中国科学: 化学 2023 年 第 53 卷 第 1 期: 19 ~ 25
SCIENTIA SINICA Chimica chemcn.scichina.com
《中国科学》杂志社 SCIENCE CHINA PRESS
专题论述 AI+合成化学专刊

Atomly.net数据平台及其在无机化学中的应用

刘淼[1,2*], 孟胜[1,2*]

1. 中国科学院物理研究所, 北京 100190
2. 松山湖材料实验室, 东莞 523808
*通讯作者, E-mail: mliu@iphy.ac.cn; smeng@iphy.ac.cn

收稿日期: 2022-08-05; 接受日期: 2022-09-21; 网络版发表日期: 2022-12-14
中国科学院(编号: ZDBS-LY-SLH007, XDB33020000, CAS-WX2021PY-0102)和科技部重点研发计划(编号: 2021YFA0718700)资助

摘要 “大数据+无机化学”是一门新兴学科, 是物质科学研究的重要工具, 也是计算材料科学的未来形态. 基于以上先进理念, 我国近期建成了一套具有自主知识产权的无机材料科学数据库, 名为Atomly材料数据库. 本文详细介绍了Atomly材料科学数据的建设理念和应用实例, 并着重介绍了数据库带来的与化学合成相关新方法, 如无机材料热力学稳定性评估、化学反应路径推测、人工智能物性预测及大数据分析, 希望Atomly数据库及相关的无机材料数据方法为领域带来广泛支撑和推进.

关键词 数据库, 无机材料, 无机化学, 人工智能

1 引言

近年来, 化学、物理、材料科学等物质科学领域涌现了若干优质行业数据库, 带给了科学家海量的基础数据和新科研工具. “大数据+科学”的科研新方式也正在潜移默化地改变物质科学研究的行为, 并有望跳跃式地提升行业效率, 变革性地改变科研范式.

在无机化学领域中, 欧美科学家借助高通量第一性原理计算, 率先开发出若干高质量数据库. 2011年, 美国的Materials Project开创了世界上第一个计算类无机材料数据库[1,2], 并研发出一套高通量计算方法和软件工具体系[3~5]. 此数据库引爆了无机化学科研的信息化浪潮. 此后, 美、欧、日各国都各自开始着手建设同类型的无机材料数据库. 例如, 德国的NOMAD[6]、瑞士的AiiDA/MaterialsCloud[7]、美国的Aflow[8]/OQMD[9]都是基于此类高通量第一性原理计算理念的相似产物. 它们的出现将物质科学推进到了大数据时代, 加速了新材料研发进程, 改变了无机化学材料的研究范式. 例如, 人类过去的70年间, 平均每年发现约3.3个氮化物晶体, 而加州大学伯克利分校的Ceder团队, 通过高通量计算等无机材料大数据方法在一年内发现了92种该类材料, 并成功实验合成其中7种[10]; 美国能源部的JCESR中心, 使用Materials Project数据库, 仅用2年时间寻找到了当时性能最优的镁离子电池正极化合物[11,12], 并获得实验验证[13].

海量行业数据也催生了数据挖掘和人工智能在物质科学中的应用. 研究人员可根据应用需求, 通过数据快速定位目标化合物的化学空间和结构空间, 从而有效筛选新材料; 还可通过人工智能算法建立目标物性的预测模型, 并由此快速遍历更广阔的无机化合物相

引用格式: Liu M, Meng S. Atomly.net materials database and its application in inorganic chemistry. *Sci Sin Chim*, 2023, 53: 19–25, doi: 10.1360/SSC-2022-0167

空间; 通过数据技术提取物理量间的隐含关联, 总结无机化学中的“构-效”关系, 带来新的认知. 例如, 通过数据和人工智能方法认知电池材料中的离子输运行为[14~16].

从我国的基础科研需求出发, 发展无机化学中的大数据方法, 如高通量计算方法、数据库等, 可以补全我国的材料研发链条, 构建无机化学研发完整的生态体系, 从而深度提升物质科学的研发能力、速度和效率.

本文着重介绍一套具有我国自主知识产权的世界级无机材料科学数据库—Atomly[17], 并详述了该数据库的技术路线及应用实例. 运用业界顶级的材料数据技术和高通量第一性原理计算, Atomly数据库将30余万个无机材料的高质量数据带到我国科研人员触手可得之处, 为业界带来无机材料数据工具平台. Atomly数据库的出现改变了我国物理、化学、材料科学等领域长期依赖西方数据库舶来品的跟随局面, 为我国物质科学发展提供了优质的基础数据, 提升了整个行业的效率及竞争力.

2 Atomly数据库简介

中国科学院物理研究所和松山湖材料实验室早在2018年初便认识到无机化合物数据库的战略重要性, 并启动了数据库开发项目. 任务目标为: 对标若干世界顶级无机化合物数据库, 建立具有我国自主知识产权的世界级数据库, 为我国物质科学领域提供优质的基础数据, 防止在材料科学基础数据领域形成“卡脖子”的态势. 经历了4年多的开发和积累, 截至2022年中期, Atomly数据库(图1)已经涵盖了32万个无机晶体结构的第一性原理计算结果, 31万个能带和态密度, 1.2万个静态介电常数张量, 1.6万个力学张量, 10万个材料的高对称点不可约表示, 2000个半导体材料的杂化密度泛函电子结构计算, 以及由此推导出的化合物热力学稳定性、电极材料性能、压电性能、电导率、载流子有效值质量等数据. 该数据库目前的化合物数据量已达到Materials Project的2倍, 电子结构数据量是Materials Project的4倍, 数据规模和质量已经跻身世界顶级水准. 用户打开网站即可查询到海量数据. 月访问量达1万人次, 累计访问量20余万人次, 用户遍布全国各省份, 被许多高校引入材料科学、物理等专业的教学中.

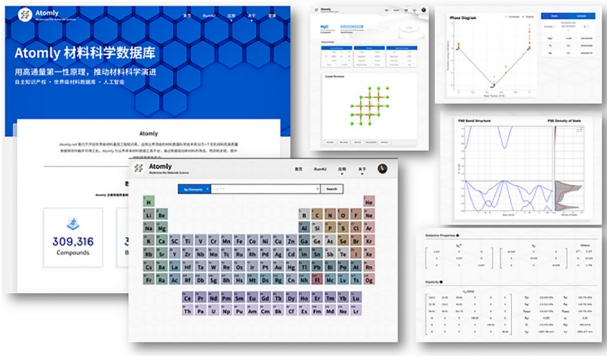

图 1 Atomly材料数据库的网站页面及数据展示(网络版彩图)

Figure 1 Website and data pages of Atomly materials database (color online).

Atomly数据库的底层逻辑与欧美同类数据库一致, 即: (1) 数据生产基于第一性原理计算, 即求解电子在原子势场中的量子力学方程, 从而推演出无机化合物的若干物性; (2) 高通量计算作业流程通过自研的软件包(IoP Materials Suite)实现, 通过高并发、全自动的第一性原理计算和结果分析作业流批量生产数据; (3) 硬件采用特殊架构的“数据增强型超级计算机”, 将高性能数据中心与针对材料科学计算优化的高性能计算机相融合. 该数据库是一个复杂的系统工程, 其最终实现了无机晶体化合物的计算、数据分析、结果推演、网站可视化, 是高度专业的领域知识和先进信息技术交叉融合的产物.

Atomly数据库有力地支撑我国物质科学研究, 并已有若干成功的用户案例. 例如, 用户通过数据库判定量子比特材料的热力学稳定性[18], 搜索二维碳材料结构[19], 依靠Atomly数据库发现新型无自旋带隙节线半导体[20], 比对和寻找量子自旋液体材料[21]等, 广泛且有效地支撑了我国物质科学研究.

3 大数据赋能无机化学研究

海量数据将给无机化学带来全新的视角, 并从方法论上带来创新. 一方面, 带来科学问题从“局部到全局”的转变, 从全局视角审视材料科学问题, 避免盲人摸象式科研; 另一方面, 引入数据科学的成熟技术(如人工智能), 通过智能工具代替人工, 并大幅提升效率.

3.1　判定无机物热力学稳定性

当我们拥有几乎所有人类已知的无机化合物的计算数据后, 便可以通过比较化合物的形成能, 勾勒出相空间的热力学稳定相边界, 并推演出任意化合物与稳定相的能量差距[22]. 图2描绘了Ti-O化学体系的热力学相图. 图中的每个点都是一个可能的无机晶体, 其形成能定义为基态Ti和O_2的反应放热. 形成能的负值代表放热, 正值代表吸热. 放热越多形成能越小, 化合物也越稳定. 由此可以通过大量的计算结果数据标定出Ti-O体系中的所有稳定相(图2中的绿点). 而非稳定相(图2中的棕点)的热力学稳定性可由其与热力学稳定相间的形成能之差定量给出, 定义为Ehull(energy above formation energy convex hull的缩写). 考虑到第一性原理计算的误差和温度的影响, 一般可以认为Ehull<50 meV的离子化物、Ehull<20 meV的合金为较稳定相, 并预示着这些无机化合物很容易被合成[23].

从上述无机化合物热力学相图判断化合物的热力学稳定性的方法是材料数据库独特的优势, 也是目前通过计算判断材料是否能被合成的最优解决方案. 相比之下, 广为使用的声子谱计算判断化合物稳定性的方法可信度极低.

借助Atomly数据库中的热力学相图, Jiang等[24]实现了新材料体系的快速搜索和预测, 筛选出25个热力学稳定的“类CsV_3Sb_5”笼目材料, 为发现新型量子材料提供了更多可能性. 其中的$CsTi_3Bi_5$已被实验合成[25]. 依据Atomly数据库, Yu等[26]从18万个无机材料中搜索得到了若干“类MgB_2”的超导材料, 并通过实验证实了$BaGa_2$极易合成并具有超导物性, 实现了材料搜索的“端到端”模式.

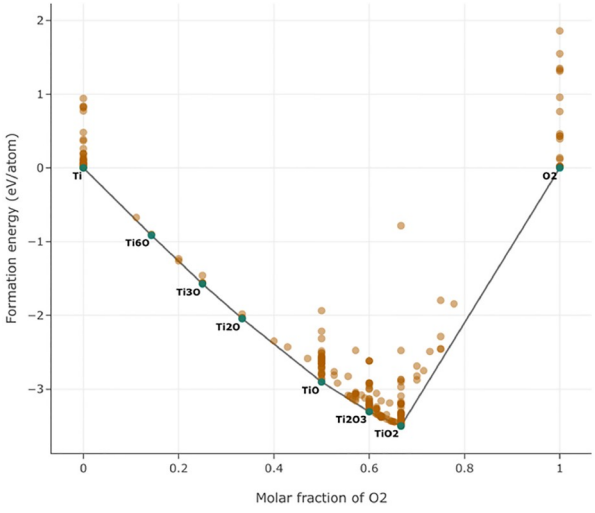

图 2　Ti-O体系的热力学相空间分布. 其中绿色点代表稳定化合物, 棕色点代表非稳定相, 黑色线条勾画出了Ti-O的形成能边界(网络版彩图)

Figure 2　Thermodynamic phase diagram of Ti-O chemical system. Green dots denote the thermodynamic stable phases, the brown dots chart many unstable phases, and the black lines show the formation energy hull of the Ti-O chemical space (color online).

3.2　推断化学反应路径

与无机化合物的稳定性类似, 也可以通过数据库推演化合物反应的方向和路径, 如电化学、界面反应和溶液pH值对反应的影响.

以Na离子电池电极材料Ga_2S_3的转化型反应为例, 可由数据库推算Na与Ga_2S_3所有可能的反应路径, 并逐步选找出所有组合中, 反应放热最多的路径, 即为该应用中的最优反应路径(图3). 通过分析, 可推断出Ga_2S_3与Na的反应, 先后经过了6步, 分别形成了GaS、

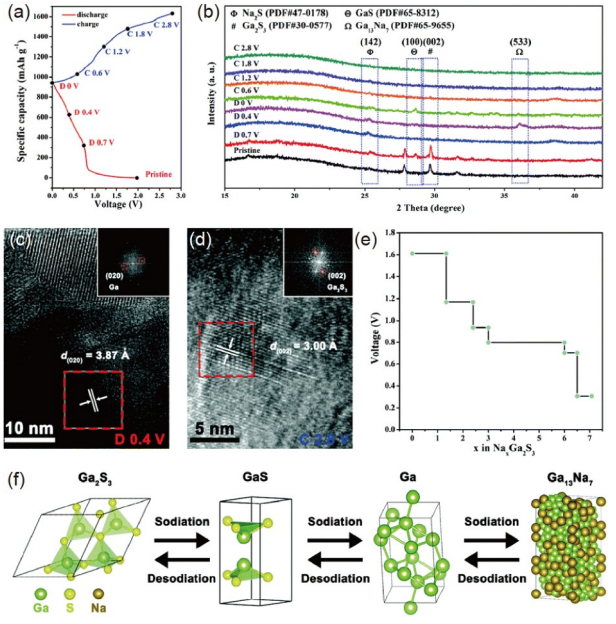

图 3　Na离子电池电极材料Ga_2S_3转化型反应的实验表征与理论计算预测对照. 通过分析数据库中各种可能的反应路径, 可以预测出充电过程中, Ga_2S_3被逐渐还原到GaS、$Na_4Ga_2S_5$、Na_3GaS_3、Ga、$NaGa_4$等中间产物, 并最终形成了Na_7Ga_{13}. 与实验观测数据相符(网络版彩图)

Figure 3　The conversion reaction of Ga_2S_3 as a viable anode material for Na ion battery. The complied reaction paths employing the database suggest the Ga_2S_3 is electrochemically reduced to GaS, $Na_4Ga_2S_5$, Na_3GaS_3, Ga, $NaGa_4$, and finally reaching Na_7Ga_{13}, which is in good agreement with experimental observations (color online).

$Na_4Ga_2S_5$、Na_3GaS_3、Ga、$NaGa_4$等中间产物，并最终形成了Na_7Ga_{13}, 印证了实验观测[27].

3.3 人工智能助力化合物物性预测

随着无机化合物数据的增长，构建准确的人工智能物性预测模型已逐步成为一个重要的科学问题，并有望形成广泛支撑行业的一种新工具. 其思路为，通过第一性原理生产数据，通过数据训练人工智能模型，进而通过模型开展大规模预测, 拓展相空间. 模型的准确性和外推本领是这个领域中的两个核心指标，决定了方法的可用性.

该领域的早期方法通常受限于的数据集的大小，因而采用较为简单的神经网络，因此精度有限. 例如，Cao等[28]利用4000余个数据点，训练了基于Magpie描述符的卷积神经网络. 数据量带来的制约，使该模型的预测精度和泛化本领非常有限. 而近年来数据数量和质量的大幅提升，并在此基础上发展出诸多新方法，如图卷积神经网络(CGCNN)[29]，甚至Transformer等算法的应用，不断提升了人工智能模型的预测精度. 比如，近期微软提出的Graphormer模型[30]，可以较为精确地预测催化反应中的吸附能.

以Atomly无机化合物数据库中个无机晶体化合物的高通量第一性原理计算数据集为基础，亦可训练得到较好的人工智能模型(图4). 例如, 近期工作显示, 从物理思维出发，以Atomly数据库中的17万数据为出发点，可构建出无机晶体形成能的高精度泛化模型，神经网络模型精度可以达到R^2=0.982，平均绝对误差(MAE)=0.072 eV atom^{-1}，优于业界知名的CGCNN、CrabNet、Roost等模型[31]. 该工作的一个亮点为提出了基于元素电负性之差的化合物形成能的特征描述符，这些描述符可有效提取出原子与邻域原子间的电负性和局域结构等信息，并精确捕捉到原子间的相互作用，进而使人工智能模型的预测精度得到了大幅提升. 同时，后续的随机森林模型进一步验证了元素电负性之差与化合物形成能的强相关性，证实了物理思维的正确性，为提升人工智能模型的可解释性提供了新思路，也为新材料搜索提供了一种高效、低成本的结合能预测手段.

该领域近期发展迅速，目前在matbench排行榜中，无机化合物的形成能预测精度已经达到甚至超越了化学精度. 例如，ALIGNN的平均绝对误差(MAE)为0.022 eV atom^{-1} [32].

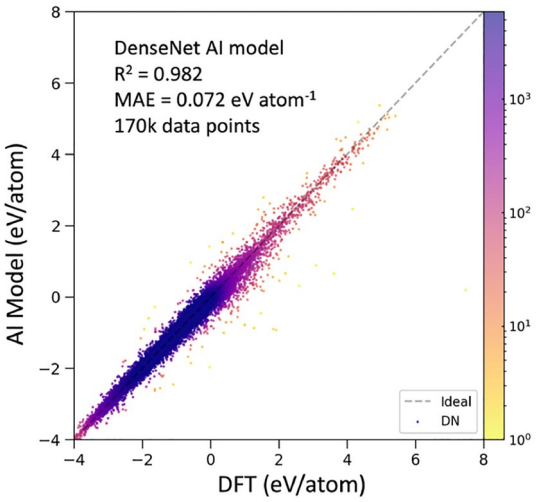

图 4 基于Atomly数据训练得到的无机化合物形成能人工智能预测模型. 该模型具有较高精度和优秀的外推本领(网络版彩图)

Figure 4 The AI model for predicting the formation energy of inorganic compounds. The model shows good accuracy and generality (color online).

该领域目前最大的瓶颈是数据. 更大、更精确的数据集势必带来模型可用性的提升. 凭借Atomly的数据量优势，其下一代人工智能模型精度已经达到MAE为0.020 eV atom^{-1}(该结果将于近期发布).

3.4 大数据揭示新认知

从原子尺度理解化合物“结构-物性”间的构效关系是物质科学领域的基本问题，深入细致地厘清物质微观局域结构的统计特征，有助于人们更好地理解物理、材料、化学等众多学科中的科学问题. 当今广泛被人们使用的物质科学基本数据大多都源自20世纪中后期, 较为陈旧. 例如, 被物质科学领域广泛使用近半个世纪的离子半径数值，源自20世纪60年代的统计数值，随着近年材料科学数据的不断积累，海量数据也将给物质科学领域带来新数值、新认知.

Jia等[33]细致地统计3万余个过渡金属氧化物的晶体信息，获得了过渡族金属离子的配位结构、价态、离子键长、原子磁矩、Jahn-Teller效应导致的局部形变、结构稳定性等物理量的详细信息，取代广泛使用近半个世纪的，仅用“离子半径”一个参数表征材料的传统做法. 数据揭示出每种过渡族金属离子都具有各

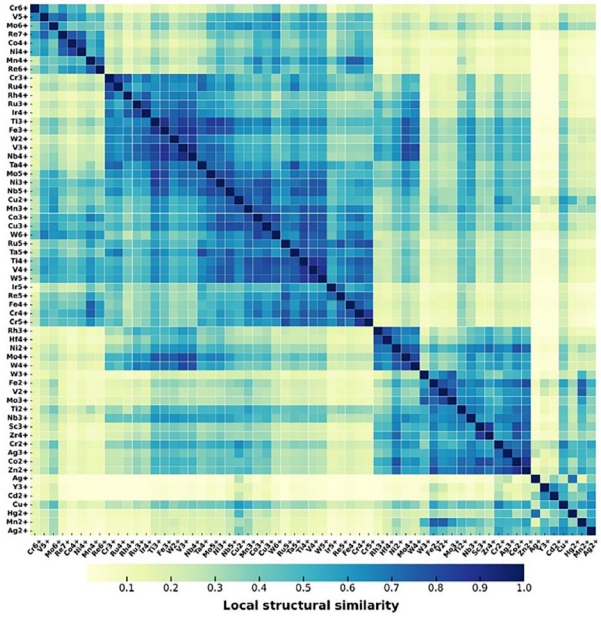

图 5　从2万个过渡金属离子化合物，10万个离子局域结构中提取出的过渡金属离子的局域结构相似性(网络版彩图)

Figure 5　The local structure similarity of transition metal cations extracted from 100,000 atomistic sites in 20,000 inorganic ionic compounds (color online).

自的“形状”、“尺寸”和“原子磁矩”，从而勾画出固体中每种过渡金属离子的“个性”画像；通过人工智能方法，进一步获取过渡金属离子的结构相似性“模板”，可用于指导新材料设计和稳定性快速评估.

依据上述离子的结构相似性知识，Atomly数据库通过离子替换扩展了人类已知的化合物空间，生成了6万余个人类未知的新结构，通过对这些结构进行高通量计算，找到了5000个稳定的新化合物(Ehull < 20 meV atom^{-1})，这些新材料很容易实验合成，有效地扩展了材料科学的化合物“版图”，为探索更广阔的无机材料相空间提出了新的思路，所有化合物信息均可在Atomly材料数据库(https://atomly.net/)中查询.

3.5　“数据+科学”的前景与挑战

自然科学已经经历了“实验观测”、“理论推理”、“计算辅助”三个形态. 近期，科学界已经普遍认识到“数据驱动”将成为推动业界变革性跃进的新范式, 也常被称为“第四范式”. 范式变革的底层逻辑是新范式将带来认知和预测能力的突飞猛进, 从而“有的放矢”地提高行业生产力和效率. 数据科学方法(如人工智能)的本质是高维参数空间的函数拟合，是自然科学领域底层数学工具的一次深刻变革，是从探索简单体系迈向探索复杂系统的必由之路，是基础科研向实际应用转变中的契机.

以计算材料科学和计算物理为例，目前的科学研究已经不满足于探索个别体系，越来越多的科学发现源自“鸟瞰”一类体系，这无疑将科学家的视野带到了新的高度. 本文的案例只是第四范式早期阶段的初步实践, 未来还有很多可能性. 例如, 但不限于: (1) 从数据出发洞悉物理量之前的隐含关联，从而为认识物质世界提供思路；(2) 通过数据建立人工智能预测模型，进而扫描广阔的相空间，快速寻找目标化合物；(3) 用人工智能模型替代目前的少参数数学模型，进而解决已有理论中的痛点问题，如密度泛函领域已有成功案例表明人工智能泛函可以更高效和精确地描述多电子体系中的交换和关联作用[34].

数据是上述可能性的基石，是本轮科学范式变革中的关键环节和挑战. 没有优质的大数据作为土壤，人工智能方法是无法生长的. 可以预见，数据将成为未来科学研究的基础资源，数据库将成为重要的科学基础设施，像大学科装置一样滋养各学科成长. 然而目前科学界虽然已经逐渐意识到数据的重要性，但未必了解科学需要什么样的数据. 我们认为科学数据需要满足以下三个关键点：结构化、高度的标准化和一致性、大数据. 数据结构化是指将数据存储为可被计算机高速读写查询的结构化数据，而非以人为对象的文件；高度的标准化和一致性要求破除数据的“批次效应”和“孤岛数据”，破除当前盲目的数据汇交的做法；数据需要密集覆盖足够的参数空间，才足以支撑人工智能建模，引发量到质的变化. Atomly数据库是以上述思考为出发点建立的，因此才能通过不断积累，形成较好的基础.

4　结论及展望

在信息技术助力下, 物质科学发展进入了“大数据+人工智能”时代. Atomly数据库的诞生, 打破了国外在此领域中的垄断地位，为广泛支撑我国物质学科的发展打造了优质基础数据及工具，并已经开始发挥效力, 切实地推进了领域发展.

目前Atomly数据库还在不断迭代和快速积累，预

计在近几年内数据数量和物性维度还会大幅提升，数量将提升到数百万量级，并将覆盖有机小分子、有机金属框架、合金等细分领域. 近期，我们将发布约120万个小分子量子化学计算数据集，50万个heusler合金计算数据集，3万个有机金属框架计算数据集. 未来还将通过人工智能及大数据方法，挑战化学合成等领域内前沿问题，并以数据为基础，支撑新材料发现、有机化学、能源材料等众多方向的研究.

Atomly的近期目标是建立科学数据的基础框架，也更侧重于能力建设. 以满足大多数科研工作者的需求为出发点，该数据库还会不断扩大数据的业务范围，并将逐步扩展到有机化学、生物等物质科学领域，最终形成可以广泛支撑行业发展的科学数据基础设施.

面对挑战，目前我国的众多机构都已经意识到了数据的重要性，预计该领域将进入一段增长期. 参考欧美数据库经验及发展规律，领域内将会涌现出多元的小规模专业数据库，通过若干专业化小数据库不断迭代，并最终沉淀得到一些精品. Atomly数据库作为先行者，已积累了技术积累和实践经验，并将继续砥砺前行, 夯实我国基础科学的数据基础.

致谢 感谢中国科学院物理所和松山湖材料实验室对本项目提供计算资源. 感谢项目的开发人员: 贾华昱、芦腾龙、梁英宗、谢帆恺、蔡光辉、喻泽、姜昱韬、王亚南、陈明威、黄一帆等.

参考文献

1 Jain A, Ong SP, Hautier G, Chen W, Richards WD, Dacek S, Cholia S, Gunter D, Skinner D, Ceder G, Persson KA. *APL Mater*, 2013, 1: 011002
2 Horton MK, Montoya JH, Liu M, Persson KA. *npj Comput Mater*, 2019, 5: 64
3 Ong SP, Richards WD, Jain A, Hautier G, Kocher M, Cholia S, Gunter D, Chevrier VL, Persson KA, Ceder G. *Comput Mater Sci*, 2013, 68: 314–319
4 Jain A, Ong SP, Chen W, Medasani B, Qu X, Kocher M, Brafman M, Petretto G, Rignanese G-, Hautier G, Gunter D, Persson KA. *Concurrency Computat-Pract Exper*, 2015, 27: 5037–5059
5 Ong SP, Cholia S, Jain A, Brafman M, Gunter D, Ceder G, Persson KA. *Comput Mater Sci*, 2015, 97: 209–215
6 Draxl C, Scheffler M. *J Phys Mater*, 2019, 2: 036001
7 Talirz L, Kumbhar S, Passaro E, Yakutovich AV, Granata V, Gargiulo F, Borelli M, Uhrin M, Huber SP, Zoupanos S, Adorf CS, Andersen CW, Schütt O, Pignedoli CA, Passerone D, VandeVondele J, Schulthess TC, Smit B, Pizzi G, Marzari N. *Sci Data*, 2020, 7: 299
8 Curtarolo S, Setyawan W, Hart GLW, Jahnatek M, Chepulskii RV, Taylor RH, Wang S, Xue J, Yang K, Levy O, Mehl MJ, Stokes HT, Demchenko DO, Morgan D. *Comput Mater Sci*, 2012, 58: 218–226
9 Saal JE, Kirklin S, Aykol M, Meredig B, Wolverton C. *JOM*, 2013, 65: 1501–1509
10 Sun W, Bartel CJ, Arca E, Bauers SR, Matthews B, Orvañanos B, Chen BR, Toney MF, Schelhas LT, Tumas W, Tate J, Zakutayev A, Lany S, Holder AM, Ceder G. *Nat Mater*, 2019, 18: 732–739
11 Liu M, Rong Z, Malik R, Canepa P, Jain A, Ceder G, Persson KA. *Energy Environ Sci*, 2015, 8: 964–974
12 Liu M, Jain A, Rong Z, Qu X, Canepa P, Malik R, Ceder G, Persson KA. *Energy Environ Sci*, 2016, 9: 3201–3209
13 Sun X, Bonnick P, Duffort V, Liu M, Rong Z, Persson KA, Ceder G, Nazar LF. *Energy Environ Sci*, 2016, 9: 2273–2277
14 He B, Ye A, Chi S, Mi P, Ran Y, Zhang L, Zou X, Pu B, Zhao Q, Zou Z, Wang D, Zhang W, Zhao J, Avdeev M, Shi S. *Sci Data*, 2020, 7: 153
15 Zhang L, He B, Zhao Q, Zou Z, Chi S, Mi P, Ye A, Li Y, Wang D, Avdeev M, Adams S, Shi S. *Adv Funct Mater*, 2020, 30: 2003087
16 He B, Mi P, Ye A, Chi S, Jiao Y, Zhang L, Pu B, Zou Z, Zhang W, Avdeev M, Adams S, Zhao J, Shi S. *Acta Mater*, 2021, 203: 116490
17 https://atomly.net
18 Li X, Zhang Y, Yang C, Li Z, Wang J, Su T, Chen M, Li Y, Li C, Mi Z, Liang X, Wang C, Yang Z, Feng Y, Linghu K, Xu H, Han J, Liu W, Zhao P, Ma T, Wang R, Zhang J, Song Y, Liu P, Wang Z, Yang Z, Xue G, Jin Y, Yu H. *Appl Phys Lett*, 2021, 119: 184003
19 Shi X, Li S, Li J, Ouyang T, Zhang C, Tang C, He C, Zhong J. *J Phys Chem Lett*, 2021, 12: 11511–11519
20 Ding G, Wang J, Chen H, Zhang X, Wang X. *J Mater Chem C*, 2022, 10: 6530–6545
21 Liu W, Yan D, Zhang Z, Ji J, Shi Y, Jin F, Zhang Q. *Chin Phys B*, 2021, 30: 107504
22 Ong SP, Wang L, Kang B, Ceder G. *Chem Mater*, 2008, 20: 1798–1807

23 Aykol M, Dwaraknath SS, Sun W, Persson KA. *Sci Adv*, 2018, 4: eaaq0148
24 Jiang Y, Yu Z, Wang Y, Lu T, Meng S, Jiang K, Liu M. *Chin Phys Lett*, 2022, 39: 047402
25 Yang HT, Zhao Z, Yi XW, Liu J, You JY, Zhang Y, Guo H, Lin X, Shen C, Chen H, Dong X, Su G, Gao HJ. arXiv: 2209.03840
26 Yu Z, Bo T, Liu B, Fu Z, Wang H, Xu S, Xia T, Li S, Meng S, Liu M. *Phys Rev B*, 2022, 105: 214517
27 Wang P, Liu M, Mo F, Long Z, Fang F, Sun D, Zhou YN, Song Y. *Nanoscale*, 2019, 11: 3208–3215
28 Cao Z, Dan Y, Xiong Z, Niu C, Li X, Qian S, Hu J. *Crystals*, 2019, 9: 191
29 Xie T, Grossman JC. *Phys Rev Lett*, 2018, 120: 145301
30 https://github.com/Microsoft/Graphormer
31 Liang Y, Chen M, Wang Y, Jia H, Lu T, Xie F, Cai G, Wang Z, Meng S, Liu M. *Sci China Mater*, 2022, doi: 10.1007/s40843-022-2134-3
32 Choudhary K, DeCost B. *npj Comput Mater*, 2021, 7: 185
33 Jia H, Horton M, Wang Y, Zhang S, Persson KA, Meng S, Liu M. *Adv Sci*, 2022, 9: 2202756
34 Kirkpatrick J, McMorrow B, Turban DHP, Gaunt AL, Spencer JS, Matthews AGDG, Obika A, Thiry L, Fortunato M, Pfau D, Castellanos LR, Petersen S, Nelson AWR, Kohli P, Mori-Sánchez P, Hassabis D, Cohen AJ. *Science*, 2021, 374: 1385–1389

Atomly.net materials database and its application in inorganic chemistry

Miao Liu[1,2*], Sheng Meng[1,2*]

[1] *Institute of Physics, Chinese Academy of Sciences, Beijing 100190, China*
[2] *Songshan Lake Materials Laboratory, Dongguan 523808, China*
**Corresponding authors (email: mliu@iphy.ac.cn; smeng@iphy.ac.cn)*

Abstract: “Big data + inorganic chemistry”, an emerging scientific discipline, creates new and important tools to futurize computational materials science. Leveraging this idea, Atomly is founded as a world-class inorganic materials database that is developed and based in China. This article presents the engineering details, demonstrates viable applications of the Atomly database, and primarily introduces the new tools and methodologies that are advancing the field of chemical syntheses, such as thermodynamic stability estimation, reaction path evaluation, artificial intelligence-guided property prediction, and big-data analysis. We hope that the Atomly database and innovative techniques can serve as useful tools to advance chemical science.

Keywords: database, inorganic materials, inorganic chemistry, artificial intelligence

doi: 10.1360/SSC-2022-0167