第2卷 第2期 2020年4月
Vol.2 No.2 Apr.2020

ISSN 2096-742X
CN 10-1649/TP

文献DOI：
10.11871/jfdc.issn.2096-742X.2020.02.006

文献PID：
21.86101.2/jfdc.2096-742X.2020.02.006

页码：78-90

开放科学标识码（OSID）

材料科学数据库在材料研发中的应用与展望

李姿昕[1,2,3]，张能[1,2,3]，熊斌[1,2,3]，胡云凤[1,2,3]，赵新鹏[1,2,3]，黄海友[1,2,3*]

1. 新材料技术研究院，北京科技大学，北京 100083
2. 北京材料基因工程高精尖创新中心，北京科技大学，北京 100083
3. 材料基因工程北京市重点实验室，北京科技大学，北京 100083

摘　要：［**目的**］随着“大数据”时代的来临，大数据技术由于可显著加速材料研发，已经成为材料科学研究者关注的热点技术之一。基于材料数据库平台的材料大数据技术更是成为“材料基因工程”的三大核心技术之一。因此，材料数据库建设对于加速新材料的研发至关重要。［**方法**］本文通过对国内外材料科学数据库的建设及应用的概括和总结，并结合材料科学数据库的发展趋势，提出了未来的研究方向。［**结果**］材料基因组（工程）理念的提出和大数据技术的快速发展，促进了国内外大量材料科学数据库的建立。相较国外而言，国内的材料科学数据库建设相对较晚。但在“十三五”国家重点研发计划专项的支持下，我国材料科学数据库平台建设有望在未来几年内取得初步成效。［**结论**］材料科学数据库的建设已经成为材料基因工程技术发展进程当中一种不可或缺的要素，但在数据库建设和应用过程中还存在很多困难亟待解决，材料科学数据库的发展仍任重道远。

关键词：数据库；大数据技术；材料信息学；材料基因工程；机器学习

Materials Science Database in Material Research and Development: Recent Applications and Prospects

Li Zixin[1,2,3], Zhang Neng[1,2,3], Xiong Bin[1,2,3], Hu Yunfeng[1,2,3], Zhao Xinpeng[1,2,3], Huang Haiyou[1,2,3*]

1. Institute for Advanced Materials and Technology, University of Science and Technology Beijing, Beijing 100083, China
2. Beijing Advanced Innovation Center for Materials Genome Engineering, University of Science and Technology Beijing, Beijing 100083, China
3. Beijing Key Laboratory of Materials Genome Initiative, University of Science and Technology Beijing, Beijing 100083, China

Abstract: [**Objective**] With the advent of the "Big Data" era, big data technology has become one of the hottest technologies attracting material science researchers because it can significantly accelerate the development of materials. The material big data technology based on the material database platform is one of the three core technologies of "Materials Genome Engineering". Therefore, the construction of a material database is very important to acceleration of the development of new materials. [**Methods**] This article summarizes the constructions and applications of material

基金项目：国家重点研发计划资助项目（2016YFB0700500，2018YFB0704300）；广东省重点领域研发计划项目（2019B010940001）；北京科技大学顺德研究生院科技创新专项资金（BK19BE030）

*通讯作者：黄海友（E-mail: huanghy@mater.ustb.edu.cn）

science databases at home and abroad, and puts forward future research directions based on the development trend of material science databases. [**Results**] The advancement of the material genome (engineering) concept and the rapid development of big data technologies have promoted the establishment of a large number of material science databases at home and abroad. Compared with developed countries, the material science database construction in China is relatively late. However, with the support of the 'Thirteenth Five-Year Plan' national key research and development program of China, the construction of China's material science database platform is expected to achieve initial results in the next few years. [**Conclusions**] The construction of material science database has become an indispensable element in the development process of material genome engineering technology. But there are still many difficulties to be resolved in the process of database construction and application. The development of a material science database remains a challenging task.

Keywords: database; big data technology; material informatics; material genome engineering; machine learning

引言

材料是科技发展的基础和先导，随着全球新一轮工业革命浪潮的掀起，加速材料的研发进程成为世界各国共同的追求。如何基于低成本、高可靠性的预测方法理性指导实验来快速获得定制性能的新材料成为与之相关的关键问题。随着“大数据”时代的到来，以机器学习等人工智能技术为代表的材料信息学领域快速发展，并迅速成为材料设计与开发的有力工具。机器学习技术已经在很多材料研究中得到了应用。例如，Xue 等通过机器学习自适应设计，仅实验合成 36 种预测成分的合金试样，就可以从包含约 800,000 种不同成分的搜索空间中找到具有极小热滞的新型多组元 Ni-Ti 基形状记忆合金 [1]。Kiyohara 等通过采用机器学习方法，仅计算不超过 0.18% 的晶体结构的偏析能即可准确得到合金元素在晶界偏析的稳定构型 [2]。Wen 等采用机器学习引导实验的策略，在机器学习反馈回路的辅助下仅通过 7 次实验便得到了高硬度高熵合金 [3]。

机器学习技术已经被证明可以有效地加速材料的研发进程。人类社会已经进入了“大数据”时代，数据资源已经得到了广大科学研究者的重视，即使是“失败”的数据，也可以用来辅助训练机器学习模型来预测成功条件 [4]。机器学习不仅能够对材料性能进行预测，同时，借助机器学习挖掘的边界条件等信息，也有助于推进对相关机理的认识。Stanev 等就是通过机器学习研究了每个超导体系中预测因子的重要性，获得了关于不同体系驱动超导性的物理机制 [5]。

然而，这种方法取决于是否有足够多的高质量的数据。但是在材料科学研究中，建立准确的机器学习模型往往需要“海量”数据进行训练。Rahaman 等建立的可对未知化学成分的钢铁材料 M_s 准确预测的机器学模型，使用了包含 2 277 条化学成分和 M_s 数据的数据库 [6]；Schmidt 等人为了通过机器学习预测立方钙钛矿体系的热力学稳定性，更是构建一个包含约 250,000 条 DFT 计算数据集 [7]。但材料科学研究面临更普遍的情况是小数据困境，即所研究的材料对象缺乏足够的高质量数据。其中一个主要原因是由于数据分散造成的，Zhou 等在采用机器学习对高熵合金进行相分类研究的过程当中，从 134 篇文献当中收集了 601 条数据来作为数据集 [8]，这大大增加了研究人员的工作量。

因此，数据库的建设成为了信息学技术在材料科学应用中的重要组成部分。美国在 2011 年奥巴马总统提出材料基因组计划时，将材料数据库作为三大基础平台之一，其建设得到了高速发展。本文首先介绍了国内外较为知名的材料数据库及其使用情况；然后，分析了数据库如何帮助机器学习技术在

材料科学研究中得到广泛应用；最后，讨论了数据库建设和应用中所面临的困难及其发展趋势。

1 数据库概述

想要实现材料基因组工程这一颠覆性研发新模式，数据共享与计算工具开发显得至关重要。数据库作为材料基因工程不可或缺的一部分，已经得到了材料科学研究者们的重视，目前，国外较为著名的材料信息数据库有加州大学伯克利分校的劳伦斯伯克利国家实验室和麻省理工学院等单位联合组建的 Materials Project[9]、杜克大学组建的 AFLOW[10] 以及美国西北大学组建的 OQMD[11-12] 等。我国在科技部、工业和信息化部等部门的大力支持下，以中国材料基因工程专用数据库为代表的材料科学数据库在快速建设当中，并且在机器学习应用领域已经取得了初步成果。

1.1 国外材料数据库建设情况

Materials Project（MP）计算材料数据库平台（https://www.materialsproject.org/），是由美国劳伦斯伯克利国家实验室（LBNL）和麻省理工学院（MIT）等单位在 2011 年材料基因组计划提出后联合开发的开放性数据库。如图 1 所示，MP 数据库存储了几十万条包括能带结构、弹性张量、压电张量等性能的第一性原理计算数据。材料体系涉及无机化合物、纳米孔隙材料、嵌入型电极材料和转化型电极材料。其中大部分的化合物都来自于 Inorganic Crystal Structure Database（ICSD）无机晶体结构数据库，数据在收录前会经过检测，所以其数据具有较高的准确性。平台中的 MP 专用计算软件也是该数据库的主要特色之一，目前已经开发完成了 Materials Explorer、Battery Explorer、Structure Predictor 等 15 个应用程序并得到了广泛应用。通过这些与数据库相关联的软件可在线对未知材料的性能进行预测，大大减少了实验量，加快了材料的开发速度。

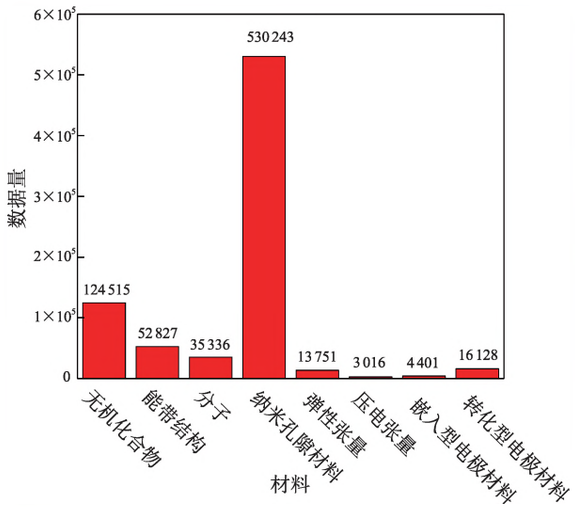

图 1 Materials Project 数据库数据量统计

Fig.1 Materials Project database statistics

AFLOW 计算材料数据库（http://www.aflowlib.org/），是由杜克大学在 2011 年开发的一个开放数据库。数据库中包含了大量第一性原理计算所得的数据，如图 2 所示，目前已存储了关于无机化合物、二元合金与多元合金等超过 557 043 524 条涉及 2 945 940 种材料的结构、性能数据，其中绝大多数数据都是预测得出的，是诸多数据库中数据含量最大的一个[13]。与 Materials Project 数据库相似，基于密度泛函理论（DFT）的量子力学计算、信息学数据挖掘和进化结构筛选策略[10]，AFLOW 计算材料数据库运用了高通量第一性原理计算，故其拥有很好的计算性能。AFLOW 数据库有 AFLOWπ、AFLOW-ML 和 PAOFLOW 等共 12 种应用程序可以有效地对材料的结构、性能等进行筛选。AFLOWπ[14] 通过引入第一性原理计算来获得材料的能带结构、态密度、声子色散、弹性特性、复介电常数、电子转移系数。以减少普通用户的技术性难题为出发点，AFLOW-ML[15] 简化了 AFLOW 的机器学习方法，提供了一个开放的 RESTful API 可访问不断更新的算法来保证各种工作流的正常运行，帮助研究人员更好地预测材料性能，推动了机器学习方法在材料中的应用。

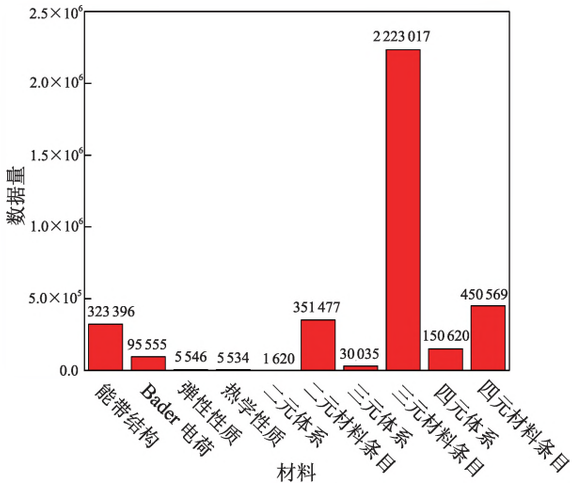

图 2 AFLOW 数据库数据量统计

Fig.2 AFLOW database statistics

Open Quantum Materials Database（OQMD） 开放量子材料数据库（http://oqmd.org/），是由美国西北大学 Chris Wolverton 团队于 2013 年建立的一个基于 DFT 计算的 637 644 种材料的热力学性质和结构的数据库，同时提供 API 接口来下载数据。正如其名，OQMD 数据库是诸多数据库中开放程度最高的一个。在 OQMD 计算平台上，研究人员不仅可以按需搜索材料的晶体结构、能带和能量等性质，还可以训练机器学习模型，用以识别潜在的新三元化合物 [16]。此外，该数据库还可以利用元素计算法给出材料的相图从而预测热力学稳定相。OQMD 计算平台的准确性也得到了大多用户的肯定，Scott Kirklin 等人 [12] 通过具体实验对比发现，运用 OQMD 计算平台可以基本准确地预测大多数元素的晶体结构与形成能。

Materials Project，AFLOW 和 OQMD 都是基于量子力学计算建设的数据库，这三个数据库计算数据所基于的晶体结构大多来自于 ICSD 数据库 [17]。ICSD 无机晶体结构数据库（http://icsd.fiz-karlsruhe.de/）的构建是由德国波恩大学无机化学研究所 Gunter Bergerhoff 教授首先提出的，自 1913 年创建以来，先后经由波恩大学、FIZ 研究所、Gmelin 研究所及美国国家标准与技术研究所（NIST）进行维护管理 [18]。该数据库建立时间较长，涵盖了金属、合金、陶瓷等非有机化合物的晶体结构信息。到目前为止，数据库中包含了超过 9 千种结构原型，共计超过 21 万种晶体结构条目，如图 3 所示，已经形成了世界最大的无机晶体结构数据库。数据库中的数据都是经过专家团队全面检查后才会上传到数据库当中。ICSD 每年都会更新两次数据，这些数据部分来源于出版期刊或实验室，还有部分来源于计算机程序生成。用户可通过参考文献、化学组成、晶胞参数、对称性以及实验和代码信息 5 种不同的方式对数据进行检索。因此，在新材料的研究过程中，ICSD 数据库被研究人员广泛应用。

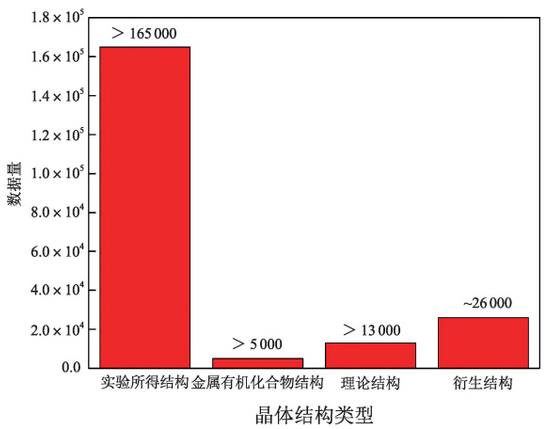

图 3 ICSD 数据库统计

Fig.3 ICSD database statistics

除了以上几个著名的材料信息数据库以外，还有一些影响力较大的数据库。由美国国家标准与技术研究所 NIST 开发的标准参考数据库系列有百余个数据库（https://www.nist.gov/srd/），其中材料类的有材料性能数据库与晶体结构数据库等，涵盖了腐蚀性能、高温超导、热力学性能、摩擦性能等内容，可按需通过分子式、分子量、化合物名称、CAS 号等途径查找，有图谱分析、同位素计算等功能。日本国立材料科学研究所开发的 MatNavi 数据库（https://mits.nims.go.jp/），涵盖了金属材料、复合材料、超导材料、聚合物、高温合金等材料种类的大

量数据内容。除基础数据库外，还包括工程数据库（如CCT曲线数据库）、四个应用与结构材料在线数据表。用户可通过搜索关键字、类别、数值等查找相应数据，有自己独创的检索系统，其输出方式也呈现多样化。NOMAD（https://www.nomad-coe.eu/）是由欧洲卓越中心（European Centre of Excellence）开发的，该数据库中的数据部分来自世界各地的研究人员与实验室，也有部分来自于其他数据库，如AFLOWlib和OQMD。该数据库的主要特色为可暂存研究人员的代码和数据，用户可以对比世界各地研究人员的计算结果，从而可以更好地研究材料的结构性能，这一特点使NOMAD数据库从众多高通量计算平台中脱颖而出。由日本科学技术公司（JST）与瑞典物相数据系统（MPDS）于1995年合作创立的PAULING FILE数据库（http://www.paulingfile.com/），主要应用于无机材料的设计与开发。该数据库在建立之初就是希望能够应用于材料数据挖掘中，能够发现可以应用于材料设计的新模式。Material Connexion

表1 主要材料科学数据库对比

Table 1 Comparison of major material science databases

数据库	材料类型	特点	网址链接
Materials Project	包括锂电池、沸石、金属有机框架等材料	数据具有较高的准确性	https://www.materialsproject.org/
AFLOW	主要为金属材料	最大的数据库	http://www.aflowlib.org/
OQMD	主要为钙钛矿材料	用户可以下载完整的数据库	http://oqmd.org/
ICSD	自1913年以来出版的已知的无机晶体结构	世界最大的无机晶体结构数据库	http://icsd.fiz-karlsruhe.de/
NIST	几乎涵盖所有材料体系	由百余个子库构成，具有严格评估标准	https://www.nist.gov/srd/
MatNavi	包括聚合物、陶瓷、合金、超导材料等材料	综合性数据库	https://mits.nims.go.jp/
PAULING FILE	主要为非有机固态材料	包括相图、晶体学数据、衍射模式和物理特性	http://www.paulingfile.com/
NOMAD	几乎涵盖所有材料体系	可暂存研究人员的代码和数据，用户可以对比世界各地研究人员的计算结果，从而可以更好地研究材料的结构性能	https://www.nomad-coe.eu/
Material Connexion	包含了碳基材料、水泥基材料、陶瓷材料、玻璃材料、金属材料、天然材料、高分子材料、材料工艺等	将材料学家与设计师直接联系起来的创新材料咨询服务机构	https://www.materialconnexion.com/
Materials Web	包括二维材料和层状体材料	在线存储二维材料电子结构为主的数据库	https://www.materialsweb.org/
Matdat	包括铝合金、钛合金等600多种材料	独特的综合在线平台，提供材料数据库和材料相关服务	https://www.matdat.com/
材料学科领域基础科学数据库	主要包括金属材料、无机非金属材料、闪烁材料、碳化硅材料、纳米材料和有机高分子材料	国内最全面的材料科学数据库之一	http://www.matsci.csdb.cn/
国家材料科学数据共享网	以钢铁材料、先进合金材料为主，也包含无机非金属材料和高分子材料	国内30余家科研单位参与建设，以整合、重构现有的、较为成熟的材料科学数据资源为基础。	http://www.materdata.cn/
MGED	以核材料、特种合金、生物医用材料、催化材料和能源材料为主，涉及几乎所有材料体系	我国最大的材料基因工程数据库平台，除数据库外，平台还拥有第一原理在线计算引擎、原子势函数库、在线数据挖掘系统等众多功能	http://www.mgedata.cn/

数据库（https://www.materialconnexion.com/）由George M.Beylerian创立，数据库中包含了碳基材料、水泥基材料、陶瓷材料、玻璃材料、金属材料、天然材料、高分子材料、材料工艺在内的超过一万条材料数据。这是一个将材料学家与设计师直接联系起来的创新材料咨询服务机构。它不仅包含大量的馆藏与数据库，还有专业的研究团队提供咨询服务与线下材料图书馆，设计师可以亲自感受、挑选各种新材料。由美国佛罗里达大学Hennig课题组创建和管理的Materials Web（https://www.materialsweb.org/）是一个以在线存储二维材料电子结构为主的数据库。用户可免费获取数据库中约700余条二维材料和1万余条层状体材料的结构、电子和热力学数据，数据库还支持生成VASP工作流和表征材料结构特征，未来还将引入机器学习工具方便科研人员进行材料科学研究。Matdat（https://www.matdat.com/）是由MATDAT LLC的创始人Robert Basan创立的一个综合平台，它包括一个材料性能数据库，一个超过12 000条信息的关于实验室、供货商、制造商名录与即将开放的研究数据储存平台。其中，材料性能数据库包括铝合金、钛合金等600多种材料的1 500多条信息。而像SpringerMaterials[19]，Materials Cloud[20]，COD[21]和ChemSpider[22]等数据库，在其相关领域也有一定的影响力。

1.2 国内材料数据库建设情况

相较于国外一些著名的材料数据库而言，我国在这方面起步较晚。为了更有效地应用和积累科学数据，在1987年，中国科学院牵头正式启动科学数据资源建设。经过多年发展，2019年全面改版的中国科学院数据云门户网站（http://www.csdb.cn/）投入运行。目前，数据库中包括1144个数据集，访问人数超过了16 000万，下载量更是高达2 000TB。这其中由中国科学院金属研究所承建的“材料学科领域基础科学数据库”（http://www.matsci.csdb.cn/）是国内最全面的材料科学数据库之一，主要包括金属材料、无机非金属材料、闪烁材料、碳化硅材料、纳米材料和有机高分子材料等子数据库。目前材料科学主题数据拥有数据总量7万余条。其中金属材料节点6万余条，无机非金属材料节点数据1万余条，涵盖了材料的热学、力学和电学等各种性能，其数据来源主要以手册、期刊文献数据为主，极大地促进了新技术与学科领域的融合发展。

我国从2001年开始逐步启动了科学数据共享工程。以国家科技部“十一五”基础条件平台项目“材料科学数据共享与服务平台建设”为依托的“国家材料科学数据共享网”（http://www.materdata.cn/）便是其中的一项重点工程。目前已整合了全国各地30余家科研单位的数据资源，其中包括了3 000种钢铁材料及材料基础的高质量数据近11万条，数据库中以材料体系划分，分为了材料基础、有色金属材料及特种合金、黑色金属材料、复合材料、有机高分子材料、无机非金属材料、信息材料、能源材料、生物医学材料、天然材料及制品、建筑材料和道路交通材料12个大类。国家材料科学数据共享网的建设为材料研究领域提供了数据共享服务与应用支撑。

2016年，由北京科技大学牵头建立的“材料基因工程专用数据库（MGED）”（http://www.mgedata.cn/）是一个基于材料基因工程的思想和理念建设的数据库和应用软件一体化系统平台。截至目前，该数据库平台包含的催化材料、铁性材料、特种合金、生物医用材料以及材料热力学和动力学数据库等各类材料数据的总量超过了76万条，累计查看量超过2万次。该平台包括了基于云计算模式的材料高通量第一性原理计算软件以及融合数据库的材料数据挖掘计算网络平台，可以实现批量作业的自动生成，并且可以对计算的结果进行自动处理、解析和数据汇交。除此之外，该平台还包含了论文信息辅助提取软件，使用人员可以使用该软件提取所阅读的论文当中的实验数据，从而可以为该平台的材料数据库填充材料数据。平台包含在线数据挖掘系统，可

直接调用数据库数据开展数据挖掘和机器学习。

除此之外，国内还建成了很多专项数据库，包括国家纳米科学中心建立的纳米研究专业数据库、北京科技大学牵头建立的国家材料环境腐蚀科学数据中心、中国科学院化学研究所承担建设的高分子材料科学数据资源节点等。这些数据库虽然使用范围相对较小，但是在特定的研究领域具有很强的针对性。

2 数据库在材料信息学领域中的应用

2.1 基于材料数据库的机器学习应用案例

如今，在“大数据”时代中，数据是进行材料科学研究的基础，而采用机器学习进行材料研究的时候，更是需要庞大数据量的支持，材料信息数据库可以非常便捷地储存和利用现有的严重碎片化的材料数据[23]。材料数据库作为材料基因工程的核心技术之一，在材料基因工程领域研究中具有不可忽视的作用，同时也为研究中数据的获取提供了便捷。数据库在机器学习研究过程当中具有不同的应用方式。

采用数据库中的数据作为训练集来训练机器学习模型，这是数据库在机器学习研究当中最广泛的应用方式。机器学习往往需要大量数据来训练模型，而数据库可以提供大量的数据支持。Tehrani 等以 Materials Project 数据库中的 3 246 个弹性模量作为训练集训练的模型，通过对晶体结构数据库中 118 287 个化合物进行预测，得到了由支持向量机回归确定的最大体模量和最大剪切模量的材料，选择典型化合物进行合成测量后发现误差小于 10%[24]。不只是理论计算类数据库在机器学习中有着重大应用，实验类的数据库也具有不可忽视的作用。Agrawal 等利用 NIMS 的数据库中的实验数据，通过对特征选择和预测建模在内的不同数据科学技术在钢材疲劳性能中的应用进行探讨，发现一些先进的数据分析技术可以在预测精度上取得显著提高，成功地证明了这种数据挖掘工具可用于按预测钢铁疲劳强度的潜力顺序对成分和工艺参数进行排名，并实际开发了相应的预测模型[25]。Stanev 等在超导临界温度的机器学习建模研究中，其数据集来自于 NIMS 创建和维护的 SuperCon 数据库，所建立的模型具有较强的预测能力，样本外推准确率约为 92%[5]。

除了作为训练集，还可以将数据库中的数据作为测试集来检验训练完成的机器学习模型的性能，采用第一性原理计算的数据训练的机器学习模型可以有效地预测晶体化合物的振动性质[26]。在这个研究中，将振动性质的预测值和 NIST 数据库中的实验值进行了对比，发现预测结果与实验结果之间的一致性是显著的。这表明该模型可以有效并且快速地预测晶体化合物的振动性质。

机器学习模型也可以对数据库中的材料进行性能预测。Cheon 等将通过三维晶体结构的原子位置训练好的机器学习模型应用于 Materials Project 数据库中的 5 万余个无机晶体材料后，可以识别出 1 173 个二维层状材料和 487 个由弱键一维分子链组成的材料。对于大多数不清楚是二维或一维材料的材料，这个模型识别材料的数量增加了一个数量级[27]。

很多数据库都内置了高通量计算框架或势库，可以间接为机器学习研究提供数据支持。在 AFLOW 数据库的高通量计算框架下，结合机器学习方法评估了大约 400 个半导体氧化物和氟化物与立方钙钛矿结构在 0、300 和 1000K 下的力学稳定性。找到了 92 种在高温下力学稳定的化合物，其中 36 种未在以往的文献中提及[28]。采用 MGED 数据库中的晶格反演势库结合机器学习，可以在大约 50 万个候选合金中快速找到具有最高相变熵变的 Cu-Al 基形状记忆合金，同时得到了部分合金元素对合金相变熵变的影响规律[29]。

数据库可以将碎片化数据整合，并不断积累，为材料研究提供数据支持。在机器学习辅助镍基单晶高温合金晶格错配度预测的研究中[30]，其数据集来源于文献摘录。而在利用机器学习算法训练实验数据预测粉末冶金材料烧结密度的研究中[31]，数据则来源于实验室积累以及文献收集。这些研究的数

据虽然来源于文献以及实验室的收集，但是为了指导未来的合金设计，都被收集在了国家材料科学数据共享网中。该数据库中的所有数据均经过所属单位和文献出处信息的验证，保证了质量的可靠性。

2.2 发展中的高通量计算软件

对于材料数据库来说，通过第一性原理等高性能、高通量的材料计算进行材料理论数据获取，并结合实验数据和经验数据，再利用信息化技术对大规模、多源异构的材料数据进行处理分析，由此才能对材料数据库所存储的数据进行充分的挖掘和利用[32]。目前，常用的高通量计算框架包括 Materials Project 和 AFLOW 等都具有较高的入门门槛。因此，高通量计算软件的发展也变得刻不容缓。

上海鞍面智能科技有限公司的 LASP 软件利用最新的高效神经网络势能面方法来进行势能面模拟计算，解决了诸如晶体结构预测、相变动力学、反应路径预测等许多复杂的反应路径及材料体系中的问题。高岩涛等人[33]基于第一性原理，利用平面波基组、赝势方法进行电子结构计算、分子动力学模拟，研发了 GPU 加速计算平台 PWMat，其比相同的 CPU 软件（例如 PEtot）的计算速度要快 20 倍左右，能够在平台上面实现 4 000 电子以上体系的模拟计算。中国科学院计算机网络信息中心的杨小渝等人研发了高通量材料计算平台 MatCloud，以及高通量材料计算数据库 MatCloudLib[34]。具有晶体结构建模、图形化界面的流程设计、性质预测、结果分析、数据提取与查询、与计算资源的集成等特色，并且可以完成对计算结果的可视化分析及展示。王宗国等人[35]以 Fe-Al 和 Al-Ti 体系为例，采用 MatCloud 的特色工作流技术快速筛选出了掺杂的稳定结构，相较于遍历筛选，计算量分别减少了 66% 和 84%。而由北京航空航天大学的孙志梅等人开发的计算平台 ALKEMIE 同样包含计算平台 MATTER STUDIO (MS) 以及数据库 DATAVAULT (DV) 两个部分，并且可以全自动地进行建模、运行以及数据分析。其中 MS 计算平台集成了第一性原理、热力学、经典分子动力学及动态蒙特卡洛模拟等计算引擎，DV 数据库当中的材料结构数据超过了 18 万条，计算完成的材料性能数据超过 1 万条。

3 存在的问题与展望

材料数据是材料科学研究的基础，随着“材料基因工程”的提出与实施，材料科学数据呈现出爆炸式增长的态势。对于材料数据库来说，其最主要的作用之一就是积累材料数据，为材料计算和实验提供数据支撑。所以已有研究数据的积累对于材料数据库的建立是十分必要的。但是国内在数据库方面的资源储备量远远不如美国、欧洲、日本等发达国家，我国的材料科学数据库处于建设初期，还无法很好地为材料科学研究者们服务，还无法满足应用的需求。

3.1 存在的问题

目前，中国材料信息数据库的建设与应用面临着很大的挑战，主要表现在以下几个方面。

（1）数据库的数据量远远不够。相较于一些发达国家而言，中国的材料科学数据库在建设方面起步较晚，数据积累量远远不足，已有的几个国家级数据库中的数据不够丰富，还处于建设初期。在“大数据”时代背景下，相比于其他领域数据量的积累速度，材料领域的数据量积累速度也较慢。

（2）数据质量评价方法与机制亟需完善。失败实验的数据依旧可以为研究工作提供其应有的价值，但是，错误的数据只会阻碍研究的进展。无论国内国外，在数据库建立之初都会将数据的质量列为重中之重。但是，错误的数据难免会存在，这就需要材料科学工作者们严格把关，将错误的数据拒之门外，为机器学习研究减少“噪音”的影响。

（3）明确数据分类。材料根据不同的分类方式

有很多类别，材料数据的分类应该根据权威的材料分类体系进行划分。同时，还应该加强年轻学生和科研工作者对材料分类的学习，在进行数据收集的时候就可以避免分类混乱，减少日后数据库的维护成本。

（4）材料数据的获取过程较为复杂。无论是材料计算数据还是实验数据，对工艺参数都显得十分敏感，往往一些工艺参数的微小变化，就可以使得同种材料的数据产生巨大差异。在进行数据收集的时候，还需要严格数据格式，明确数据来源以及数据的生产条件。

（5）数据的共享程度仍有待提高。在现在这个“大数据”时代，已经有很多科研机构和生产单位意识到了数据的重要性。不同的研究单位往往都拥有自己的数据库，但是，这些数据库的共享程度非常低，并且很多都是单一性能或者单一材料体系，无法形成一个系统的综合类材料信息数据库。而且数据格式也具有其自身的特色，这也影响了其共享程度。

（6）数据知识产权问题依旧严峻。这也是造成数据共享程度较低的一个主要原因。“大数据”时代，数据是一笔很大的“财富”，而对这笔“财富”的知识产权属性和保护还没有一个明确的法律界定，很多研究工作者也不愿意无偿贡献数据，尤其是一些生产单位的数据，更是涉及到了其商业机密。

（7）生产数据的收集有很大的困难。一些生产数据会涉及到生产单位的核心技术或者商业机密。但是部分不涉密数据的收集力度依旧不大，很多数据库在这方面存在很大空白。

（8）数据的收集、更新，与数据库的维护需要专业人员监管。现在一些数据库的数据收集、更新与数据库的维护是由青年学生和研究工作者完成的，但是部分学生和研究工作者对材料科学领域的知识理解得不够深刻和系统，在进行数据库建设的时候往往会造成很多失误，影响了数据库中的数据质量和数据库的建设进度，所以需要专业人员进行监管。

3.2 未来的发展方向

材料信息数据库的建设刻不容缓，由于近年来“材料基因工程”的提出与发展，数据库的建设与发展也受到了极大的关注。中国在材料科学研究领域已经积累了大量的数据，但是，这些数据还没有很好地被收集起来，加大材料数据收集和共享力度显得十分重要。而在收集数据的过程当中应该对数据质量严格把关，对数据格式严格要求，对数据知识产权问题加强管理，提高数据库中数据的质量和共享程度。高通量材料计算和高通量制备与表征是“大数据”时代补充材料信息数据库数据量的有效手段，发展高通量计算平台、高通量制备技术和高通量表征技术可以有效缓解数据收集困难的情况，同时也可以降低材料数据收集过程的复杂程度，增加相同工艺参数下材料的数据量。中国现在材料信息数据库的建设属于“边建设边使用”，数据库的建设是一项长期的工作，应该优先建立一些热门材料体系的专题材料数据库，优先解决国家科技重大专项和国防建设急需数据研究的情况。由于数据库建设是材料基因工程领域中重要的一环，同时中国数据库的发展与发达国家相比还有较大差距，因此，中国的材料数据库建设还具有很大的发展空间。

4 结束语

本文对国内外材料信息数据库的建设情况和使用情况进行了简单的介绍。总的来说，材料基因组工程领域作为一个新兴的科学研究领域，已经取得了初步成效。材料基因工程作为颠覆性技术，想要实现新材料研发周期缩短一半、研发成本降低一半的目标就离不开数据库的支撑。在过去约 10 年间的发展中，材料科学数据库的发展情况呈现出“百家争鸣”的态势，众多材料学研究者都认识到了数据的重要性。因此，未来几年中国材料科学数据库在建设和应用上将迎来一个快速发展时期。

利益冲突声明

所有作者声明不存在利益冲突关系。

参考文献

[1] D. Xue, P.V. Balachandran, J. Hogden, J. Theiler, D. Xue, T. Lookman. Accelerated search for materials with targeted properties by adaptive design[J]. Nature communications 7 (2016) 11241.

[2] S. Kiyohara, T. Mizoguchi. Effective search for stable segregation configurations at grain boundaries with data-mining techniques[J]. Physica B: Condensed Matter 532 (2018) 9-14.

[3] C. Wen, Y. Zhang, C. Wang, D. Xue, Y. Bai, S. Antonov, L. Dai, T. Lookman, Y. Su. Machine learning assisted design of high entropy alloys with desired property[J]. Acta Materialia 170 (2019) 109-117.

[4] P. Raccuglia, K.C. Elbert, P.D. Adler, C. Falk, M.B. Wenny, A. Mollo, M. Zeller, S.A. Friedler, J. Schrier, A.J. Norquist. Machine-learning-assisted materials discovery using failed experiments[J]. Nature 533(7601) (2016) 73.

[5] V. Stanev, C. Oses, A.G. Kusne, E. Rodriguez, J. Paglione, S. Curtarolo, I. Takeuchi. Machine learning modeling of superconducting critical temperature[J]. npj Computational Materials 4(1) (2018) 1-14.

[6] M. Rahaman, W. Mu, J. Odqvist, P. Hedström. Machine Learning to Predict the Martensite Start Temperature in Steels[J]. Metallurgical and Materials Transactions A 50(5) (2019) 2081-2091.

[7] J. Schmidt, J. Shi, P. Borlido, L. Chen, S. Botti, M.A. Marques. Predicting the thermodynamic stability of solids combining density functional theory and machine learning[J]. Chemistry of Materials 29(12) (2017) 5090-5103.

[8] Z. Zhou, Y. Zhou, Q. He, Z. Ding, F. Li, Y. Yang. Machine learning guided appraisal and exploration of phase design for high entropy alloys[J]. npj Computational Materials 5(1) (2019) 1-9.

[9] A. Jain, S.P. Ong, G. Hautier, W. Chen, W.D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation[J]. Apl Materials 1(1) (2013) 011002.

[10] S. Curtarolo, W. Setyawan, G.L. Hart, M. Jahnatek, R.V. Chepulskii, R.H. Taylor, S. Wang, J. Xue, K. Yang, O. Levy. AFLOW: an automatic framework for high-throughput materials discovery[J]. Computational Materials Science 58 (2012) 218-226.

[11] J.E. Saal, S. Kirklin, M. Aykol, B. Meredig, C. Wolverton. Materials design and discovery with high-throughput density functional theory: the open quantum materials database (OQMD)[J]. Jom 65(11) (2013) 1501-1509.

[12] S. Kirklin, J.E. Saal, B. Meredig, A. Thompson, J.W. Doak, M. Aykol, S. Rühl, C. Wolverton. The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies[J]. npj Computational Materials 1 (2015) 15010.

[13] S. Curtarolo, W. Setyawan, S. Wang, J. Xue, K. Yang, R.H. Taylor, L.J. Nelson, G.L. Hart, S. Sanvito, M. Buongiorno-Nardelli. AFLOWLIB. ORG: A distributed materials properties repository from high-throughput ab initio calculations[J]. Computational Materials Science 58 (2012) 227-235.

[14] A.R. Supka, T.E. Lyons, L. Liyanage, P. D'Amico, R.A.R. Al Orabi, S. Mahatara, P. Gopal, C. Toher, D. Ceresoli, A. Calzolari. AFLOW π : A minimalist approach to high-throughput ab initio calculations including the generation of tight-binding hamiltonians[J]. Computational

Materials Science 136 (2017) 76-84.

[15] E. Gossett, C. Toher, C. Oses, O. Isayev, F. Legrain, F. Rose, E. Zurek, J. Carrete, N. Mingo, A. Tropsha. AFLOW-ML: A RESTful API for machine-learning predictions of materials properties[J]. Computational Materials Science 152 (2018) 134-145.

[16] B. Meredig, A. Agrawal, S. Kirklin, J.E. Saal, J. Doak, A. Thompson, K. Zhang, A. Choudhary, C. Wolverton. Combinatorial screening for new materials in unconstrained composition space with machine learning[J]. Physical Review B 89(9) (2014) 094104.

[17] A. Belsky, M. Hellenbrandt, V.L. Karen, P. Luksch. New developments in the Inorganic Crystal Structure Database (ICSD): accessibility in support of materials research and design[J]. Acta Crystallographica Section B: Structural Science 58(3) (2002) 364-369.

[18] M. Hellenbrandt. The inorganic crystal structure database (ICSD)—present and future[J]. Crystallography Reviews 10(1) (2004) 17-22.

[19] https://materials.springer.com/.

[20] https://materialscloud.org/discover/.

[21] http://crystallography.net/.

[22] http://www.chemspider.com/.

[23] 尹海清, 刘国权, 姜雪, 张瑞杰, 曲选辉. 中国材料数据库与公共服务平台建设[J]. 科技导报, 2015,33(10) : 50-59.

[24] A. Mansouri Tehrani, A.O. Oliynyk, M. Parry, Z. Rizvi, S. Couper, F. Lin, L. Miyagi, T.D. Sparks, J. Brgoch. Machine learning directed search for ultraincompressible, superhard materials[J]. Journal of the American Chemical Society 140(31) (2018) 9844-9853.

[25] A. Agrawal, P.D. Deshpande, A. Cecen, G.P. Basavarsu, A.N. Choudhary, S.R. Kalidindi. Exploration of data science techniques to predict fatigue strength of steel from composition and processing parameters[J]. Integrating Materials and Manufacturing Innovation 3(1) (2014) 8.

[26] F. Legrain, J.s. Carrete, A. van Roekeghem, S. Curtarolo, N. Mingo. How chemical composition alone can predict vibrational free energies and entropies of solids[J]. Chemistry of Materials 29(15) (2017) 6220-6227.

[27] G. Cheon, K.A.N. Duerloo, A.D. Sendek, C. Porter, Y. Chen, E.J. Reed. Data mining for new two-and one-dimensional weakly bonded solids and lattice-commensurate heterostructures[J]. Nano letters 17(3) (2017) 1915-1923.

[28] A. van Roekeghem, J. Carrete, C. Oses, S. Curtarolo, N. Mingo. High-throughput computation of thermal conductivity of high-temperature solid phases: the case of oxide and fluoride perovskites[J]. Physical Review X 6(4) (2016) 041061.

[29] X.P. Zhao, H.Y. Huang, C. Wen, Y.J. Su, P. Qian. Accelerating the development of multi-component Cu-Al-based shape memory alloys with high elastocaloric property by machine learning[J]. Computational Materials Science 176 (2020) 109521.

[30] X. Jiang, H.Q. Yin, C. Zhang, R.J. Zhang, K.Q. Zhang, Z.H. Deng, G.Q. Liu, X.H. Qu. An materials informatics approach to Ni-based single crystal superalloys lattice misfit prediction[J]. Computational Materials Science 143 (2018) 295-300.

[31] Z. Deng, H. Yin, X. Jiang, C. Zhang, K. Zhang, T. Zhang, B. Xu, Q. Zheng, X. Qu. Machine leaning aided study of sintered density in Cu-Al alloy[J]. Computational Materials Science 155 (2018) 48-54.

[32] 王卓, 杨小渝, 郑宇飞, 雍岐龙, 苏航, 杨才福. 材料基因组框架下的材料集成设计及信息平台初探[J].科学通报, 2013, 58(35):3733-3742.

[33] 高岩涛, 贾伟乐, 王龙, 汪林望. 超软赝势密度泛函分子动力学计算中的若干优化算法[J].科研信息化技术与应用, 2015,6(4) :47-53.

[34] 杨小渝, 王娟, 任杰, 宋健龙, 王宗国, 曾雉, 张小丽, 黄孙超, 张平, 林海青. 支撑材料基因工程的高通量材料集成计算平台[J].计算物理, 2017, 34(6) :697-704.

[35] Z. Wang, X. Yang, L. Wang, J. Wang, M. Zhang, X. Zhao, J. Ren, Z. Zeng. CE Screen: An energy-based structure screening automatic workflow[J]. Computational Materials Science 143 (2018) 55-62.

收稿日期：2020 年 2 月 16 日

李姿昕，北京科技大学新材料技术研究院，在读研究生，主要研究方向为机器学习。

本文承担工作为：数据库平台概述，以及论文的写作与修改。

Li Zixin is a master student at Institute for Advanced Materials and Technology, University of Science and Technology Beijing. Her main research interest is machine learning.

In this paper she undertakes the following tasks: organizing the review on the development of database, as well as the writing and revision of manuscript.

E-mail: s20191363@xs.ustb.edu.cn

张能，北京科技大学新材料技术研究院，在读研究生，主要研究方向为基于机器学习方法的 Cu-Al 合金断裂性能研究。

本文承担工作为：数据库应用的分析与讨论，以及论文的写作与修改。

Zhang Neng is a master student at Institute for Advanced Materials and Technology, University of Science and Technology Beijing. His main research interest is machine learning based research on fracture properties of Cu-Al alloys.

In this paper he undertakes the following takes: analysis and discussion of database applications, as well as the writing and revision of manuscript.

E-mail: s20191421@xs.ustb.edu.cn

熊斌，北京科技大学新材料技术研究院，在读研究生，主要研究方向为机器学习方法在形状记忆合金马氏体相变研究中的应用。

本文承担工作为：文献的整理、查阅。

Xiong Bin is a master student at Institute for Advanced Materials and Technology, University of Science and Technology Beijing. His main research interest is machine learning based research on martensite transformation of shape memory alloys.

In this paper he undertakes the following tasks: collecting and summarizing references.

E-mail: g20189313@xs.ustb.edu.cn

胡云凤，北京科技大学新材料技术研究院，在读研究生，主要研究方向为高弹热效应形状记忆合金以及机器学习。

本文承担工作为：文献的整理、查阅。

Hu Yunfeng is a master student at Institute for Advanced Materials and Technology, University of Science and Technology Beijing. Her main research interests are focusing on high elastocaloric effect shape memory alloys and machine learning.

In this paper she undertakes the following tasks: collecting and summarizing references.

E-mail: s20181326@xs.ustb.edu.cn

赵新鹏，北京科技大学新材料技术研究院，硕士，主要研究方向为第一性原理计算，机器学习以及高弹热效应形状记忆合金。

本文承担工作为：数据库在机器学习中的应用分析与讨论以及全文统筹。

Zhao Xinpeng, master, is studying at Institute for Advanced

Materials and Technology, University of Science and Technology Beijing. His main research interests are First-principles calculations, machine learning and high elastocaloric effect shape memory alloys.

In this paper he undertakes the following tasks: conceptualizing and organizing the review on the application of database in machine learning.

E-mail: g20179183@ustb.cn

黄海友，北京科技大学新材料技术研究院，工学博士，副研究员。在 Applied Physics Letters，APL Materials，Scripta Materials 等期刊上发表学术论文 61 篇。授权发明专利 5 项；参编著作 3 部；2017 年获教育部自然科学奖二等奖。主要研究方向包括材料基因工程数据库与大数据技术，基于数据驱动的新材料研发和形状记忆合金等。

本文承担工作为：总体架构和主要思路。

Huang Haiyou, D.E, is an associate research fellow at Institute for Advanced Materials and Technology, University of Science and Technology Beijing. He has published more than 60 academic papers in Applied Physics Letters, APL Materials, Scripta Materials and other journals. He also has authorized 5 invention patents, edited 3 books and won the second prize of the Natural Science Award of the Ministry of Education in 2017. His main research interests are materials genome engineering database, big data technology, data-driven development method of new materials and shape memory alloys, etc.

In this paper he undertakes the following tasks: constructing manuscript structure and making up main ideas.

E-mail: huanghy@mater.ustb.edu.cn

引文格式：李姿昕,张能,熊斌,等.材料科学数据库在材料研发中的应用与展望[J]. 数据与计算发展前沿,2020,2(2):78-90.DOI:10.11871/jfdc.issn.2096-742X.2020.02.006.PID:21.86101.2/jfdc.2096-742X.2020.02.006.

Li Zixin, Zhang Neng, Xiong Bin, et al..Materials Science Database in Material Research and Development: Recent Applications and Prospects[J].Frontiers of Data & Coputing,2020,2(2): 78-90.DOI:10.11871/jfdc.issn.2096-742X.2020.02.006.PID:21.86101.2/jfdc.2096-742X.2020.02.006.