyu Liu[1], Pin Chen[1,2], Yutong Lu[1,2*]

[1]School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou Guangdong
[2]Nation Supercomputer Center in Guangzhou, Sun Yat-sen University, Guangzhou Guangdong

Received: Mar. 6th, 2022; accepted: Apr. 5th, 2022; published: Apr. 12th, 2022

*通讯作者。

文章引用: 刘瑾瑜, 陈品, 卢宇彤. CSPTH: 基于天河二号的晶体结构预测软件框架[J]. 计算机科学与应用, 2022, 12(4): 866-878. DOI: 10.12677/csa.2022.124088

Abstract

Crystal structure is the critical information for understanding the physical and chemical properties of materials. Therefore, theoretical prediction of crystal structures only with chemical composition and external conditions is significant. It has become a trend to design new materials through high-performance computing clusters or even using supercomputers. In this paper, we developed an open source framework for crystal structure prediction (CSP) based on Tianhe-2 supercomputer, named CSPTH. When designing the algorithm in CSP, we chose the most efficient genetic algorithm in our framework and adopted numerous technologies to improve prediction accuracy. Specifically, we used a multi-process parallel method to generate the trying structures. The space group restriction is introduced to reduce the searching space and improve the structural diversity in population. We utilized a crystal fingerprint to eliminate the similar structures, which can avoid the problem of "gene drift". In particular, considering the characteristics of crystal structure algorithm and the system environment of Tianhe-2, we optimize the crystal structure prediction algorithm from two aspects: task management and data management. In terms of task management, we designed a multi-layer task scheduling management module to distribute fine-grained tasks (within a node) and coarse-grained tasks (multi-nodes) according to the system size of tasks to improve the efficiency of employing resources. In our data management module, the data of each computing task is temporarily cached in the RAMDISK of the computing node, and the useful information is extracted and later stored in the MongoDB database, which can avoid a large number of small files stored in the public storage. CSPTH has been used to predict the structures of 15 known element, binary and ternary systems. Experimental results show that CSPTH can predict all the corresponding stable structures with the only known of chemical composition and external pressure.

Keywords

Crystal Structure Prediction, Genetic Algorithm, High-Performance Computing

1. 引言

在过去的几十年中，利用计算机进行晶体结构预测的方法，使得系统地进行新材料设计已成为了可能[1] [2] [3] [4] [5]。结构预测的本质就是在只给定化学组分和外界条件下，确定全局能量最低结构的原子排列方式。具体地说就是在势能面上寻找全局能量最低点。尽管原理上很简单，然而从理论上确定物质结构，是物理、化学和材料研究领域的长期难题。其根本原因在于物质势能面的高度复杂性，随着单胞内原子数目的增加，体系可能的结构数目成指数增加；并且精确地评估结构的能量也需要大量的计算，理论结构预测就是在如此复杂的势能面中搜索全局能量最低的结构，具有巨大挑战性。在材料的结构搜索领域，已报道的具代表性的方法有模拟退火法[6] [7]、能谷跳跃[8]、极小值跳跃[9]、随机法[10] [11]、遗传算法[12] [13] [14] [15]、元动力学法[16]和粒子群优化算法[17] [18]。表 1 列举了当前主流的结构搜索计算软件，其中 USPEX [14] [19] [20]和 CALYPSO [21] [22] [23]是目前报道的用户以及论文引用最多的软件。

Table 1. Summary of structure search software
表 1. 结构搜索软件情况

软件名	方法	应用领域	开发语言	开源许可	发布时间
Upack	随机结构，聚类	有机	Fortan	开源	1995 [24]
AIRSS	随机结构搜索	无机	-	开源(GPL-2)	2006 [10] [11]
USPEX	进化算法	无机/有机	Matlab	开源	2006 [14] [19] [20]
CALYPSO	粒子群优化	无机	Fortan	不开源	2010 [17] [18]
Xtalopt	进化算法	无机	C++	开源(BSD)	2011 [25]
DMACRYS	元动力学	有机	-	商业	2010 [26]
GASP	遗传算法	无机	JAVA	开源(GPL v3)	2013 [27]
GRACE	基于 DFT	无机	-	商业	2013 [28] [29] [30]
IM^2ODE	多目标拆分优化	无机	Fortan	开源(LGPL-3)	2015 [31]
GATor	遗传算法	有机	Python	开源(BSD-3)	2018 [32]
MAISE	神经网络、进化算法	无机	C	开源(GPL-3)	2020 [33]
CrySPY	贝叶斯优化	无机	Python	开源(MIT)	2021 [34]

相关晶体结构预测的算法是典型的计算密集性和数据密集型应用。首先，基于进化算法的晶体结构方法需要使用准确的第一性原理的方法去计算种群中的每个结构的能量，而此第一性原理方法非常耗时，使得晶体结构预测的效率比较低；其次，相关算法都需要产生大量的种群结构数，而使用第一性原理的方法评价每一个结构将产生批量的作业，并且需要迭代进行多轮计算，需要消耗大量的计算资源；最后，由于调用的是第三方程序对种群中的每个结构进行优化，在不更改第三方程序源码的情况下，难以对每个结构的输出进行统一管理，将会产生大量的临时文件数据。近年来，随着高性能计算领域的快速发展以及材料设计领域需求的提升，通过高性能计算集群甚至利用超级计算机来加速晶体结构预测已逐步成为趋势。然而相关的软件并没有针对高性能集群进行并行或者高通量管理的设计，不能充分地发挥高性能集群的性能。基于此，我们基于天河二号超级计算机开发了一套开源的晶体结构预测软件框架，命名为 CSPTH。在任务管理上，我们设计了多层任务调度管理模块，根据计算任务的规模大小的调整细粒度作业(节点内)以及粗粒度作业(跨节点)，提升资源的利用率；在数据管理上，我们将每个计算任务的数据都临时储存于计算节点的 RAMDISK，提取有效信息后再存储于 MongoDB 数据库，避免大量小文件存储于公共存储。在算法层面，我们基于当前效率最高的遗传算法进行晶体结构预测，并采用了多种技术提升结构预测效率，包括并行化生成种群结构；引入空间群限定，减少自由度搜索，提升结构多样性；引入晶体指纹进行相似性算法，排除相似结构干扰，避免“基因漂变”(genetric drift) [35]的问题。目前这一软件框架已在多种已知体系上进行了实验，实验结果表明 CSPTH 能根据给定的组分及外部压力准确找到相应的稳定结构。

2. 晶体结构预测算法

2.1. 遗传算法

我们基于遗传算法设计了晶体结构预测算法。如图 1 所示，该算法主要包含四个步骤：1) 初始化种群结构：基于对称性限定随机生成结构；2) 结构优化：通过第一性原理计算对结构进行局域优化；3) 生成新结构：通过结构交叉、晶格突变、原子交换这三种变异操作及随机方法生成候选结构。图 2 为三种

变异操作的示意图；4) 挑选下一代种群结构：基于晶体结构的能量构造适应度函数，挑选低能量结构进入下一代种群。重复寻优过程，直到找到稳定的结构。

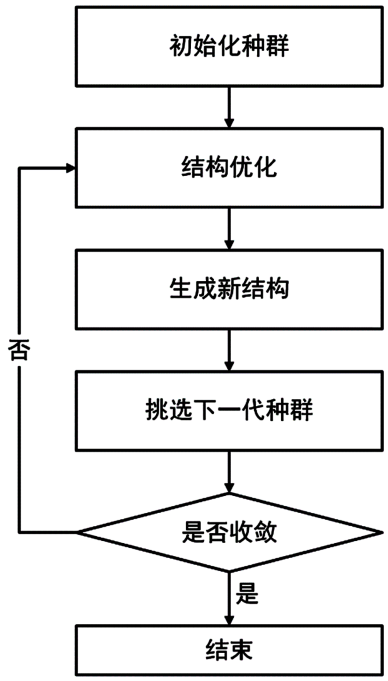

Figure 1. Flowchart of genetic algorithm
图 1. 遗传算法流程图

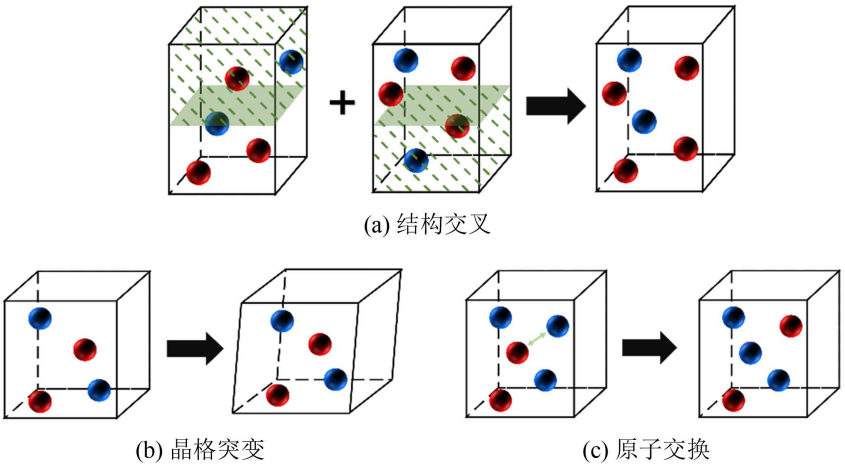

Figure 2. Diagram of variation operations
图 2. 变异操作示意图

2.2. 基于对称性限定的结构生成优化

晶体结构的对称性由 230 种空间群决定。根据空间群的不同，可以将结构划分为 7 大晶系，不同空间群的晶体结构差异较大。然而，纯随机生成的晶体结构通常不具备对称性或仅有较低的对称性(空间群为 P1)，这可能导致整个遗传算法对于化学空间的探索局限在空间群为 P1 的结构集合内。由于自然界中低能量结构通常具有一定的对称性，纯随机生成的空间群为 P1 的结构几乎都是无效结构。为了避免生成

大量无效的低对称性结构，我们采用基于对称性限定的随机方法生成结构。首先，在 230 个空间群中随机选择一个。为了尽可能地探索化学空间，每一次挑选空间群时，都尽可能采用未使用过的空间群。随后，依据空间群构造晶格及原子坐标：根据空间群对应的布拉菲晶格确定部分晶格常数，其余晶格常数随机生成；通过随机组合空间群对应的维科夫(Wyckoff)位置生成原子坐标。

2.3. 结构生成并行化优化

在遗传算法中，种群中结构的数量影响着算法整体对化学空间的探索程度，使用较大的种群数可以快速采样化学空间，有助于算法尽快收敛到全局最优解。在现有的晶体结构预测软件中，结构生成是串行的。由于结构生成本身需要大量的尝试才能生成有效结构，尤其是在加入了多种条件限制后，在种群数较大的情况下，结构生成会成为整个算法的效率瓶颈。因此，我们针对结构生成(包含基于对称性限定的随机结构生成、结构交叉、晶格突变、原子交换)部分进行并行化优化。针对每一个结构生成操作提交相应的任务，通过 HTCondor 分发到天河节点上。为了充分利用天河节点 24 核的计算资源，提高计算效率，采用单节点多任务模式，在一个节点上并行多个结构生成任务。

2.4. 基于晶体指纹排除相似结构

相似的结构经过局域结构优化后，通常会落入势能面的同一个能谷中，即收敛为同一个低能量结构。因此，种群中的相似结构对于寻找全局最低能量的结构并没有帮助，相反地，会造成计算资源的浪费。我们采用局部径向分布函数[36]计算晶体指纹，通过比较晶体指纹的相似度消除种群中的相似结构。

局部径向分布函数 $g_{\alpha\beta}(r)$ 表示与类型为 α 的原子相距 r 的薄壳区域(薄壳厚度为 dr)内存在 β 类原子的概率，其计算公式如下：

$$g_{\alpha\beta}(r)=\frac{N_\beta}{V}dn_{\alpha\beta}(r)V_{shell}=\frac{N_\beta}{V}dn_{\alpha\beta}(r)4\pi r^2dr\ \left(V_{shell}=\frac{4}{3}\pi(r+dr)^3-\frac{4}{3}\pi r^3\simeq 4\pi r^2dr\right)$$

其中，N_β 为 β 类原子的数量，V、V_{shell} 分别为整个结构的体积、薄壳区域的体积，$dn_{\alpha\beta}(r)$ 表示与类型为 α 的原子相距 r 的薄壳区域(薄壳厚度为 dr)内存在 β 类原子的数量。

基于局部径向分布函数，可以通过下式计算得到每一个结构的晶体指纹 fp：

$$fp(i,j,r)=g_{a_i,a_j}(r),\quad i=1,\cdots,S;\quad j=1,\cdots,S;\quad r=1,2,\cdots$$

其中，S 为原子种类数。

通过下式计算结构 i 和结构 j 的晶体指纹之间的余弦距离，如果余弦距离小于设定的阈值，则认为两个结构相似，消除两者中能量较高的结构。

$$d_{ij}=\frac{1}{2}\left(1-\frac{fp_i\,fp_j}{\|fp_i\|\|fp_j\|}\right)$$

3. CSPTH 框架介绍

晶体结构预测算法是典型的计算密集性和数据密集型应用。晶体结构预测算法依赖准确的第一性原理的方法提供每个结构的能量信息，而第一性原理计算非常耗时。如果缺乏针对高性能集群设计的并行管理，多次批量地进行第一性原理方法计算会降低晶体结构预测算法的效率，也不能充分地发挥高性能集群的性能。在调用基于第一性原理方法的第三方程序进行结构优化时，会产生大量的临时文件数据，数据存储、数据移动以及数据分析将产生瓶颈，造成共享文件系统负担过重。基于此，我们设计了针对天河二号集群的晶体结构预测框架 CSPTH，为晶体结构预测算法提供任务管理与数据管理。

CSPTH 框架如图 3 所示，包含晶体结构预测算法、局域结构优化、任务管理以及数据管理四个子模块。下面将对局域结构优化、任务管理以及数据管理三个子模块进行介绍。

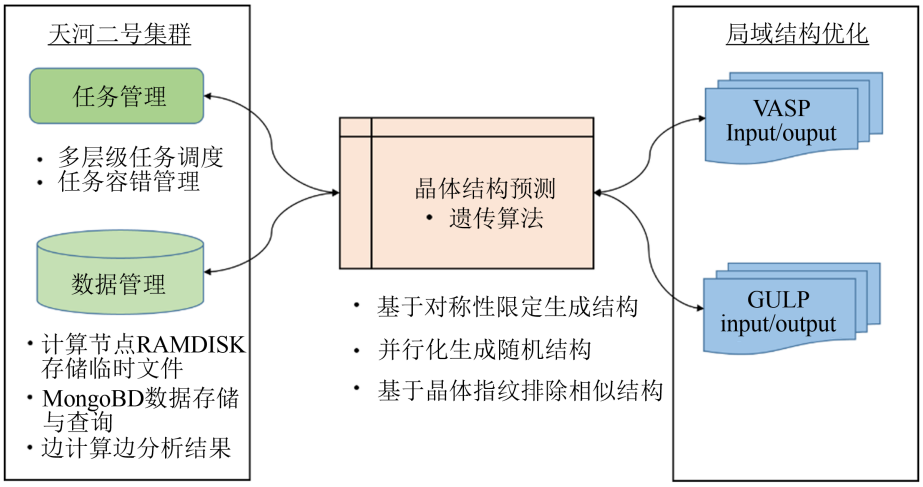

Figure 3. Framework of CSPTH
图 3. CSPTH 框架图

3.1. 局域结构优化

CSPTH 目前支持第一性原理软件包或者基于力场的分子动力学程序作为接口来实现晶体结构的几何优化。这些软件通过调整晶体结构中各个原子的位置，可以得到一系列的结构，这些结构所具有的能量组成了一个势能面，几何优化计算的目标就是找到能量的最小值点对应的晶体结构。目前的结构优化技术常用算法包括共轭梯度算法、最速下降、线性最小化和拟牛顿法等。这些极小值点将会为产生下一代结构提供更物理的结构信息。因此局域优化对于结构预测的成功具有非常重要的作用。其中第一性原理方法是进行结构优化最准确的方法，但同时这一方法非常耗时。

3.2. 任务管理

CSPTH 和同类型的进化算法软件类似，通过调用第三方的软件对当前种群的每个结构进行结构评价，这些程序中既有并行效率高的程序(如 LAMMPS [37]、VASP [38])，也有只能在单节点内多线程并行的程序(如 Gaussian [39])，并且具体任务所用资源的大小也和计算体系的规模有关，因此需要支持灵活多样的任务分发与管理。然后大多数传统的大规模高性能集群支持的作业队列管理系统如 SLURM [40]，PBS [41] 和 SGE [42]等，都配置为粗粒度的计算资源。当作业数扩展到一千以上时，容易出现作业失败的问题。这一问题主要来自两个方面，一个是输入文件的参数问题，另一方面是计算集群的节点故障(如：节点内存溢出、输入输出阻塞、网络延迟等)，如果手动处理这些作业，将会影响预测晶体结构的效率。

因此，在任务管理模块中，我们针对天河二号系统的环境，设计了多级任务调度和作业容错两个子模块。在天河二号集群上，采用的是 SLURM 作业管理系统，每个节点配置 24 个 CPU 物理核。在多级任务调度子模块中，我们通过一级的 SLURM 作业管理系统申请一定数量的资源，然后通过二级的 HTCondor [43]对资源进一步细化，并进行任务的分发与管理。通过这种方式解决不同任务对于调度粒度的需求，同时保证了对计算资源的统一管理。对于容错，我们主要考虑作业失败后自动处理的能力。对于输入文件参数导致的问题，我们针对应用软件的常见错误，建议一一对应的处理方法；对于系统节点故障问题，我们首先通过作业管理系统排除指定节点，然后将作业重新提交。

3.3. 数据管理

由于调用的是第三方程序进行结构的评估，在不更改源程序的前提下，难以对每个结构的输出进行统一管理，当计算任务扩展到成千上万以后，将会产生大量的临时文件数据，数据存储、数据移动以及数据分析将产生瓶颈。如果直接访问和存储文件将产生大量的输入输出(Input/output, IO)操作，容易导致共享文件系统负担过重。数据库为存储和数据查询提供了一个有吸引力的解决方案。在 CSPTH 的框架中，我们通过使用 MongoDB [44]和本地 RAMDISK 来避免使用共享文件系统(如图 4 所示)。在计算过程中产生的输出文件临时地存储在计算节点的 RAMDISK 中，然后将有效的信息提取出来直接存储到 MongoDB 中。MongoDB 作为数据存储引擎，它是一个高性能、高可用、高可扩展的、开源的非结构化数据库。这种架构适用于稀疏和类文档的数据存储。通过使用 MongoDB“索引”，可以轻松地对分子进行查询和排序。此外，MongoDB 使用“分片”(一种在多台机器上分布数据的方法)来支持以高吞吐量方式部署大型数据集，随着数据库的增长平衡查询负载来提高计算性能。最后，MongDB 接受高达 16 MB 的大数据，足以存储常规的输出文件。第三方应用软件涉及处理大量的纯文本文件。在不修改应用程序源代码的情况下，必须通过在共享文件磁盘上移动来处理大量的临时文件。因此，我们充分利用计算节点中的 RAMDISK 用于临时存储应用程序所需的 IO 文件(如图 4 所示)。RAMDISK 提供高速、低延迟的 IO 操作来处理大量的小文件，同时将 MongDB 部署在共享存储上。通过在 MongDB 和 RAMDISK 之间移动数据，有效缓解了共享文件存储的 IO 压力。

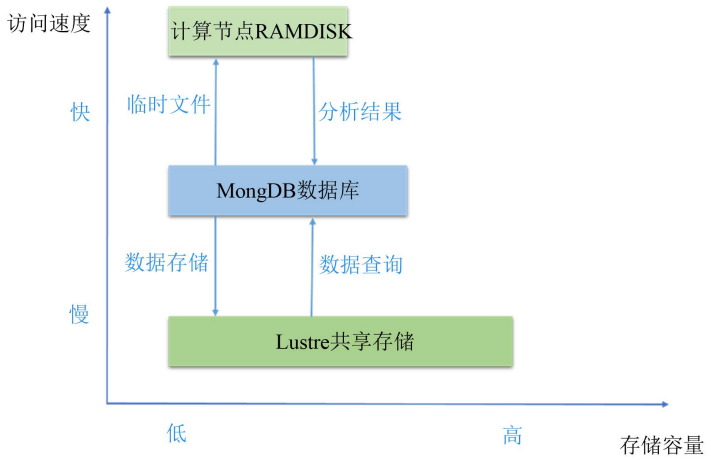

Figure 4. Diagram of dataflow
图 4. 数据流示意图

对于种群中结构的评价，需要根据每个结构的自由能进行排序，挑选出适合的结构进行下一代结构的生成。考虑到共享文件存储的高性能计算系统，需要避免大量小文件导致的 IO 过载问题。因此，分析共享存储磁盘上的输出文件是不明智的。在 RAMDISK 中完成计算后，将分析输出文件，并根据自由能以及相应的结构存储到数据库中进行排序。当小文件数量急剧增加时，此方法可最大限度地减少 IO 压力。

4. 实验结果与分析

我们参考了 USPEX [14]以及 CALYPSO [17]测试过的体系进行 CSPTH 的性能测试。如表 2 所示，CSPTH 在单质，二元以及三元体系中的共 15 个一定压强下的结构全部被成功预测出来，成功率为 100%。其中单质测试案例包括 Li、C、Mg 以及 Si，CSPTH 在第一代就预测出了空间群为 $P6_3/mmc$ 的单质 Li 材料，对于单质 Mg 也只在第二代预测出来。对于双元体系材料，我们对 SiO_2、TiH_2 以及 Al_2O_3 三种结构

进行预测，对于 SiO_2 在不同压强下的三种结构均在第一代预测出来；TiH_2 的两个已知结构也在第一和第四代中预测出来。CSPTH 对于所含原子数较小的单质、二元化合物所在的化学空间表现出了较强的搜索能力，这主要归功于基于对称性限定的结构生成方法。引入对称性的限定，使得遗传算法在探索中能同时关注不同空间群下的多样化的结构，迅速锁定可能的低能量结构所在的势能面的能谷。对于较复杂的三元体系 $MgSiO_3$ 以及 $MgAlO_4$ 也分别在第十三代和第十五代预测出来。实验结果表明，CSPTH 不仅具有较高的预测准确度，同时也具有较高的预测效率，均在前 15 代内预测出了结构。同时，我们使用 CSPTH 预测出了一些新颖的结构，如图 5 所示。此外，我们还在天河二号系统上进行了压力测试，同时申请了 500 节点进行所有结构的计算，最大的作业并发数为 260，总共产生了 43,980 个临时文件。我们在三天的时间内完成了所有的计算和数据分析。

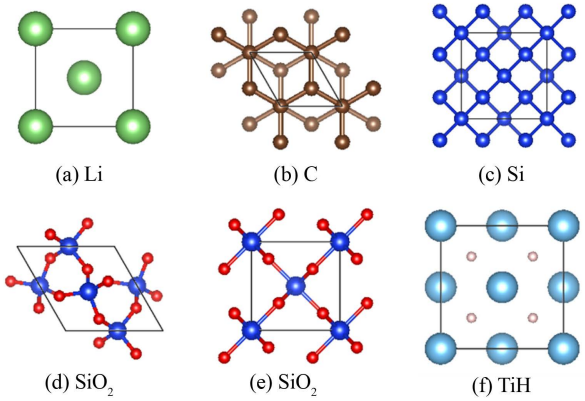

Figure 5. Structures predicted by CSPTH
图 5. CSPTH 成功预测的结构

Table 2. CSPTH tested on known structural systems
表 2. CSPTH 预测的已知结构体系

系统	压强	空间群	体积	找到目标结果所需代数	种群大小
Li [45]	0 GPa	$P6_3/mmc$	42.69	1	30
	0 GPa	$Im\bar{3}m$	42.55	3	30
C [46]	0 GPa	$P6_3/mmc$	35.29	7	30
Mg [47]	0 GPa	$P6_3/mmc$	46.15	2	30
Si [48] [49]	0 GPa	$Fd\bar{3}m$	160.19	6	30
	0 GPa	$Fd\bar{3}m$	159.85	11	30
SiO_2 [50] [51]	0 GPa	$P3_212$	113.24	1	30
	20 GPa	$P4_2/mnm$	46.51	1	30
	271 GPa	Pa3	60.69	1	30
TiH_2 [52]	0 GPa	$Fd\bar{3}m$	87.00	4	30
	0 GPa	I4/mmm	43.87	1	30
Al_2O_3 [53] [54]	0 GPa	$R\bar{3}c$	84.50	1	30
	300 GPa	$R\bar{3}c$	254.93	9	80
$MgSiO_3$ [55]	120 GPa	Cmcm	120.34	13	100
$MgAlO_4$ [56]	100 GPa	Pnam	234.46	15	100

详细案例

算例硅的配置文件如图 6 所示。基本设置为原子类型为硅(Si)，对应的硅原子数量设置为 8，外部压力为 0 Gpa，种群大小为每代 30 个结构。作业相关的设置为该算例最多使用 45 个节点，每个 vasp 计算使用 2 个节点。根据配置文件，CSPTH 首先会通过 SLURM 作业管理系统申请 45 个计算节点。在 CSPTH 运行过程中，由二级的 HTCondor 对这 45 个计算节点进行任务的分发与管理。在遗传算法中，主要会产生两类计算任务：结构生成任务及结构局域优化任务。对于结构生成任务，HTCondor 会为其分配 4 个核的计算资源。对于结构局域优化任务，HTCondor 会根据该任务的计算精度为其分配 16~64 个核的计算资源。结构局域优化的计算精度由其 K 点网格密度决定，我们设置粗、中等、细这三种网格密度，分别对应 16、32、64 个核的计算资源。在这两类任务执行过程中会产生大量的临时文件，这些文件都会临时存储在计算节点的 RAMDISK。当结构局域优化任务完成时，CSPTH 会从这些临时文件中读取优化后的结构信息以及结构对应的能量，并将它们写入 MongDB 数据库。在每一代种群完成结构生成、结构局域优化、结构挑选后，CSPTH 会从 MongDB 数据库中获取该代所有结构的详细信息，写入到文本文件 gen_x_info 中(其中 x 表示种群代数)。图 7 展示了 Si 算例中第 11 代种群的信息汇总文件。文件中记录了三部分内容：1) 种群结构性质：每一个结构的信息，包括能量、空间群、该结构最早产生的代数；2) 最低能量列表：从第 1 代到第 11 代种群的最低能量列表；3) 结构来源：每一个结构的生成方式，如果该结构来自于变异操作，会记录父本结构。

```
[basic]
atype = Si
anum = 8
pressure = 0
tsize = 30

[job]
total_node = 45
vasp_job = 2
```

Figure 6. Input file of Si case
图 6. Si 算例中的输入文件

```
Generation 11:
struct 2 :          enthalpy= -43.282028      entyhalpy(eV/atom)= -5.4102535      nsym = 227      gen = 11
struct 35 :         enthalpy= -43.280859      entyhalpy(eV/atom)= -5.410107375    nsym = 227      gen = 6
......
the best is 2, its enthalpy= -43.282028

low enthalpy list:
[-41.44115, -42.20226, -42.20226, ....., -43.28268, -43.28268, -43.28268, -43.28268, -43.28268]

Structures generation infos:
struct 2 :          type= heredity    parents = ['6_32', '1_50']
struct 35 :         type= Random      parents = None
......
```

Figure 7. The detail information of the 11th population in Si case is recorded named gen_11_info
图 7. Si 算例中第 11 代种群的信息汇总文件 gen_11_info

根据种群的信息汇总文件，可以还原出遗传算法的搜寻稳定 Si 结构的过程，如图 8 所示。随着种群的演化，每一代中的最优结构的能量逐步下降，意味着 CSPTH 找到的结构逐渐趋于稳定。在第 6 代和 11 代分别对应较低能力与最低能量出现的时刻，这两代很可能首次发现了不同的稳定 Si 结构。随后，通过比较晶体指纹相似性，确定已在第 6 代和第 11 代分别找到已知的稳定结构。

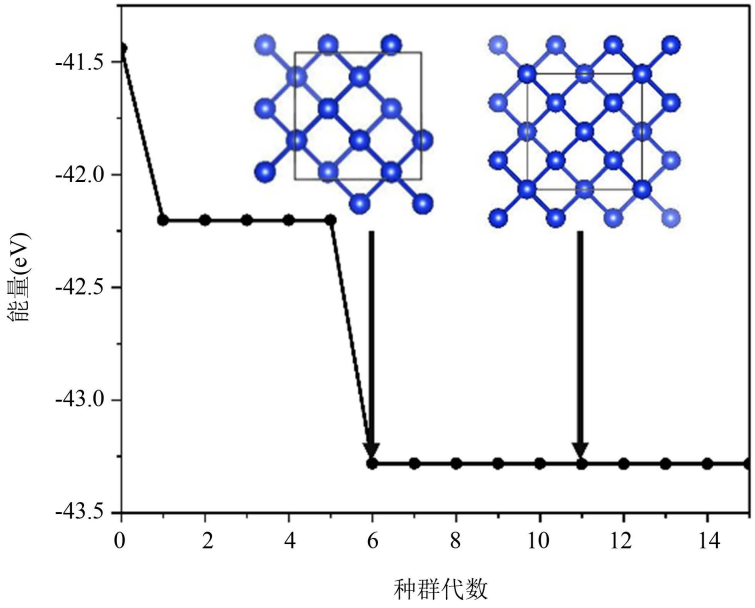

Figure 8. The lowest enthalpy against generation in Si case
图 8. Si 算例中最低能量随种群代数变化

5. 结语

在本文中，我们介绍了 CSPTH 框架的实现，该框架可用于在给定化学成分和外部压力的情况下预测晶体结构。CSPTH 实现了结构生成、结构评估和结构选择的自动化，使实验科学家也能够进行晶体结构预测。特别地，我们针对天河二号超级计算机的软硬件环境进行优化，包括任务管理以及数据管理，提高软件的效率、鲁棒性以及可扩展性，我们初步测试了最高 260 个并发任务，调用 500 个节点进行高通量计算，在三天的时间内完成了 9 个材料体系共 15 个算例的结构预测，预测出全部已知实验观察的结构，成功率为 100%。CSPTH 的高成功率和高性能证明了其作为晶体结构预测工具的可靠性和应用前景。

基金项目

广东省重点领域研发计划(2019B010940001)；广州市科技计划项目(201604016005)。

参考文献

[1] Liu, Y., Wang, R., Wang, Z., Li, D. and Cui, T. (2022) Formation of Twelve-Fold Iodine Coordination at High Pressure. *Nature Communications*, **13**, Article No. 412. https://doi.org/10.1038/s41467-022-28083-4

[2] Luo, D., Qiao, X. and Dronskowski, R. (2021) Predicting Nitrogen-Based Families of Compounds: Transition-Metal Guanidinates $T(CN_3)$ (T=V, Nb, Ta) and Ortho-Nitrido Carbonates T'2CN4 (T'=Ti, Zr, Hf). *Angewandte Chemie International Edition*, **60**, 486-492. https://doi.org/10.1002/anie.202011196

[3] Luo, W., Nakamura, Y., Park, J. and Yoon, M. (2021) Cobalt-Based Magnetic Weyl Semimetals with High-Thermodynamic Stabilities. *npj Computational Materials*, **7**, Article No. 2. https://doi.org/10.1038/s41524-020-00461-w

[4] Liu, X., Niu, H. and Oganov, A.R. (2021) COPEX: Co-Evolutionary Crystal Structure Prediction Algorithm for Complex Systems. *npj Computational Materials*, **7**, Article No. 199. https://doi.org/10.1038/s41524-021-00668-5

[5] Kvashnin, A.G., Tantardini, C., Zakaryan, H.A., Kvashnina, Y.A. and Oganov, A.R. (2020) Computational Search for New W-Mo-B Compounds. *Chemistry of Materials*, **32**, 7028-7035. https://doi.org/10.1021/acs.chemmater.0c02440

[6] Pannetier, J., Bassas-Alsina, J., Rodriguez-Carvajal, J. and Caignaert, V. (1990) Prediction of Crystal Structures from Crystal Chemistry Rules by Simulated Annealing. *Nature*, **346**, 343-345. https://doi.org/10.1038/346343a0

[7] Schön, J.C. and Jansen, M. (1996) First Step towards Planning of Syntheses in Solid-State Chemistry: Determination

of Promising Structure Candidates by Global Optimization. *Angewandte Chemie International Edition in English*, **35**, 1286-1304. https://doi.org/10.1002/anie.199612861

[8] Wales, D.J. and Doye, J.P. (1997) Global Optimization by Basin-Hopping and the Lowest Energy Structures of Lennard-Jones Clusters Containing up to 110 Atoms. *The Journal of Physical Chemistry A*, **101**, 5111-5116. https://doi.org/10.1021/jp970984n

[9] Goedecke, S. (2004) Minima Hopping: An Efficient Search Method for the Global Minimum of the Potential Energy Surface of Complex Molecular Systems. *Journal of Chemical Physics*, **120**, 9911-9917. https://doi.org/10.1063/1.1724816

[10] Pickard, C.J. and Needs, R.J. (2006) High-Pressure Phases of Silane. *Physical Review Letters*, **97**, Article ID: 045504. https://doi.org/10.1103/PhysRevLett.97.045504

[11] Pickard, C.J. and Needs, R.J. (2011) *Ab Initio* Random Structure Searching. *Journal of Physics*: *Condensed Matter*, **23**, Article ID: 053201. https://doi.org/10.1088/0953-8984/23/5/053201

[12] Woodley, S., Battle, P., Gale, J. and Catlow, C.A. (1999) The Prediction of Inorganic Crystal Structures Using a Genetic Algorithm and Energy Minimisation. *Physical Chemistry Chemical Physics*, **1**, 2535-2542. https://doi.org/10.1039/a901227c

[13] Abraham, N.L. and Probert, M.I. (2006) A Periodic Genetic Algorithm with Real-Space Representation for Crystal Structure and Polymorph Prediction. *Physical Review B*, **73**, Article ID: 224104. https://doi.org/10.1103/PhysRevB.73.224104

[14] Oganov, A.R. and Glass, C.W. (2006) Crystal Structure Prediction Using *Ab Initio* Evolutionary Techniques: Principles and Applications. *Journal of Chemical Physics*, **124**, 201-419. https://doi.org/10.1063/1.2210932

[15] Trimarchi, G. and Zunger, A. (2007) Global Space-Group Optimization Problem: Finding the Stablest Crystal Structure without Constraints. *Physical Review B*, **75**, Article ID: 104113. https://doi.org/10.1103/PhysRevB.75.104113

[16] Martoňák, R., Laio, A. and Parrinello, M. (2003) Predicting Crystal Structures: The Parrinello-Rahman Method Revisited. *Physical Review Letters*, **90**, Article ID: 075503. https://doi.org/10.1103/PhysRevLett.90.075503

[17] Call, S.T., Zubarev, D.Y. and Boldyrev, A.I. (2007) Global Minimum Structure Searches via Particle Swarm Optimization. *Journal of Computational Chemistry*, **28**, 1177-1186. https://doi.org/10.1002/jcc.20621

[18] Laio, A. and Parrinello, M. (2002) Escaping Free-Energy Minima. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 12562-12566. https://doi.org/10.1073/pnas.202427399

[19] Glass, C.W., Oganov, A.R. and Hansen, N. (2006) USPEX—Evolutionary Crystal Structure Prediction. *Computer Physics Communications*, **175**, 713-720. https://doi.org/10.1016/j.cpc.2006.07.020

[20] Oganov, A., Lyakhov, A. and Valle, M. (2011) How Evolutionary Crystal Structure Prediction Works—And Why. *Accounts of Chemical Research*, **44**, 227-237. https://doi.org/10.1021/ar1001318

[21] Wang, Y., Lv, J., Zhu, L. and Ma, Y. (2010) Crystal Structure Prediction via Particle Swarm Optimization. *Physics Review B*, **82**, Article ID: 094116. https://doi.org/10.1103/PhysRevB.82.094116

[22] Wang, Y., Lv, J., Zhu, L. and Ma, Y. (2012) CALYPSO: A Method for Crystal Structure Prediction. *Computer Physics Communications*, **183**, 2063-2070. https://doi.org/10.1016/j.cpc.2012.05.008

[23] Zhang, Y., Wang, H., Wang, Y., Zhang, L. and Ma, Y. (2016) Computer-Assisted Inverse Design of Inorganic Electrides. *Physical Review X*, **7**, Article ID: 019903. https://doi.org/10.1103/PhysRevX.7.019903

[24] Van Eijck, B., Mooij, W.T. and Kroon, J. (1995) Attempted Prediction of the Crystal Structures of Six Monosaccharides. *Acta Crystallographica Section B*: *Structural Science*, **B51**, 99-103. https://doi.org/10.1107/S0108768194009651

[25] Lonie, D.C. and Zurek, E. (2011) $X_{TAL}O_{PT}$: An Open-Source Evolutionary Algorithm for Crystal Structure Prediction. *Computer Physics Communications*, **182**, 372-387. https://doi.org/10.1016/j.cpc.2010.07.048

[26] Price, S.L., Leslie, M., Welch, G.W., Habgood, M., Price, L.S., Karamertzanis, P.G. and Day, G.M. (2010) Modelling Organic Crystal Structures Using Distributed Multipole and Polarizability-Based Model Intermolecular Potentials. *Physical Chemistry Chemical Physics*, **12**, 8478-8490. https://doi.org/10.1039/c004164e

[27] Tipton, W.W. and Hennig, R.G. (2013) A Grand Canonical Genetic Algorithm for the Prediction of Multi-Component Phase Diagrams and Testing of Empirical Potentials. *Journal of Physics*: *Condensed Matter*, **25**, Article ID: 495401. https://doi.org/10.1088/0953-8984/25/49/495401

[28] Neumann, M., van de Streek, J., Fabbiani, F., Hidber, P. and Grassmann, O. (2015) Combined Crystal Structure Prediction and High-Pressure Crystallization in Rational Pharmaceutical Polymorph Screening. *Nature Communications*, **6**, Article No. 7793. https://doi.org/10.1038/ncomms8793

[29] Neumann, M. and van de Streek, J. (2018) How Many Ritonavir Cases Are There Still out There? *Faraday Discussions*, **211**, 441-458. https://doi.org/10.1039/C8FD00069G

[30] Mortazavi, M., Hoja, J., Aerts, L., Quéré, L., van de Streek, J., Neumann, M.A. and Tkatchenko, A. (2019) Computational Polymorph Screening Reveals Late-Appearing and Poorly-Soluble form of Rotigotine. *Communications Chemistry*, **2**, Article No. 70. https://doi.org/10.1038/s42004-019-0171-y

[31] Zhang, Y.Y., Gao, W., Chen, S., Xiang, H. and Gong, X.G. (2015) Inverse Design of Materials by Multi-Objective Differential Evolution. *Computational Materials Science*, **98**, 51-55. https://doi.org/10.1016/j.commatsci.2014.10.054

[32] Curtis, F., Li, X., Rose, T., Vazquez-Mayagoitia, A., Bhattacharya, S., Ghiringhelli, L.M. and Marom, N. (2018) GAtor: A First-Principles Genetic Algorithm for Molecular Crystal Structure Prediction. *Journal of Chemical Theory and Computation*, **14**, 2246-2264. https://doi.org/10.1021/acs.jctc.7b01152

[33] Hajinazar, S., Thorn, A., Sandoval, E.D., Kharabadze, S. and Kolmogorov, A.N. (2021) MAISE: Construction of Neural Network Interatomic Models and Evolutionary Structure Optimization. *Computer Physics Communications*, **259**, Article ID: 107679. https://doi.org/10.1016/j.cpc.2020.107679

[34] Yamashita, T., Kanehira, S., Sato, N., Kino, H., Terayama, K., Sawahata, H., *et al*. (2021) CrySPY: A Crystal Structure Prediction Tool Accelerated by Machine Learning. *Science and Technology of Advanced Materials*: *Methods*, **1**, 87-97. https://doi.org/10.1080/27660400.2021.1943171

[35] Mosley, J.W., Operskalski, E.A., Tobler, L.H., Andrews, W.W., Phelps, B., Dockter, J., *et al.* (1987) Genetic Algorithms and Their Applications. *Proceedings of the* 2*nd International Conference on Genetic Algorithms*, L. Erlbaum Associates Inc., 1-8.

[36] Seko, A., Togo, A. and Tanaka, I. (2018) Descriptors for Machine Learning of Materials Data. In: Tanaka, I., Ed., *Nanoinformatics*, Springer, Singapore, 3-23. https://doi.org/10.1007/978-981-10-7617-6_1

[37] Thompson, A.P., Aktulga, H.M., Berger, R., Bolintineanu, D.S., Brown, W.M., Crozier, P.S., *et al.* (2022) LAMMPS—A Flexible Simulation Tool for Particle-Based Materials Modeling at the Atomic, Meso, and Continuum Scales. *Computer Physics Communications*, **271**, Article ID: 108171. https://doi.org/10.1016/j.cpc.2021.108171

[38] Kresse, G. and Furthmüller, J. (1996) Efficient Iterative Schemes for *Ab Initio* Total-Energy Calculations Using a Plane-Wave Basis Set. *Physical Review B*, **54**, 11169-11186. https://doi.org/10.1103/PhysRevB.54.11169

[39] Frisch, M.J., Trucks, G.W., Schlegel, H.B., Scuseria, G.E., Robb, M.A., Cheeseman, J.R., *et al*. (2016) Gaussian-16 Revision B.01. Gaussian Inc., Wallingford, CT.

[40] Yoo, A.B., Jette, M.A. and Grondona, M. (2003) SLURM: Simple Linux Utility for Resource Management. *Workshop on Job Scheduling Strategies for Parallel Processing*, Seattle, 24 June 2004, 44-60. https://doi.org/10.1007/10968987_3

[41] Bode, B., Halstead, D.M., Kendall, R., Lei, Z. and Jackson, D. (2000) The Portable Batch Scheduler and the Maui Scheduler on Linux Clusters. 4*th Annual Linux Showcase & Conference* (*ALS* 2000), Atlanta, 10-14 October 2000, 27-34.

[42] Gentzsch, W. (2001) Sun Grid Engine: Towards Creating a Compute Power Grid. *Proceedings First IEEE/ACM International Symposium on Cluster Computing and the Grid*, Brisbane, 15-18 May 2001, 35-36. https://doi.org/10.1109/CCGRID.2001.923173

[43] Thain, D., Tannenbaum, T. and Livny, M. (2005) Distributed Computing in Practice: The Condor Experience. *Concurrency and Computation*: *Practice and Experience*, **17**, 323-356. https://doi.org/10.1002/cpe.938

[44] Banker, K., Garrett, D., Bakkum, P. and Verch, S. (2016) Mongodb in Action: Covers Mongodb Version 3.0. Simon & Schuster, New York.

[45] Barrett, C.S. (1956) X-Ray Study of the Alkali Metals at Low Temperatures. *Acta Crystallographica*, **9**, 671-677. https://doi.org/10.1107/S0365110X56001790

[46] Trucano, P. and Chen, R. (1975) Structure of Graphite by Neutron Diffraction. *Nature*, **258**, 136-137. https://doi.org/10.1038/258136a0

[47] Raynor, G. and Hume-Rothery, W. (1939) A Technique for the X-Ray Powder Photography of Reactive Metals and Alloys with Special Reference to the Lattice Spacing of mg at High Temperatures. *Journal of the Institute of Metals*, **65**, 477-485.

[48] Wu, H., Hartman, M., Udovic, T., Rush, J., Zhou, W., Bowman, R. and Vajo, J. (2007) Structure of the Novel Ternary Hydrides Li_4Tt_2D (*Tt* =Si and Ge). *Acta Crystallographica. Section B, Structural Science*, **B63**, 63-68. https://doi.org/10.1107/S0108768106046465

[49] Elliot, A. (2010) Structure of Pyrrhotite 5*C* (Fe_9S_{10}) *Acta Crystallographica. Section B, Structural Science*, **B66**, 271-279. https://doi.org/10.1107/S0108768110011845

[50] H. d'Amour, W.D. and Schulz, H. (1979) Structure Determination of α-Quartz up to 68 x 10^8 Pa. *Acta Crystallographica. Section B, Structural Science*, **B35**, 550-555. https://doi.org/10.1107/S056774087900412X

[51] Kuwayama, Y., Hirose, K., Sata, N. and Ohishi, Y. (2005) The Pyrite-Type High-Pressure Form of Silica. *Science*, **309**,

923-925. https://doi.org/10.1126/science.1114879

[52] Irving, P.E. and Beevers, C.J. (1971) Some Metallographie and Lattice Parameter Observations on Titanium Hydride. *Metallurgical Transactions*, **2**, 613-615. https://doi.org/10.1007/BF02663362

[53] Pauling, L. and Hendricks, S.B. (1925) The Crystal Structures of Hematite and Corundum. *Journal of the American Chemical Society*, **47**, 781-790. https://doi.org/10.1021/ja01680a027

[54] Duan, W., Wentzcovitch, R.M. and Thomson, K.T. (1998) First-Principles Study of High Pressure Alumina Polymorphs. *Physical Review B*, **57**, 10363-10369. https://doi.org/10.1103/PhysRevB.57.10363

[55] Oganov, A.R. and Ono, S. (2004) Theoretical and Experimental Evidence for a Post-Perovskite Phase of $MgSiO_3$ in Earth's D'' Layer. *Nature*, **430**, 445-448. https://doi.org/10.1038/nature02701

[56] Fang, C.M. and de With, G. (2002) Crystal Structure and Chemical Bonding of the High-Pressure Phase of mgal2o4 from First-Principles Calculations. *Philosophical Magazine A*, **82**, 2885-2894. https://doi.org/10.1080/01418610208240072